Exploiting Wikipedia Semantics for Computing Word Associations

by

Shahida Jabeen

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Computer Science.

Victoria University of Wellington
2014

Abstract

Semantic association computation is the process of automatically quantifying the strength of a semantic connection between two textual units based on various lexical and semantic relations such as hyponymy (`car` and `vehicle`) and functional associations (`bank` and `manager`). Humans have can infer implicit relationships between two textual units based on their knowledge about the world and their ability to reason about that knowledge. Automatically imitating this behavior is limited by restricted knowledge and poor ability to infer hidden relations.

Various factors affect the performance of automated approaches to computing semantic association strength. One critical factor is the selection of a suitable knowledge source for extracting knowledge about the implicit semantic relations. In the past few years, semantic association computation approaches have started to exploit web-originated resources as substitutes for conventional lexical semantic resources such as thesauri, machine readable dictionaries and lexical databases. These conventional knowledge sources suffer from limitations such as coverage issues, high construction and maintenance costs and limited availability. To overcome these issues one solution is to use the *wisdom of crowds* in the form of collaboratively constructed knowledge sources. An excellent example of such knowledge sources is Wikipedia which stores detailed information not only about the concepts themselves but also about various aspects of the relations among concepts.

The overall goal of this thesis is to demonstrate that using Wikipedia for computing word association strength yields better estimates of humans' associations than the approaches based on other structured and unstructured knowledge sources. There are two key challenges to achieve

this goal: first, to exploit various semantic association models based on different aspects of Wikipedia in developing new measures of semantic associations; and second, to evaluate these measures compared to human performance in a range of tasks. The focus of the thesis is on exploring two aspects of Wikipedia: as a formal *knowledge source*, and as an informal *text corpus*.

The first contribution of the work included in the thesis is that it effectively exploited the knowledge source aspect of Wikipedia by developing new measures of semantic associations based on Wikipedia hyperlink structure, informative-content of articles and combinations of both elements. It was found that Wikipedia can be effectively used for computing noun-noun similarity. It was also found that a model based on hybrid combinations of Wikipedia structure and informative-content based features performs better than those based on individual features. It was also found that the structure based measures outperformed the informative-content based measures on both semantic similarity and semantic relatedness computation tasks.

The second contribution of the research work in the thesis is that it effectively exploited the corpus aspect of Wikipedia by developing a new measure of semantic association based on asymmetric word associations. The thesis introduced the concept of asymmetric associations based measure using the idea of directional context inspired by the free word association task. The underlying assumption was that the association strength can change with the changing context. It was found that the asymmetric-association based measure performed better than the symmetric measures on semantic association computation, relatedness based word choice and causality detection tasks. However, asymmetric-associations based measures have no advantage for synonymy-based word choice tasks. It was also found that Wikipedia is not a good knowledge source for capturing verb-relations due to its focus on encyclopedic concepts specially nouns.

It is hoped that future research will build on the experiments and discussions presented in this thesis to explore new avenues using Wikipedia for finding deeper and semantically more meaningful associations in a wide range of application areas based on humans’ estimates of word associations.

Acknowledgment

"The difference between a successful person and others is not a lack of strength, nor a lack of knowledge, but rather a lack of will". Three years ago, this *will* motivated me to leave my family, my job and my country to pursue my career in research. Three precious years of my life in New Zealand have been full of experiences that taught me the value of hard work in life. This thesis is an outcome of the support, help and sacrifice of many priceless people who have had a great influence on my life and studies over the last three years. I greatly admire my supervisors, Dr. Sharon Gao and Dr. Peter Andreae, for believing in me. I am indebted to them for their outstanding support and mentoring. Our weekly meetings have undoubtedly been key in keeping me focused and giving me the courage to achieve my goals. I would also like to express my gratitude to Prof. Mengjie Zhang, for being very supportive and encouraging. His success celebrations and Christmas BBQ's have always been wonderfully exciting. I wish to thank ECRG group for creating an insightful and friendly research environment. I am also grateful to Peter D. Turney (National Research Council of Canada) for providing me datasets for the word choice task. I also want to thank Thomas K. Landauer (LSA and NLP Lab, University of Colorado) for providing the TOEFL dataset.

Last but not the least, I want to thank my family for their continuous support, help and prayers and my beloved husband for his unconditional love. It would not have been possible for me to complete my journey without his support and encouragement.

List of Tables

List of Figures

List of Abbreviations

APRM - Asymmetry-based Probabilistic Relatedness Measure

CFP - Context-Filtered Profile

COPA - Choice of Plausible Alternatives

CPRel - Context profile based Relatedness

DCRM - Directional Context based Relatedness Measure

GP - Gaussian Process

KS - Knowledge Source

L-1-O CV - Leave-One-Out Cross Validation

LE - Link Estimation

LP - Link Probability

NTF - Normalized Term Frequency

OSR - Overlapping Strength based Relatedness

PMI - Pointwise Mutual Information

SVM - Support Vector Machines

WLM - Wikipedia Link based Measure

Contents

Chapter 1

Introduction

1.1 Motivation

Thinking is a dynamic process of human cognition and is based on human knowledge and experience of the real world. Furthermore, knowledge is a product of human interpretation of objects and phenomena of reality and manifest itself in words. Humans can infer implicit relationships between words based on their knowledge about the world and their ability to reason about that knowledge. Humans' word associations can be interpreted by the principle of learning by contiguity. According to this principle, objects once experienced together tend to be linked to each other in the human imagination, so that if one of the objects is thought of then the other object is likely to be thought of also [218]. Association by contiguity exists when objects are situated close together in time or space e.g. *king and queen*. Humans can compare and judge semantic associations of different words even if they do not know the formal definition and boundaries of associations. For instance, they can tell that `Potato` and `Chips` have stronger association than `Potato` and `Pen`. Moreover, dissimilar or opposite words can also be strongly associated. So if a person is told a word `day`, his first response word might be `night` which is opposite in meaning to the word `day` but is still closely associated. Associating such words is generally an easy task for humans but imitating the same phenomenon using automated approaches is limited by shallow and restricted knowledge, poor ability to infer hidden relations and lack of sufficient understanding of the complex relations that exist between real word objects.

There is no formal definition of semantic association but generally it is referred to as the semantic connection between two textual units (words, phrases, n-grams, sentences or even documents) [76, 129]. Accordingly, semantic association computation is the task of automatically quantifying the strength of a semantic connection between two textual units based on different kinds of semantic relations. It takes two pieces of text as input and produces a real number representing the strength of the semantic connection between them as output.

With the exponential growth of the World Wide Web and ever increasing importance of retrieving relevant information from the web, semantic association computation has become a critical research area. Semantic association computation is a key component of many applications belonging to a multitude of fields such as computational linguistics, cognitive psychology, information retrieval and artificial intelligence. Examples of such applications include information extraction [209], text summarization [11], opinion mining [155], question answering [146], text classification [57], query expansion [212], topic identification [78, 18, 39], automatic keyphrase extraction [139], topic indexing [127], word sense disambiguation [2, 158, 132], document clustering [193, 78, 195] and spelling correction [22, 91].

The process of semantic association computation quantifies the association strength between two concepts by identifying a chain of possible lexical and semantic connections between them. It requires an understanding of the implicit relations of concepts based on deeper level of world knowledge about the entities in consideration. For instance, consider the following pair of text fragments.

- `Teeth`
- `Crime`

To correctly assess that these two concepts are related, background knowledge on `Forensic Science` is required but this knowledge can not be found in the text fragments themselves. The task of computing semantic associations of natural language concepts relies on the knowledge of a broad range of real world concepts and their implicit as well as explicit relations [225]. A major limiting factor for computing semantic associations is undoubtedly the requirement of a good background knowledge source.

The choice of background knowledge source plays a critical role in the performance of the semantic association measure relying on it. Background knowledge sources vary from domain specific thesauri to lexical dictionaries and from hand crafted taxonomies to knowledge bases. A knowledge source should have the following three attributes to effectively support semantic association computation.

- **Coverage:** The coverage of a knowledge source is the extent of its collected knowledge. It should be broad enough to provide information about all relevant concepts. Also, it should have sufficient depth to discover the hidden chains of various lexical and semantic connections among related concepts.

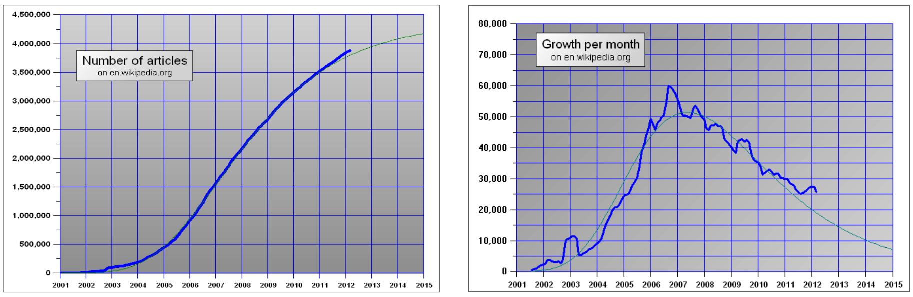
- **Quality:** The information contained by a knowledge source should be of high encyclopedic quality. Moreover, the information provided by a knowledge source should be accurate, authentic and up-to-date.

- **Semantic Richness:** A knowledge source should be semantically rich. This involves encoding many of the classical as well as non-classical semantic relations among concepts, implicitly or explicitly.

Various knowledge sources are used to backup semantic measures in computing semantic associations. Based on the differences in the structural organization of information, existing knowledge sources range from unstructured text corpora to semi-structured and fully structured knowledge sources. Unstructured corpora are simple to create and cheap to maintain. Semantic approaches based on unstructured or informal corpora usually rely on statistical or distributional features at word level. The choice of corpus and its size have direct impact on the performance of such approaches. On the other hand, semi-structured and fully-structured knowledge sources are expensive to build but more effectively encode the relations which are only implicit in the unstructured text corpora.

Researchers have started to exploit the web-originated resources as substitutes for conventional lexical semantic resources such as dictionaries, thesauri and lexical databases. These conventional knowledge sources suffer from certain limitations such as coverage issues, high construction and maintenance costs and limited availability. To overcome such issues, one solution is to exploit the *wisdom of crowds* [228] in the form of collaboratively constructed knowledge sources. Existing research has shown the semantic capabilities of collaboratively constructed knowledge sources in various natural language processing tasks [216, 197, 227, 140]. An excellent example of the knowledge sources relying on the *wisdom of crowds* is Wikipedia, which stores a great deal of information not only about concepts themselves but also about various aspects of their relationships. It has many properties in common with conventional knowledge sources (such as dictionaries, thesauri, wordnets and encyclopedia) but it represents and combines them in a unique way. Wikipedia sufficiently demonstrates all the necessary aspects of a good knowledge source, as given below.

- **Coverage:** As a collaboratively constructed knowledge source, the coverage of Wikipedia evolves as the human knowledge does. Since its inception in January 2001, Wikipedia is growing continuously and has become the largest freely available knowledge source with more than 4 million articles[1]). Its growth has been exponential in the number of key edits and articles [219]. Wikipedia occupies 6th position among the most visited websites on the Internet[2]. As an online encyclopedia, free from page limits, unrestricted by weight and volume, it satisfies the notion of *comprehensiveness*. It is a multilingual encyclopedia (available in 285 languages[3]) constructed by collaborative efforts of thousands of volunteers. The English version of Wikipedia is the largest of all. Its growth rate and the number of articles are shown in Figure 1.1. This model is related to the size of Wikipedia, which is always growing as there will always be new events and people to be described in the future.



(a) Actual number of articles (b) Article growth per month

Figure 1.1: The growth statistics of English Wikipedia [214].

As a knowledge source Wikipedia not only focuses on general vocabulary but also covers a large number of named entities and domain specific terms. The specialty of Wikipedia is that each of its articles is dedicated to a single topic with an additional benefit of heavy linking between articles. Wikipedia also covers specific senses of a word, synonymy, spelling variations, abbreviations and derivations. Im-

[1] `http://en.wikipedia.org/wiki/Wikipedia:Size`
[2] `http://www.alexa.com/topsites`
[3] `http://en.wikipedia.org/wiki/List_of_Wikipedias`

portantly, it has a semantically rich network relating categories that cover different types of lexical semantic relations such as hyponymy and hypernymy[4]. Wikipedia provides a knowledge base for extracting information in a more structured way than a search engine and with a coverage better than other knowledge sources [129].

- **Quality:** Wikipedia is an example of global collective intelligence. It paved the way for rapid expansion by letting its users modify any existing article or create new articles. Wikipedia articles are continually edited and improved over time, leading to an upward trend of quality as indicated by the Figure 8.1(b). Due to popularity and extensive use of Wikipedia, a number of studies were conducted to assess the quality of Wikipedia based on many important metrics such as the number of edits [219], unique editors [114], factual accuracy [45, 20, 93], credibility [33], revert time [211] and formality of language [47]. These studies have shown that the quality of Wikipedia articles continue to increase on average, as the number of collaborates and the number of edits increase [219, 220]. The quality of Wikipedia is comparable to other hand crafted encyclopedias. Research has demonstrated that the quality of Wikipedia content is found comparable to that of Encyclopedia Britannica [93]. Wikipedia articles reflect changes according to time as well as occurrence of events and are kept quite up-to-date. It is true that few articles are of high quality from the start. However, the articles pass though a long process of discussion, debate and argument before setting into a balanced representation of knowledge. Hence, Wikipedia can be utilized in artificial intelligence and natural language processing applications in the same way as manually created knowledge resources, which is an invaluable benefit of Wikipedia.

- **Semantic Richness:** In comparison with other knowledge sources,

[4] please see glossary for detail

> Wikipedia is a semantic resource that combines certain semantic properties in a unique way. Different structural components of Wikipedia reflect various aspects of a good semantic resource. Wikipedia has various semantic elements which implicitly as well as explicitly encode the lexical semantic information. With extensive *hyper-linking* between semantically related *articles*, Wikipedia could be considered as a huge semantic network free of size constraints. This semantic network is a rich source for explicitly labeled links between articles. Another source of lexical information are *redirects* which cover a whole range of various elements like synonymy, surface forms, abbreviations and derivations. The *category network* of Wikipedia is a good source for identifying implicit relations of concepts. *Disambiguation pages* not only represent various senses of a word but also categorize these senses into different groups explicitly.

All these attributes indicate the potential of Wikipedia to be used as a promising resource for mining semantic connections to augment word association computation.

1.2 Problem Statement

Despite many research efforts, computing semantic associations comparable to human judgments is still considered a hard problem as no single optimal solution is able to estimate all kinds of semantic associations correctly. Over the past few years, semantic association computation has received great attention due to its critical role in the performance of natural language semantics based applications. However, there are certain factors that hindered the performance of automatic approaches from computing semantic associations. Most prominent of these factors are inherent complexity of natural language semantics, poor estimates of common sense and the requirement for a background knowledge with sufficient depth and breadth to decode hidden semantic connections among various concepts. While the success of existing approaches addressing the first two factors has been slow, the third factor has been investigated vigorously due to the availability of huge repositories of online knowledge organized in unstructured to semi-structured and fully-structured knowledge sources over the last two decades. Sections 2.5 and 2.6 present a detailed survey of different kinds of knowledge sources and the association computation research based on those knowledge sources.

Existing research on semantic association computation has attempted to compute the strength of semantic associations by taking into account various aspects such as statistical features, syntactic and lexical dependencies, rhetorical relations and semantic indicators derived from various sources of world knowledge such as thesauri, lexical databases, dictionaries, ontologies and encyclopedias. Consequently, this problem is studied in a variety of fields. The research community has witnessed a steady transformation of information retrieval from syntactic information extraction (where statistical and distributional methods were employed to determine the physical presence or absence of a word or a concept) to semantic information retrieval (where much deeper understanding of implicit relations between concepts is needed to retrieve the desired information).

Existing approaches have attempted to solve the problem of computing the semantic association strength by taking into account various aspects (such as distributional similarity, lexical relations, textual content and rhetorical relations) of external sources of world knowledge such as dictionaries, thesauri, lexical databases and encyclopedias. Consequently, this problem is studied using various design models based on these aspects. The approaches to computing semantic associations can be grouped into three fundamental models based on their underlying framework [19]. The *geometric model*, also known as *spatial model*, represents each concept in an n-dimensional space and semantic association is derived from the inverse of distances between concepts in this space. The *feature-based model*, also called *combinatorial model*, represents each concept as a set of features and uses the commonalities and differences between the feature sets for computing semantic associations between concepts. The *structural model*, often referred to as *network model* or *graph-based model*, represents concepts as nodes in huge structures such as semantic networks, taxonomies and graphs. The edges between nodes indicate relationships between concepts. Semantic associations are then computed using various techniques such as graph matching, activation spreads and random walks.

While the focus of existing approaches is on improving the overall performance of semantic association computation, very little effort has been put to understand, analyze and compare the underlying dynamics of semantic association computation. These dynamics include a number of factors such as the choice of background knowledge, utilization of certain aspects of a semantic resource, coping with a multitude of semantic relations and the choice of design models of semantic associations. There is a need for better interpretation and analysis of these factors influencing semantic association computation. A detailed analysis of these factors will provide an insight into understanding the underlying mechanism of semantic association computation. This will also help the later semantic association measures to build upon a balanced combination of these factors customized according to the situation and need of an application.

1.3 Research Goals

The goal of the thesis is to demonstrate that using Wikipedia for computing word association strength gives better estimates of humans' associations than the approaches based on other structured and unstructured knowledge sources. There are two key challenges to achieve this goal: first, to exploit various semantic association models based on different aspects of Wikipedia in developing new measures of semantic associations; and second, to evaluate these measures compared to human performance in a range of tasks. The focus of the thesis is on exploring the effectiveness of two aspects of Wikipedia: as an unstructured *text corpus*, and as a well-structured *knowledge source*. Following the literature in these two research directions, the thesis presents a performance investigation of the underlying design models used for computing semantic word associations. For this purpose, the thesis demonstrates an exploitation of various semantic elements of Wikipedia to augment a semantic association measure with deeper understanding of classical and non-classical relations between concepts. The thesis focuses on understanding the dynamics of computing semantic associations by investigating and analyzing various factors that control and affect the semantic bias of semantic association measures. Based on the findings, the thesis presents an exploitation of Wikipedia semantics for improving the performance of various Natural Language Processing (NLP) applications that critically rely on good estimates of semantic associations. In order to fulfill the overall goal and to find answers to related research questions, the thesis establishes a set of research objectives corresponding to the following research questions.

i) *How can various aspects of Wikipedia be effectively used for computing semantic associations of words?*

   The research community has analyzed various perspectives of the factors that play a key role in the performance of semantic association measures. Some approaches focused on statistical aspects of informal text corpora while others borrowed structural semantics from well-structured knowledge sources such as lexical dictionaries, knowledge bases and word thesauri. To answer the question above, the thesis explores two aspects of Wikipedia: as a knowledge source and as a corpus. This involves developing different approaches based on the two aspects of Wikipedia and understanding their focus on various lexical and semantic relations of concepts. Existing research shows that Wikipedia is an excellent resource for semantic mining and can be effectively used for various tasks based on natural language semantics interpretation [127]. Semantically rich elements of Wikipedia implicitly as well as explicitly encode different types of both classical and non-classical semantic relations.

Objective 1: *To leverage Wikipedia aspects in developing new approaches for computing semantic associations of words.*

The fundamental objective of the thesis is to harness semantic capabilities of Wikipedia for computing semantic associations of words. This research presents an exploitation of Wikipedia semantics in two directions. In the first research direction, the *knowledge source* aspect of Wikipedia is used to identify semantic connections of various concepts. Wikipedia is a well-structured semantically rich formal knowledge source with many structural elements that can be effectively used for computing semantic associations. Two such elements—the hyperlink structure and informative-content of Wikipedia articles—are used to develop new measures of semantic associations. In the second research direction, Wikipedia is viewed as an informal *text corpus* for scaffolding semantic associations of words with asymmetric word associations. Wikipedia, being a huge repository of dedicated information on all kind of topics, has an enormous coverage that is continuously growing. This feature makes it an excellent resource to be used as an unstructured text corpus for calculating probabilistic word associations. The analysis of Wikipedia as an unstructured text corpus will identify the semantic bias and limitations of this aspect of Wikipedia. This research is an explicit attempt at experimenting and investigating both aspects of Wikipedia on the task of computing semantic associations of words.

ii) *What is the impact of using various semantic association measures based on different design models?*

The classification of various design models presented in the Section 1.2 is fundamental in nature but is too simplistic to model all the complex relations that might exist between various concepts. This calls for developing new combinations of these models and investigating their impact on computing semantic associations. hence, the second research objective is to develop new semantic association measures based on different design models and their combinations for computing semantic association strengths. It is also important to compare and analyze strengths and weaknesses of the underlying design model of semantic association measures.

Objective 2: *To develop and compare new semantic association measures based on various design models.*

The second research objective is to develop new semantic association measures using combinations of various underlying design models. The thesis investigates the strengths and limitations of each underlying model in various scenarios using Wikipedia as the background knowledge source. For this purpose, new semantic association measures based on various design models and their combinations will be presented and compared on the task of semantic association computation. Based on the knowledge source aspect of Wikipedia, three new measures relying on combinatorial model, geometric model and hybrid model will be discussed. Using the corpus aspect of Wikipedia, a new model based on asymmetric probabilistic associations will be developed. An in-depth analysis is conducted to identify limitations of various semantic association measures founded on different design models.

iii) *How well does a Wikipedia-based semantic association measure perform in a natural language semantics-based application?*

There are two methods for evaluating the performance of a semantic association measure. One way is to estimate the correlation of automatically computed scores with manual judgments using publicly available benchmark datasets. This method of direct evaluation is straightforward but suffers from certain limitations: first, the size of the dataset used for this kind of evaluation is usually small and creating larger datasets are expensive and time consuming [5]. The other limiting factor is the nature of benchmark datasets which are not general enough to cope with all kinds of classical and non-classical relations. A large proportion of some datasets only contain semantically similar word pairs while other datasets focus mainly on the collocation-based word pairs. Hence, a semantic association measure performing extremely well on one dataset might not produce good results on other datasets due to bias in their underlying model towards certain type of semantic relations. In order to avoid these limitations, the performance of a semantic association measure is indirectly evaluated by employing it in solving a natural language processing task that critically relies on good estimates of semantic associations. Semantic association computation is the key component of many NLP tasks such as lexical chaining, information retrieval, text summarization, web page clustering and text mining. Hence, the performance and suitability of a semantic measure should be investigated in an independent application setup.

Objective 3: *Task-oriented evaluation of new Wikipedia-based semantic association measures.*

There are two main factors that profoundly influence the performance of a semantic association measure: first, the choice of the underlying knowledge source; and second, the underlying model of computing semantic associations. These two factors play a decisive role in determining the semantic bias of a semantic association measure. The thesis presents various approaches for developing new Wikipedia-based semantic measures and exploiting them in two task-oriented applications chosen according to the underlying design model and

[5] This problem is somewhat overcome only recently by using MTURK workers [167, 71].

the nature of the knowledge source. The first application employs the Wikipedia-based measure for solving word choice problems, which is a well known application for the task-oriented evaluation of semantic measures. The second application involves using the Wikipedia-based semantic association measures for detecting causal connections between textual units using COPA evaluation [66]. These applications are useful for evaluating the effectiveness of semantic association measures on different semantic levels.

1.4 Thesis Outline

The thesis presents a comprehensive study aimed at exploiting various semantically rich elements of Wikipedia for computing semantic associations of words. Thesis layout and coherence of the main chapters is shown in figure 3.

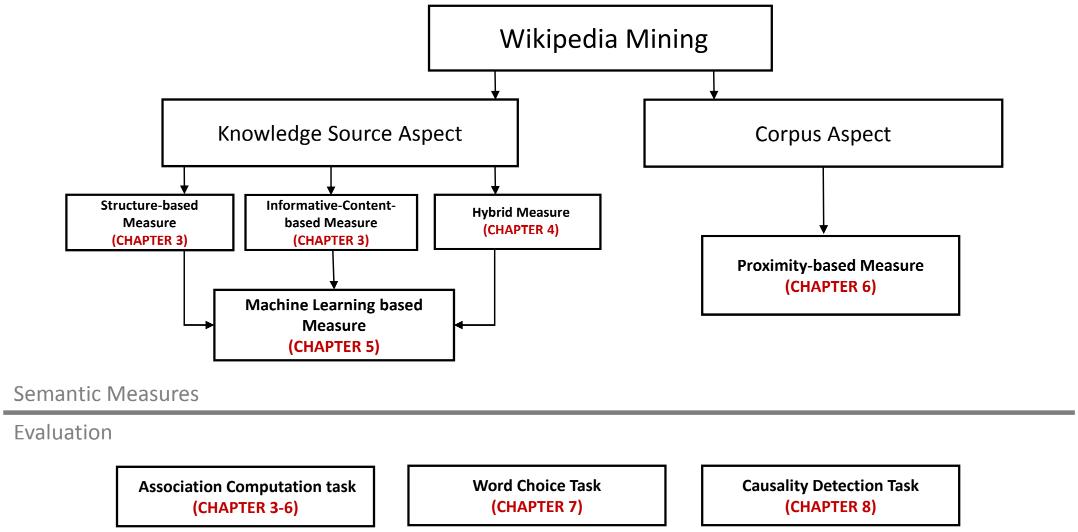

Wikipedia Mining
Knowledge Source Aspect
Corpus Aspect
Structure-based Measure (CHAPTER 3)
Informative-Content-based Measure (CHAPTER 3)
Hybrid Measure (CHAPTER 4)
Proximity-based Measure (CHAPTER 6)
Machine Learning based Measure (CHAPTER 5)
Semantic Measures
Evaluation
Association Computation task (CHAPTER 3-6)
Word Choice Task (CHAPTER 7)
Causality Detection Task (CHAPTER 8)

Figure 1.2: Thesis layout and connectivity of main chapters.

Chapter 2 outlines the existing manual and automated methods of estimating semantic associations. It details the evaluation metrics commonly used for measuring and comparing the performance of semantic association measures. It also details the pervasiveness of semantic associations in various applications belonging to a multitude of fields and identifies the limitations and contributing factors of semantic approaches in various research streams. It then categorizes the existing research on computing semantic associations of words and surveys various techniques for computing word associations in each category.

Chapter 3 utilizes two elements of Wikipedia—the hyperlink structure and informative-content of Wikipedia articles—in computing three new semantic association measures. The strengths and limitations of each measure are identified by performing both domain specific and domain independent evaluations.

Chapter 4 presents a new method of semantic association computation that exploits both structural and informative-content based aspects of Wikipedia. Inspired by humans' asymmetric associations, the new method is founded on the idea of directional contexts borrowed from asymmetric free word associations. The chapter identifies the fundamental properties of free word associations using a discrete association task and develops a new hybrid measure that uses these properties as the basic assumption of its underlying design model. The measure is evaluated on both symmetric and asymmetric association computation task. This follows a discussion on the difference between the focus of humans' semantic associations and knowledge source based automatic semantic associations.

The empirical analysis and discussion in the first four chapters leads to the argument that association computation is a complex multi-dimensional problem that can be effectively handled by considering the complementary features based on various semantic aspects of one or more knowledge sources. To support this assertion, Chapter 5 presents a hybrid model that combines various design models of semantic computation using Wikipedia. It involves developing both supervised and unsupervised versions of the hybrid model and demonstrates that the hybrid models based on multiple features outperform the ones based on individual features.

Chapter 6 extends the idea of free word associations from document level to corpus level by developing a new probabilistic associations based measure of semantic associations. The chapter identifies the limitations of previously presented semantic association measures and overcomes them by computing the associative probabilities of words at corpus level based on the proximity assumption. It also provides a detailed comparison of semantic association measures based on probabilistic asymmetric associations with other symmetric measures.

Chapter 7 presents a task-oriented evaluation of a Wikipedia-based measure of semantic associations—Asymmetry-based Probabilistic Relatedness Measure (APRM). The measure is employed in two subtasks of the word choice problem task: relatedness-based word choice problems and synonymy-based word choice problems. This involves a detailed discussion of the strengths and limitations of APRM measure in each scenario.

Chapter 8 presents another task-oriented evaluation of the APRM measure. For this purpose, the choice of plausible alternatives task is used. Under the same experimental setup, the performance of APRM is compared with other Wikipedia-based association measures. This follows an investigation of various factors affecting the performance of causality detection systems.

Chapter 9 summarizes the key contributions and concludes the thesis by suggesting some generalization of the findings. This is followed by a discussion on the future research directions founded on the contributions of the thesis.

This follows a list of appendices consisting of a list of publications based on the research work presented in the thesis, a glossary of the terms used through out the thesis and process flow diagrams.

Chapter 2

Background

The work described in this chapter introduces the fundamentals of using knowledge sources for computing semantic associations and presents a detailed survey of existing semantic association measures, categorized according to their underlying sources of background knowledge. The first part of the chapter introduces the fundamental of semantic associations, different terms used for referring to semantic associations and different kinds of background knowledge sources. Later sections of the chapter present a classification of the existing approaches to semantic computation into two broad categories and their grouping into different classes according to the underlying design model of semantic associations. The strengths and limitations of various approaches belonging to each research stream are discussed for comparative analysis and various performance indicators of semantic similarity are also identified.

2.1 Fundamentals of Semantic Associations

A semantic connection between any two concepts indicates the presence of a semantic relation between them. Consider for instance, the word pairs `(car,automobile)` and `(bird,kiwi)`. The words in these two pairs are connected through classical taxonomic relations such as synonymy (car and automobile are synonyms) and hyponymy (kiwi is a bird). Such relations are called *classical* relations. However, many words share more complex relations which can not be easily defined and mapped through lexical relations, for instance (Freud,Psychology), (lemon,sour) and (car,carbon emission). The relations of such word pairs cannot be easily detected with a shallow semantic or lexical analysis and are called *non-classical* relations [144]. It is difficult to tell the exact number of semantic relations as the human language continues to evolve. However, Cassidy [28] identified 400 types of semantic relations in the semantic network, FACTOTUM, which was based on the 1911 edition of Roget's thesaurus [96].

Classical relations exist in linguistic and lexical resources such as monolingual and bi-lingual machine readable dictionaries and thesauri. Morris and Hearst [144] pointed out that linguistic resources such as WordNet cover only the *classical* relations such as hypernymy (vehicle, car), hyponymy (bird, kiwi), troponymy (move, run), antonymy (rise, fall), meronymy (car, steering) and synonymy (tool, implement)[1]. Such word pairs are characterized by an overlap of some defining features and fall in the same syntactic class. For instance, the words car and truck share a set of features related to their structure (wheels,steering,brakes ,engine) as well as function (driving) and both words belong to the same class (vehicle). They used the term *classical* relations to refer to those relations which share properties of classical categories.

2.1.1 Terms and Definitions

There are three terms widely used in the literature of natural language semantics to refer to semantic associations: semantic similarity, semantic relatedness and semantic distance.

Semantic similarity consists of semantic connections between two con-

[1] Please see glossary for more detail on each of these terms

cepts that have similar nature, composition or attributes. Examples of semantic similarity are synonymy, hypernymy and hyponymy relations. Sanchez [183] defined semantic similarity as a taxonomic proximity of two words. For instance, `car` and `truck` are semantically similar because both are `vehicles` and share a set of similar features. According to Budanitsky [23], semantic similarity is a special aspect of semantic relatedness. Semantic similarity covers most of the previously discussed *classical* relations.

Semantic relatedness is closely connected with semantic similarity but is a more general term involving a multitude of *classical* and *non-classical* relations [171]. It not only covers semantic similarity but also relates those concepts which apparently do not have similar nature, composition or attributes but are strongly associated. For instance `vaccine` and `immunity` are not semantically similar by their very nature but have a strong `Cause -Effect` relationship. Semantic relatedness covers various types of lexical relations such as antonyms `(day,night)`,meronyms `(wheel,auto-mobile)` holonymy `(tree,bark)` and so on. It also maps functional associations, relating those words which apparently do not have lexical relation such as `(manager,bank)` as well as collocational semantics–those words which make a new concept when they occur together such as `credit card`. The higher the semantic similarity or relatedness the stronger the semantic association between the concepts. Resnik [171] elaborated the difference between semantic similarity and semantic relatedness by giving an example of two word pairs. In case of word pair `(car,gasoline)`, though both words by their intrinsic nature and composition are not similar, still they are closely related. According to him, the words in another word pair `(car,bicycle)` are considered more similar to each other rather than related due to their categorical nature.

Semantic distance is used in a similar but opposite context. It is the inverse of both semantic similarity and relatedness. It indicates how distant two words are, in terms of their meanings. The more related or similar the two words, the smaller the semantic distance between them i.e. the increase in semantic distance correlates with the decrease in semantic similarity or relatedness. In applications where the measure of similarity or relatedness is important, the semantic distance is usually converted into a similarity score using a transformation function such as those introduced by Gracia and Mena [68] and Harispe *et al.* [73].

Semantic distance is the most ambiguous of all three terms that are used for referring to semantic associations. It is generally used while talking about similarity as well as relatedness but this usage is not always correct particularly in the case of antonyms, where two words are quite distant in terms of similarity but still are strongly connected semantically. Examples of the three terms of semantic associations are shown in Table 2.1.

Table 2.1: Correlation of the association strength of three terms of semantic associations.

Word1	Word2	Semantic Similarity	Semantic Relatedness	Semantic Distance
Day	Sunday	High	High	Low
Day	Sunlight	Low	High	Low
Day	Night	Low	High	High
Day	Cartoon	Low	Low	High

2.2 Applications of Semantic Measures

Semantic measures are used as core components in a growing number of applications that critically rely on good estimates of semantic associations. The scope of semantic measures is multidisciplinary, ranging from computational linguistics to artificial intelligence and from cognitive psychology to information retrieval. A multitude of artificial intelligence applications are essentially founded on semantic association measures. The following list highlights some of these applications and how they use semantic association measures:

Automatic keyphrase extraction is the task of systematically extracting the topic-representative keywords and phrases from a text document with minimal or no human intervention [205]. Keyphrase extraction approaches fundamentally rely on the concept of contextual relatedness as candidate words can be potential keywords if they are related to the main topic of the document.

Word sense disambiguation task requires the automatic identification of the correct sense of a word in a given context [147]. Word sense disambiguation applications largely rely on the understanding of the context in which an ambiguous word occurs in a piece of natural language text. Semantic measures play a critical role in such applications in finding out the relation of each sense of a target ambiguous word with its surrounding context words to pick the correct sense.

Paraphrasing involves identification, generation and extraction of phrases, sentences and longer textual units that represent almost the same information [6]. For instance, two text fragments *"X made Y"* and *"Y is the product of X"* are paraphrases of each other. These paraphrases are identified by their property of being semantically close to a given piece of text and having the same context around them. The input to a *paraphrasing recognition application* is a pair of text fragments and output is a judgment score indicating whether the members of the input pair are paraphrases or not. The aim of paraphrase recognition application is to achieve maximum agreement to the human scores on the task. Among various techniques for paraphrase recognition, co-occurrence based vector space models of association and lexical similarity across a dictionary or thesaurus are commonly used approaches [51].

Automatic text summarization is the task of producing an informative summary from a single document or multiple documents [79, 11]. The summary is a collection of coherent sentences that are either extracted from the document or generated using machine learning techniques. A query based summarization task first identifies the text segments in a document that are semantically relevant to the query and then generates a summary based on the most relevant text segments. A critical factor in query-based summarization is to choose those sentences of the text that are semantically closest to the user generated query. A variant of text summarization is query based summarization–that produces a summery consisting of most salient points that suit a user's information need (expressed in the form of a query) [201]. Graph-based semantic approaches and probabilistic models are generally used for solving this task [79, 11, 63, 60].

Machine translation is the task of automatically translating a piece of text from one natural language to another language. A machine translation system takes as input a text fragment from a source language and translates it into a text fragment in the target language with minimal or no loss of generality [103]. For this purpose, various design models are used for substituting words or phrases from source language to the target language. This substitution requires interpretation of complex linguistics and semantic context of words in two different languages [118].

Contextual Spelling error detection and correction is the task of identifying the words in the documents that are spelled incorrectly, determining the candidates for misspelled words and replacing them with the candidate words that are semantically relevant to the context of the text [83]. Generally, spell checkers cannot flag such contextually incorrect words because they consider words in isolation. For instance, the sentence using a word `bar` instead of the word `car` is logically incorrect but the spell checker can not pick this type of errors. To detect such errors, surrounding context is of critical importance. Finding semantic relatedness of a word to its neighboring words leads to spell error identifications. To fix such errors, both semantic and probabilistic information are used in literature [92].

Automatic query expansion is the process of reformulating a query to improve the performance of information retrieval process [27]. Performance of information retrieval systems is largely affected by the poorly- formulated queries representing a user's needs. To cope with this issue, user generated query is automatically augmented with more semantically relevant words in the same context. The process of query expansion involves identifying and evaluating query words, finding out other semantically relevant words in the query area and adding those words to the existing query to retrieve more relevant documents. A special case of query expansion is *interactive query refinement*, in which the user is given the choice to select the appropriate words for reformulation of queries.

Opinion mining, also known as *sentiment analysis*, is the task of automatically determining the attitude (opinion, appraisal, emotions) of the people regarding entities and their attributes [116]. With the massive growth of social media, such as web blogs, reviews, forum discussion and social networking on the web, opinion mining has become a critical area. A fundamental aim of opinion mining approaches is to determine the polarity of a text fragment. Words and sentences are assigned negative, positive or neutral polarity according to the writer's mood. Advance levels of sentiment analysis involves automatically estimating star ratings. A specialized task is the aspect-based sentiment analysis that involves analyzing opinions regarding certain attributes of an entity. Semantic measures play an important role in identifying the important concepts in an unstructured piece of text and detecting their relations to the aspect of an entity.

Web search results clustering is the process of organizing search results by topics into a small number of coherent groups or clusters [26]. In response to a user query, a ranked result set is returned by a search engine. This list consists of web pages on different subtopics or meanings of the query. The user keeps visiting the links one by one until the required information is found. To speed up this process of information access and to satisfy the user's needs in a better way, an automatic clustering approach provides a clustered view of the results. It takes as input a set of web documents retrieved by a search engine and outputs a set of labeled clusters of these documents. Performance of clustering approaches critically rely on the precise estimation of the semantic similarity or semantic distance between a pair of web results [195].

Thus, semantic measures play a major role in the good performance of many applications. Most of these applications do not require to determine the nature or exact type of semantic relation between two concepts but only the strength of their semantic association.

2.3 Can Humans Estimate Semantic Associations?

Humans are good at distinguishing between related and unrelated concepts. For example, it is easy for humans to tell that the word `tomato` is more related to the word `Sauce` than to the word `Bottle`. But, can humans correctly estimate the strength of association of two words on a particular scale (for instance similar words get a score of 1, dissimilar words get a score of 0 and associated words get a score somewhere between 0 and 1)? How strongly do they agree or disagree to each other on estimates of word associations? Will agreement of the same group of people vary over different sets of words? Various attempts have been made to explore the answers to such questions by creating datasets and measuring the agreement of human raters on the datasets.

One of the earliest works that tried to find answers to these questions was a quantitative experiment conducted by Rubenstein and Goodenough [178] in 1965. In their experiment, 51 human judges were given 65 noun word pairs, each pair written on a separate card and were asked to arrange them in the decreasing order of similarity. Then, an averaged continuous similarity value on the scale of [0-4] was assigned to each card based on all judgments. The same experiment was repeated after two weeks with the same human subjects and the new human judgments had a Pearson's correlation γ of 0.85 with the old judgments. This dataset, known as R&G dataset [178], is a collection of 65 English noun words pairs. In this dataset, the word pairs that were given high scores consisted of semantically similar words. However, the dataset had no word pairs that were semantically related but dissimilar.

In a similar study, Miller and Charles [136] conducted an experiment on general English based dataset consisting of 30 noun term pairs taken from the R&G dataset. In this experiment, 38 human judges were asked to rate the word pairs in this dataset according to similarity and the judgments of all participants for each term pair were averaged to get a similarity score on a scale of [0-4], where 0 means unrelated and 4 means synonyms. These ratings were found to have a high correlation ($\gamma = 0.97$) with the mean ratings of R&G dataset. This dataset is generally referred to as M&C dataset. Resnik [172] replicated the same experiment on this dataset and reported an inter-annotator agreement of 0.90.

In order to analyze human estimates on verb similarity, Resnik and Diab [173] conducted a similar experiment in which they split 10 human subjects into *context* and *no-context* groups. The subjects in the *context* group were given 48 verb pairs with example sentences illustrating the intended sense of each verb in the given pair and were explicitly asked to rate them according to semantic similarity rather than relatedness on a scale of [0-5]. The same experiment with same verb pairs without example sentences was repeated for the *no-context* group. They found that the inter-rater agreement[2] of *context* group was 0.79 whereas that of *no-context* group was 0.76. Yang and Powers [223] conducted an experiment in which they collected the human ratings on a larger verb similarity dataset consisting of 130 verb pairs. Both these datasets are particularly useful for assessing the ability of a semantic relatedness measure to estimate the verb relatedness. This dataset is known as YP-130 dataset.

Finkelstein *et al.* [53], created a larger dataset of 353 word pairs (including the 30 noun pairs from M&C dataset with 0.95 inter-rater agreement with Resnik) scored by 13 to 16 human judges on a scale of [0-10]. This dataset, referred to as WordSimilarity-353 (WS-353), consists of word pairs that are quite diverse in nature ranging from semantically similar word pairs such as `(King,Queen)` to semantically dissimilar but related word pairs such as `(Maradona,Football).` It also includes proper nouns `(Michael Jackson),` associative term pairs `(Wednesday News)` and abbreviations `(FBI)` and `(OPEC),` making it a challenging dataset for evaluating the performance of semantic measures.

The annotation guidelines of WS-353 dataset did not distinguish between semantic similarity and relatedness, but most existing semantic measures are semantically biased towards either similarity computation or relatedness computation. Hence, to analyze the impact of similarity and relatedness on the system performance, Agirre *et al.* [1] experimented with two subsets of WS-353 dataset focused on similarity and relatedness. To create those subsets, two human judges grouped all pairs in WS-353 dataset into three subsets (similarity pairs, relatedness pairs and unrelated pairs) and then created two new datasets: WS-similarity (the union of similarity pairs and unrelated pairs) and WS-relatedness (the union of related pairs and unrelated pairs). The inter-rater agreement for their experiment

[2] please see glossary for details

was 0.80 with a Kappa score[3] of 0.77.

Another effort to provide additional data for evaluating term relatedness was done by Randinsky *et al.* [167], who created a new dataset, MTURK-287, consisting of 287 word pairs. Each word pair was rated by 23 MTURK workers on a scale of [0-5].

In an attempt to create a larger relatedness-based dataset (MTURK-771), 20 MTURK workers assigned relatedness score to 771 English word pairs on a scale of [0-5] [71]. The inter-rater agreement was found to be 0.89 with extremely small variance. This is by far the largest available dataset of word relatedness. Similar efforts were made in languages other than English such as German datasets [69, 229].

Details of benchmark datasets used for evaluation of semantic similarity and relatedness computation approaches are summarized in Table 2.2. The inter-annotator agreement indicates that the humans are good at estimating the noun-noun relatedness but find it harder to agree when they are presented with a combination of part of speech word pairs. This table also suggests that predicting semantic similarity is easier than predicting semantic relatedness. It is also worth mentioning that the annotators were presented with words having multiple senses (ambiguous or polysemous words), hence the participants had to estimate the similarity or relatedness of the most closely related senses. For example, in the word pair `(crane,implement)`, both words have multiple senses, however humans assigned an overall high relatedness score to this pair based on the closeness of their senses `crane machine` and `tool`. Overall, these experiments achieved the goal of proving that humans can actually estimate semantic associations of words. Moreover, the datasets used in these experiments became the benchmarks for evaluating the performance of semantic measures by comparing the automatically computed scores with manual human judgments. This method will be used for evaluating the

[3] A measure for assessing the degree to which two or more raters agree on assigning data to different categories

performance of various new measures in the later chapters of the thesis.

Table 2.2: Statistics of Benchmark datasets used for evaluation of semantic association measures.

Dataset	Year	Language	No. of Pairs	POS	No. of Subjects	Scale	Annotator Agreement
R&G	1965	English	65	N	51	[0-4]	0.85
M&C	1991	English	30	N	38	[0-4]	0.90
WS-353	2002	English	353	N,V,A	13-16	[0-10]	0.87
YP-130	2005	English	130	V	6	[0-4]	0.76 and 0.79
Gurevych	2005	German	65	N	24	[0-4]	0.81
Zesch *et al.*	2006	German	350	N,V,A	8	[0-4]	0.69
Zesch and Gurevych	2006	German	222	N,V,A	21	[0-4]	0.49
WS-Similarity	2009	English	203	N,V,A	2	[0-10]	0.80
WS-Relatedness	2009	English	252	N,V,A	2	[0-10]	0.80
MTURK-287	2011	English	287	N,V,A	23	[0-5]	-
MTURK-771	2012	English	771	N,V,A	20	[0-5]	0.89

2.4 Evaluating Semantic Measures

According to Budanitsky and Hirst [23, 227], there are three methods for evaluating the performance of semantic measures: *comparison with human judgments*, where manual human judgments are used as gold standard[4] for evaluation; *task-oriented evaluation*, where the measure is applied in a real world application and tested indirectly; and *Mathematical analysis*, where formal properties of relatedness measure are assessed. The first two methods are most widely used evaluation method. The mathematical analysis is seldom used.

[4] a manually annotated solution set.

2.4.1 Correlation with Manual Judgments

To compare automatically computed scores with human judgments, a correlation value between the automatic and manual scores is computed. A correlation value indicates two things: the strength of association between two variables; and the direction of the relation (increasing / decreasing or negative / positive). The strength of the correlation is the magnitude of the association between two variables. Depending on the discipline, this magnitude $|c|$ is interpreted or assessed by the the following general guidelines [37]:

- $0.1 < |c| < 0.3$.... weak correlation
- $0.3 < |c| < 0.5$.... moderate correlation
- $0.5 < |c|$.......... strong correlation

The direction of the association between two variables is interpreted by sign of the correlation coefficient: -1 means perfectly negative relationship; 0 indicates no relationship at all; and +1 shows perfect positive relation between the variables. Two types of correlation metrics are commonly used in the literature. Details of both metrics are as follows:

- **Pearson Product-momentum correlation coefficient** , also known as *Pearson correlation* (denoted by r), is a measure of strength of linear association between two variables. Pearson's correlation indicates the deviation of data points from the best fit line (a line that fits all the data points of two variables). The Pearson correlation assumes a value between -1 to +1, depending on whether the association of two variables is negative and positive. The stronger the association of two variables, the higher the correlation value. A value of $r = 0$ means that there is no association of two variables. A value of $r > 0$ indicates that the association of two variables is positive (the value of one variable increases if the value of another variable is increased).

A value $r < 0$ means a negative association (increasing one variable decreases the other variable). Pearson's correlation is computed as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.1}$$

where n is the sample size and X_i, Y_i are the scores of the two variables being compared.

- **Spearman's rank order correlation coefficient** , also called *Spearman's rho* (denoted by ρ or r_s), is a non-parametric measure of association strength of two variables. It measures how well the association strength of two variables can be described by using a monotonic function. It is used when the distribution of data makes Pearson correlation coefficient misleading or undesirable [77]. Fundamentally, Spearman's correlation coefficient is a special case of Pearson's correlation, in which data values are converted to ranks before computing the correlation. It does not require the assumption that the relation of two variables should be linear nor does it requires the data of those variables to be on interval scale. Spearman's correlation is also less sensitive to outliers than Pearson's correlation.

  Spearman's ρ is computed by two methods, depending on whether there are rank ties in the data or not. If there are no rank ties in the data then the Spearman's correlation is computed by the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2.2}$$

  where n is the sample size and d is the difference of ranks of two variables. In case of ties, the Pearson's formula is used on variable ranks rather than their actual values, given as below:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3}$$

where n is the sample size and raw scores X_i, Y_i are converted to ranks x_i, y_i.

Existing work on computing semantic similarity often used Pearson's correlation as the evaluation metric. However, there are certain limitations of using Pearson's correlation:

- Pearson's correlation is highly sensitive to outliers. Outliers have a large impact on the best fit line and can lead to very different conclusions regarding the data. This can be visualized by the scatter plot in Figure 2.1, which shows the effect of an outlier on the Pearson's correlation (r) computation.
- Pearson's correlation measures the strength of linear association of two variables. Hence, the results of Pearson's correlation of two variables could be misleading when their scores are non-linearly correlated.
- A basic assumption of Pearson's correlation computation is that the data of two variables should be normally distributed. When values of the variables are not normally distributed, Pearson's correlation is a poor choice.

In contrast to these limitations, Spearman's correlation is considered more robust than Pearson's correlation. Spearman's correlation does not restrict the variable values to be on an interval scale, rather the data values could be on ordinal scale as well. According to Agirre *et al.* [1], Pearson's correlation is less informative and suffers much when the scores of two variables are non-linearly correlated so a transformation function is usually used in such cases to map the system's output to new values that correlate well with human judgments, whereas Spearman's correlation is independent of such data-dependent transformations and does not require any data normalizations. In comparison with Pearson's correlation, Spear-

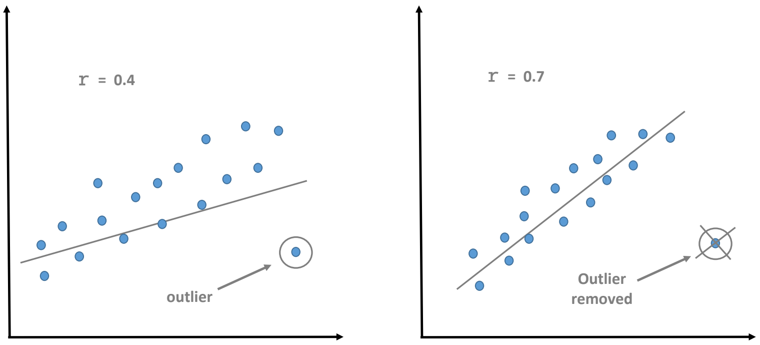


Figure 2.1: Effect of outlier on the Pearson's correlation computation

man's correlation is much less sensitive to outliers [80]. Additionally, by considering the ranks of the two variables in comparison, Spearman's correlation disregards any assumptions about their data distributions. However, Spearman's correlation coefficient is also not without limitations. In the case of many tied ranks, it tends to give higher correlation values than Pearson's correlation. For applications that rely on thresholding relatedness scores, a measure that produces high Spearman's correlation but with very small differences in the actual relatedness scores would be of little use [228]. However, this problem could be addressed by scaling up the relatedness scores.

Under similar conditions, both Spearman's and Pearson's correlations may produce very different results, hence cannot be directly compared. Care must be taken in comparing and interpreting the results. For comparative analysis of the methods presented in this thesis, both correlation metrics are used to compare the results with those approaches who have reported their results using either or both correlation variables.

2.4.2 Task-Oriented Evaluation

The small size of manually annotated datasets limits the reliability of the direct evaluation of a semantic computation approach. To support the performance claims, such approaches can be evaluated indirectly by investigating their usefulness in various applications. Existing approaches to compute semantic relatedness can be broadly categorized into two classes: approaches relying on the well structured *knowledge sources*; and approaches relying on the unstructured *text corpora*. Table 2.3 shows a list of applications based on natural language semantics and the examples of various approaches categorized according to their underlying knowledge sources used in these applications.

Table 2.3: Examples of approaches using knowledge-sources and text corpora in various applications.

Application	**Using Knowledge Sources**
Keyphrase Extraction	Zhang *et al.* [232]
Word Sense Disambiguation	Sidharth *et al.* [158], Navigli [147]
Coreference Resolution	Ng [150], Lee *et al.* [107]
Named Entity Recognition	Richman and Schone [174]
Document Clustering	Huang *et al.* [89], Hu *et al.* [88]
Query Expansion	Collins *et al.* [38], Carpineto and Romano [27]
Lexical Chaining	Barzilay and Elhadad [11], Erekhinskaya and Modovan [48]
Application	**Using Text Corpora**
Automatic Thesaurus Generation	Brussee and Wartena [21], Zohar *et al.* [235]
Text Summarization	Henning [79]
Paraphrasing	Bolshakov and Gelbukh [17], Androutsopoulos and Prodromos [6]
Spelling Correction	Islam and Inkpen [92], Hirst and Budanitsky [83]
Machine Translation	Marton *et al.* [122], Koehn [103]
Language Modeling	Zhai [231]
Word Sense Separation	Schutze [186], Levin *et al.* [112]
Temporal Information Retrieval	Alonso *et al.* [5], Strötgen *et al.* [196]
Topic Identification	Clifton *et al.* [36]

Semantic measures play a major role in the good performance of many applications. Most of these applications do not require determining the nature or exact type of the semantic relation between two concepts but only the strength of their semantic association. While some of the semantic measures between words can be directly extended to measure the relatedness of larger units of text such as phrases, sentences and even documents, other measures may need minor or major adaptations to be used in a specific scenario. Generally, every semantic measure has its own advantages and disadvantages, hence carefully choosing an appropriate semantic measure for a particular application leads to remarkable improvements in its performance.

A recent trend in the area of semantic association computation is to combine the features of both types of resources— knowledge sources and text corpora. Such approaches follow a hybrid approach and are shown to perform better than the approaches using single knowledge sources [1, 71].

2.5 Background Knowledge Sources

Semantic associations are based on the background information that implicitly or explicitly supports the relationship of the concepts. Computing a measure of semantic association requires extracting information from some source of background knowledge, also referred to as *semantic proxy* [73]. Various kinds of knowledge sources are grouped into two broad classes: *informal knowledge source*, which have implicit semantic connections; and *formal knowledge sources*, which encode the semantic connections explicitly. The informal semantic proxies or knowledge sources include the semantic connections in the form of distributional context and co-occurrence patterns in the unstructured text, whereas the formal background knowledge sources are used to draw the semantic connections that are often encoded in the form of semantic graphs or taxonomies in the structured and semi-structures knowledge sources. The remainder of this section discusses the attributes of both categories of knowledge sources frequently used in the task of semantic association computation.

2.5.1 Informal Knowledge Sources

Knowledge sources that support the knowledge-lean approaches of semantic association computation are often referred to as the *unstructured* or *informal* knowledge sources (Informal-KS). The type of knowledge sources belonging to this category do not organize words or concepts in a structured way, hence the lexical and semantic connections among words are not explicitly defined. Unlike structured knowledge sources, the unstructured knowledge sources do not provide sense tagging of words. The approaches using unstructured knowledge sources are generally based on distribution of words in a huge collection of documents [54, 74]. Consequently, the background information is collected at the word level rather than at the concepts level.

One class of Informal knowledge sources consists of *text corpora*. A text corpus is constructed by compiling a large number of electronically processed and stored text documents. Each text corpus is a document collection on topics belonging to a single domain or combined from multiple domains. Text Corpora are generally big in size; some corpora have millions of documents. These corpora are used for statistical analysis, validating linguistic rules and building distributional models of word associations [36]. This also involves modeling the word usage patterns.

In order to effectively analyze the content of a text corpus, various kinds of preprocessing and annotations are performed before using it. Examples of annotations include part-of-speech tagging (POS-tagging), where the information about POS class (verb, noun, adjective etc) of each word is added to the corpus content in the form of POS tags, and lemmatization or stemming[5], where the lemma or base form of each word is included. Advanced levels of structured analysis involve parsing the content of a text corpus. Such parsed corpora are often called *Treebanks* [121]. However, these corpora are usually smaller in size. Other forms of cor-

[5] Please see glossary for detail

pora analysis include annotations for syntactic and dependency parsing and pragmatic and morphological analysis. Corpora are used as the main source of knowledge in certain fields such as computational linguistics, machine translation, corpus linguistics, cognitive psychology and natural language processing [10].

Another type of informal knowledge sources is the *web* [101]. The web is an excellent knowledge source in terms of coverage. It is a huge collection of documents that are used in an informal way to compute distributional models or usage patterns. Typically, there are two ways in which the web is used as an informal text corpora: first, the snapshot of the web is used as a collection of documents; and second, using the text snippets (a small summary of retrieved web page) retrieved as a result of generating a query. Consequently, some approaches use the web snapshot as an unstructured collection of documents, while others use search results to compute word associations [92, 180]. The approaches using the web as a text corpora take benefit from its shear size due to its continuously growing huge coverage of knowledge whereas the approaches based on text snippets are computationally inexpensive as they do not require preprocessing of the web.

The approaches using informal knowledge sources do not rely on explicit definitions and identifications. Hence, being knowledge-source independent, these approaches are easier to adapt to different applications and are easily scalable. Table 2.4 lists the commonly used informal knowledge sources[6].

2.5.2 Formal Knowledge Sources

Formal Knowledge sources (Formal-KS), also called *structured* knowledge sources, explicitly organize the information in structural elements. The

[6] Further details can be found on `http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml` and `https://catalog.ldc.upenn.edu/byproject`.

Table 2.4: Some well-known text corpora used by corpus-based approaches.

Knowledge Source	**Word Size***	**Domain**
Brown Corpus	3M	American English Prose
British National Corpus	100M	British Spoken & Written English
Project Gutenberg	4.2M	English E-books
Penn Tree Bank	4.5M	American English
Wall Street Journal	30M	English News
Web 5-grams	1T	English word N-grams**
American National Corpus	20M	American Text
Weblog Personal Stories	10M	Weblog stories
English Gigaword	4B	Collection of Multiple Corpora

* M - Million, B - Billion, T - Trillion

** N-grams are the text phrases up to a certain size

most commonly used formal knowledge sources are dictionaries, thesauri, lexical databases, encyclopedias and ontologies. The fundamental differences between all these formal knowledge sources lie in the type and amount of knowledge that they contain and the way this knowledge is organized in a specific structure. Earlier approaches of semantic association computation used the first two types of knowledge sources–dictionaries and thesauri. However, with the overwhelming growth of web content and the endeavors to understand the semantic needs of web users and to provide them faster access to desired content, different kinds of knowledge have been organized in the form of online dictionaries, thesauri, encyclopedias and ontologies. Most of these knowledge sources are general purpose and some are multilingual (such as WordNet, Wikipedia and Wiktionary). However, to meet the needs of users concerned with specific domains, domain-specific ontologies have been constructed. This section highlights the structural attributes of various well known formal knowledge sources.

Dictionaries are one of the earliest sources of knowledge that were used in computational linguistics. A dictionary is an alphabetical list of words, where each word is accompanied by a definition, POS classes, derivations and sometime examples of word usage. It can be mono-lingual (with content in one language) or multi-lingual (to support translation between multiple languages). When a dictionary is organized in a hierarchy (sub type-super type), it is called a *Taxonomy*. An example of a dictionary is Wiktionary, which is a collaboratively constructed, freely available, online dictionary. It is multilingual (available in 172 languages) and consists of approximately 3.5 million entries [230]. In comparison with a standard dictionary such as Oxford English Dictionary[7], the Wiktionary offers a wide range of semantic and lexical relations, hence is often called as the *knowledge base* or the *thesaurus* [230]. Wiktionary has many features in common with WordNet. It has an article page for every word which lists various word classes. Each word class corresponds to a concept. Each concept in turn is accompanied by a short definition (like a WordNet gloss) followed by usage examples. Following WordNet, Wiktionary also defines lexical semantic relations such as part-of-speech, pronunciation, synonyms, collocations, examples, sample quotations, usage and translation into other languages [230]. It also includes words from all parts of speech but lacks the factual and encyclopedic knowledge as contained by Wikipedia. Unlike Wikipedia, where the links are not explicitly annotated, Wiktionary explicitly encodes the lexical semantic relations in the structure. However, like Wikipedia, Wiktionary also has a massive linking of various types: intra-linked entries refer to other entires that exist within the same language version of Wiktionary; inter-linked entries correspond to entries in other languages of Wiktionary; and external links connect the Wiktionary entries to external resources such as Wikipedia, WordNet and other web-based resources.

A *thesaurus* is a way of organizing the words into groups centralized

[7] `http://www.oed.com/`

by the idea which they express [176]. It is a catalog of semantically similar words organized into verbs, nouns, adjectives, interjections and adverbs. It is used as a source or repository of knowledge. Unlike a dictionary, which is an alphabetic list of words, a thesaurus is structured around concepts. While dictionaries are used to find meaning or definitions of words, thesauri are used to find the best word in a particular context [96]. These words are synonyms or nearly similar in meaning to a given word. The first English language thesaurus was created in 1905 and is known as Roget's Thesaurus [176]. It organizes the ideas expressed by a language into six classes: *abstract relations, material world, space, intellect, sentiments, volitions and moral powers* [96]. These classes are further divided into sections which include almost 1000 headings. This can be viewed as a huge tree having many branches. This organization of concepts remained amazingly unaffected by the changes over time as it easily accommodated new concepts that emerged with the passage of time, without much changes in the overall structure.

Lexical databases are lexical resources that provide computerized access to their content. In these databases, lexical information is represented by synonyms of a target word, its relation with other words and its lexical categorization. This information is usually organized in the form of a taxonomy. The most popular example of online lexical databases is WordNet[8], which is inspired by psycholinguistic theories of human lexical memory. It is an electronic dictionary as well as lexical database. The fundamental idea was to provide support for searching dictionaries conceptually rather than just alphabetically [134]. It organizes verbs, nouns and adjectives into sets of synonyms called *Synsets*. Each synset represents the underlying lexical concepts. Additionally, a synset comprises of a gloss (a brief definition of a term) and various usage of the synset members in one or more short sentences [8]. Word forms having several contexts are represented by several distinct synsets. Thus, WordNet has unique form-

[8] Freely available at `http://wordnet.princeton.edu/wordnet/`.

meaning pairs for each word [213]. In WordNet, synsets are interlinked to other synsets through lexical relations and conceptual-semantics, resulting into a network of concepts which can be navigated with a browser [134]. WordNet is not just an online thesaurus or an ordinary dictionary. It offers far more lexical semantics than both of those.

Ontologies organize the knowledge in explicit taxonomic structures that represent various relations among the concepts. Since they are manually or semi-automatically generated, their content is semantically rich and is of high quality. Ontologies store the knowledge in complex structures known as *knowledge-bases*. This knowledge is then retrieved by some inference mechanism that is capable of reasoning about facts. This inference engine uses a set of rules and other forms of logic to deduce new facts. Among different kinds of lexical knowledge, the most general and widely applicable type is the knowledge about everyday world that is possessed by all people. This knowledge is called *common sense knowledge* [117]. For instance, *"green apples are sour"* and *"you feel happy when you get a gift"* are common sense issues. To encode and utilize this type of knowledge, Cyc [110] was developed as a large common sense knowledge base [9]. The primary goal behind the construction of Cyc was to build a knowledge base that was suitable for variety of reasoning and problem solving tasks in different domains [124]. Cyc is a mature knowledge base but is still growing. To describe 250,000 terms, it contains more then 2.2 million manually crafted assertions in the form of rules and facts stated in the nth-order predicate calculus [170]. In general, Cyc knowledge base is divided into three layers: *Upper*, *Middle* and *Lower* ontologies. Each ontology represents a different level of generality of information contained within it. There are three main components of Cyc system: knowledge base, inference engine and the natural language system [125]. The knowledge base includes assertions describing various concepts interrelated by predicates. The information in the knowledge base is arranged in a hierarchical graph of

[9] `http://www.cyc.com/`

micro-theories, also called *reasoning contexts*. The Cyc inference engine is responsible for extracting the information from the knowledge base to determine the truth of a sentence. The natural language component includes lexicons (for mapping words to Cyc concepts), parsers (for translating English text into the language of Cyc (CycL)) and generation subsystems.

An *encyclopedia* compiles a summary of information from all fields or from a particular branch of knowledge. It is considered as a compendium of human knowledge. An encyclopedia is like a dictionary in that it contains entries for terms organized alphabetically and gives definition for each term but it has two main differences: first, it contains more details on the topics as compared to the dictionaries, which mainly include the definitions and part-of-speech details; and second, the focus of an encyclopedia is on factual information, whereas the focus of a dictionary is on linguistic information. Although encyclopedias offer lexical details as in dictionary but unlike dictionaries, they seldom label the relationships explicitly. Encyclopedias are classified as general-purpose (such as Encyclopedia Britannica[10] and Wikipedia[11]) and domain-specific (such as Medline Plus[12] and MeSH[13] on medical information).

Wikipedia

An example of such an encyclopedia is Wikipedia, which is a collaboratively constructed, multilingual and freely available online encyclopedia [230]. Wikipedia compiles detailed information on each topic in the form of *articles*. Although the first paragraph of a Wikipedia article implicitly introduces the definition of the topic, it does not explicitly include any gloss or definition section for that. Wikipedia groups similar information in three types of collections: *categories*, *lists* and *info boxes*. Each article

[10] `http://www.britannica.com/`

[11] `http://en.wikipedia.org/`

[12] `http://www.nlm.nih.gov/medlineplus/`

[13] `http://www.nlm.nih.gov/mesh/`

belongs to one or more categories and each *category* in turn contains multiple articles. Wikipedia articles are connected to other semantically related articles though *hyperlinks* which are not type-annotated. This makes Wikipedia a huge graph of semantically connected articles. Wikipedia includes *redirect* pages to an article representing a collection of its aliases or tightly coupled-synonyms. Another source of synonymy is Wikipedia *labels* or *anchor texts*, which are loosely-coupled synonyms, as they include other lexical relations such as hypernymy and hyponymy as well. Wikipedia tends to manually resolve the issue of polysemy by providing the users with a *disambiguation page* corresponding to a query. A *disambiguation page* lists the links to the articles of all possible senses of the query word, thus helps the user to select the intended sense article. Wikipedia also includes a template of factual information for similar kinds of articles in the form of *info boxes*.

Wikipedia offers many advantages over other knowledge sources such as WordNet and Wiktionary. Most important of all is its excellent coverage of concepts specially of proper nouns. It provides vast amount of domain specific knowledge which makes it an attractive resource. Milne *et al.* [137] conducted research to investigate the coverage of Wikipedia in the domain of food and agriculture. They showed that Wikipedia provides good coverage of agricultural topics and their relations, approaching the coverage of a professional thesaurus. In contrast, the coverage of WordNet is limited with little or no coverage of domain specific vocabularies and limited coverage of proper nouns. For instance, given the word `Jaguar`, WordNet contains only the dictionary-oriented sense, `Jaguar Panther` and does not contains any information on `jaguar cars`, whereas given the same word, Wikipedia retrieves almost 40 senses of jaguar including `Jaguar cars`, `Jaguar(band)`, `Jaguar(novel)` and `Jaguar(Computer).` In order to investigate the topical coverage of Wikipedia, Halavais and Lackaff [70] compared the coverage of Wikipedia against printed books. They found that the topical knowledge contained by Wikipedia is generally good. They also concluded that the article length and the denser connections among Wikipedia article and categories are two indicators of rich semantic information. Recently, Wikipedia has been used as a knowledge source in many applications. For instance text wikification [139], topic identification [185], ontology learning [197], text categorization [57], text clustering [9] and information extraction [221].

2.6 Existing Approaches to Computing Semantic Associations

Prior work on semantic association computation can be divided into three main research streams according to their usage of background knowledge source: *Knowledge-lean approaches*, using informal knowledge sources as the underlying corpora; *knowledge-rich approaches*, based on formal knowledge sources; and *hybrid approaches*, which combine multiple features of one or more knowledge sources (including both formal-KS and informal-KS). Over the years, a number of approaches have been proposed in these three research streams. The models underlying these approaches originated from a range of fields including information retrieval, geometry, statistics, probability theories and information theory. Section 1.2 of the thesis introduced three fundamental models of semantic associations: the *geometric model*, also known as *spatial model*, relying on the *Vector Space models* (VSM) ; the *feature-based Model*, also called *combinatorial Model*, relying on commonalities and differences between the feature sets; and the *structural model*, often referred to as *network model* or *graph based model*, relying on the formal structures of concepts. Figure 2.2 gives an overview of the background knowledge sources, their semantic elements and the underlying design models of various classes of approaches relying on these knowledge sources.

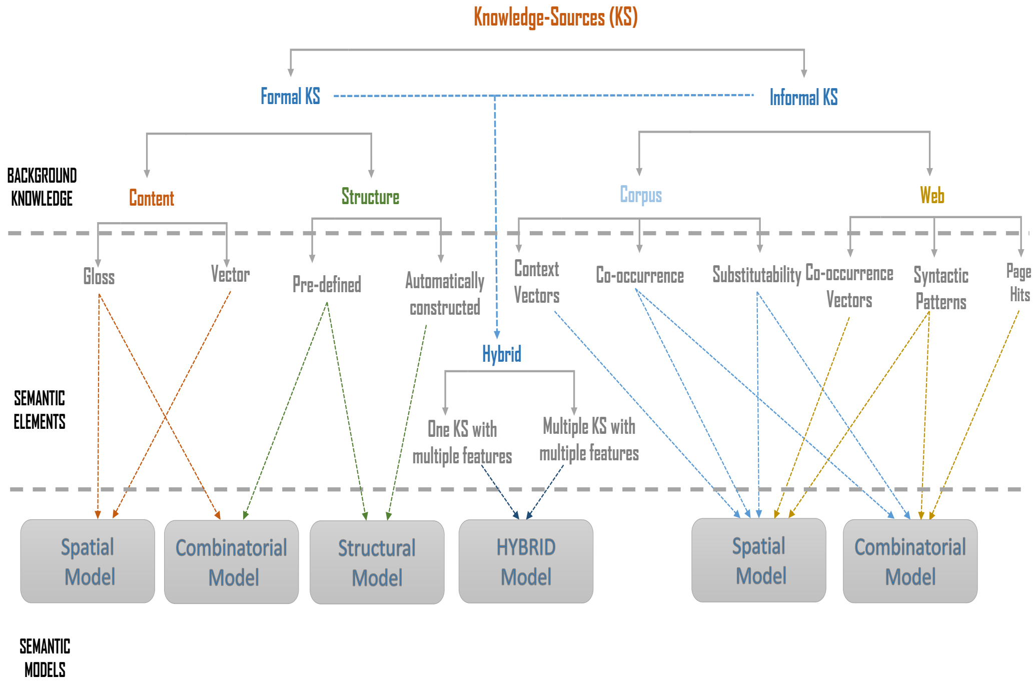


Figure 2.2: The landscape of semantic association computation approaches

2.6.1 Knowledge-Lean Approaches

Informal knowledge source based approaches are sometimes referred to as *distributional approaches* as these are inspired by a well known maxim coined by Firth [54] which says, *"You shall know a word by the company it keeps"*.

Weed [217] defined two words to be distributionally similar if: the two words occur in each other's contexts; they occur in similar contexts; or the plausibility does not change if one word is substituted by the other word. In general the context of a word is the set of words around a given word. However, different researchers used different definitions of *"around"*: some used a fixed size window of neighboring words; others used the containing sentence, paragraph, document or syntactically related neighboring words.

Informal knowledge source based approaches to computing semantic associations are based on three main assumptions: *topicality* assumption, *proximity* assumption and *parallelism* assumption [81].

According to the *topicality* assumption, also known as the distributional hypothesis [54, 74], words found in similar context tend to be semantically similar and two words that are similar are likely to be used in similar contexts. For instance, the words `bird` and `Pigeon` have an overlapping set of context words {`fly`, `eggs`, `nest`, `wings`, `warm-blooded`, `two-legged`, . . .}. Approaches based on *topicality* assumption use statistics of the relative frequency with which a word appears near other words. Although the implementation specifics may vary between the approaches based on the *topicality* assumption, such as the proximity of words counted as near may vary from 3 words in Random Indexing [181] to 300 words in Latent Semantic Indexing (LSA) [104], such approaches are included in this category based on the overall assumption of *topicality*. Examples of topicality-based approaches include LSA [104], Random Indexing [181] and second order co-occurrence vector-based approaches such as [91, 116].

The *Proximity* assumption, like the *topicality* assumption, relies on the word co-occurrences within a proximity window but the way it measures the co-occurrences is different from the *topicality* assumption. It states that two words that are semantically associated tend to co-occur near each other rather than having similar context words. For instance, the words `fish` and `water` are semantically associated as they frequently occur near each other. The proximity assumption underlies the syntagmatic word associations (words that co-occur more frequently than expected by chance). Examples of approaches based on *proximity* assumption include PMI (Pointwise Mutual information) [34], PMI-IR (PMI-Information Retrieval) [206] and LC-IR (Local Context-Information Retrieval) [81]. Proximity assumption applies to the semantic relatedness as well as semantic similarity.

The final assumption is based on grammatical *parallelism*, in which similar words tend to have similar grammatical frames. This assumption considers the frequency with which words are linked to other words by grammatical relations. Such approaches are based on the grammatical functions such as `(subject-verb)` and `(verb-object)` as well as selectional properties of verbs. It also involves considering paradigmatic relations of words (a relation in which words can substitute each other without affecting the meaning). Approaches based on this assumption include [115, 1, 16].

There are two main streams of research on knowledge-lean semantic association computation: *corpus-based* approaches and *web-based* approaches. Both of these research streams build on top of the three distributional assumptions.

Corpus-based Approaches

Early research work on computing semantic associations used unstructured text corpora as the underlying knowledge source. Such approaches applied simple syntactic techniques on text corpora to identify word associations [1, 187, 119]. Generally, two types of design models of association are used by corpus-based approaches. The first type of design model—the spatial model—involves generating vectors of word co-occurrences and using measures such as cosine similarity[14] to compute the distance between the two vectors in high dimensional space. The second type of design model—the combinatorial model—computes word associations by considering the overlaps and differences of the word occurrences.

Latent Semantic Analysis (LSA) [105] is a well known high-dimensional vector space model based approach for computing word similarity based on frequency of word occurrences. LSA was proposed as a high dimensional linear association model, where latent concepts are represented by most prominent dimensions in the data using *Singular Value Decomposition*

[14] Please see glossary for details

(SVD). It begins by constructing the co-occurrence matrix in which rows indicate the terms and columns represent the documents. The cell entries in this term-by-document Matrix represent the frequency of occurrences of the terms in each document. It expects similar words to have similar vectors based on the *topicality* assumption. To cope with the data sparseness issue, SVD is used as a dimensionality reduction technique. This results in reducing the dimension of the term-by-document matrix to approximately 300 highly related dimensions. Inspired by LSA, Hoffman proposed Probabilistic LSA [87] that constructs a low dimensional concept space based on statistical latent class model known as *aspect model*.

There are other approaches which used a similar design model for computing similarity [187, 119], however with certain differences. These approaches used sliding window approach over the corpus rather than computing co-occurrences at document level. Consequently, these approaches constructed a term-by-term matrix rather than a term-by-document matrix. In this term-by-term matrix, the entries were the co-occurrences of the terms within a window of reference rather than the term occurrences in the documents. The remaining method is the same as of LSA. Cosine similarity is computed between the vectors representing the words.

Weed and Weir [217] analyzed the plausibility of substituting two words in noun-verb syntactic relations in BNC corpus [109]. A prerequisite of their approach was the requirement of POS-tagged corpora, which prevented the application of their approach on unprocessed corpora.

Lin *et al.* [115] used a distributional similarity measure based on the distributional patterns of words. In order to bootstrap the semantics from a corpus, they parsed a 64 million words corpus (consisting of Wall Street Journal, APN Newswire and San Jose Mercury). From the parsed corpus they extracted 56.5 million triples and computed word similarity using the distributional patterns based on the *parallelism* assumption. Using this distributional similarity measure, they constructed a thesaurus that was similar to WordNet.

Another statistical technique used for relatedness computation is *Latent Dirichlet Allocation* (LDA) [14]. LDA represents a document as a mixture of words where each word is attributable to one of the document topics. Sun *et al.* [198] used LDA-based Fisher Kernel for text segmentation.

Agirre *et al.* [1] examined both *topicality* and *parallelism* assumptions on the 1.6 Tera-word Web corpus. They produced context vectors based on extracted syntactic patterns (containing a word w centered at the window) using a context window based approach (to test the *parallelism* assumption) and produced the context vectors based on neighboring words in a window of reference (to test *topicality* assumption). Finally they computed the cosine similarity between the vectors representing the given words.

Islam and Inkpen [91] used the BNC corpus [109] to get the frequencies of co-occurrence of a given pair of terms in the window of size $2\alpha + 1$ (where α is the number of words around a target word on either side) and used these frequencies to compute the PMI scores of all co-occurring terms. Based on these scores they sorted the list and selected the top n words. Then they computed the PMI summation function of each term by adding up the PMI scores of all the terms in the vectors that have non-zero PMI scores with the other given term. In other words, they considered only those terms that occurred in the context of both input terms. Then they added the normalized PMI summation scores of both given words to get the final similarity score.

Liu *et al.* [116] also constructed the second order co-occurrence vectors of biomedical corpora to compute the similarity of biomedical words. They computed the bigram term-by-term co-occurrence matrix (with term collocation frequencies as the matrix entries). Then they constructed the definitions of given words based on UMLS corpus and WordNet. The relatedness between two words is computed as the cosine of angle between the centroid of their definition vectors.

Web-based Approaches

The web is a heterogeneous collection of documents. Its dynamic nature, domain independence, universality and huge knowledge coverage make it an ideal resource for extracting background knowledge [68]. Various methods have analyzed the use of the web as a corpus [101, 206, 35, 180]. Web-based approaches cope well with the problem of data sparseness, in particular for the words that rarely occur. Keller *et al.* [100] conducted a study in which they found that web frequencies correlate with word frequencies in the carefully annotated BNC corpus [109]. They also found that web frequencies can be used to reliably predict the human plausibility. The web is used as a corpus in many NLP applications such as automatic thesaurus construction [31], word sense disambiguation [133], extraction of lexical relations [49], synonymy identification [206] and word clustering [123].

In general the web-based approaches utilize various features extracted from web for computing semantic associations. Some well known features include page counts, co-occurrence vectors and syntactic patterns derived from text snippets and web pages. Consequently, web-based measures can be grouped into *co-occurrence* based measures relying on *proximity* assumption, *vector space* based approaches using *topicality* assumption and *lexical pattern* based approaches using *parallelism* assumption.

Peter Turney [206] proposed an unsupervised approach to identify the synonyms from the web. He queried a web search engine Alta Vista to get the hits for given words and use these statistics to compute the similarity score based on the distributional measure PMI-IR. Pointwise Mutual Information (PMI) [34] is a measure of information overlap between two variables , commonly used in information theory and statistics, and is given as below:

$$PMI(x, y) = log\frac{f(x, y)}{(f(x)f(y))}$$

where $f(x)f(y)$ is the probability of co-occurrence when two variables x

and y are independent, and $f(x,y)$ exceeds $f(x)f(y)$ when they are dependent. Hence, PMI is the degree of statistical dependence between two variables. It is a measure of how much one word tells about the other word. This information gain could be positive as well as negative. The log of this ratio indicates the amount of information about the presence of one variable when the other variable is observed. A major problem with PMI-based approaches is that the PMI score is high for rare words, which does not necessarily mean that there is a noteworthy dependence between the words.

Cilibrasi and Vitanyi [35] proposed *Normalized Google Distance* (NGD), as a distance measure between words using the number of hits returned by Google search engine for a given pair of words. They used the tendency of two words to co-occur in the web pages as a measure of semantic distance of word and phrases. NGD used the World Wide Web as the corpus and Google page counts to get the frequency of word occurrences. NGD is well-founded on information distance and the Kolmogrov complexity theories [35]. The formula for NGD is as follows:

$$NGD = \frac{max(logf(x), logf(y)) - logf(x,y)}{logM - min(logf(x), logf(y))}$$

where *f(x)* denoted the number of web pages containing word x and *f(x,y)* represents the number of web pages containing both words x and y.

Ruiz-Casado *et al.* [179] retrieved the search text snippets containing first the word from web results, extracted the context around the word, replaced it with the second word and checked for the existence of this new pattern. Similarity of two words is then computed as the percentage of the sentences in which a word can be replaced by another word.

Sahami *et al.* [180], computed the similarity of short text snippets using web search results. A text snippet consists of a title, a short description and a link to that web page. For each of the given words, corresponding text snippets were converted into weighted vectors and the similarity between two words is computed as the inner product between their centroids. This approach can be used to compute similarity of two text snippets, even when the snippets have no overlapping words. They assumed that the text snippets could be single words, multi-word phrases or entire text snippet. Their approach was based on the *topicality* assumption.

Matsuo *et al.* [123] used the statistics collected from search engines to compute the distributional similarity. Their method queried the given words individually as well as when put together to get the number of page hits. They used these statistics to compute Pointwise Mutual Information (PMI) measure.

Chen *et al.* [30] analyzed the web as a live corpus by presenting various association measures based on the web. They used a model of web search with double checking to get statistics from web snippets and used them to compute various adapted co-occurrence measures. For two given words P and Q, they collected text snippets retrieved by a search engine. In the snippets of Q, they computed the occurrences of word P and vice versa. These values were combined in a non-linear way to compute the similarity between P and Q using the *Co-occurrence Double-Checking* measure as follows:

$$CODC(P,Q) = \begin{cases} 0 & \text{if } f(P@Q) = 0 \; or \; f(Q@P) = 0 \\ e^{log\left(\frac{f(P@Q)}{f(Q)} \times \frac{f(Q@P)}{f(P)}\right)^{\alpha}} & \text{otherwise} \end{cases}$$

where $f(P@Q)$ denotes the number of occurrences of Q in the top-ranked text snippets of query P in Google search engine and α is a controlling parameter. This method critically relies on the ranking of search engines to retrieve the top-ranked relevant snippets. A search engine considers many other factors in result ranking. Hence, it is difficult to find the occurrences of the word P in the top ranked text snippets of query Q if the snippets are not relevant. This is evident from their results, where most of the word pairs get a zero similarity score. Also, their assumption disregards asymmetric association of words. Due to asymmetry, it is possible that the word P occurs in the top snippets of the word Q but not vice versa.

In such cases, the CODC measure assumes that the given two words are unrelated, hence assigns them zero similarity score.

Gracia and Eduardo [68] computed semantic similarity by generalizing the *Normalized Google Distance* (NGD) measure to exploit the web as a source of knowledge. They transformed NGD into a generalized word relatedness measure that could be used with any web search engine. They transformed the NGD formula so that the distance scores it computes are bounded in a new range of [0-1] rather the original range [0-∞]. Hence they proposed the following transformation to NGD (which they called Normalized Web Distance (NWD) rather than NGD).

$$relWeb(x,y) = e^{-2NWD(x,y)}$$

Glason *et al.* [62] used a web-based similarity measure that used the search engine counts to compute similarity of word pairs as well as groups of words. They used the web search count by measuring the decline in the number of hits as more words are appended in the query using AND operator.

Bollegala *et al.* [16] presented a measure based on learning an optimal combination of the page counts of the web (as the global context) and pattern extraction from web snippets (as the local context) returned by a search engine for a given word pair. They implemented four co-occurrence-based measures, namely the Jaccard measure [94], the Dice [43] measure, the Overlap measure [188] and the PMI measure [34]. They extracted the syntactic patterns containing given two words in the proximity window of seven words, sorted them according to their frequency of occurrence, clustered them according to the semantically similar relations and computed the Cosine similarity between cluster centroids and the sorted vector of patterns. Finally, they trained a two-class SVM on these features to classify a word pair as synonym or non-synonym. Their approach is based on a combination of proximity and *parallelism* assumptions.

2.6.2 Knowledge-Rich Approaches

According to the nature of various aspects of underlying knowledge source, the approaches based on formal knowledge sources are broadly categorized into two main streams: *structure-based approaches* and *content-based approaches*. Each research direction has its own advantages and disadvantages (discussed in the Section 2.7 of the thesis).

Structure-based Approaches

Structure-based approaches rely on the organization of the human intellect into semantically rich and well-defined semantic networks or graphs. Literature review indicates that there are two directions of research in structure-based approaches.

The first research direction includes the semantic association computation approaches based on predefined structures or taxonomies of existing knowledge sources such as the IS-A hierarchy of WordNet or hyperlink structure of Wikipedia. Since these structures are predefined, their strengths lie in the huge coverage of knowledge and explicit semantic encoding in a structured way. On the other hand, their downside is the structural inflexibility of the predefined knowledge sources in explicitly controlling the semantics. In the structure-based exploration of semantic similarity, Rada *et al.* [166] computed the number of edges between two terms in the hierarchy of MeSH which is a controlled vocabulary thesaurus providing conceptual hierarchy of medical terms that permits searching at various levels of specificity. Agirre and Rigau [2] proposed a dictionary-based measure by using the conceptual density and depth of dictionary to identify the shortest path distances between the concepts in the set structure. Hirst and Onege [84] considered the types of relations encoded by the edges of a path between two concepts. They theorized that if a path between two concepts consists of many edges that belong of different lexical relations then the two concepts are semantically distant concepts. They used the path length and number of direction changes in the taxonomy to compute relatedness between two concepts. A similar approach was adopted by Leacock and Chodoro [106], who proposed an edge-based[15] measure for computing semantic similarity using WordNet. They introduced the shortest path measure for WordNet. Milne and Witten [138] used Wikipedia hyperlink structure to compute semantic relatedness based on various features derived from Wikipedia hyperlinks. One class of such approaches are known as *path-based approaches*. Other examples of such approaches include the work of [29, 185, 138].

The second research direction consists of approaches that automatically generate huge semantic networks or graphs from various existing knowledge sources. These approaches have an advantage of controlling flexibility of the way the semantics are encoded in their structure but are computationally expensive to generate and suffer from scalability issues. Some worth mentioning approaches in this research direction are [113, 148, 225, 224, 169, 90]. Sonya and Markovitch [113] proposed *Compact Hierarchical Explicit Representation* in which they converted the Wikipedia category network into a multiple inheritance hierarchy. They assumed that two concepts are considered related if their hierarchical representations with respect to structure are similar. Navigli and Ponzetto [148] proposed a graph-based multilingual approach to compute semantic relatedness. They constructed and used BabelNet, a multilingual lexical knowledge source, to construct subgraphs for a word pair in different languages and computed semantic relatedness based on the subgraph intersection. Yeh *et al.* [225] constructed Wikipedia-based semantic graph and applied *Random Walk* with *Personalized Page Ranks* to compute semantic relatedness for words and texts. Ramagae *et al.* [169] constructed a semantic graph from WordNet and used the *Random Walk* algorithm to get the distribution of two textual units. Finally, they computed the similarity of two distributions using various semantic association measures. Yazdani and

[15] An edge is a link between two nodes on a path

Popescu-Belis [224] constructed a semantic network of concepts extracted from Wikipedia hyperlink structure and article content and used the *Random Walk* algorithm to compute the textual distance. Iosif and Potamianos [90] followed an unsupervised approach to construct a semantic network using co-occurrence or context similarity features extracted from the web-corpus.

Content-based Approaches

Content-based approaches make an effective use of the textual content at the concepts, documents or knowledge source level. There are various types of content-based approaches: *vector* based approaches, which are some variants of the *Vector Space Model* (VSM); and the *gloss* based approaches, which rely on the word overlap between the glosses of the concepts.

Vector Space Model (VSM), a well known algebraic model, represents input words as weighted vectors and computes the deviation of the angles between the vectors to compute the similarity [182]. In vector-based approaches, Baeza-Yates and Ribeiro-Neto [7] used bag of words model and compared the text fragments in the vector space. Their approach was quite simple to implement but performs sub-optimally when the text fragment to be compared shares very few words or when concepts are expressed by their synonyms rather than their actual words. Gabrilovich and Markovich [58] proposed *Explicit Semantic Analysis* (ESA) to incorporate human knowledge into relatedness computation by constructing concept vectors and comparing them using Cosine similarity. Hassan and Mihalceae [76] introduced *Salient Semantic Analysis* (SSA) by modeling frequently co-occurring words in the contextualized profile for each word. They only used words with high saliency or relevance to the document. Their approach works for relatedness computation of both word pairs and text pairs. *Temporal Semantic Analysis* (TSA) [167] was proposed to incorporate temporal dynamics to enhance text relatedness models. TSA represented each input word as a concept vector and extended static representation with temporal dynamics. Halawi *et al.* [71] proposed constrained learning of relatedness in which they learned a suitable word representation in a latent factor space. Content-based approaches specially those relying on the *Vector Space Models* (VSM) are known to perform better than structure-based approaches but are still biased towards more frequently occurring words.

The idea of gloss vectors was introduced by Lesk [111] for solving word sense disambiguation task. The Lesk algorithm compared the gloss of every context word around a target ambiguous word with the gloss of every sense of the target word. A sense sharing maximum word overlap with the glosses of the context words was selected as the intended sense. The idea was intuitive but suffered from the length of the gloss which is fairly short in dictionaries, hence does not provide sufficient vocabulary [8]. Banerjee and Pederson [8] extended the idea of gloss overlap by including glosses of other concepts related to a given concept. They observed that the synset of two related words are also related and have common words in their glosses. Hence, they extended the view of a relatedness between two words by considering various explicit semantic relations defined in WordNet such as hyponyms, hypernyms, meronyms, holonyms and troponyms. Other approaches which used the idea of gloss overlaps include [210, 129, 230, 69].

2.6.3 Hybrid Approaches

In *hybrid approaches*, multiple aspects of one or more knowledge source(s) are combined to compute word associations. Hybrid approaches usually combine the strengths of different features extracted from one or more knowledge sources which could be formal or informal. Hybrid approaches using multiple features generally lead to better performance than the approaches using single features as indicated by Agirre [1]. However, this performance improvement is usually achieved at the expense of higher computational costs. It is also possible that the hybrid method also inherit the limitations of the individual features or the underlying knowledge source(s) as indicated by [233]. Again, there are two research streams in the category of hybrid approaches.

First research stream consists of hybrid approaches that combine multiple aspects of the same knowledge source. Bollegala *et al.* [16] proposed a semantic association computation approach that extracted two features from web as a knowledge source: page counts based co-occurrences and clustered lexical pattens. They used support vector machine-based supervised algorithm to compute semantic similarity based on combining these two features. Similarly, Ponzetto and Strube [164] adapted path-based, information-content-based and taxonomy-based measures of WordNet to Wikipedia category network and combined them using SVM for computing word similarity. Mohammad *et al.* [199] proposed a semantic relatedness measure using various Wikipedia features such as articles, category graph and redirects. They combined these features in a system and used various similarity measures such as Dice, Simpson, Jaccard and Cosine similarity to compute the similarity between words. Other example of similar approaches that combined various aspects of a single knowledge source include [199, 224].

In the second research stream, hybrid approaches combine various aspects of multiple knowledge sources. Resnik [171] combined corpus statistics with a lexical taxonomy. The fundamental idea was to identify a concept in the lexical taxonomy that includes any information shared by the two concepts for similarity. He proposed the concept of *Information Content (IC)* based on the shared sub-sumer. *Information content* refers to the specificity of a concept [159]. Higher values of information content refer to more specific concept. For instance, the IC value of a general term `idea` will be lower than that of a more specific term `coin`. For a concept c in the WordNet hierarchy, the information content is defined as the negative log of probability of observing c (in terms of frequency of occurrences) and is given by,

$$IC(c) = -logP(c)$$

where $P(c)$ is the probability of occurrence of the concept c. The smaller the value of information content of two concept the more similar they are. Resnik [171] used IC to refer to the amount of information shared between two concepts and computed it using the *Least Common Subsumer* (LCS) or the *lowest super-ordinate* concept in the WordNet hierarchy. The more specific the LCS concept, the less distant the concepts. The formula for computing semantic distance of two concepts is given by,

$$Resnik(c_1, c_2) = IC(LCS(c_1, c_2))$$

Resnik measure is simple but suffers when different concept pairs have same LCS. Lin [115] attempted to refine the Resnik measure by using the IC of individual concepts and computed the similarity of two concepts as,

$$lin(c_1, c_2) = \frac{2 \times Resnik(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

Jiang and Conrath [97] extended the same concept of combining the corpus statistics with lexical taxonomy. Their approach combined the information content measure with the edge count of noun terms in the IS-A hierarchy. They augmented the Resnik measure by using the IC of individual concepts given by,

$$JCN(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times Resnik(c_1, c_2))}$$

However, the IC-based methods can only be computed for nouns and verbs in WordNet as these are organized in hierarchies in WordNet. However, these hierarchies are separate, thus the IC-based methods can only be applied to the same POS class word pairs (noun-noun or verb-verb pairs) [159]. Similarly, Agirre *et al.* [1] proposed a supervised approach to compute semantic similarity and relatedness using various features including personalized Page Ranks based on WordNet graph and proximity window based approaches using text corpus of 1.6 Terawords. Zesch *et al.* [230] used multiple knowledge sources namely, Wikipedia, Wiktionary and WordNet and computed path length-based and concept vector-based measures of similarity on them. Other examples of hybrid approaches include [131, 226].

2.7 Limitations of Existing Approaches

The survey of various approaches of semantic association computation presented in this chapter provides an insight into the behavior of the approaches in each research direction. There are certain unique features of informal knowledge source based approaches that make them an attractive option for computing semantic associations: first, these measures are applicable in resource poor languages; and second, these measures are able to mimic both semantic similarity as well as relatedness depending on the design model that they use. Similarly, the semantic measures that belong to the class of formal knowledge sources are able to capture semantics at the concept level rather than just at the word level. Such approaches make a good use of the rich semantics implicitly and explicitly encoded in the semantic knowledge resources. However, it is important to identify limitations of various approaches in each research stream. Understanding the shortcomings of using various design models and knowledge sources will be helpful in selecting suitable semantic measures in various application settings.

2.7.1 Limitations of Knowledge-Lean Approaches

Corpus-based approaches enjoy the advantage of flexible adaptation to other informal knowledge sources. This flexibility in the choice of background knowledge source makes these approaches more applicable in various domains than the other approaches which are limited by the coverage, completeness and structure of the formal knowledge sources. Corpus-based approaches rely on *distributional hypothesis*, which is criticized for certain weaknesses. Distributional approaches assume that words which co-occur or appear in the same context are highly related which is not always true. For instance, consider the term pair `bread,butter`. These words frequently co-occur in a corpus and are assigned relatedness score higher than synonyms due to existence of a common phrase *bread and butter*. In general, corpora follow Zipf's law [234] for word coverage which is usually skewed [16]. In other words, regardless of the size of the corpus, rarely used words have very low frequency of occurrence, which leads to unreliable *Vector Space Models* based on distributional hypothesis. Hence, distributional approaches perform poor on rarely used words. Also, distributional approaches do not explicitly distinguish between semantic similarity and semantic relatedness. Coupled with the fact that there are obvious differences in the semantic similarity and relatedness, distributional approaches are not suitable for applications which make a distinction in these two semantic types. Moreover, distributional approaches suffer from huge computational costs of pre-processing the entire corpus to discover the co-occurrences and lexical patterns of words in the corpus. Also, distributional approaches compute semantic associations at word level rather than at concept level, hence suffer from word sense conflation.

There are certain limitations of web-based approaches that use search engine results for computing semantic associations [101]. The data returned by a search engine is always truncated. The number of result hits returned by a search engine are often rounded up estimations, hence are approximate. The page count having occurrence of a word varies corresponding to the search engine load and other factors. Hence, the statistics returned by a search engine are not reliable. Also, the search engines

[16] Please see the glossary for the detail

are frequently updated, so the results are not reproducible. Another limitation of web-based approaches is that the systems using result snippets returned by a search engine are limited by the maximum number of documents returned per query (typically around a thousand) [1]. The text snippets do not present enough context for computing word associations. For instance, each text snippet returned by the Google search engine contains around 10 words on average, which might not be sufficient to represent a document and extract the context. Moreover, the approaches based on document snippets are limited by the query syntax. The information retrieval systems and search engines critically depend on the query to be sufficiently well-structured to retrieve the required information. Poorly formatted queries can lead to irrelevant documents which adversely affect the performance of statistical and distributional methods of word associations. Text snippets based approaches also suffer from context quality. The search engines do not retrieve results from linguistics perspective (for instance word class). The documents having search terms in their title and headings are often ranked high, hence occupy the top ranks of the result set. This disregards the semantics of a query by only looking at the shallow syntactic features.

2.7.2 Limitations of Knowledge-Rich Approaches

Structure-based approaches are theoretically simple and their implementation and adaptation to other structures is quite straight forward. However, these approaches rely heavily on the underlying knowledge source, thus are sensitive to the taxonomic structure. Such approaches critically rely on the knowledge coverage, degree of completeness, density and depth of the underlying structure [12]. For instance, Pederson *et al.* [160] reported the difference in the performance of the same approach on different depths and specificities of the underlying taxonomies. Similarly, al-Mubaid *et al.* [4] showed that the performance of their system on MeSH taxonomy degraded when the same approach was applied to a different taxonomic structure, SNOMED-CT. Similar observations were reported by Strube and Ponzetto [129] and Zasch and Gurevych [228] when adapting WordNet-based measures on Wikipedia and Wiktionary. Wang *et al.* [215] highlighted another issue with structure-based approaches that the path-based approaches usually consider a single path using a single type of lexical relation (such as Is-A hierarchy). This limits their capability by overlooking other types of semantic relations. Also, the background information used by structure-based approaches in not sufficient to represent complex semantic relations.

IC-based measures suffer from similar limitations as the structure-based approaches due to their critical reliance on the underlying taxonomy. IC-based methods also ignore many semantic relations due to the limitation of underlying hierarchy. In general, both structure-based and IC-based methods are more suitable for estimating semantic similarity rather than semantic relatedness, because of their focus on the taxonomic relations [233].

Gloss-based approaches are computationally inexpensive as compared to the other semantic approaches. However, they gather the background information from the glosses, which does not always provide sufficient semantic evidence. From this perspective, such approaches also suffer from the knowledge acquisition bottlenecks of the underlying knowledge source. This problem is partially handled by augmenting the glosses from multiple knowledge sources to construct *pseudo-glosses* that cover more semantic evidences [69, 230, 228].

Vector-based approaches mostly differ in the way the concept/word vectors are computed. The remaining algorithmic details are the same in most of the approaches. Depending on the underlying method of constructing the vectors, these approaches can be adapted to different knowledge sources easily. Zesch te al [228] showed this by adapting various vector-based approaches across different knowledge sources. However, these approaches are also not without limitations. Vector-based approaches are usually computationally expensive. Many of them require the preprocessing of huge knowledge sources before generating the vectors [42, 58, 71]. Moreover, these approaches often suffer from the issue of data sparseness and require using the dimensionality reduction techniques such as SVD.

2.8 Performance Indicators

This section highlights certain important factors that affect the performance of a semantic measure. These factors are listed as follows:

1. **Choice of Knowledge Source:** The literature review reveals that the choice of knowledge source is the most important factor that controls and guides other factors. The existing approaches surveyed in the chapter were also categorized according to this factor in an attempt to identify the orientation of semantic approaches belonging to each category of knowledge sources. The approaches based on a corpus have a natural tendency to find distributional associations of words. Such approaches are ideal for certain kinds of applications such as automatic thesaurus generation, lexical chaining and semantic relation identification. Similarly, formal knowledge source based approaches rely on explicitly defined semantics, hence would be a good choice for applications such as query expansion, topic modeling and opinion mining.

2. **Choice of Design Model:** The second most important factor is the choice of underlying design model. It is clear from the previous discussion that using the same knowledge sources with different design models results in different performance and semantic bias. For instance, a semantic association computation approach based on text corpus and a combinatorial model would be a good approach for es-

timating collocations but might not be a good choice for synonymy detection. On the other hand, if a semantic association computation approach is based on the geometric model of association then the focus of such approach will be on finding words having a similar context, thus leading to better estimates of synonymy. Similarly, a semantic association measure using a formal knowledge source and the structural model of association might be good at predicting semantic similarity due to its reliance on the taxonomic structure. In contrast, approaches based on formal knowledge sources and the geometric model would be good at judging the semantic relatedness of words because these do not rely on the overlap of directly related features.

3. **Types of Semantic Relations:** The selection of a background knowledge source also affects the orientation of a semantic measure towards *classical* or *non-classical* relations. Generally, the approaches relying on taxonomy or structural semantics are not good at predicting *non-classical* relations which are not explicitly encoded in the taxonomic structures. Such approaches are also not good at finding the cross-POS relations. Their strength lies in excellent estimates of the lexical-semantic relations between words. The choice of knowledge source and the underlying design model contribute to the performance of a semantic measure on estimating various types of semantic associations such as semantic similarity, semantic relatedness and distributional similarity (which could further be categorized into co-occurrences, collocations, synonymy, and asymmetric associations).

4. **Nature of Datasets:** Finally, the performance of a semantic measure is affected by the choice of dataset. Existing datasets on word associations focus on one or more semantic classes such as POS-classes or various types of semantic associations. The approaches having a bias towards a specific semantic class will perform better on that kind of

dataset but might produce worst result on other datasets. Hence, evaluation of a semantic association computation approach on such datasets could lead to unreliable conclusions about the performance of that approach. The nature of a dataset should be investigated from different perspectives of term relations such as POS classes, cross-POS relations and types of semantic relations. It is observed that there is not a single dataset that is suitable for evaluating the performance of all kinds of approaches on all types of semantic relations. Each of the existing datasets provides quite a restrictive view of the capabilities of semantic approaches. Hence, it is important for future approaches to build a unified framework for supporting a comparative evaluation testing different perspectives of semantic associations.

2.9 Chapter Summary

This chapter introduced the fundamentals of semantic association computation and discussed various approaches for computing semantic associations from *the usage of underlying knowledge source* point of view. There are two parts of this chapter. The first part introduced the concepts of semantic associations in general and various types of semantic associations. This follows the discussion on pervasiveness of semantic measures in various natural language processing applications. Then, it detailed the experiments reported on estimating the human judgments on semantic associations and showed that the inter-rater agreement of semantically similar word pairs was found higher than that of semantic relatedness, which indicated the inherent complexity in correctly identifying and estimating the semantic relatedness. The second part of this chapter identified two main categories of background knowledge sources and detailed the attributes of various knowledge sources in each category. It also classified the existing approaches according to their underlying knowledge sources and discussed the strengths and limitations of each approach. Finally, based on the survey, this chapter identified the limitations of existing approaches and important factors that directly or indirectly affect the performance of semantic associations measures.

Chapter 3

Mining Wikipedia for Computing Semantic Relatedness

The work described in this chapter presents an exploitation of the semantic richness of a collaboratively built structured knowledge source, Wikipedia, in computing semantic associations of words. Wikipedia is a knowledge source that offers rich semantics defined by collective human intelligence (as opposed to lexical or statistical estimations) in a collaborative environment, resulting in a huge resource with a continuously growing coverage of knowledge. The chapter develops new semantic association measures based on Wikipedia and investigates their semantic capabilities on the task of computing semantic associations. Section 3.1 details semantically rich structural elements of Wikipedia; Section 3.2 presents two new semantic association measures based on structural semantics mined from Wikipedia. Section 3.3 details a new semantic measure based on semantic mining from the informative-content [1] of Wikipedia articles; Sections 3.4 details the evaluation of Wikipedia-based semantic association measures on domain-independent datasets. This follows Section 3.5 presenting the domain-specific evaluation of the semantic measures on biomed-

[1] *Informative-content* is different from the term *information content* coined by Resnik [171]. It refers to the actual textual content of Wikipedia articles.

ical datasets. Finally, Section 3.6 concludes the chapter with a brief summary of the key contributions.

3.1 Wikipedia as a Semantic Knowledge Source

Wikipedia is a free encyclopedia available in more than 250 languages[2]. The English version is the largest of all with more than 4 million articles. Wikipedia is a multifaceted knowledge source utilized diversely in a number of researches. It has multiple perspectives of consideration including an encyclopedia, a corpus, a database, a thesaurus, an ontology, and a network structure [126]. This section details the semantic role of various structural elements of Wikipedia that can be exploited for computing semantic associations. An overview of these structural elements is as follows.

Articles: An article is the main entity of information in Wikipedia[3]. An article is a piece of free text written about a single topic[4]. Wikipedia articles are a result of continuous and collaborative effort of thousands of free contributors who followed a set of rules, called *Wikipedia Manual of Style*, to write an article. This manual ensures a consistent format of every Wikipedia article, giving it a steady and readable look as well as connecting it to other articles having the same context [132]. Every Wikipedia article has a fairly predictable layout including certain required elements [126]. The beauty of Wikipedia as an encyclopedia is the fact that each Wikipedia article is focused on one particular topic which not only includes the dictionary like definition of the topic but also discusses the topic with respect to its super-concepts (hypernyms), sub-concepts (hyponyms) and related-concepts, hence offers much richer semantics than the mare

[2] Available at: `http://en.wikipedia.org`

[3] articles and concepts are used interchangeably in the thesis and are used to refer to a Wikipedia article or one of its possible senses

[4] Wikipedia articles also include videos, pictures, graphs and audios but the subject matter is mainly text

structure.

Article Titles: Each article has a title in the form of a single word or a well formed phrase, sufficiently depicting the informative-content of that article. It often includes additional information about the scope of the article. For instance, the concept `Rooster(Zodiac)` indicates that the article is about the zodiac sign rooster rather than rooster bird. Wikipedia article titles are unique identifiers and can be used as descriptors in thesauri and URI's in ontologies [126].

Redirects: A redirect is a page having no text but a directive in the form of a redirect link to the target article [126]. Redirects represent the name variations of articles. They are a good source of mapping synonyms, plurals, closely related words, abbreviations, alternative spellings and other variations in the surface form[5] to the titles of the target articles. For instance, redirects of the Wikipedia article `heart` include `Heart(anatomy)`, `Heart(organ)`, `heart(biology)`, `cardiac physiology`, `cardiac`, `Human heart` and `four chambered heart`. All of these redirects refer to the same concept `heart`. Wikipedia Redirects are a good source of resolving synonymy issues without the need of an external thesaurus.

Hyperlink Structure: Wikipedia is a huge network of densely connected articles in which every article refers to other related articles through hyperlinks. A hyperlink is a reference to a Wikipedia article and can be followed by clicking on it. The article having a link is referred to as the *anchor article*; and the article pointed by the link is called *target article*. Like the web, Wikipedia hyperlinks provide a reader with immediate access to semantically relevant information on other pages and embody essential information about relatedness. However, a Wikipedia hyperlink is different from an ordinary web link, where a hyperlink connects any two web pages regardless of the context [99]. Wikipedia hyperlink from article `A` to article `B` shows that article `B` is semantically related to the content of article `A`.

[5] please see the glossary for description

Anchor Texts: Each Wikipedia article has a number of anchortexts, also referred to as **labels**. An anchortext is a single word or phrase used for labeling a link in an *anchor article*. Wikipedia labels are a very good source of encoding synonyms and other variations of the title of a *target article*. These variations range from tightly coupled synonyms such as `rooster` and `cock` to various lexical relations such as hyponymy (`skin` to `animal skin`), meronymy (`skin` to `skin cell`) and related concepts (`skin` to `skin care`). They are an extremely useful component of Wikipedia because Wikipedia contributors modify them according to their own culture and customs in addition to the context of the article in which they are used. For instance, a label for `United States` is `Yankee-land`, which is a British slang for referring to *United States of America*. They not only encode the synonyms and surface forms but also the polysemy and the likeliness of each sense [139]. For instance, a link labeled with `Tree` has 92.82% likeliness of linking to the target article `Tree(plant)` and only 2.57% likeliness of linking to the target article `Tree(Dataset Structure).`

Categories: Wikipedia articles are organized into generic categories. Wikipedia category network is a taxonomic structure that emerged from collaborative tagging [230]. A Wikipedia article may belong to multiple categories. For example, the article `Happiness` belongs to categories like `Positive mental attitude`, `Concepts in ethics`, `Philosophy of love`, `Positive psychology`, `Emotions` and `Pleasure`. Similarly a category may include anywhere from a single to hundreds of articles. For instance the category `Positive mental attitude` includes 37 articles in total. The Wikipedia category network is a directed acyclic graph starting from a single category called the *root category*, which is further divided into 12 main branches. All Wikipedia articles belong to one or more of these categories. This way the top categories represent the the most general generalizations of a topic while the deeper level of categories represent more specific topic generalizations. These categories can be mined to extract various kinds of relations ranging from generic relations such as hypernyms and hyponyms to more specific relations such as holonymy, meronymy and homonymy [126].

Disambiguation Page: A disambiguation page is created and maintained to represent a polysemous word (a word with multiple senses). Whenever a Wikipedia user searches for a term with multiple senses or contexts, a disambiguation page is presented to the user. For example, the word `Mars` leads a user to the disambiguation page having multiple senses like `Mars(astrology),Mars(chocolate bar),Mars(Band)` and `Mars(mythology)`. The context of the word `Mars` is so diverse that this word can not be used correctly without considering its appropriate sense. Due to this reason, every disambiguation page in Wikipedia contains a list of possible senses for a given word. These senses are further grouped according to some generic categories such as music, politics, law, geography, persons and places.

Infoboxes: An infobox is a special type of reusable templates to display an article's relevant useful information summary in a structured form [126]. For example, for each Wikipedia article about a specific university, there is an infobox displayed on the top right side of the article having a list of factual data about that university such as the establishment date, location, size and website.

3.2 Mining Structural Semantics from Wikipedia

The semantic relatedness literature points to the exploitation of two main Wikipedia structures for computing semantic associations: the hyperlink structure and the category structure. The thesis uses the hyperlink network for computing semantic associations. Although Wikipedia category graph is also used in literature for computing semantic associations, it suffers from lack of good structural connectivity, uneven density and less semantic information due to the presence of many administrative categories as compared to the hyperlink structure. Also, the number of Wikipedia categories per article are much less than the number of hyperlinks. Hence, the focus of the structure-based measures is on using the Wikipedia hyperlink structure for computing semantic associations.

A Wikipedia hyperlink is a connection between two Wikipedia articles sharing some context. Articles having links referring to a specific article a_i are called its *in-link* articles. Similarly, articles which are referred to by the article a_i are called its *out-link* articles, as indicated by Figure 3.1.

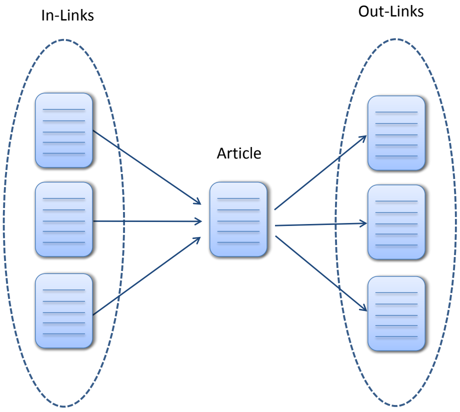


Figure 3.1: Wikipedia hyperlink structure

The structural semantic mining is done by exploiting the hyperlink structure that intrinsically encodes a variety of semantic and lexical relations among Wikipedia articles.

In this section, two new association measures based on Wikipedia's hyperlink structure are presented: The first measure is called *WikiSim* and the second measures is called *Overlapping Strength-based Relatedness* (OSR). Both of these measures make effective use of different features based on Wikipedia hyperlinks for computing semantic associations. The first measure uses the hyperlink overlap, whereas the second measure considers the averaged relative strength of relatedness of all elements in the set of shared links.

3.2.1 WikiSim

The first presented measure WikiSim, is a Simpson coefficient [16] inspired measure of the relatedness of two terms. WikiSim takes into account the proportion of links shared by corresponding articles of the two given terms. The approach starts with matching input terms t_1 and t_2 to their corresponding Wikipedia articles a_1 and a_2 respectively. For a_1, a link set (LS_1), consisting of its all distinct in-link and out-link articles, is constructed and compared with the link set (LS_2) of a_2, to find out the link overlap set. This link overlap set is then used to compute the relatedness as follows.

$$WikiSim(t_1, t_2) = \begin{cases} \frac{|LS_1 \bigcap LS_2|}{\min(|LS1|,|LS_2|)} & \text{if } |LS_1 \bigcap LS_2| \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

In the above formula, $|LS_1 \bigcap LS_2|$ is the set of all links shared by both articles. This overlapping is normalized by the length of the smaller link set to avoid any bias in the results due to larger size of either link set. This is particularly useful in case of term relatedness, where the size of a link set is unknown in advance and the link sets have unequal size. *WikiSim* generates the final scores on a scale of [0-1], where 0 means unrelated and 1 means highly related (e.g synonym).

3.2.2 Overlapping Strength based Relatedness (OSR)

WikiSim considers all shared links equally useful. However, some links might be equally important to both input terms, whereas some others might be semantically oriented more towards either input term but not both as shown in Figure 3.2. Certain Wikipedia concepts are closer to either of the given concepts `Car` or `Gasoline`, whereas some concepts like `Economics of automobile usage` have symmetric semantic orientation towards both concepts.

The fact that all shared links between two concepts are not equally useful, highlights the need of identifying the semantic orientation and useful-

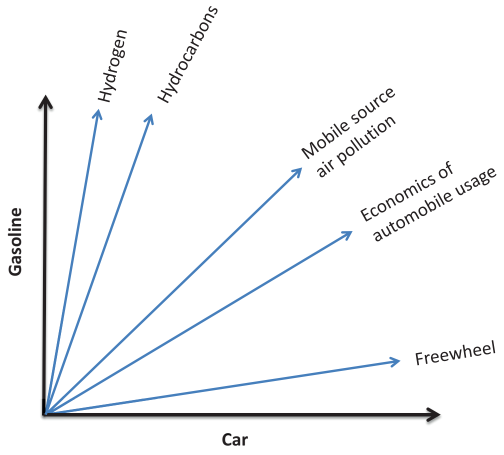


Figure 3.2: Semantic orientation of the *shared associates* of two concepts `car` and `Gasoline`.

ness of their *shared associates* in order to compute the final relatedness of two concepts.

Overlapping Strength-based Relatedness (OSR) is the second proposed measure that uses the idea of semantic orientation of shared links for computing semantic association of a given term pair. OSR takes into account the combined association strength of shared links of both input terms and uses it to compute the final association score of a given term pair.

Following WikiSim, the approach matches each a pair of terms t_1 and t_2 to the corresponding Wikipedia articles $article_1$ and $article_2$ and extracts the link sets of both articles. The link sets are then compared to get an overlap set consisting of all the link articles that are common to both sets. Each article in the overlap set is referred to as a *shared associate*. The association strength of each *shared associate* a_s is then computed as the product of its relatedness strength (SR) with $article_1$ and $article_2$ as follows:

$$AssocStrength(a) = SR(a, article_1) \times SR(a, article_2) \tag{3.2}$$

where SR is computed using *Article Comparer* of *Wikipedia Miner Toolkit*

[140]. *Article Comparer* is a machine learning based algorithm for computing the relatedness strength of any two Wikipedia articles. It extracts multiple features of the given Wikipedia articles and learns the optimal relatedness scores based on these features. The features used by *Article Comparer* are shown in Figure 3.3.

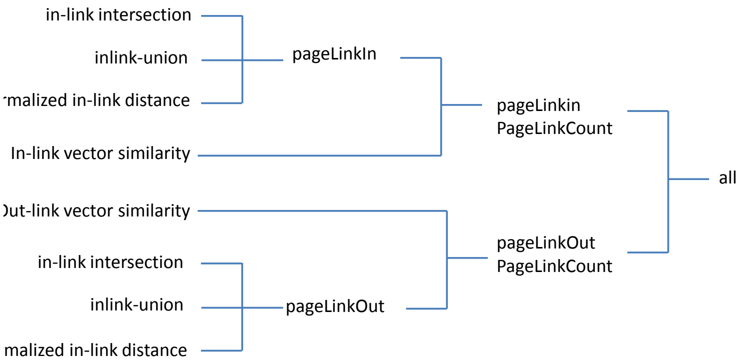


Figure 3.3: A set of Wikipedia features used by *Article Comparer* [140]

Two simple features are the *intersection size* and *union size* of the individual link sets of both Wikipedia articles. The third feature, *Normalized Link Distance (NLD)*, is modeled after *Normalized Google Distance (NGD)* [35] and is based on the in-link set of both Wikipedia articles, given as:

$$NLD(a,b) = \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min|A|,|B|)} \quad (3.3)$$

where a and b are the articles and A and B are sets of all articles that link to a and b respectively. The last feature *Link Vector Similarity (LVS)* is modeled after *Vector Space Model (VSM)* [182] and is based on the out-link set of both articles. It computes a link occurrence-based vector called *lf-iaf* rather than *tf-idf* vectors[6]. The link frequency (lf) signifies the importance of a link in an article and is computed as:

[6] Please see glossary for detail

$$lf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.4}$$

where $n_{i,j}$ is the number of times a_j links to l_i and k refers to the total number of outlinks of a_j. The *inverse article frequency (iaf)* measures the general significance of a link and is given as:

$$iaf = \log\left(\frac{|W|}{|\{a : l_i \epsilon\, a\}|}\right) \tag{3.5}$$

where W is the total number of Wikipedia articles and the denominator refers to the total in-link articles of l_i. The link weights are computed using the product $lf \times iaf$ for each link in the corresponding vectors of input term pair. Finally, the Cosine similarity of both vectors is computed to get the final relatedness score of an article pair. The strength of relatedness of two terms depends on the overlapping strength of the *shared associates* of their corresponding Wikipedia concepts. The association strength of any

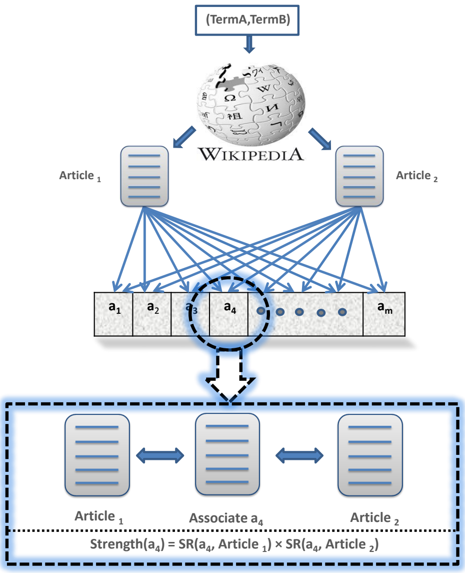


Figure 3.4: The framework for computing the *Overlapping Strength-based Relatedness (OSR)* score for a given term pair.

two terms t_1 and t_2 is computed as follows:

$$OSR(t_1, t_2) = \frac{\sum_{a_s \in C_s} AssocStrength(a_s)}{|L|} \tag{3.6}$$

where $|L|$ represents the total number of links of both concepts and C_s is set of *shared associates*. Equation 3.6 rewards all those links which are important to both input articles. So, the less useful a link is for either input articles, the lower the association strength of that link. The association strength of a link is zero if the link is not related to either of the input articles even if it is very closely related to the other input article. The framework to compute OSR-based association scores is shown in Figure 3.4.

3.3 Mining Informative-Content based Semantics from Wikipedia

A traditional way of computing word associations is to represent the context of individual words in a multidimensional space and compute the distance between their corresponding vectors. This section introduces a new measure of word associations called *Context Profile-based Relatedness (CPRel)* using the informative-content of Wikipedia articles. CPRel improves the vector representation of words by constructing a context profile of each concept corresponding to the input word based on certain features derived from the informative- content of the Wikipedia articles.

3.3.1 Context Profile-based Relatedness (CPRel)

Context Profile based Relatedness (CPRel) exploits the semantic richness of Wikipedia articles by using their informative-content which encode implicit information about the lexical relations (such as synonyms, hyponyms, hypernyms) that are typically present in the Wikipedia articles.

CPRel extracts the informative-content of a Wikipedia article corresponding to an input term. The informative-content includes many words

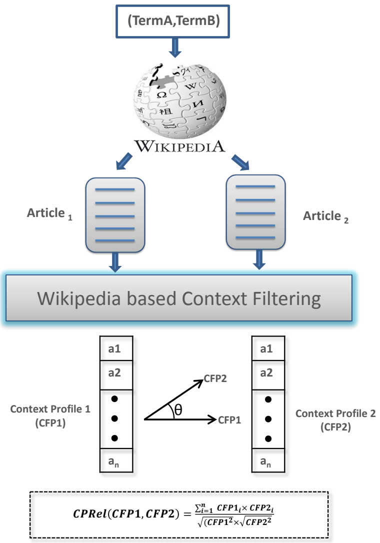

(TermA,TermB)
WIKIPEDIA
Article $_1$
Article $_2$
Wikipedia based Context Filtering
a1
a2
a$_n$
Context Profile 1
(CFP1)
Context Profile 2
(CFP2)
CFP2
CFP1
θ
$CPRel(CFP1, CFP2) = \frac{\sum_{i=1}^{n} CFP1_i \times CFP2_i}{\sqrt{(CFP1^2} \times \sqrt{CFP2^2}}$

Figure 3.5: The framework for computing the CPRel score of a given term pair.

that are contextually related to each input word. It goes through a number of filtering passes to get rid of noisy words and to extract a set of candidate contextual concepts of a given word. The set of candidate concepts is filtered further to get rid of all noisy labels that do not constructively contribute to the context of a given word. The resultant set is referred to as *Context-Filtered Profile* (CFP) of an input word. The weight of each label in a CFP is computed using three weighting schemes based on features extracted from Wikipedia. The filtered context profiles are then compared using Cosine similarity to compute semantic association score of a word pair. The process of computing the semantic associations of words using CPRel is shown in Figure. 3.5.

The Informative-content of both articles are preprocessed to convert them from *MediaWiki* format to plain text. After article matching and the preprocessing phase, the informative-content of each article goes through a number of filtering steps to eliminate unnecessary words. The rest of this section details the context filtering and weighting schemes used for computing semantic associations of words.

Context Filtering

The aim of context filtering is to weed out all common English words that are not helpful in context identification and matching. For this purpose, a list of common English words[7] is used. All words with length less than 3 are also discarded (a heuristic used by [76] as an approximation of stop word removal).

N-grams phrases (up to 5-grams) are extracted from the informative-content of an article. Clearly, many n-grams would be of no help in supporting the context of a given word. Useful n-grams are the ones that represent a semantically meaningful concept. A good source of the phrases that represent meaningful concepts is the set of Wikipedia labels (anchor-text in links), which provide a wide range of different ways of referring to a concept in Wikipedia. These labels not only include Wikipedia article titles but also synonymy, hyponymy and hypernymy. Matching an n-gram with Wikipedia labels could be helpful in judging whether it is useful or not. Therefore the set of n-grams is pruned by removing each n-gram that does not match with a Wikipedia label.

The pruning of n-grams alone is not sufficient because Wikipedia contains such a large collection of articles that some labels are not very useful e.g. *is*,*of* and `for`. In order to find out the usefulness of Wikipedia labels, *Link Probability* (LP) [132] is used. LP is a proven measure to signify *"keyphraseness"* of a word. It is the defined as an estimate of the probability of a keyword being used as a link in Wikipedia. It is computed as the ratio of the number of Wikipedia articles having a keyword as a link to the number of Wikipedia articles in which that keyword occurs in any form (as a link or a word) and is given by,

[7] available at http://www.db-net.aueb.gr/gbt/resources/stopwords.txt

$$P(keyword|w) = \frac{count(Dkey)}{count(Dw)} \tag{3.7}$$

where $count(Dkey)$ represents the number of articles having a word w as a label and $count(Dw)$ is the number of articles in which the word appears. In general, the more generic the Wikipedia label, the lower the link probability. So a label `car` gets a lower LP value (0.01) than a label `sports car`, which in turn gets a low LP value (0.17) than `Ferrari` (0.27). Moreover, the most generic labels get extremely low LP value, such as the label `the` gets an LP value of 8.7×10^{-6}.

CPRel uses Wikipedia labels and the link probability for context filtering, which is performed in two phases: in the first filtering pass all words which are not valid Wikipedia labels are discarded, leaving only those keywords that match with Wikipedia labels; in the second pass, all labels having LP values below a certain cutoff threshold α are discarded. So, $k \epsilon CFP_w$ if $k \in Candidates(w)$ and $LP(k) > \alpha$ where, k is a candidate label and CFP_w is the set of filtered labels representing the context of a word w. All labels with LP values above the threshold α are used to generate the CFP of each input word. A CFP represents the context of an input word in the high dimensional space of Wikipedia labels.

Weighting Schemes

Each label in the CFP of each input word is stemmed and assigned a hybrid weight based on three weighting schemes. After stemming, the root/stem word r of each label is assigned a weight w based on combined weights of its derivational/inflectional word set $\{w_1, w_2....w_n\}$.

- **Term Frequency-based weighting**: Term Frequency (TF) is a common heuristic used to find out the importance of a term in a document. In general, frequently occurring words in an article contribute more towards the context of the article and are considered important.

Based on this assumption, *Normalized Term Frequency (NTF)* is computed as the ratio of sum of frequencies of all the inflectional forms of a root word to the sum of frequencies of all the root words in an article and is given as

$$NTF(r) = \frac{\sum_{i=1}^{k} TF(d_i)}{\sum_{j=1}^{m} TF(r_j)} \tag{3.8}$$

where k represents the total number of derivational words d of a root word r and m represents the total number of root words.

- **Link Estimation based weighting**: In general, *NTF* is good at finding frequently occurring contextual concepts in an article but its not always helpful. Some of the labels may still exist in an article with high frequency count but of not much specificity and relevance. To counter such keywords, *Link Estimation (LE)* of a root word *r* is used to signify the importance of a word as a label in the Wikipedia. If a word frequently occurs as a label in Wikipedia then it is considered a significant keyword. Based on this assumption, *Link Estimation (LE)* is defined as the ratio of sum of link article count (number of articles where the word occur as a link, $count(Dkey_{w_i})$) of each inflectional form to the sum of total article count (number of articles where a word occurs at all, $count(Dw_{w_i})$) of them. *LE* is computed as below:

$$LE(r) = \frac{\sum_{i=1}^{k} count(Dkey_i)}{\sum_{i=1}^{k} count(Dw_i)} \tag{3.9}$$

where *k* represents the number of inflectional forms of a root word. This measure penalizes all unwanted common words which succeed in passing through the *stop word filter* and the *label filter* and have high *NTF*.

- **Hybrid weighting**: Previous two weighting schemes are combined to compute the weight of all labels in each CFP.

$$w(r) = \frac{2 \times LE(r) \times NTF(r)}{LE(r) + NTF(r)} \tag{3.10}$$

> This hybrid measure creates a balance between local significance of a term within a Wikipedia article and its global importance as a label in Wikipedia encyclopedia. Any label having high normalized term frequency but low link estimation will be penalized and vice versa.

Finally, the context of each input word is represented in a high dimensional space of Wikipedia labels. The deviation of angle between the two $CFPs$ is computed as the Cosine similarity represents the extent of association between the words $w1$ and $w2$ and is computed as follows.

$$CPRel(w1, w2) = \frac{\sum_{i=1}^{n} CFP1_i \times CFP2_i}{\sqrt{\sum_{i=1}^{n}(CFP1_i)^2)} \times \sqrt{\sum_{i=1}^{n}(CFP2_i)^2)}} \tag{3.11}$$

where $CFP1$ is the context profile of the word $w1$ and $CFP2$ is the context profile of the word $w2$. The Cosine of $0°$ is 1, which means the two words are the same. Other values of Cosine similarity below 1 refer to various degrees of association between the given words.

3.4 Domain-Independent Evaluation

For performance evaluation of the semantic association measures detailed in the previous section of the chapter, the direct evaluation method is adopted using both Pearson's and Spearman's correlation metrics. Three datasets are used in the experiments: M&C, R&G and WS-353 datasets (detailed in Section 2.3 of the thesis). Some word pairs in the datasets did not have corresponding Wikipedia articles and are referred to as *missing* word pairs in the experiment. Other word pairs that match with corresponding Wikipedia articles are called *non-missing* word pairs. *Missing* word pairs are problematic for the new measures since the measures depend on Wikipedia articles corresponding to each word. To cope with such word pairs, two versions of each dataset were created: datasets with all term pairs; and datasets with only *non-missing* term pairs (removing missing term pairs from each dataset resulted in 24 word pairs in M&C, 58 word pairs in R&G and 314 pairs in WS-353 datasets). Since the semantic measures assign zero scores to missing word pairs, the experiments used *Wikipedia Link-based Measure* (WLM) [138] for computing the scores of missing term pairs. Therefore, the performance of the semantic association measures were investigated in three experiments. The first experiment applied the semantic measures to all term pairs (assigning zero weights to missing term pairs); the second experiment applied the approaches to *non-missing* term pairs only (to analyze the actual performance of the semantic measures); and the third experiment applied the measures combined with WLM (used for scoring the missing term pairs) on all term pairs.

Many terms in each dataset are ambiguous and have multiple meanings. Ambiguity is certainly a non-trivial issue and can greatly affect the performance of a semantic association measure. A wrong disambiguation can lead to misleading correlations with manual judgments on a term pair. To address this issue and analyze the performance of the semantic measures from the perspective of association computation, term pairs in each dataset were manually matched to the corresponding Wikipedia articles[8]. However, to analyze the impact of disambiguation on the overall performance of a semantic measure, later experiments have also compared the performance of the presented semantic measures on both manually-disambiguated and automatically-disambiguated versions of each dataset.

For association computation using a semantic measure, the version of Wikipedia released in July 2011 is used. It contains 33GB of uncompressed XML markups, which corresponds to more than four million articles. To easily draw upon the content of Wikipedia, the latest version of *Wikipedia-Miner Toolkit* is used [140].

[8] http://www.nzdl.org/wikipediaSimilarity/

3.4.1 Statistical Significance Test

For computing the correlation and testing the statistical significance, the IBM SPSS Statistics 20 toolkit is used. For finding out whether the association of the automatically computed scores with the manual judgments is statistically significant, SPSS default hypothesis test is used. The hypothesis can be explained in two different ways depending on whether the significance test is two-tailed or one-tailed.

- The hypothesis for the two-tailed significance test is as follows:
  - Null Hypothesis $H_0 : \rho = 0$ the population correlation coefficient is 0 (no association)
  - Alternate Hypothesis $H_1 : \rho \neq 0$ the population correlation coefficient is not 0 (the association is statistically significant)
- The hypothesis for the one-tailed significance test is as follows:
  - Null Hypothesis $H_0 : \rho = 0$ the population correlation coefficient is 0 (no association)
  - Alternate Hypothesis $H_1 : \rho > 0$ the population correlation coefficient is greater than 0 (the association is positive) or
  - Alternate Hypothesis $H_1 : \rho < 0$ the population correlation coefficient is not 0 (the association is negative)

where ρ is the population correlation coefficient. If a specific directional association[9] between the variables (in comparison) is not hypothesized then the two-tailed significance test is used. By default SPSS computes statistical significance at alpha=0.01 and alpha=0.05. The two-tailed statistical significance hypothesis is used in the experiments of Chapters 3 to 6.

[9] please see Section 2.4.1 for details on directionality

3.4.2 Results and Discussion

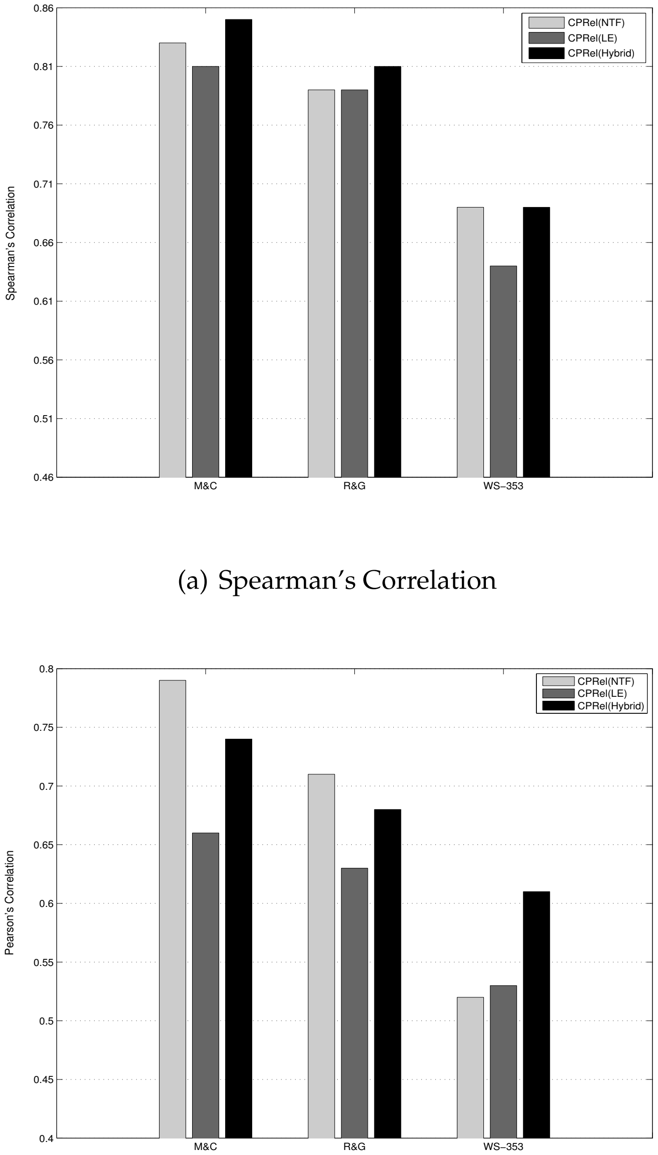
This section details the empirical analysis of Wikipedia-based semantic measures and compares the performance of the association computation measures based on Wikipedia structure and informative-content–CPRel, WikiSim and OSR. The bold values in each following table indicate the best correlation-based performance on a specific dataset.

The first experiment analyzed the performance of CPRel using three weighting schemes for computing semantic associations to select the best performing weighting scheme: CPRel-NTF, CPRel-LE and CPRel-Hybrid. Both Spearman's correlation (ρ) and Pearson's correlation (r) are used to compare automatically computed results with human judgments on all datasets.

Figure 3.6 compares the performance of the three variants of CPRel on *non-missing* versions of all datasets. It is clear from the figure that CPRel-Hybrid surpassed the other two methods on both M&C and R&G datasets and performed best equal with CPRel-NTF on the WS-353 dataset (on Spearman's correlation (ρ)). These results show that a combination of normalized term frequency and link estimation is a good indicator of semantic associations. Hence, we decided to use CPRel-Hybrid as the informative-content based measure in later experiments.

In the next experiment, three measures based on Wikipedia hyperlink structure and informative-content are compared. Table 3.1 shows the correlation-based performance of the presented measures on *all-terms* versions of three datasets.

Clearly, WikiSim outperformed the other two measures on all datasets using Spearman's correlation and on the M&C dataset using Pearson's correlation. On the other hand, CPRel-Hybrid outperformed the other two measures on R&G and WS-353 datasets using Pearson's correlation. However, the correlation-based performance of the presented semantic measures on all datasets is not high in general. The reason for these low correlation values is that all the semantic measures assigned zero score to a term



(a) Spearman's Correlation

(b) Pearson's Correlation

Figure 3.6: Performance comparison of three variants of CPRel on *non-missing* versions of three datasets: M&C, R&G and WS-353 datasets.

Table 3.1: Performance comparison of the semantic measures on *all-terms (All)* versions of M&C, R&G and WS-353 datasets.

Aspect	Approach	Correl.	Datasets		
			M&C (All)	R&G (All)	WS-353 (All)
Structure	WikiSim	ρ	**0.78**	**0.77**	**0.64**
		r	**0.58**	0.60	0.38
	OSR	ρ	0.74	0.73	0.63
		r	0.51	0.54	0.35
Contents	CPRel-Hybrid	ρ	0.66	0.61	0.59
		r	0.54	**0.69**	**0.55**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

pair if either of the terms did not match with its corresponding Wikipedia article. However, this does not reflect the actual performance of the semantic measures. In order to analyze their actual performance, the semantic measures were compared on non-missing versions of all datasets, as shown in Table 3.2.

When the performance of the three Wikipedia-based semantic measures on the non-missing version of each dataset was investigated, a noteworthy rise in the corresponding correlation values was observed on all datasets as shown in Table 3.2. Again, WikiSim outperformed the other two approaches on all datasets using Spearman's correlation while CPRel performed the best on all three datasets using Pearson's correlation. This shows that the presented measures are actually good at estimating the semantic association strength as long as the input words match with corresponding Wikipedia articles.

In order to report the performance of the semantic measures on *all-terms* versions of all datasets, a hybrid approach was followed for computing the automatic scores, which computed the semantic association

Table 3.2: Performance comparison of the semantic measures on *Non-Missing (NM)* versions of M&C, R&G and WS-353 datasets.

Aspect	Approach	Correl.	Datasets		
			M&C (NM)	R&G (NM)	WS-353 (NM)
Structure	WikiSim	ρ	**0.87**	**0.87**	**0.70**
		r	0.71	0.65	0.44
	OSR	ρ	0.84	0.84	0.69
		r	0.63	0.58	0.33
Contents	CPRel-Hybrid	ρ	0.85	0.81	0.69
		r	**0.74**	**0.68**	**0.61**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

Table 3.3: Performance comparison of semantic measures with WLM measure on *all-terms (All)* versions of M&C, R&G and WS-353 datasets.

Aspect	Approach	Correl.	Datasets		
			M&C (All)	R&G (All)	WS-353 (All)
Structure	WikiSim+WLM	ρ	**0.83**	**0.83**	**0.66**
		r	0.62	0.62	0.41
	OSR+WLM	ρ	**0.83**	0.80	0.64
		r	0.64	0.68	0.49
Contents	CPRel-Hybrid+WLM	ρ	**0.83**	0.74	0.62
		r	**0.84**	**0.77**	**0.57**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

scores of missing term pairs using WLM measure [138] and combined them with scores of each semantic measure on the *all-terms* version of each dataset. The results of this experiment are reported in Table 3.3. Clearly, on *all-terms* versions of all datasets, a consistent behavior of presented measures was observed. The Spearman's correlation of WikiSim+WLM was again highest when compared with other two measures on all three datasets. Similarly, on Pearson's correlation CPRel+WLM was found to top the other two approaches on all three datasets.

3.4.3 Impact of Disambiguation

In order to investigate the impact of term disambiguation on the overall performance of a semantic measure, another experiment compared the performance of the semantic association measures on two versions of all datasets: one version with manually disambiguated terms; and another version with automatically disambiguated terms. For automatic disambiguation, the *Label disambiguator* of Wikipedia-Miner toolkit [140] was used. *Label disambiguator* is a machine learning algorithm that takes a two words as input and generates their best-related sense pair as output. Depending on the disambiguation confidence, the output of the disambiguator can be one or more sense pairs. The experiment used only the top-ranked sense pair for each input term pair. For a fair comparison, all term pairs that could not be matched with corresponding Wikipedia articles were excluded from both versions of each dataset. The result of this comparison is shown in Table 3.4.

A cursory look at the table shows that the performance of the semantic measures on the task of semantic relatedness computation was adversely affected when the automatic disambiguation was used. Analysis of the results revealed that the automatic disambiguator assigned many sense pairs which were very different from the original term pair. For instance, for the term pair (`Crane`, `Implement`), a sense pair (`Crane(machine)`, `List of agricultural machinery`) was returned as the output sense pair by *Label disambiguator*, which resulted in a different semantic association score. Consequently, the automatic disambiguation produced low

Table 3.4: Performance comparison of the semantic measures on M&C, R&G and WS-353 datasets using manual and automatic disambiguations.

Aspect	Approach	Correl.	Automatic Disambiguation			Manual Disambiguation		
			M&C	R&G	WS-353	M&C	R&G	WS-353
Structure	WikiSim	ρ	0.75	**0.77**	**0.67**	0.80	**0.90**	0.70
		r	0.59	0.55	0.35	0.68	0.64	0.39
	OSR	ρ	**0.78**	0.76	0.64	**0.83**	**0.90**	**0.71**
		r	**0.73**	**0.67**	**0.54**	0.80	**0.81**	**0.61**
Contents	CPRel-Hybrid	ρ	0.73	0.67	0.59	0.82	0.81	0.66
		r	0.72	**0.67**	**0.54**	**0.82**	0.80	0.60

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

correlation values on each dataset when compared with that of the manual disambiguation. This also indicates that the automatic disambiguation of term pairs still needs improvement. This is yet an open research topic to explore. There is a big performance difference between the automatic and manual disambiguations. Hence, the performance comparison of the Wikipedia-based semantic association measures was based on the manual disambiguations. Please note that these measures can also be used with automatic disambiguation.

3.4.4 Further Analysis

The correlation of the Wikipedia-based semantic measures on WS-353 was not as high as on the other two datasets. There are three main reasons for the performance drop on the WS-353 dataset. First, on analyzing the dataset, it was found that most of word pairs in this dataset share a context only when put together but are not semantically related otherwise. For instance, in the word pair `(secret,weapon)`, both words have different contexts, which might not be strongly related but when put together as `secret weapon`, both words are strongly related in a new context. A further analysis revealed that there were 90 word pairs in this dataset which fall into this category. For such word pairs, the presented measures did not produce realistic association scores. Second, the words having collocational relations such as `(Disaster,Area)` get a high score only if they co-occur within a proximity window i.e. if they are placed lexically close in the text. However, all the presented Wikipedia-based semantic measures focused on the semantic closeness which does not require the words to co-occur in close proximity, thus yielded low correlations. Third, the focus of WS-353 dataset is on both semantic similarity as well as on semantic relatedness [228]. Semantic similarity is easier to predict than relatedness which is considered broader as well as more complex than semantic similarity [23]. In this particular dataset, even humans have low agreement on many word pairs and assigned different scores to the same word pair. There are cases where the inter-rater agreement on this dataset falls below 0.65 as compared to the M&C and R&G datasets where the inter-rater agreement remained between 0.88 and 0.95 [204].

The presented Wikipedia-based semantic measures are computationally inexpensive as none of them require extensive pre-processing of the whole knowledge source. However, an obvious limitation is the mapping of each input term to the corresponding Wikipedia article, which might not always be available. Existing knowledge sources suffer from coverage limitation to varying extents. However, due to the continuously growing nature of Wikipedia, knowledge acquisition from Wikipedia will hopefully not be a major issue in future. By extending the context of given terms to the corpus level, this mapping can be avoided and this is addressed in Chapter 6 of the thesis.

The presented research capitalizes on word-to-word semantic association computation; however, with small modifications, the presented measures could be adapted to the task of association computation of large textual units such as sentences and documents.

3.4.5 Similarity Vs. Relatedness

The annotation guidelines for the WS-353 dataset did not distinguish between similarity and relatedness as mentioned in [1]. Evaluating semantic measures on WS-353 dataset, which is a combination of both kinds of associations, is problematic because different semantic association measures are appropriate for measuring similarity or measuring relatedness. Hence, to address this issue, Agirre *et al.* [1] reused the existing human judgments on the WS-353 dataset and created two new gold standard datasets: WS353-Similarity, consisting of 203 term pairs (unrelated term pairs and similarity term pairs); and WS353-Relatedness with 252 term pairs (unrelated term pairs and related but not similar term pairs). To analyze the impact of similarity and relatedness on the performance of the semantic measures, we experimented with both subsets of the WS-353 dataset and reported the results in Table 3.5 and 3.6. In the reported results, WS353-Similarity is represented as WS353-(S) and WS353-Relatedness is represented as WS353-(R).

Table 3.5 shows the performance comparison of the semantic measures on *all-terms* versions of both subsets. On the *all-terms* version of WS353-(S), OSR outperformed the other two measures on Spearman's correlation, whereas CPRel performed the best on WS353-(S) using Pearson's correlation. In general, all three approaches produced higher correlations on the similarity-based dataset than on the relatedness-based dataset, which indicates the intrinsically complex nature of semantic relatedness.

Table 3.6 indicates the performance comparison of the semantic measures on the *non-missing* versions of both subsets of the WS-353 dataset. Since the *non-missing* version of every dataset represents the actual performance of each approach, we have reported the results in a separate table. Clearly, on the *non-missing* version of the WS-353-(S) dataset, OSR outperformed the other two approaches using the Spearman's correlation and CPRel-Hybrid showed best performance using the Pearson's correlation. On the WS353-(R) dataset with *non-missing* term pairs, WikiSim outper-

Table 3.5: Performance comparison of the semantic measures on *all-terms* versions of similarity (WS-353-(S)) and relatedness (WS-353-(R)) based subsets of the WS-353 dataset.

Aspect	Approach	Correl.	Dataset	
			WS-353-(R) (All)	WS-353-(S) (All)
Structure	WikiSim	ρ	0.59	0.69
		r	0.49	0.66
	OSR	ρ	**0.63**	**0.73**
		r	**0.52**	0.61
Contents	CPRel-Hybrid	ρ	0.51	0.67
		r	0.47	**0.67**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

formed the other two approaches using both Spearman's and Pearson's correlations.

3.5 Domain-Specific Evaluation: Biomedical Data

Section 3.4 of the chapter presented a domain-independent evaluation of the Wikipedia-based semantic measures as all datasets used in the previous experiments were general-English based word pairs. However, the number of datasets used for direct evaluation of semantic relatedness computation approaches is limited. In order to extensively investigate the performance of the Wikipedia-based measures, the following experiment used domain-specific datasets. The purpose of this experiment is two fold: first, to analyze the effectiveness of the semantic association measures on domain-specific data; and second, to investigate semantic capabilities of Wikipedia on the task of domain-specific term relatedness.

Table 3.6: Performance comparison of the semantic measures on *Non-Missing (NM)* versions of similarity (WS-353-(S)) and relatedness (WS-353-(R)) based subsets of the WS-353 dataset.

Aspect	Approach	Correl.	Dataset	
			WS-353-(R) (NM)	WS-353-(S) (NM)
Structure	WikiSim	ρ	**0.68**	0.75
		r	**0.57**	0.68
	OSR	ρ	0.66	**0.76**
		r	0.56	0.62
Contents	CPRel-Hybrid	ρ	0.57	0.72
		r	0.51	**0.70**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

3.5.1 Datasets

The following biomedical datasets are used in the domain-specific evaluation of the semantic measures.

- **MiniMayo dataset:** A biomedical dataset consisting of 30 medical term pairs annotated by 3 physicians and 9 medical index experts [160] on a four point scale: practically synonyms, related, marginally related and unrelated. The difference of judgments between medical experts and physicians stems from the professional training and activities of the two groups. Medical experts are trained to use the hierarchical classifications of medical concepts while physicians are trained to diagnose and treat patients. The reported inter-rater agreement of physicians' scores was 0.68, whereas that of experts was 0.78. The agreement across both groups was found to be 0.85.

- **MeSH Dataset:** A biomedical dataset [85] consisting of 36 term pairs derived from MeSH ontology, which is a taxonomic hierarchy of medical concepts. Scores of 8 human experts on 36 MeSH term pairs were averaged to provide similarity scores on a scale of [0 - 1].

- **MayoMeSH Dataset:** Both previously mentioned domain-specific datasets are combined to generate a dataset of 65 medical term pairs. Humans scores on both datasets were normalized on a scale of [0 - 4] where 0 means unrelated and 4 means exactly the same. This dataset is created for performance analysis of the semantic measure on a larger set of biomedical term pairs.

All term pairs in MeSH and MiniMayo datasets matched with their corresponding Wikipedia articles, hence the *non-missing* and WLM versions of these datasets were not needed in the experiments.

3.5.2 Results and Discussion

Following domain-independent evaluation, we used the Spearman's rank-order correlation coefficient (ρ) and the Pearson's correlation coefficient (r) to compare our results with human judgments on these datasets.

Table 3.7 shows the performance of the Wikipedia-based semantic measures on biomedical datasets. Overall, WikiSim performed consistently well on all datasets using both Spearman's and Pearson's correlations. On the MeSH dataset, CPRel outperformed the other two measures using both correlation metrics. However, on other four datasets, WikiSim outperformed the other two measures using Spearman's correlation. On both versions of MiniMayo and MayoMeSH datasets, an opposite trend was seen in the correlation of the structure-based and the informative-content based approaches. The results of this experiment show that the structure-based approaches correlated well with physicians' judgments whereas the content-based approach produced higher correlation with medical experts.

Table 3.7: Performance of the semantic measures on three biomedical datasets: MiniMayo, MeSH and MayoMeSH datasets.

Aspect	Approach	Correl.	Datasets				
			MeSH	MiniMayo	MiniMayo	MayoMeSH	MayoMeSH
				(Experts)	(Physicians)	(Experts)	(Physicians)
Structure	WikiSim	ρ	0.84	**0.79**	**0.79**	**0.80**	**0.78**
		r	0.84	0.67	**0.82**	**0.81**	0.79
	OSR	ρ	0.74	0.66	0.73	0.72	0.71
		r	0.60	0.65	0.64	0.62	0.64
Contents	CPRel-Hybrid	ρ	**0.85**	0.78	0.72	0.77	0.75
		r	**0.87**	**0.85**	0.79	0.80	**0.80**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

We expected structure-based measures to correlate better with the experts' judgments than with the physicians because the experts' judgments are essentially based on their knowledge of hierarchical structure. Despite having a well-organized structure, Wikipedia hyperlinks are not based on tightly-bounded lexical relations as arranged in classical taxonomies. It is based on linking those Wikipedia articles that share some related context, hence offers far more rich semantics than the classical structures such as WordNet taxonomy. We hypothesize this to be the reason for better correlation of the structure-based association measures with the physicians' judgments. However, further experimentation is required to support this argument. Despite the larger size, the overall best performance is achieved on the MeSH dataset using both Spearman's and Pearson's correlations because this dataset includes term pairs having classical lexical and semantic relations between the terms such as `Migraine` is a hyponym of `Headache`. The results of experiments have also shown that Wikipedia is a valuable knowledge source not only for computing general-English based terms but also for computing relatedness of biomedical terms.

3.5.3 Further Analysis

In order to understand the performance difference of the presented measures on both correlation metrics i.e. Spearman's correlation coefficient and Pearson's correlation coefficient, the scatter plots of the performance of all three measures on the MeSH dataset were drawn and compared as shown in Figure 3.7. It can be seen in the Figure 3.7 that there was approximately linear association between their automatically produced scores and the manual scores which resulted in very strong Pearson's correlation values. However, the presence of an outlier in the score produced by the OSR measure pulled down the value of Pearson's correlation to 0.59 which is much lower than the Pearson's correlation values of the other two measures. These results resonate with the discussion on the effect of outlier on the performance of Pearson's correlation coefficient (Please see Section 2.4.1 for details). This analysis also reveals that the reason of low performance of any measure on Pearson's correlation in general and the OSR measure in specific is the linearity assumption of the Pearson's correlation.

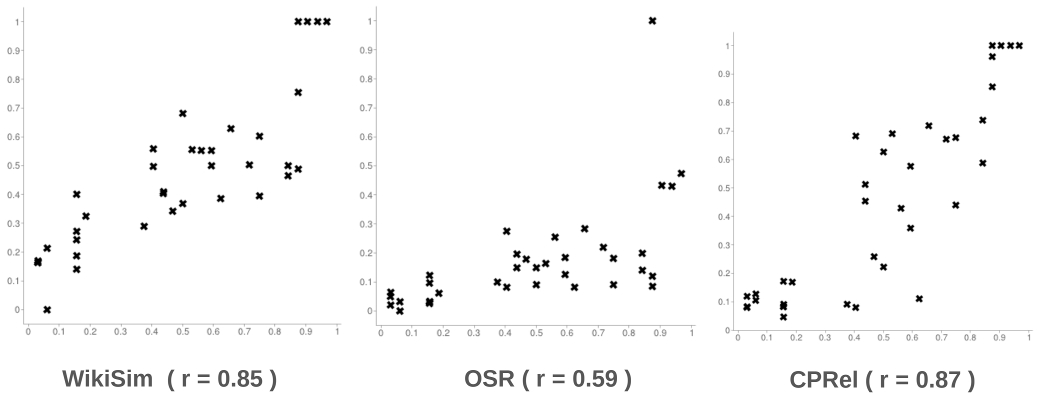


Figure 3.7: Effect of linearity on the performance of three measures on MeSH dataset.

3.6 Chapter Summary

This chapter addressed the problem of semantic association computation based on semantic mining of Wikipedia. The structural and informative-content based features of Wikipedia were used in formulating new measures for semantic association computation of words. This chapter highlights two main contributions of this work. First, an exploitation of semantic capabilities of Wikipedia, based on developing three association measures: *WikiSim*, a Wikipedia hyperlink structure based combinatorial measure, which uses the overlap of *shared associates* as an indicator of semantic associations; *OSR*, a word association measure that uses the semantic orientation of shared associates between two Wikipedia articles for computing the semantic association strength, and is based on the combinatorial model of associations; and *CPRel*, a vectorial association measure based on features extracted from encyclopedic informative-content of Wikipedia articles. The second main contribution is the exploitation of the coverage and semantic richness of Wikipedia as a knowledge source on both domain-independent as well as domain-specific word association computation task. The experimental analysis reveals that each semantic measure has its own strengths and weaknesses. For instance, WikiSim is an association measure based on the combinatorial model of associations. Similarly, the OSR measure mines the semantics of shared concepts by computing the semantic orientation of shared features. This measure is intriguing because it does not considers all features equally useful and rewards only those features that are related to both input words. This feature can be used for context filtering by those approaches which are based on feature overlap sets. Finally, the informative-content-based approach exploits the detailed encyclopedic-information as the context of a specific word from its corresponding Wikipedia article. This is a vectorial measure that implicitly covers the in-links along with the extra semantics mined from the informative-content of an article. CPRel-hybrid performed consistently well on all datasets using Pearson's correlation, which gives an insight into the performance behavior of this measure. The research in this chapter has empirically shown that Wikipedia is an invaluable knowledge source for the semantic association computation task and can be effectively used as for the association computation of biomedical terms.

Chapter 4

Directional Context Helps!

The work described in this chapter explores how the process of semantic association computation can be guided by the idea of asymmetric associations from the free word association task[1]. The chapter presents an investigation and evaluation of the assertion that asymmetric associations can be useful for computing semantic relatedness of words. Section 4.1 introduces the concept of word associations and their applications in a multitude of domains. Section 4.2 details the motivation for adapting semantics from the free word association task. Section 4.3 presents a new semantic association measure based on directional context mining. Section 4.4 discusses the evaluation setup and Section 4.5 reports experimental results on the performance evaluation of the new measure and presents a detailed analysis of the results. This also includes the experiment on analyzing the performance of the semantic measure on asymmetric association computation and comparing the automated asymmetric associations with humans asymmetric associations to investigate the focus of both. Finally, Section 4.6 concludes the chapter with a brief summary.

[1] The idea of *free word associations* discussed in this chapter is different from the generic term *semantic word association*. The notion of word association (explicitly differentiated from similarity and relatedness based on the idea of asymmetry) is used in this chapter as a fine grain concept (defined in Section 4.1 of the chapter).

4.1 Word Associations

The human brain is a natural classifier of concepts into their respective conceptual sub-spaces. It associates ideas and concepts based on past experiences and produces relationships between the objects and phenomena of the real world using these associations. There are various types of human associations (each linked with a specific sensory organ) such as visual, gustatory, olfactory, tactile and auditory associations [56]. All these types are inter-connected, hence stimulation of any of these senses results in associations and vice versa. For instance, if a person sees an object, he immediately associates the name of that object with its observed characteristics. Similarly when he reads or hears the name of the object, he quickly recalls the characteristics of that object. As an example, when a person looks at a lemon, he associates the word `lemon` with some physical attributes such as `yellow color` and `round shape`. When he tastes the lemon, he associates the word `lemon` with `sour taste`. Later, when he reads or hears the word `lemon` somewhere, he might feel the `tangy taste` of lemon in his mouth or visualize the `yellow color` and `round shape` immediately.

Word association—a type of human associations— is the stimulation of an associative pattern by a word[2]. It is one of the basic mechanisms of memory and has been widely researched to investigate the underlying lexical-semantic models of human memory [202]. Word association research emerged as a psychological science with Francis Galton [59], who believed that there might be a link between a person's Intelligence Quotient (IQ) and his word associations. Since then, it has been studied in a range of disciplines.

Nelson *et al.* [149] defined free word associations–a special case of word associations–as a task that requires participants to produce the first word, that comes to their mind that is related in a meaningful way to a presented

[2] http://dictionary.reference.com/

word. This procedure is also called *discrete associations* task, as the participants are required to produce a single response word for each given word. In free word association tests, the given word is referred to as the *stimulus* or *cue* word and the resulting word is called *response* word. The response could be a single word or a list of many words associated with a single *cue* or *stimulus* word.

4.1.1 Applications of Word Associations

The idea of word associations is used in different disciplines including cognitive psychology, computational linguistics and artificial intelligence. This section gives an overview of various applications of word associations.

The earliest work in Psychology using word associations was on how to model the behavior of subconscious mind by Francis Galton [59], who investigated the mental imaginary to find out a relation between a person's intelligence and his word associations. Word associations are extensively used in psychology for personality prediction and assessment [50, 128, 46, 141]. The associative responses of the human subjects were interpreted by the principle of learning by contiguity. According to this principle, objects once experienced together tend to be linked to each other in the human imagination, so that if one of the objects is thought of then the other object is likely to be thought of also [218]. Blueler [15] identified the association trouble in schizophrenia as a fundamental symptom leading to a number of secondary disorders. In a follow-up study, the semantic association trouble in schizophrenia was explored through a computational semantic network as a conceptual tool to compute inter-word links and to make queries about different semantic levels of the responses of schizophrenic patients.

In cognitive psychology, Teversky and Hemenway [207] used word associations to generate the *environmental scenes*, which refer to the settings in which human actions occur. They developed a taxonomy of various kinds of *environmental scenes* and identified the basic level as well as the most useful levels of the taxonomy for other domains of knowledge concerned with the environment. A similar research used word association tests to generate an ontology of geographic categories [190]. They founded their experiments on the Prototype theory [177], which states that for most people, some objects are better representative of a category than others and there is high inter-rater agreement of human subjects on what constitute a good or bad example. For instance, `chair` is a better representative of the category `furniture` than `stool`. They assessed the human cognitive categories, which constitute of a radial structure with central members surrounded by a periphery of less typical members.

The free word association test is also used as a tool for constructing lexical repositories such as ontologies, taxonomies and thesauri [151, 192]. The focus of such approaches was on eliciting the most frequently associated words to a given stimulus word for the purpose of generating term hierarchies, which reflect the behavior of end-users in organizing vocabulary around a central concept. WordNet, an online lexical reference system, was inspired by psycholinguistic theories of human cognition and was designed to aid in searching dictionaries conceptually rather than just alphabetically [135]. For constructing WordNet, English speaking participants were asked to tell the first word they thought of, in response to highly familiar words drawn from different syntactic categories such as verbs, nouns, adjectives and adverbs. Their responses were then used to represent various lexical relations of the concepts under each category.

The word association methodology has been found useful in identifying the use of the language of a specific user group. Nielson [153] used the word association method to improve user interaction and access to information retrieval system. He aimed at making information retrieval more intuitive and user-friendly by adapting to the user's search behavior. He used the word association methods to catch colloquial vocabulary of end-users and integrated this colloquial vocabulary and the word relations according to the specific language of the user. The word association test was also incorporated into the design of the hierarchical displays in information retrieval systems [192]. This theoretical framework was based on analyzing the cue-response inter-relations and incorporating them within the thesaurus display. In another research, Nielson [152] evaluated the user associations in the construction of a *corporate thesaurus*, which is a special kind of thesaurus developed according to the specific needs of a work context. He argued that word association test is a valuable method to identify a set of terms related to the mental model of the employees within the work domain and can be used to construct the corporate thesaurus.

Based on the assumption that the ability to associate concepts leads to creativity, word association has also been used in the task of computational creativity. Toivonen *et al.* [203] focused on the task of automatic poem generation based on linguistics corpora. They used a background corpus for generating word association network to control semantic coherence and topics. Klebanov and Flor [102] generated word association profiles to assess the quality of writing and used these profiles to improve a system for automatically scoring essays. They identified classes of word pairs in the content vocabulary based on their associative strengths. Ferret [52] used word associations to perform topic analysis for identifying topics in a text, delimiting topic boundaries and identifying the relations between various segments. He combined the word repetition with lexical cohesion of a collocation network for topic segmentation and link detection.

Another domain effectively utilizing the concept of word associations is association rule mining [82]. Association rule mining aims at discovering strong associations of item-sets in a database using different measures of interestingness [3]. An association rule is an expression that involves two item sets X and Y such that $X \implies Y$. This rule expresses that whenever a transaction T, in a database D, contains X than T probably contains Y also. Agarwal *et al.* [3] proposed a method for mining association rules

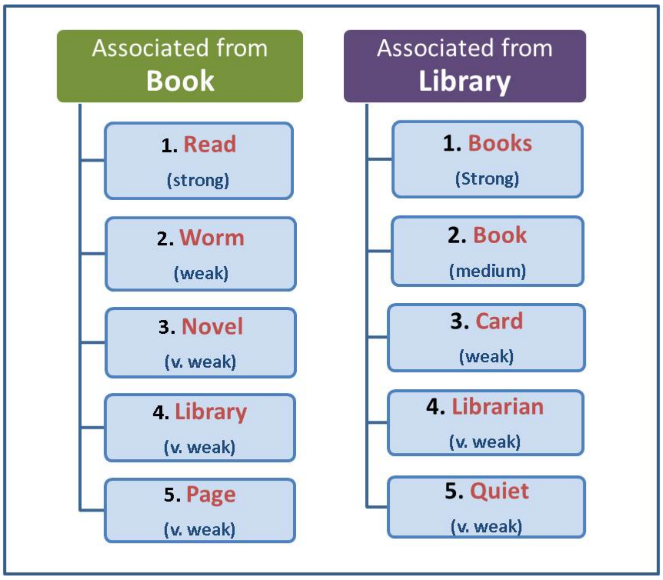


Figure 4.1: A comparison of top five *Response words* of given *Cue words* `Book` and `Library`.

from a large database of customer transactions. Tamir and Singer [200] proposed a new interestingness measure inspired by human word associations and used it for association rule discovery and scoring.

4.2 Semantics of Free word Associations

The semantics of free word associations could be useful for computing the strength of semantic association of words. Consider two *stimulus* words `book` and `library`[3]. Figure 4.1 presents the top five *response words* produced by human participants and ordered by the strength of their associations with the *stimulus* words. It is interesting to note that the word `Library` occupies rank 4 in the *response word* list of the stimulus `book`, whereas in the *response word* list of the stimulus `library`, the top two

[3] These statistics are collected from a huge database of free word associations on `http://wordassociation.org/`

ranks are occupied by the words `book` and `books`. This example demonstrates certain properties of word associations, which we have identified from the discrete association task to build a new model of association computation.

- **Impact of Context:** The same word pair assumes different association strengths in different contexts. For instance, in the context of *study*, the association strength of a word pair `book` and `library` is higher than the word pair `book` and `bookshop`, whereas in the context of *publishing and marketing*, the association strength of a word pair `book` and `bookshop` is higher than the word pair `book` and `library`. In this example, two different contexts lead to two different association strengths. However, this multiplicity of contexts is not limited to only two contexts and could lead to multiple and potentially different association strengths for the same word pair.

- **Directional Context and Asymmetry:** In the absence of an explicit context, a *cue word* is used to determine the context in which the association of a cue-response pair is determined. For a given word pair (`X`, `Y`), when the *cue word* is `X` and *response word* is `Y`, the association of `X` and `Y` is computed in the semantic space of the word `X`. On the other hand, when the *cue word* is `Y`, the association of `X` and `Y` is computed in the semantic space of the word `Y`. Hence, it is the cue word that provides a directional context, in which the association of a word pair is computed. This means that the association strength is asymmetric if words are presented in order such that the first word determines the context.

4.2.1 Guiding Semantic Association Computation

In conventional approaches to semantic relatedness computation, the relatedness of two words is assumed to be symmetric, such that $rel(a, b)$ =

$rel(b, a)$, where a and b are two given words. This assumption essentially disregards the contexts in which these words could be associated, hence considers their relation in a context-independent way. In real scenarios, such as in web document retrieval and microblog clustering, where the context is quite diverse and changes rapidly, computing realistic scores of a word pair according to the topical context in which it appears, is critical. Hence, context consideration is important for computing semantic relatedness.

There are three types of contexts: *local context*, *topical context* and *directional context*. A set of words that occur in the proximity of a given word is called its *local context*. This *local context* is taken into account on document level. There is a stream of research on semantic relatedness which used this local context to compute semantic word associations [104, 181, 91, 116]. The second type of context, *topical context*, refers to an explicit and wider context in which the strength of association of two given words is computed. For instance, when relating two words `cloud` and `rain` in the context of *weather forecasting*, the word pair assumes a different association than when relating them in the context of *Sunbathe* or *picnic*. The third type of context, directional context, assumes that the topical context is determined by the first word of a given word pair in the absence of an explicit topical context. The task of measuring word relatedness involves presenting a word pair without any explicit context to the relatedness computation approach. Thus, following free word associations task, we use the individual words of a given word pair as stimuli for determining the directional contexts in the absence of a given topical context. The focus of this research is on using the idea of directional context for computing the semantic relatedness of a given word pair. Since the context is derived for each input word in a given word pair, the association of a word pair is based on only two directional contexts.

Inspired by the asymmetry in humans associations, this chapter presents a novel approach to computing semantic relatedness guided by the idea of directional contexts. The approach is unique because it adapts the idea of asymmetric associations and uses it for semantic association computation, which (to the best of our knowledge) has not been explicitly done before[4].

The directional context identification is important for finding association of two terms. To extract the directional context of each input word, Wikipedia is used as a knowledge source. Context mining is based on Wikipedia hyperlink network, which encodes not only many lexical relations but also many sophisticated semantic relations such as *cause-effect* and *functional associations*. The chapter presents a new approach to computing the semantic association strength of a given word pair in the directional context of each input word and combines these asymmetric association strengths into a symmetric relatedness score. It is worth mentioning that we compute the asymmetric associations of terms differently from the free word association task. The difference lies in the use of well structured and semantically rich context rather than just considering the usage-based collocation statistics for word association computation.

4.3 Methodology

Figure 4.3 presents the framework for computing word relatedness using Wikipedia-based asymmetric associations. It is clear from the figure that the approach for computing word relatedness is divided into three phases: The first phase, *Directional Context Mining*, extracts the directional context of each term in a given term pair. This phase makes use of Wikipedia as a knowledge source to identify the relevant context of each term. The second phase, *Inverted Index Generation*, constructs an inverted index of the mined context. The third phase, *Semantic Relatedness Computation*, com-

[4] It is worth mentioning that the idea of semantic measures based on asymmetric associations itself is not new. There are a few measures such as Kullback Leibler divergence, α Skew divergence and Co-occurrence Retrieval Models (CRM's) which capitalize on asymmetric associations of words [143].

putes the bi-directional strength of a term pair and combines its forward and backward association strengths to get the final relatedness score. Although the relatedness computation is based on asymmetric associations, the final relatedness score is symmetric and is normalized on a scale of [0 - 1].

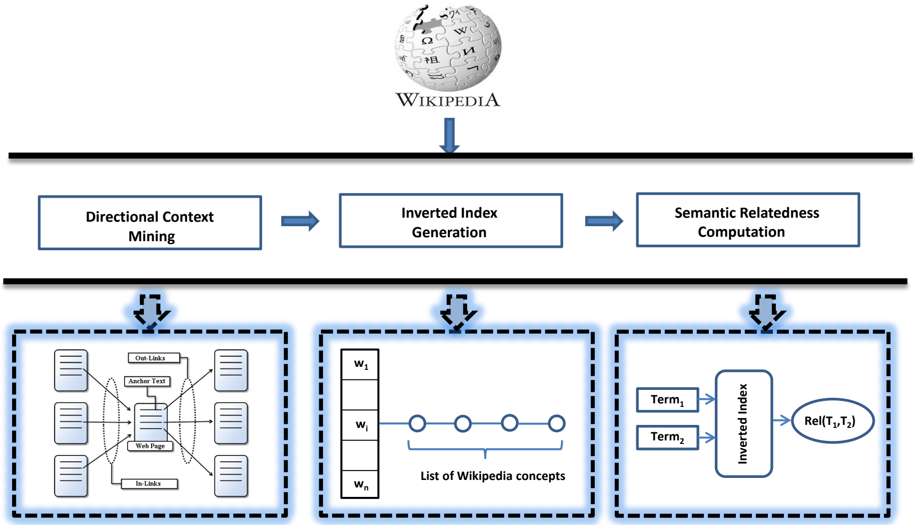

WIKIPEDIA
Directional Context Mining
Inverted Index Generation
Semantic Relatedness Computation
Out-Links
Anchor Text
Web Page
In-Links
w_1
w_i
w_n
List of Wikipedia concepts
$Term_1$
$Term_2$
Inverted Index
$Rel(T_1,T_2)$

Figure 4.2: The framework for computing semantic relatedness based on directional contexts.

4.3.1 Directional Context Mining

Asymmetric relatedness computation is highly dependent on semantic richness of the extracted context. The more semantically relevant the context is, the better the performance of the semantic relatedness measure. We opt to use Wikipedia for directional context extraction because it is a semantically rich knowledge source. Since each Wikipedia article represents a single concept and is linked to many other semantically related articles, the directional context of a term consists of the Wikipedia articles that are semantically related to the corresponding article of a given term. This research makes effective use of Wikipedia hyperlink structure for directional context mining.

Each term in an input term pair (t_i, t_j) is mapped to its corresponding Wikipedia article a_1 and a_2 respectively. For each matched Wikipedia article, its inlinks and outlinks are extracted to construct a context vector C_i. This context vector represents the respective directional context of an input term. The focus of this research is on computing the semantic relatedness based on the directional context of each given term.

4.3.2 Inverted Index Generation

To have a faster access to various lexical and statistical features of the articles in each context vector, the second phase constructs an inverted index of the context vector of each term. An inverted index is a data structure that is keyword-centric (keyword→ articles) rather than data-centric (article→ keywords). It allows faster search responses because instead of searching the informative-content of Wikipedia articles directly, only the index is searched. This is equivalent to retrieving the pages related to a keyword by searching the index at the end of the book rather than searching the content of each page for the keywords directly. To obtain informative content for indexing, the Wikipedia article of each context vector are preprocessed. This involves a series of steps which convert the MediaWiki format[5] of each Wikipedia article in the context vector to plain text; all redirects are mapped to their target articles; overly specific articles having less than 100 non-stop-words or less than 5 inlinks and outlinks are also discarded; stemming and stop word removal are not performed at this point. Finally, an inverted index of each context vector is constructed[6].

[5] The Wikipedia Miner Toolkit [138] is used for this.

[6] Using Apache Lucene open source Java library for indexing and searching.

4.3.3 Semantic Relatedness Computation

Given a term pair (t_1, t_2), the third phase computes the final symmetric relatedness score based on the asymmetric directional association strength of each term to the other term. The directional association strength of a term to another term is the association strength of the term pair in the context of the first term. When the association strength of input term pair t_1 and t_2 is computed in the context of term t_1, we assume the association strength is in the*forward direction*. Whereas, when the association strength of the same term pair is computed in the context of term t_2, it is considered as the association strength in the *backward direction*. Hence, the directional association strength of the same term pair assumes different association scores depending on the context of input term in consideration.

From the inverted index, a set of all articles C having both input terms t_i and t_j occurring within a proximity window of size $2\epsilon + 1$ are extracted, where ϵ is set to 20 after preliminary experiments. For each article $a \in C$, its relatedness to the articles a_i and a_j corresponding to input terms t_i and t_j are computed. Then, the association strength of input term pair (t_i, t_j) is computed by the following formula:

$$Association\_Strength(t_i \rightarrow t_j) = \frac{\sum_{a \in C} rel(a, a_i) \times rel(a, a_j)}{|C_i|} \tag{4.1}$$

where C_i is the context vector of each term and $rel(a, a_i)$ and $rel(a, a_j)$ are computed using *Article Comparer* of Wikipedia Miner Toolkit [138]. The *Article Comparer* is a machine learning based algorithm for computing semantic relatedness of Wikipedia articles. It extracts multiple features of given Wikipedia articles and learns the optimal relatedness scores based on the features.

The asymmetric association strength indicates the directional association of an input term pair. Thus, the directional association strengths of both input terms are linearly combined to get the final *Directional Association based Relatedness Measure* (DCRM) score for the input term pair as follows:

$$DCRM(t_1, t_2) = \lambda \times Association\_Strength(t_1 \rightarrow t_2) \\ + (1 - \lambda) \times Association\_Strength(t_2 \rightarrow t_1) \quad (4.2)$$

where λ is a coefficient of association such that $0 \leq \lambda \leq 1$ and its value is chosen to be 0.5, to give equal importance to both directional association strengths.

4.4 Evaluation Setups

For performance evaluation of the DCRM measure, a direct evaluation method is used with Spearman's and Pearson's correlation as the evaluation metrics. Three datasets are used in the performance evaluation of the DCRM measure: M&C, R&G and WS-353 datasets. Apache Lucene[7] is used for inverted index construction and various Lucene API's are used for computing the statistics from the inverted index. We evaluated the performance of three variants: DCRM on all term pairs; DCRM+WLM on all term pairs; and DCRM on non-missing term pairs. The first variant applies the DCRM measure on all term pairs (with missing term pairs assigned zero scores). The second variant, DCRM+WLM uses the WLM measure [138] for computing the scores of only missing term pairs and combining it with the DCRM measure. The third variant, DCRM (NM) applies the DCRM measure on term pairs excluding the missing pairs. The reminder of the experimental setup is the same as used in the previous chapter.

4.5 Results and Discussions

The performance comparison of DCRM with the previously best performing measure, WikiSim (Chapter 3), on the M&C, R&G and WS-353 datasets is shown in Table 4.1. The bold values in each following table indicate the best correlation-based performance on a specific dataset. A cursory look

[7] `http://lucene.apache.org`

Table 4.1: Performance comparison of the DCRM measure with the previously best measure WikiSim on domain independent datasets. Bold values indicate the best correlation-based performance of any measure on a specific dataset.

	M&C		R&G		WS-353	
Method	ρ	r	ρ	r	ρ	r
DCRM (All)	0.58	**0.61**	0.60	**0.71**	0.61	**0.50**
WikiSim (All)	**0.78**	0.58	**0.77**	0.60	**0.64**	0.38
DCRM (Non-Missing)	0.85	**0.73**	0.84	**0.73**	**0.70**	**0.61**
WikiSim (Non-Missing)	**0.87**	0.71	**0.87**	0.65	**0.70**	0.44
DCRM+WLM (All)	**0.84**	**0.77**	0.77	**0.63**	**0.69**	**0.60**
WikiSim+WLM (All)	0.83	0.62	0.83	0.62	0.66	0.41

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

at Table 4.1 shows that like the WikiSim measure, DCRM also produced low correlations on datasets with all term pairs due to many zero scoring term pairs, but computing the correlation on a subset of each dataset excluding missing term pairs resulted in a significant increase in the correlation values of both evaluation metrics on all datasets. This demonstrates the true performance of the DCRM measure and indicates its effectiveness on the task of semantic relatedness computation as long as given terms match with corresponding Wikipedia concepts. In comparison with previously best performing measure, WikiSim, all three versions of DCRM consistently produced higher Pearson's correlation than WikiSim on all datasets. Combining DCRM with WLM (for computing scores of missing term pairs) resulted in a slight gain of Spearman's correlation over WikiSim measure on the M&C dataset (from 0.83 to 0.84) and a significant gain on the WS-353 dataset (from 0.66 to 0.69). These results show that on the task of computing semantic relatedness, which covers various types of classical and non-classical relations, DCRM performed better than the WikiSim measure. The results also indicate the semantic orientation of both semantic measures. While, the WikiSim measure showed good performance on estimating semantic similarity by producing high correlation on the M&C and R&G datasets, the DCRM measure showed good performance on the task of semantic relatedness computation by producing high correlation on WS-353 datasets. Overall, DCRM produced higher Pearson's correlation than WikiSim on all datasets.

4.5.1 Asymmetric Relatedness Evaluation

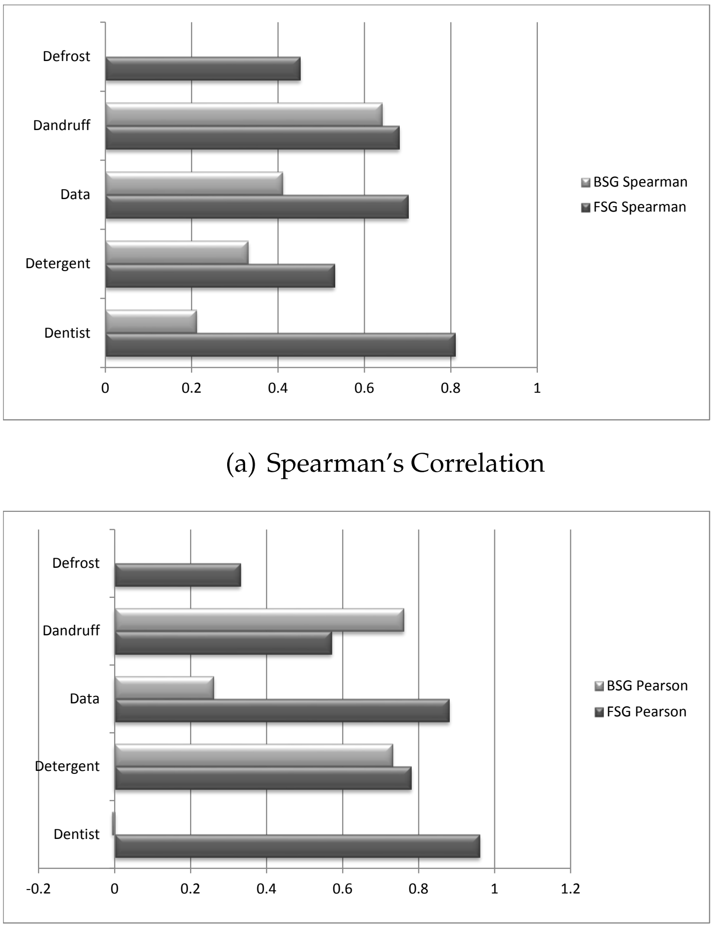
To examine the asymmetric nature of the DCRM measure, another set of experiments was conducted on a subset of Free Association Norms (FAN) dataset [149]. FAN is a database of free word associations. To construct FAN, human raters were presented with a stimulus word and were required to write the first response word that came to their mind and was meaningfully related or strongly associated with the stimulus word. More than 6000 participants generated nearly three quarters of a million responses to 5019 stimulus words. FAN dataset includes a number of statistics corresponding to each term pair. We were interested in the 6th field (FSG) representing *forward strength* of each term pair and 7th field (BSG) representing *backward strength* of each term pair in the database. Both FSG and BSG for FAN dataset are defined as the number of participants that wrote a particular *response* word to a *stimulus* word divided by the total number of participants in that group. For asymmetric relatedness evaluation, we used a noun subset of FAN dataset consisting of 43 noun term pairs. This subset consists of five stimulus words followed by a number of responses for each stimulus word ranging from 6-12.

In order to compare humans asymmetric associations with automatically generated asymmetric associations, directional associations of each stimulus-response pair were computed using Equation 4.2. Overall, Spearman's correlation on forward strength of FAN subset was 0.51 and Pearson's correlation was 0.26 whereas, the Spearman's correlation was 0.62 and Pearson's correlation was 0.39 on backward strength.

To gain further insight into the nature of stimulus-response relations and the difference between human and automatic associations, another experiment was conducted on the noun subset of FAN, in which the DCRM-based forward and backward association strengths were compared with human generated forward and backward association strengths for each stimulus word separately and the results are reported using both Spearman's correlation and Pearson's correlation, as shown in Figure 4.3.

There is a clear variation on both correlation values particularly on the stimulus word *Dentist*, where the Pearson's correlation ranges from -.007 (slightly visible to the left side of the y-axis in the graph) on backward association to 0.96 on forward association. Also, the correlation on backward association strength of the stimulus word *Defrost* is 0 (not visible in the graph).

The results of this experiment highlight the definitive difference between humans asymmetric associations and automatic asymmetric association. Humans asymmetric associations are generally directional and are based on collocations and co-occurrences, whereas automatic asymmetric association not only covers lexical and distributional properties of words but also considers fine grained semantics which are generally not covered by the conventional association-based measures. Humans usually give high scores to those terms which frequently co-occur or are lexically used together. The overall correlation of automatically computed directional associations is not high as the DCRM measure is not entirely based on distributional or lexical statistics. When individual term pairs were analyzed, it was found that human have scored the term pair `(dentist,pain)` higher than `(dentist,orthodontics)`, whereas the DCRM measure assigned a higher score to the latter pair.



(a) Spearman's Correlation

(b) Pearson's Correlation

Figure 4.3: A comparison of Spearman's correlation and Pearson's Correlation on stimulus words of FAN subset.

Clearly, human judgments are based on their everyday experience of having pain during a dental treatment, whereas automatic asymmetric associations are backed up by the encyclopedic knowledge of Wikipedia, which finds `dentist` semantically more closer to `orthodontics` than to `pain`. We argue that a word pair could be related from different perspectives. Hence, it is possible that the fundamental criteria of relating two words is well motivated but is derived differently by humans and the automated approaches to arrive at diverging judgments. Based on this experiment, we make another observation that semantic relatedness measures relying on the encyclopedic knowledge for automatically computing asymmetric associations do not correlate well with human judgments on association-based tasks. The reason is the difference of perspectives from which such judgments are made.

4.6 Chapter Summary

This chapter guided the task of semantic relatedness computation by asymmetric word associations. The chapter presented a new semantic measure, DCRM, based on a hybrid model that combines the structural features based combinatorial model with the proximity-based distributional model. This hybrid model relies on the strength of Wikipedia structure and its informative-content for computing asymmetric associations. Empirical results have shown that the DCRM measure based on directional word associations is a good indicator of relatedness. Another experiment analyzed the asymmetric behavior of the DCRM measure by comparing with the patterns of asymmetric associations of humans. The results of this experiment conclude that the perspective of DCRM-based semantic associations is clearly distinct from that of humans asymmetric word associations. DCRM, being a knowledge source-based measure, identifies word associations using the encyclopedic or factual background knowledge, whereas manual word associations produced by humans focus more on their personal experiences than on the factual data. This aspect of human symmetry is interesting and could be explored further based on the asymmetric associations. We believe that in comparison with the approaches based on semantic or lexical relations extracted from knowledge sources, statistical approaches based on word co-occurrences could be more effective on association-based tasks.

Chapter 5

Hybrid Model for Semantic Association Computation

The semantic association measures presented in the previous two chapters focused on individual aspects of Wikipedia: WikiSim and OSR (Chapter 3) are based on Wikipedia's hyperlink structure to get shared associates; CPRel (Chapter 3) focused on the informative-content of Wikipedia articles for constructing the contextual profiles; and DCRM (Chapter 4) focused on computing distributional asymmetric associations based on both structure and the informative-content of Wikipedia. Different aspects of a knowledge source cover different kinds of semantic and lexical relations. Based on this assumption, this chapter presents a new hybrid model for learning semantic associations of words by combining the features extracted from various aspects of Wikipedia. The focus of the chapter is on finding the optimal feature combination(s) that enhances the performance of semantic association computation.

Section 5.1 presents an investigation and comparison of the focus of features based on different elements of Wikipedia and establishes an argument in the favor of hybrid features over single features. Section 5.2 details a new hybrid model for learning semantic association computation. Section 5.3 details the evaluation setup for the hybrid model. Section

5.4 reports the results of the hybrid model and presents a comparative analysis of two different versions of hybrid models with previously best performing measures. Finally, Section 5.5 presents a brief summary of the chapter.

5.1 Hybrid Features Vs. Single Features

Existing approaches to computing semantic associations focus on specific aspects of a corpus or a formal knowledge source such as distributional features, structural features and content-based features. Such approaches lead to better predictions of some semantic or lexical relations but perform poorly on others because of their semantic bias and limitations. For instance, taxonomy or structure based approaches are generally good at computing the similarity between two concepts that exist in the same hierarchy and are connected through a path. Consequently, such measures are good at finding semantically similar concepts but are limited by the coverage and organization of concepts in the structure of the underlying taxonomy.

On the other hand, approaches based on distributional profiles of two concepts focus more on co-occurrence based association of concepts. Such approaches are good at judging associations of the concepts that co-exist frequently but might not be strongly related. For example, distributional approaches assign an association score higher than synonyms to a term pair `(bread,butter)` due to a common phrase *bread and butter*, as indicated by Yih *et al.* [226]. Distributional approaches suffer from the limitation of their underlying text corpora having skewed coverage of words according to Zipf's law [1] [234], thus are biased towards more frequent words. Consequently, regardless of the corpus size, such approaches produce unreliable results for rarely used words.

Content-based approaches focus on concepts rather than just words.

[1] Please see the glossary for the detail

Such approaches make use of the informative-content or glosses of concepts and use the *Vector Space Models* (VSM) to compute association scores. Such approaches are also biased towards the size of the informative-content and suffer from limited knowledge coverage of their underlying knowledge source(s).

A recent trend in semantic association computation research is to base the semantic association measure on multiple aspects of a single knowledge source or multiple knowledge sources. Such approaches combined various aspects of a knowledge source(s) and shown that association measures utilizing multiple aspects of a knowledge source(s) perform better than those which rely on individual aspects [23, 184, 216, 1, 162, 199]. This makes sense because the strength of one feature complements weaknesses of other features. Combining multiple aspects leads to improved correlation with human judgments due to the complementary coverage of a multitude of semantic and lexical relations by various features. This chapter presents a new model based on multiple aspects of the Wikipedia knowledge source and demonstrates that a model of semantic association computation using features based on multiple aspects shows significant improvement over the models using features based on individual aspects.

5.2 Hybrid Features based Model

The chapter presents a new hybrid model that uses a regression function for combining three new Wikipedia-based features to predict final association scores. There are two main components of the hybrid model: *feature generation* and *validation and learning semantic associations*. The first component generates the three new features using multiple aspects of the Wikipedia knowledge source: first feature relies on the Wikipedia hyperlink structure; second feature is based on the informative-content of Wikipedia articles; and third feature combines the former two aspects of Wikipedia. The second component takes as input the three features and uses a classifier to learn the regression model that maximizes the correlation of predicted scores with human judgments.

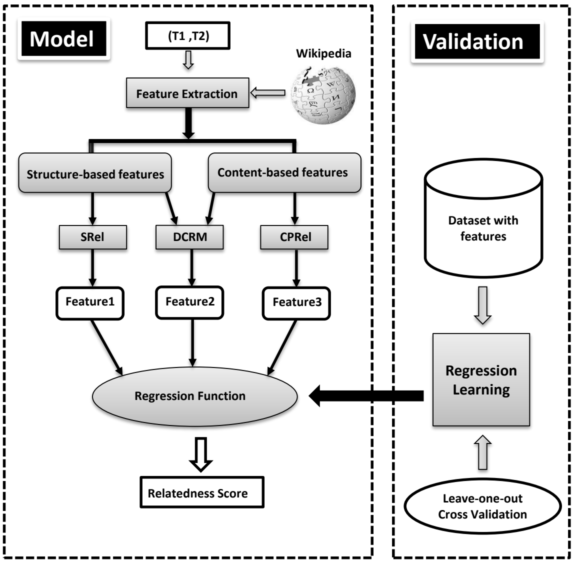


Figure 5.1: Hybrid features based model for computing semantic associations.

5.2.1 Feature Generation

When asked to relate two terms, human intrinsically relate their corresponding concepts in a wider context rather than just relating two lexical forms of those concepts. Hence, the features presented in this chapter are based on representing the words by their corresponding Wikipedia articles. In Wikipedia, each article is dedicated to one particular concept[2] and is connected to many other concepts through hyperlinks. Structural semantic mining is done by exploiting the Wikipedia hyperlink structure which intrinsically encodes a variety of semantic and lexical relations

[2] The terms *article* and *concept* are used interchangeably.

among Wikipedia articles.

The hybrid model uses the following three features (each of which refers to the semantic association measure presented in the previous chapters of the thesis):

- **Feature 1:** The first feature combines the structure-based measures of associations (presented in the sections 3.2.1 and 3.2.2 of the thesis): WikiSim, which is based on the proportion of Wikipedia hyperlinks shared by two concepts; and OSR, which is based on the normalized semantic orientation of the link overlap set. These two measures are linearly combined to get a new feature, SRel (Structure-based Relatedness), as follows.

  $$SRel(t_1, t_2) = \frac{1}{2} \times [WikiSim(t_1, t_2) + OSR(t_1, t_2)]$$

  The focus of this new feature is on semantic similarity as well as on semantic relatedness as it accounts for both structure as well as semantic orientation.

- **Feature 2:** The second feature, CPRel, is a vectorial[3] approach to computing associations using the informative-content of Wikipedia articles. This approach is focused on semantic relatedness. For details of this feature, please refer to the Section 3.2.3 of the thesis.

- **Feature 3:** The third feature, DCRM, is based on a combination of both hyperlink structure and the informative-content of Wikipedia. This feature uses the structural overlap of given words as a conceptual space and computes the co-occurrence based distributional associations of given word pairs in this conceptual space for computing semantic associations. The focus of this feature is on co-occurrence-based relatedness. Detailed description of this feature is included in Section 4.3 of the thesis.

[3] Inspired by *Vector Space Model*

5.2.2 Validation and Learning of Semantic Associations

The model automatically computes the semantic association score of each test bed term pair using a regression model that optimally combines the three features. To generate optimal supervised regression model, four different classifiers are used: Gaussian Process (GP), with the Pearson VII function-based universal kernel (PUK), which implements Gaussian processes for regression without hyper-parameter-tuning; SMOreg (SMO), with the Pearson VII function-based universal kernel (PUK), which implements the Support Vector Machine for non-linear regression; Linear Regression (LR), which finds a regression line that best fits all data points; and Multi-Layer Perceptrons (MLP), which uses back propagation to classify instances. All classifiers are used with default parameter settings of Weka [72]. Each classifier learns how to most effectively combine various features to maximize the correlation of predicted scores with human judgments.

To test the performance of features combinations, three single features were used to generate various feature combinations and the effect of each feature combination on the performance of semantic association computation was analyzed. In the following experiments, the individual features SRel, DCRM and CPRel are referred to as f_1, f_2 and f_3. Based on these three single features, three 2-feature combinations, (f_{12}), (f_{13}) and (f_{23}) and one 3-feature combination (f_{123}) were generated, resulting in seven features in total.

Since the datasets are small in size, cross validation is used to avoid over fitting. The learning process is validated using two different classifier evaluation algorithms: 10-fold cross validation and leave-one-out cross validation. Leave-one-out cross validation is the same as a K-fold cross-validation with K representing the number of observations in the original dataset.

5.3 Evaluation

The experiments in the chapter also followed the standard procedure for direct evaluation as used in the previous chapters. Hence, three benchmark datasets of association computation were used: M&C, R&G and WS-353 datasets. The performance of the hybrid model is reported using both evaluation metrics: Spearman's correlation coefficient (ρ) and Pearson's correlation coefficient (γ). The bold values in each following table indicate the best correlation-based performance on a specific dataset.

5.4 Results and Discussion

Table 5.1 reports the correlation-based performance comparison of the four classifiers using the hybrid model with 3-feature combination (f_{123}). Bold values indicate the highest correlation on a particular dataset using either correlation measure.

Table 5.1: Performance comparison of four classifiers using the hybrid model based on the 3-feature combination (f_{123}).

Classifier	Correl.	L-1-O CV			10-fold CV		
		M&C	R&G	WS-353	M&C	R&G	WS-353
SMOReg	ρ	0.76[0.55-0.8]	**0.89**[0.82-0.93]	0.67[0.60-0.72]	0.73[0.50-0.86]	0.83[0.73-0.89]	0.66[0.59-0.71]
	γ	**0.85**[0.70-0.92]	0.89[0.82-0.93]	0.68[0.61-0.73]	**0.83**[0.67-0.91]	0.85[0.76-0.90]	0.66[0.59-0.71]
GP	ρ	**0.77**[0.56-0.88]	**0.89**[0.82-0.93]	**0.71**[0.65-0.75]	**0.74**[0.51-0.86]	**0.84**[0.74-0.90]	**0.70**[0.64-0.74]
	γ	**0.87**[0.70-0.92]	**0.9**[0.84-0.93]	**0.70**[0.64-0.74]	0.82[0.65-0.91]	**0.87**[0.79-0.92]	**0.68**[0.61-0.73]
LR	ρ	0.72[0.48-0.85]	0.82[0.72-0.88]	0.64[0.57-0.69]	0.72[0.48-0.85]	0.8[0.69-0.87]	0.63[0.56-0.68]
	γ	0.78[0.58-0.89]	0.84[0.74-0.89]	0.53[0.45-0.60]	0.74[0.51-0.86]	0.81[0.70-0.88]	0.51[0.42-0.58]
MLP	ρ	0.72[0.48-0.85]	0.82[0.72-0.88]	0.64[0.57-0.69]	0.65[0.37-0.81]	0.83[0.73-0.89]	0.65[0.58-0.70]
	γ	**0.85**[0.70-0.92]	0.87[0.79-0.91]	0.64[0.57-0.69]	0.67[0.40-0.82]	**0.87**[0.79-0.91]	0.64[0.57-0.69]

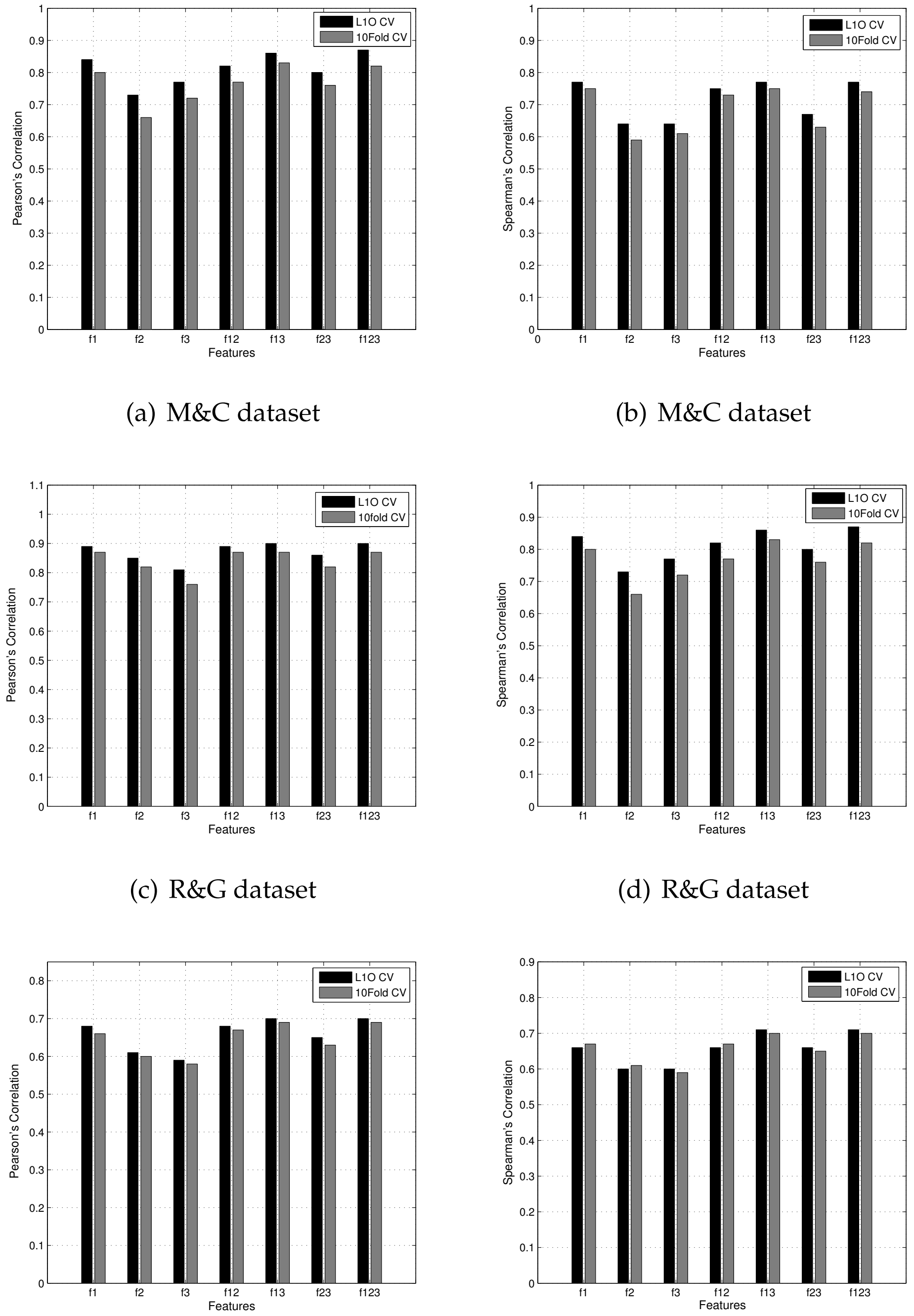
Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any classifier on a specific dataset.

The reported results were statistically significant at 95% confidence level (with $\alpha = 0.05$). The correlation values are accompanied by their corresponding confidence intervals. Confidence intervals signify that the true mean of the population is expected to lie within this interval. The smaller datasets usually have wider confidence intervals. Also, the higher the correlation value, the narrower the confidence interval. Overall, the performance of the GP classifier was consistently good across all datasets. Hence, this classifier is used for performance analysis of the hybrid model in comparison with the previously best performing approaches. Also, the Spearman's correlation based performance of the hybrid model using leave-one-out cross validation was higher than the same model using 10-fold cross validation because of more training instances.

5.4.1 Performance Comparison of Features

Figure 5.2 shows performance comparison of all feature combinations using leave-one-out and 10-fold cross validations on all three datasets. The figure depicts the performance of the GP-based hybrid model with different feature combinations using Pearson's and Spearman's correlations. Clearly, the 3-feature combination (f_{123}) outperformed other features in most of the cases. This demonstrates that a combination of structure and content-based features is a good indicator of semantic associations. Also, the feature f_{13} performed comparable to the best performing feature (f_{123}). Among single features, the structure-based feature f_1 outperformed other two single features on all datasets. It is clear from the figure that in most of the cases, the hybrid model based on leave-one-out cross validation outperformed the one based on 10-fold cross validation. However, the performance difference of both models was obvious on smaller dataset but became negligible on largest dataset WS-353, due to increased number of training instances for 10-fold cross validation.

To further analyze the overall performance of the features and their combinations, averaged_ranks of the features are computed for selecting the overall best performing feature(s). Algorithm 1 describes the pseudo



(a) M&C dataset

(b) M&C dataset

(c) R&G dataset

(d) R&G dataset

(e) WS-353 dataset

(f) WS-353 dataset

Figure 5.2: Performance comparison of hybrid models based on individual features and their combinations on three benchmark datasets using 10-fold and Leave-one-out cross validations.

code for optimal feature(s) selection based on averaged feature ranking. F_{D_i} represents the set of features with their correlation values on the i^{th} dataset. The function $Sort(F)$ sorts a set of all features F and the function $rank(f)$ returns the rank of a feature f in the sorted set of features.

Algorithm 1 Feature Selection based on Averaged Ranking

```
$Best\_Features \leftarrow \phi$
**for all** $i \in (F_{D_i})$ **do**
    $F_{D_i} \leftarrow Sort(F_{D_i})$
**end for**
$avg\_rank \leftarrow \frac{\sum_{f \in F_D}(rank(f))}{|F_D|}$
$F \leftarrow avg\_rank$
$F \leftarrow Sort(F)$
$min \leftarrow rank(f_1)$
**for all** $f \in F$ **do**
    **if** $rank(f) = min$ **then**
        $Best\_Features \leftarrow f$
    **end if**
**end for**
**return** $Best\_Features$
```

To select the optimal feature(s), the following steps are performed:

- All features are sorted and ranked according to their correlation values on each dataset. Low ranks correspond to high correlation values and vice versa.

- Ranks of each feature on all datasets are averaged to get its averaged_rank score.

- The features are sorted and ranked again according to their average rank scores.

- The feature(s) with lowest rank are selected as the best performing feature(s).

Table 5.2: A comparison of averaged feature ranking on all datasets. The highest rank corresponds to the lowest correlation-based performance of a feature and vice versa.

Feature	Rank Number			
	L-1-O	L-1-O	10-fold	10-fold
	(ρ)	(γ)	(ρ)	(γ)
f_1	3	2	1	3
f_2	5	5	5	5
f_3	6	6	6	6
f_{12}	3	3	3	3
f_{13}	2	1	1	1
f_{23}	4	4	4	4
f_{123}	1	1	2	2

A comparison of averaged feature ranking is presented in Table 5.2. The aim of this experiment is to analyze the overall rank based performance of each feature. The best feature on leave-one-out cross validation was f_{123} and on 10-fold cross validation was f_{13}. It is intuitive to note that the first single feature f_1, which is a combination of two structure-based measures, outperformed the other two single features, while the third single feature f_3, which is the informative-content based association measure produced the lowest correlation of all. However, when the structure-based measure is combined with the informative-content based measure using a regression function, this new feature f_{13} surpassed all other feature combinations. These results demonstrate that distinct features complement each other in various association scenarios, resulting in significant performance improvement. That is why the feature combination f_{13} outperformed the individual features f_1 and f_3.

5.4.2 Hybrid_Model with Maximum Function

To gain further insight into the behavior of the hybrid model, the regression function is replaced by the maximum function and association scores are computed again for all datasets. In this experiment, given three features for a term pair, the hybrid model picks the maximum of the three as the final association score. Table 5.3 compares the performance of the maximum_function based and GP-based hybrid models with the best performing measures of the previous chapters.

Table 5.3: Performance comparison of both hybrid models with previously best performing approaches on three benchmark datasets: M&C, R&G and WS-353.

	Datasets			Datasets		
	M&C	R&G	WS-353	M&C	R&G	WS-353
Method	(ρ)	(ρ)	(ρ)	(γ)	(γ)	(γ)
WikiSim+WLM	0.83	0.83	0.66	0.62	0.62	0.41
DCRM+WLM	**0.84**	0.77	0.69	0.77	0.63	0.60
Hybrid_Model (GP)	0.77	0.89	0.71	**0.87**	**0.90**	0.70
Hybrid_Model (max)	0.80	**0.91**	**0.79**	0.85	0.86	**0.81**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

In all but one case, both hybrid models outperformed the previously best Wikipedia-based approaches. On both M&C and R&G datasets, the GP-based hybrid_model yielded the highest Pearson's correlation. The hybrid_model using maximum function also performed comparable to the GP-based model on these datasets using the Pearson's correlation. However, on WS-353 dataset, the hybrid_model based on maximum function outperformed all other measures on both correlation metrics. It also surpassed all other approaches on the R&G dataset using the Spearman's correlation.

It is surprising to discover that the results of hybrid model with maximum function were superior to more sophisticated supervised regression learning using the Spearman's correlation (in all cases) and on the largest dataset, WS-353, using the Pearson's correlation. Section 5.2 of the chapter indicated that the three features are based on different aspects of Wikipedia. Consequently, these features are good at detecting different kinds of term relations. For instance, if the structure-based feature is unable to find out a strong relation of two terms in the Wikipedia structure, it does not mean that the relation between the terms does not exist at all. It could be the inability of that feature to detect the relation of the given term pair due to its underlying aspect—Wikipedia hyperlink structure. Hence, if the informative-content based feature finds out a stronger association of the same term pair then the low associations predicted by other features should be overshadowed by this feature's indication of a stronger relation. The maximum function represents these semantics by selecting the kind of relation that is strongest. Surprisingly, the supervised regression appears to be unable to represent and learn these semantics better than the maximum_function based hybrid_model. This experiment also shows the poor ability of the classifiers used in the supervised regression learning to cope with such scenarios due to lack of their internal mechanism in handling the maximum likelihood of individual features while converging to the final regression function.

The experiment demonstrates that using the hybrid model based on multiple features improves the performance of semantic association computation by providing complementary coverage of lexical and semantic relations. Thus, leading to high correlation with human judgments.

5.5 Chapter Summary

This chapter addressed the problem of semantic association computation using a supervised machine learning based hybrid model. The research contributions of this chapter are two-fold: first, it presented a new hybrid model based on three features generated from multiple aspects of Wikipedia for learning semantic association computation; second, it used a correlation-based feature ranking to select an optimal feature combination(s). The experiments demonstrate the effectiveness of the hybrid model on computing semantic associations of words. The chapter investigated the impact of individual features and their combinations on learning semantic association computation and empirically showed that the best performing feature combinations are the ones that are based on multiple semantic elements of Wikipedia. This supports the basic assumption of the research that combining multiple aspects based semantics leads to better estimates of semantic associations. Empirical results have also shown that the maximum function based hybrid model performed well, exceeding the GP-based hybrid model on all datasets using the Spearman's correlation. The reason was the preferential selection of a feature with maximum association score over those features that yielded poor estimates of semantic associations. It was also found that the regression model using leave-one-out cross validation marginally outperformed the ones using 10-fold cross validation on smaller datasets but this margin decreased with increasing dataset size due to sufficient training instances for learning an optimal regression model.

Chapter 6

Probabilistic Associations as a Proxy for Semantic Relatedness

The idea of asymmetric association based similarity measure (presented in Chapter 4) is intriguing due to directionality of context as well as effective when evaluated on multiple similarity and relatedness datasets. However, this approach requires the mapping of input terms to their corresponding Wikipedia articles which, if they do not exist, could lead to performance degradation. Other semantic association measures presented in the previous chapters of the thesis also suffered from the same limitation. Hence, this chapter focuses on coping with this limitation without compromising the effectiveness of association computation. For this purpose, an approach for computing a new measure of semantic associations is developed that takes into account the text corpus aspect of Wikipedia. This approach is based on computing asymmetric association based probabilities and combining them to get symmetric word association scores. The focus of the semantic association computation approach is on computing associations of words rather than concepts.

Section 6.1 introduces the probabilistic associations. Section 6.2 presents an approach for computing a new probabilistic association based semantic relatedness measure. Section 6.3 details the evaluation setup, datasets and evaluation metrics used in the experiments. Section 6.4 presents a detailed empirical analysis of the new semantic association measure. Finally, section 6.5 concludes the chapter by discussing the strengths and future prospects of the new semantic measure.

6.1 Probabilistic Associations

The idea of guiding semantic relatedness computation based on directional word associations was introduced in the previous chapter. The focus of this chapter is on considering the context of each word at a corpus level, thus avoiding the requirement to map the input words to the corresponding Wikipedia articles. Asymmetry-based probabilistic associations are co-occurrence based probabilities of word associations in the context of a given word. Probabilistic associations for computing semantic relatedness rely on distributional properties of words in an unstructured text corpus. Hence, the semantic relatedness computation approach uses the Wikipedia corpus for extracting co-occurrence based associative probabilities of words.

Probabilistic associations are used to develop two types of relatedness measures: an asymmetric association based relatedness measure, and several standard symmetric relatedness measures. The asymmetric association based relatedness measure relies on asymmetry-based probabilistic word associations. The symmetric relatedness measures include the Dice coefficient, the Simpson coefficient, Adapted Normalized Google Distance (ANGD) and Adapted Pointwise Mutual Information (APMI). The performance of the asymmetric association based relatedness measure is compared with the symmetric relatedness measures (used as baseline measures).

A limitation of co-occurrence based measures is the poor ability to predict strong relatedness scores for synonymous words. Wikipedia, being a semantically rich knowledge source, includes various implicit and explicit indicators of semantic connections between different concepts. To handle synonymous words, probabilistic associations are augmented with these indicators of synonymy derived from Wikipedia.

6.2 Asymmetry-based Probabilistic Relatedness Measure

The approach for computing novel asymmetry-based Probabilistic Relatedness Measure (APRM) is divided into two phases: first phase constructs an inverted index using Wikipedia articles; and the second phase extracts the directional context of two terms and computes their relatedness based on asymmetry-based probabilistic associations.

6.2.1 Inverted Index generation

The probabilistic measures of relatedness discussed in the chapter are based on word co-occurrences extracted from a large corpus. To get these word co-occurrence probabilities, Wikipedia is used as an unstructured text corpus. For this purpose, an inverted index of Wikipedia articles is constructed. The informative-content of each Wikipedia article is preprocessed before indexing. This involves a series of steps which convert the MediaWiki format[1] of each Wikipedia article to plain text; all redirects are mapped to their target articles; overly specific articles having less than 100 non-stop-words or less than 5 inlinks and outlinks are also discarded; informative-content of each article is lemmatized for indexing. Finally, an inverted index of Wikipedia articles is constructed[2] as in Chapter 4 of the thesis.

[1] The Wikipedia Miner Toolkit [140] is used for this.

[2] Using Apache Lucene open source Java library for indexing and searching.

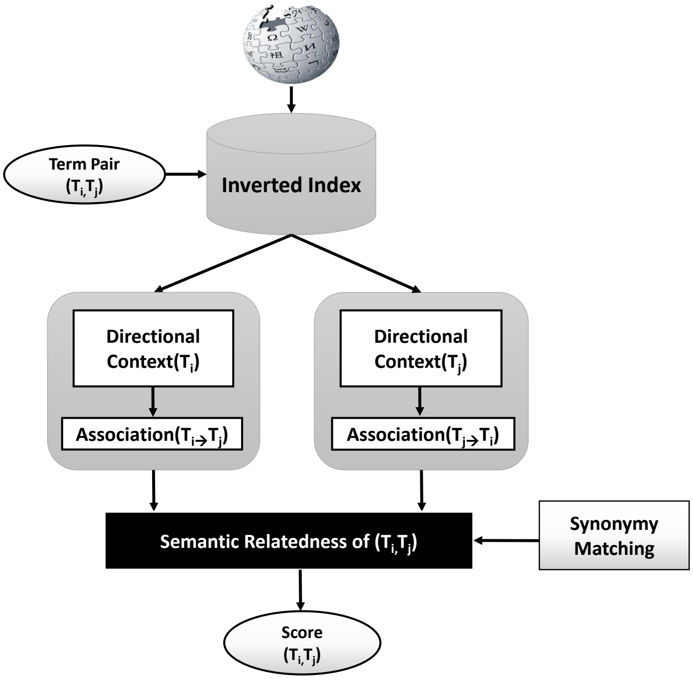


Figure 6.1: The framework for computing term relatedness using the APRM measure.

6.2.2 Asymmetry-based Relatedness Measure

Asymmetry-based Probabilistic Relatedness Measure (APRM) is the new measure that computes and combines directional association strengths of given two terms to get their relatedness score. When the association strength of a given term pair (T_i, T_j) is determined in the context of the first term T_i, it is referred to as *forward association strength* and is denoted by $Association(T_i \rightarrow T_j)$. Similarly, when association strength of the same term pair is computed in the context of the second term T_j, it is called *backward association strength*, denoted by $Association(T_j \rightarrow T_i)$. The APRM measure linearly combines these asymmetric association strengths to get the final symmetric relatedness score. The framework for computing the APRM measure is shown in Figure 6.1.

The context of a term refers to a set of those Wikipedia articles in which that term appears. Given a term pair (T_i, T_j), the context of each term is extracted from the inverted index. The directional association strength is then computed in the context of each term as the probability of co-occurrence of both terms within a proximity window of size $2w + 1$, where w refers to the number of words on either side of an occurrence of a target word:

$$Association(T_i \to T_j) = \frac{p\,(T_i\ near\ T_j)}{p\,(T_i)}$$

where $p(T_i\ near\ T_j)$ is the fraction of Wikipedia articles having the given two terms within a proximity window and $p(T_i)$ is the fraction of Wikipedia articles containing the term T_i. The association strength of a term pair in two different contexts is asymmetric in nature and changes with the change of context for the same term pair. The APRM measure linearly combines the asymmetric forward and backward association strengths of the given term pair (T_i, T_j) to compute the final symmetric relatedness score as follows:

$$APRM(T_i, T_j) = (1 - \lambda) \times Association(T_i \to T_j) + \lambda \times Association(T_j \to T_i)$$

where λ is the coefficient of association and is set to 0.5 to give equal importance to both directional association strengths.

Probabilistic association computation is a proximity assumption based method for capturing and using word associations to estimate the strength of their relatedness. However, there are certain cases where methods based on proximity assumption fail to cope effectively. For instance, if an input term pair consists of synonymous words such as `(construct,build)` then it is less likely that these synonyms co-occur in close proximity very often in the corpus. In such cases, synonymous word pairs get lower scores than collocational words pairs. APRM, being a co-occurrence based measure also suffers from the same limitation. To cope with this, the second component of the semantic association computation approach exploits the knowledge source aspect of Wikipedia. Two structural features of Wikipedia are used as explicit indicators of synonymy: *redirects* and *senses*.

For a given term pair, both terms are matched with corresponding Wikipedia articles. For each matched article, its redirects are collected. Redirects are various surface forms of an article title that are used to refer to the same article and signify the synonyms of that article. For instance, the set of redirects of `United States` includes `U.S.`, `America`, `The States`, `United States Of America` and `US`. Generally, there are much fewer redirects of an article than labels and they represent more tightly-coupled synonyms of an article as compared to labels. Wikipedia labels are loosely coupled synonyms, as these synonyms are tailored according to the language and culture of the referring authors. For a given term pair, if a match is found in the redirect sets of both terms then the term pair is considered as a synonymous pair and is assigned maximum score. If either term does not match with a corresponding Wikipedia article then all senses of its Wikipedia label are retrieved and searched for a match, as in the case of redirects. If there is a synonymy relation between given terms, either match results in producing maximum score for a term pair.

6.2.3 Symmetry-based Relatedness Measures

In this research, four baseline symmetric relatedness measures, based on co-occurrence probabilities, are implemented and compared with APRM.

The first symmetric measure is **Adapted Normalized Google Distance Measure (ANGD)**. Cilibrasi *et al.* [35] used the tendency of two terms to co-occur in web pages by proposing NGD as a measure of semantic distance of words and phrases. NGD used the World Wide Web as the corpus and Google page counts to get the frequency of word occurrences. NGD is well-founded on the information distance and Kolmogrov complexity theories [35][3]. The formula for NGD measure is as follows:

$$NGD = \frac{max(logf(x), logf(y)) - logf(x,y)}{logM - min(logf(x), logf(y))}$$

where $f(x)$ denoted the number of web pages containing word x and $f(x,y)$ represents the number of web pages containing both words x and y. Gracia and Mena [68] noted that NGD is a generalized measure that could be used with any web search engine. They transformed the NGD formula so that the distance scores it computes are bounded in a new range of [0-1] rather than the original range [0-∞]. Following their transformation, ANGD is given by:

$$ANGD(T_i, T_j) = e^{-2NGD(T_i,T_j)}$$

ANGD uses Wikipedia as the underlying corpus rather than the web and counts the number of Wikipedia articles having both terms in a different way from the original formula. It uses the proximity window of reference approach to get the number of documents having both terms in them. Different window sizes were tested to choose an optimal size and results on this parameter are discussed in the Section 6.4.2 of the chapter.

The second symmetric measure is **Dice coefficient** [43], which is a well known measure used in information retrieval. Given the two terms x and y, the Dice coefficient is computed as:

$$Dice(x,y) = \frac{2 \times f(x,y)}{f(x) + f(y)}$$

where $f(x)$ denotes the number of Wikipedia articles containing word x and $f(x,y)$ denotes the number of articles containing both words x and y within a proximity window.

The **Simpson coefficient** [188], often called **Overlap coefficient** is another symmetric relatedness measure that computes the overlap between

[3] Please see the glossary

two sets. The formula for the Simpson coefficient is given below:

$$Simpson(x, y) = \frac{log(f(x, y))}{min(log(f(x)), log(f(y)))}$$

The document overlap set of two words is normalized by the size of the smaller set to avoid any bias introduced by the larger set size.

The last symmetric relatedness measure is **Adapted PMI** (APMI), which is inspired by the PMI measure [34] and is computed as follows:

$$APMI(x, y) = \sqrt{\frac{log(f(x, y))}{log(f(x)) * log(f(y)))}}$$

Since the equation is symmetric, the amount of information acquired about the presence of the first term when the second term is observed, is the same as when observing the first term for the presence of second term, which explains the term mutual information.

Both asymmetric and symmetric relatedness measures are implemented using the same experimental settings. The underlying corpus for all measures is Wikipedia and synonymy matching is used by all measures to augment their scores.

6.3 Evaluation

The direct evaluation of the relatedness computation approach is based on computing the ranks of automatically computed scores and comparing them with that of human judgments using Spearman's rank order correlation coefficient. The results are reported on all the datasets discussed in the Section 2.3 of the thesis.

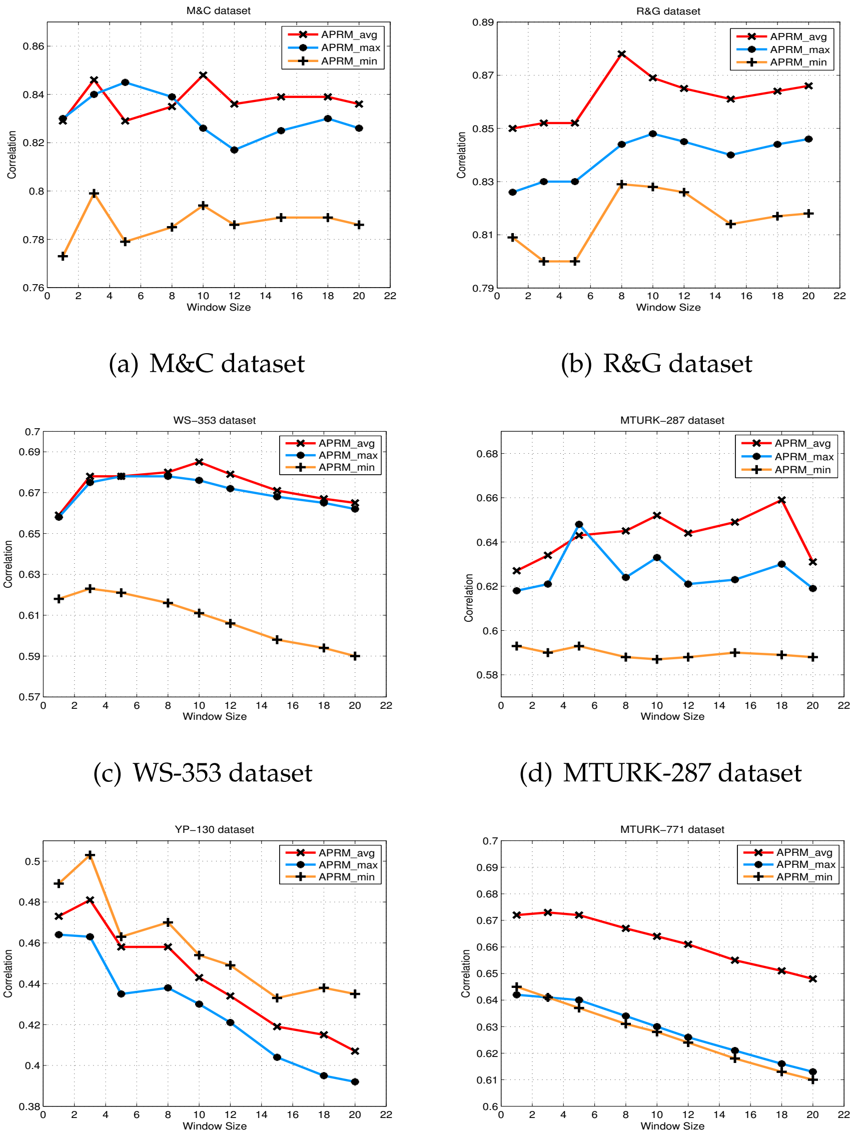
6.4 Experimental Results and Discussion

The experiment used three versions of APRM and baseline measures on eight datasets and nine different window sizes (ranging from 1 to 20). Results of the experiment were evaluated by measuring the correlation of automatically computed scores with the human judgments using Spearman's correlation coefficient. Finally, the results of APRM with the optimal version and window size are compared with other state-of-art measures. The bold values in each following table indicate the best correlation-based performance on a specific dataset.

The first part of this section analyses the impact of three different ways of combining the forward and backward association strengths for computing semantic relatedness. The second part analyses the system performance over various window sizes in order to find out a window size that optimizes the performance of the APRM measure. The third part of the section compares the performance of the APRM measure with four baseline measures implemented using the same experimental setup and resources. The final part of the section analyzes the performance of the APRM measure by comparing it with state-of-art similarity and relatedness measures on various datasets followed by a detailed discussion on strengths and weaknesses of the APRM measure.

6.4.1 Combining Directional Association Strengths

To analyze the impact of combining directional association strengths on the overall system performance, three variants of APRM were explored. APRM_max combines the directional association strengths by selecting the maximum of the two association strengths for each term pair. APRM_avg linearly combines the two association strengths using a coefficient of association λ. APRM_min takes the minimum of the two association strengths into account. Logically, it makes sense to test the maximum and average combinations of the directional association strengths because the aim of similarity measures is to maximize the closeness of the automatically computed similarity score with the manual judgment. However, the similarity and relatedness datasets generally consist of term pairs with related-



(a) M&C dataset

(b) R&G dataset

(c) WS-353 dataset

(d) MTURK-287 dataset

(e) YP-130 dataset

(f) MTURK-771 dataset

Figure 6.2: Performance comparison of three variants of APRM on different window sizes using all datasets.

ness scores ranging from strongly similar/related-to-unrelated term pairs. Hence, to have an insight into the impact of weakly related-to-unrelated term pairs on overall performance, it turns out to be useful to test the minimum score based combination also. A comparison of these three combinations on various datasets is shown in Figure 6.2.

On all datasets, a comparison of three approaches for combining the forward and backward association strengths is reported on various window sizes. Overall, APRM_avg performed better than the other two methods. Hence, we opt to use it in later comparisons with other state-of-art and baselines measures. There was no optimal window size on which each version always performed the best. The highest correlation was achieved by each version of APRM using different window sizes on each dataset which is due to the difference in the nature of term pairs included in each dataset. This factor is investigated in the next section in detail.

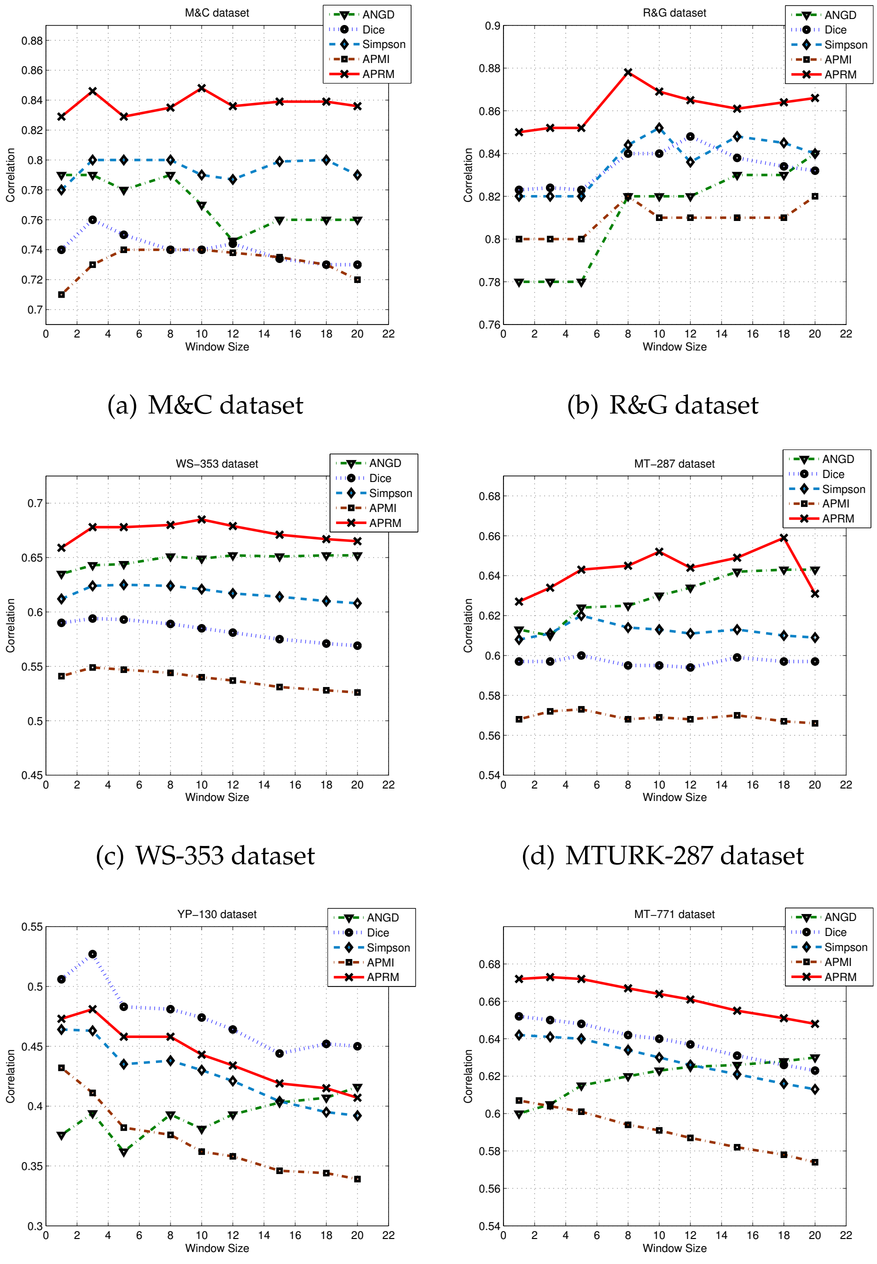
On all but one dataset, the lowest performing variant of APRM was APRM_min, as expected. However, on the verb similarity dataset (YP-130 dataset), APRM_min surpassed the other two methods on all window sizes and achieved the highest value on proximity window with $w = 3$. On YP-130 dataset, the correlation is found to have an overall decreasing trend with the increasing window size. The wider the window size the lower the correlation value. Consequently, on this dataset, looking for verbs in close proximity produced more realistic similarity scores that correlated better with human judgments than the scores of wider window sizes. The results also revealed that humans assigned low similarity scores to the verb pairs that is why the APRM_min correlated well with human judgments. However, the maximum correlation value achieved on verb similarity dataset is still the lowest correlation achieved on any dataset. This indicates that the Wikipedia-based method is not good at handling verb relations. This is in accordance with Zesch's finding [228] that classical verb similarity is better modeled in WordNet (which is a linguistic resource) than in Wikipedia (which is a collaboratively-made crowd resource).

6.4.2 Effect of Window Size

Both symmetric and asymmetric approaches discussed in the previous section used a proximity window of size $2w+1$ to compute the co-occurrence based probabilities. This proximity window is based on selecting w number of words on either side of a target word. Changing the size of the proximity window affects the overall performance, so it is critical to analyze the behavior of APRM on different window sizes before actually deciding on an optimal window size. To understand the behavior of the APRM measure on various window sizes, nine different windows sizes were tested as shown in the Figure 6.3. APRM is presented by a solid line in each graph. For each window size, the relatedness scores of term pairs in each dataset are computed and correlated with manual scores. Figure 6.3 shows that the difference in correlation values is very small which demonstrates that APRM is not very sensitive to the window size: the difference of maximum and minimum correlation produced by APRM_avg was not greater than 0.026 on all datasets except YP-130 datasets, where this difference increased to 0.066. Overall, using a proximity window with $w = 10$ produced consistently good results. The Spearman's correlation of APRM_avg using $w = 10$ on M&C and R&G datasets is 0.85 and 0.87 respectively. However, on verb similarity dataset, it was 0.44 which means that APRM_avg is not a measure of choice for computing verb similarity. On MTURK-287, which is a complicated relatedness-based dataset, APRM_avg performed well with Spearman's correlation of 0.65. The following section compares APRM using proximity window with $w = 10$ with state-of-art approaches on various datasets.

6.4.3 Symmetry Vs. Asymmetry

Figure 6.3 also compares the performance of APRM_avg with that of four symmetric measures: Adapted Normalized Google distance (ANGD), the Dice coefficient, the Simpson coefficient and Adapted PMI (APMI). It is



(a) M&C dataset

(b) R&G dataset

(c) WS-353 dataset

(d) MTURK-287 dataset

(e) YP-130 dataset

(f) MTURK-771 dataset

Figure 6.3: Effect of changing window size on the performance of various relatedness measures on all datasets.

clear from the figure that asymmetry-based measure, APRM_avg, has surpassed all the symmetric measures on all datasets except YP-130 dataset. The other asymmetry-based measure APRM_max has also performed better than all symmetric measures on three dataset and equal to the best symmetric measure on two datasets. Among symmetric measures, there is no clear winner. The Simpson coefficient outperformed the other symmetric measures on two datasets (M&C and R&G), while achieved second best correlation on other four datasets. Similarly, ANGD surpassed others symmetric measures on WS-353 and MTURK-287 datasets while the Dice coefficient outperformed other symmetric measures on YP-130 and MTURK-771 datasets. Note that on the YP-130 dataset, the Dice coefficient turned out to be the overall best performing measure. This experiment clearly highlights the advantage of using asymmetric measures over symmetric measures for semantic relatedness computation.

6.4.4 Similarity Vs. Relatedness

Table 6.1: Performance comparison of symmetric and asymmetric association based measures on two subsets of WS-353: WS353-Sim and WS353-Rel using proximity window with $w = 10$.

Aspect	Measure	Datasets		
		WS353-Sim	**WS353-Rel**	**WS353-All**
		(ρ)	(ρ)	(ρ)
Symmetric Relatedness	ANGD	0.72	0.64	0.623
	Dice	0.70	0.61	0.64
	Simpson	**0.73**	0.63	0.63
	APMI	0.65	0.58	0.591
Asymmetric Associations	APRM_avg	**0.73**	**0.65**	**0.69**

Note: All values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

Since the WS-353 dataset consists of both semantically similar and related word pairs, it is desirable to compare the performance of all measures on the similarity and relatedness-based subsets of this dataset separately. The results on similarity and relatedness subsets of WS-353 dataset are shown in Table 6.1. On the similarity subset, WS353-Sim, APRM_avg preformed comparable to the Simpson coefficient. However, on the relatedness subset, WS353-Rel, APRM_avg performed better than other measures. This shows that asymmetric association based APRM is a better indicator of relatedness and can be a good choice for detecting semantically related words but still performs as well as the other measures on semantic similarity based datasets. Overall, on WS353-All dataset, consisting of both similarity and relatedness based term pairs, APRM_avg outperformed all other measures by a good margin.

6.4.5 Comparison with State-of-art Approaches

Table 6.2 compares the performance of APRM_avg with the existing state-of-art approaches using a proximity window with $w = 10$ on the MTURK-771 dataset. The performance of APRM_avg is compared with five state-of-art approaches. *Distributional Similarity (DS)*, formulated by Dagan *et al.* [40] and Lee [108], is a proximity assumption based vectorial approach. It generates vectors of weighted word frequencies with which the vector words co-occur with input words (within a proximity window of size 3) in a corpus and computes the cosine similarity of these co-occurrence vectors. *Explicit Semantic Analysis (ESA)* [58] incorporates human knowledge into relatedness computation by constructing concept vectors and comparing them using Cosine similarity. *Temporal Semantic Analysis (TSA)* [167] incorporated temporal dynamics to enhance text relatedness models. TSA represented each input word as a concept vector and extended the static vectorial representation of words with temporal dynamics. *Latent Dirichlet Allocation (LDA)* [14] is a generative topic model that finds the latent topics underlying a document based on word distributions. It assumes each document as a mixture of topics where each topic is attributed by a distribution of words. *Constrained LEArning of Relatedness (CLEAR)* [71] is a machine learning based approach for computing word relatedness constrained by the relatedness of known word pairs. CLEAR learns the word relatedness based on word co-occurrences extracted from three different text corpora. The optimization process is constrained by a set of synonymy-based related word pairs extracted from WordNet. Using the co-occurrence statistics, CLEAR represents each input word as a vectors of latent vectors and computes Cosine similarity between the vectors of the input word pair. The results of these approaches on MTURK-771 dataset were originally reported in [71]. MTURK-771 is the largest and the most recent of all similarity and relatedness datasets, hence not reported much in the literature. Moreover, MTURK-771 is a relatedness-based dataset.

Table 6.2: Performance comparison of APRM_avg measure with existing state-of-art measures on the MTURK-771 dataset.

Measures	Source	Datasets
		MTURK-771
		(ρ)
DS	Corpus	0.57
ESA	Wikipedia	0.60
TSA	Wikipedia & Corpus	0.60
LDA	Corpus	0.61
CLEAR	WordNet & Corpus	**0.72**
APRM_avg ($w = 10$)	Wikipedia & Corpus	0.65*
APRM_avg (best)	Wikipedia & Corpus	0.66*

* values are statistically significant at $\alpha = 0.01$ level (two-tailed) and p-value $< .001$.
Bold values indicate the best correlation-based performance of any measure on a specific dataset.

On the MTURK-771 dataset, APRM_avg surpassed all other approaches except CLEAR [71] which produced highest correlation on this dataset. CLEAR combines the distributional statistics from multiple informal text corpora and learns the relatedness of words constrained by the known related word pairs extracted from WordNet synsets. However, two aspects of the MTURK-771 dataset construction are worth mentioning: first, this dataset is based on extracting related word pairs which are linked with graph distances between 1-4 in the WordNet graph; and second, the word pairs used in this dataset are all nouns. Clearly, the first point indicates a bias in CLEAR as it learns its relatedness constrained by the related word pairs extracted from WordNet which is the same knowledge source used to create the MTURK-771 dataset. Although CLEAR achieved the highest performance, it depends on preprocessing of three huge text corpora to generate and use training data for learning and optimization of model parameters, which is the downside of this approach. On the other hand, APRM_avg, without requiring huge computational resources or training data, outperformed most of the state-of-art approaches in predicting relatedness scores. The good performance of APRM_avg on the MTURK-771 dataset conforms that it is quite good at detecting the relatedness of noun-noun term pairs.

6.5 Chapter Summary

This chapter presented a method for computing term relatedness using associative probabilities based on the idea of directional contexts. The evaluation of APRM on all the publicly available benchmark datasets of term similarity and relatedness demonstrated superior performance to most other state-of-art existing approaches. The APRM measure is found to be particularly useful in predicting relatedness of noun term pairs. However, it did not perform well on verb relatedness dataset. The reason of this low performance is partly the nature of the underlying knowledge source, Wikipedia, which being an encyclopedia does not focus on verbs as much as on nouns and partly because humans estimates of verb similarity are generally lower than that of noun similarity.

Chapter 7

Task-driven Evaluation

An indirect evaluation of a semantic association measure involves exploiting it in an application that critically relies on good estimates of semantic associations. The performance improvement of such an application reflects a direct impact of employing a good semantic association measure. The literature review suggested the usefulness of investigating the performance of a semantic association measure by an indirect evaluation that employs the measure in a benchmark-independent application setup. This kind of evaluation does not suffer from the size limitation of the datasets and is free of any bias as observed in manual judgments[1]. A well known application of word association measures is to use them in the word choice task [206, 199, 142]. The more word choice questions a semantic association measure solves correctly, the better the performance.

Section 7.1 presents a discussion on the limitations of previously presented Wikipedia-based measures in order to select a suitable measure for solving the word choice task. Section 7.2 introduces the fundamentals of

[1] Such a bias is found in the WordSimilarity-353 dataset, where some word pairs such as (`Arafat,terror`) were given higher ratings than some other words pairs such as (`Arafat,peace`) due to geopolitical bias of the human judges of that dataset. Reproducing human judgments on such word pairs in some other part of the globe can give very different human ratings as well as inter-rater agreements.

the word choice task. Section 7.3 includes a survey of previous attempts to solve this task using various techniques. Section 7.4 details the evaluation setup including the datasets and evaluation metrics. Section 7.5 reports the performance of the APRM measure on both relatedness and synonymy-based word choice questions. Section 7.6 concludes the chapter by presenting a brief summary.

7.1 Selection of an Association Measure

The thesis presented different measures of semantic association computation in the previous chapters. The strengths and limitations of each measure were also analyzed. Although the knowledge source based semantic measures performed well on the task of ranking semantic associations, they required mapping of input words to corresponding Wikipedia articles, which may or may not exist, thus leading to knowledge acquisition bottleneck. Moreover, due to the encyclopedic nature of Wikipedia, it largely focuses on noun concepts. This is both the strength as well as the weakness of Wikipedia as a knowledge source. There are many applications that target nouns such as word sense disambiguation, keyphrase extraction and topic modeling. However, the formal knowledge source based semantic measures suffer from coverage limitation when used in certain applications where the target words belong to part-of-speech classes other than nouns. Also, the inherent ambiguity in natural language text also limits the performance of such measures. The performance of such measures drops down when the textual data involves frequent ambiguous concepts, due to the inaccuracy of the automatic disambiguation process. This is evident from the comparison of automatic and manual disambiguations (Section 3.4 of the thesis), where the performance of semantic measures suffered from inaccurate automatic disambiguations.

On the other hand, the corpus based measure APRM does not require matching of input terms with corresponding Wikipedia articles. It overcomes this limitation by considering the context of words at the corpus level rather than at the article level. Unlike knowledge source based measures, APRM is a knowledge source independent measure and can be easily adapted to any text corpora. Also, it is not semantically biased towards any POS-class and is observed to work well on estimating semantic association of cross-POS word pairs because it computes word co-occurrence-based probabilities from unstructured text of the Wikipedia corpus rather than explicitly considering Wikipedia concepts. Hence, it is not sensitive to cross-POS words. The previous chapter has shown the effectiveness of the APRM measure on computing semantic association. Hence, this chapter presents a task-driven evaluation of the semantic association measure APRM in solving word choice questions on a range of datasets used for this task.

7.2 The Word Choice Task

A word choice question consists of a given problem word and a list of candidate words or phrases. The goal is to select the most closely related candidate word to the given problem word. A word choice question is given as:

gynecology

a) `female health`

b) `habits`

c) `athletics`

d) `brain`

To perform the task automatically, the relatedness between each candidate word and the problem word is computed using an association measure and a candidate word having maximum relatedness with the problem word is selected as the correct option. There is always one correct candidate, which in the case of above example is `female health`. A tie is considered among candidate words if two or more candidate words are equally related to the problem word. If one of tied candidate words is the correct answer then the question is considered to be correctly solved. However, the score of the question is penalized by the number of tied candidate words for that question. So, a problem q which is correctly solved without a tie is assigned a score of 1 but with two or more ties will be scored according to the following formula:

$$Score_q = \begin{cases} \frac{1}{\#\ of\ tied\ candidates} & \text{if tied candidates} \geq 2 \\ 1, & \text{if one correct candidate} \end{cases} \tag{7.1}$$

The overall score of a semantic association measure based system for solving the word choice task on a specific dataset with Q number of questions is the sum of individual score of all questions.

$$Score = \sum_{q \in Q} (Score_q) \tag{7.2}$$

A semantic association measure based system might not be able to compute scores for certain questions due to the coverage limitation of its underlying knowledge source, in which case those questions are not considered as attempted.

7.3 Related Work

The word choice task is a well known application used for performance evaluation of semantic association measures.

Launder and Dumais [104] used word choice problems for empirical evaluation of their technique, *Latent Semantic Analysis* (LSA), based on vectorial semantics. LSA relies on collecting the relative frequencies with which a word co-occurs with other words and used them to generate word vectors. LSA computed the cosine of the angle between two word vectors as the word similarity score. The performance of LSA was evaluated using a dataset of 80 word choice questions selected from the synonymy portion of *Test Of English as a Foreign Language* (TOEFL) provided by *Educational Testing Service* (ETS). LSA achieved an accuracy of 65% on the TOEFL dataset.

Turney [206] used an unsupervised learning algorithm based on the idea of *Pointwise Mutual Information and Information Retrieval* (PMI-IR) to mine the web for synonymy. PMI-IR analyzed statistical data collected by a search engine from the web. Statistical approaches to synonymy detection (which is a special case of the word choice task) are based on word co-occurrences [120]. There is a subtle difference between collocation and co-occurrence. Word collocations are grammatically bound, ordered occurrences of words, whereas co-occurrences are more general phenomenon of words that are likely to occur in the same context [206]. Like LSA, PMI-IR is also based on word co-occurrences. Based on the proximity assumption, PMI-IR computed the score of a candidate word as the conditional probability of the problem word given the candidate word. PMI-IR was evaluated on a new dataset for synonymy detection. This dataset consists of 50 word choice questions collected from the tests for students of *English as a Secondary Language* (ESL). Turney evaluated his approach on both TOEFL and ESL datasets and showed that PMI-IR performed better than LSA on the TOEFL dataset with 73.70% accuracy and achieved 74% accuracy on the ESL dataset.

Jarmasz and Szpakowicz [95] defined the degree of synonymy to be the semantic similarity and computed it using the shortest path distance between the sets of references of given two words in Roget's thesaurus taxonomy. They used their measure for synonymy detection using two existing datasets on the word choice task and created a new dataset RDWP, of 300 word choice problem collected from *Reader's Digest Word Power* (RDWP) game. In the RDWP game, readers are required to pick the candidate that is closest in meaning to the given problem word. Following [104, 206], Jarmasz and Szpakowicz reported their results using the percentage of correctly answered word choice questions. Their approach using Roget's thesaurus achieved 78.75% accuracy on TOEFL, 82% accuracy on ESL and 74.33% accuracy on the RDWP dataset.

Inspired by PMI-IR, Higgins [81] proposed *Local Context-Information Retrieval* (LC-IR), a statistical method for computing word similarity. Their approach was based on the assumption of absolute adjacency-based parallelism, which takes into account the count of bi-gram patterns of given two words. They reported the results on the TOEFL, ESL and RDWP datasets using accuracy. LC-IR was found to perform better than Roget's thesaurus-based approach on both TOEFL (with 81.3% accuracy) and ESL (with 78% accuracy) datasets.

Mohammad *et al.* [142] proposed cross lingual distributional profiles of concepts for computing semantic associations and evaluated their measure in solving word choice questions. They also created a German dataset, consisting of 1008 word choice questions collected from 2001 to 2005 issues of the German language edition of Reader's Digest. They argued that the evaluation metric $Score$ by itself does not show the true performance picture of a semantic measure due to the number of ties, which could be high for many measures. Therefore, they used precision, recall and F score as the evaluation metrics. Their definitions of precision ($P = Score/Attempted$), recall ($R = Score/Total$) and F-measure ($F = \frac{2 \times P \times R}{P+R}$) were different from the classical precision, recall and F-measure used by information retrieval approaches.

Zesch *et al.* [230] compared generalized path-based measures and concept vector based measures using three different knowledge sources: WordNet, Wikipedia and Wiktionary. They evaluated the measures using both English and German RDWP datasets. However, they computed recall as $R = |A|/n$, where $|A|$ is the number of attempted word choice questions and n represents the total number of word choice questions. Following Zesch *et al.* [230], Weale *et al.* [216] used the PageRank algorithm over the Wiktionary category graph to compute a number of semantic similarity metrics. Those metrics were evaluated on the task of word choice questions using the TOEFL, ESL and RDWP datasets. They used overall percentage *Raw* (correct number of guesses over all question) and precision-based-percentage *Prec* (correct number of guesses by attempted questions) as the evaluation metrics for their approach.

Pilehvar *et al.* [161] proposed a unified approach to computing semantic similarity using a sense-based probability distribution and evaluated their approach in three semantic tasks. They tested their approach on synonymy detection using the TOEFL dataset and reported the results using accuracy as the evaluation metric. Their method achieved 96.25% accuracy on the TOEFL dataset.

Mohammad *et al.* [199] proposed a hybrid semantic association measure using various Wikipedia-based features. They used their measure for solving the word choice task using the RDWP dataset. They followed the same evaluation criteria as used by Zesch [228] and reported the results using accuracy, coverage and H measure.

7.4 Evaluation

The APRM_avg measure is used to compute the pairwise relatedness of each candidate word and the problem word of each question. A candidate word with maximum APRM score is selected and matched with the correct answer for that question. If the automatic answer matches the manual answer then the question is assumed to be solved and is assigned a score of 1. If more than one candidate word gets the same APRM score with the target word then the score is penalized by the number of ties. In case of an incorrect answer, zero score is assigned to that question. The remainder of this section details the performance metrics and datasets used in the evaluation and comparison of the APRM_avg measure (using proximity window with $w = 10$) with the state-of-art approaches on the word choice task.

7.4.1 Performance Evaluation Metrics

To evaluate the performance of semantic association measures on the word choice task, different evaluation criteria have been used. The simplest of the performance metrics used for evaluation of semantic measures on the word choice task is accuracy. However, later approaches identified the limitations of existing evaluation metrics and proposed new evaluation metrics.

Jarmasz and Szpakowicz [95] used the overall $Score$ (Equation 7.2) for performance evaluation of an association measure on the word choice task. However, Zesch [228] found this evaluation criteria problematic. He argued that this evaluation parameter favors those approaches that attempt more questions based on just random guessing. Mohammad *et al.* [142] used precision, recall and F-measure as the evaluation metrics. However, they defined recall as $Recall = \frac{Score}{|Q|}$, where $Score$ is the total score of a semantic association measure on a dataset consisting of $|Q|$ questions. This recall returns 0 if the $Score = 0$, regardless of the number of questions attempted. This is different from the recall used for evaluation of information retrieval systems, where if a system retrieves all documents, its recall is one, regardless of the relevance of the retrieved documents. For the word choice task, recall is equal to one if and only if all the questions are answered correctly, hence, it is of very limited value. Thus, following Zesch [228] we decided not to use precision and recall but evaluate our approach using a combination of accuracy and coverage. Accuracy shows how many of the attempted questions are correct. It is defined as:

$$Acc = \frac{Score}{|A|} \tag{7.3}$$

where $Score$ is the overall score of a semantic measure on a word choice dataset and $|A|$ represents the number of questions attempted by a semantic association measure. Coverage indicates how many of the word choice questions are attempted. It is defined as:

$$Cov = \frac{|A|}{|Q|} \tag{7.4}$$

where $|A|$ is the number of questions attempted and $|Q|$ is the total number of questions in a given dataset. The overall performance evaluation of a semantic measure is based on calculating the harmonic mean of the accuracy and coverage. Thus, the evaluation metric H-measure is defined as:

$$H = \frac{2 \times Acc \times Cov}{Acc + Cov} \tag{7.5}$$

H-measure is analogous to F-measure (which is the harmonic mean of precision and recall) in information retrieval. H-measure balances both evaluation metrics by considering the accuracy with respect to the proportion of word choice questions covered.

7.4.2 Datasets

The following datasets were used in the experiments solving the word choice task.

- **rd_100 dataset** consists of 100 word choice questions collected by Jarmasz and Szpakowicz [95]

- **rdCANADA_200 dataset** consists of 200 word choice questions from Reader's Digest collected by Turney [206].

- **rdNOUN_153 dataset** consists of a combination of noun word choice questions from both rdCANADA_200 and rd_100 datasets.

- **RDWP dataset** consists of 289 word choice questions collected by Jarmasz and Szpakowicz [95] from the *Word Power* game of the Canadian edition of Reader's Digest from year 2000-2001 [95].

- **ESL dataset** consists of 50 synonymy-based word choice questions found in the *English as a Second Language* (ESL) tests collected by Turney [206].

- **TOEFL dataset** consists of 80 synonymy-based word choice questions found in the *Test of English as a Foreign Language* (TOEFL) collected by Launder and Dumais [104].

7.5 Results and Discussion

This section reports the performance of the APRM measure on two types of word choice questions: relatedness-based word choice questions, and synonymy based word choice questions. The first experiment uses APRM in solving relatedness-based word choice questions. Since APRM focuses on computing semantic relatedness due to its underlying proximity assumption, its performance on solving relatedness-based word choice questions is discussed in detail. However, to investigate the behavior of the APRM measure on synonymy, the second experiment uses APRM in solving synonymy-based word choice questions.

7.5.1 Relatedness-based Word Choice Problems

The performance of APRM is reported on three relatedness-based Reader's Digest datasets in Table 7.1.

It is clear from the table that on all three datasets APRM performed well. The highest accuracy is achieved using a proximity window with $w = 5$ on the rdCANADA_200 dataset, which is the largest of the datasets reported in Table 7.1. However, the highest coverage and the best H-measure were achieved on the rdNOUN_153 dataset, which consist of nouns words. This again conforms the effectiveness of using a Wikipedia-based measure on computing noun-noun associations. On all datasets,

Table 7.1: The Performance of APRM on various word choice datasets. The best performance on each dataset is shown in bold.

Dataset	Window	Total	Attempted	Score	# of Ties	Acc	Cov	H-measure
rd_100	5	100	79	53	2	0.67	0.79	0.73
	10	100	83	57	2	0.69	0.83	0.75
	20	100	84	58	0	0.69	0.84	**0.76**
rdCANADA_200	5	200	162	125	0	0.77	0.81	**0.79**
	10	200	172	124	0	0.72	0.86	**0.79**
	20	200	176	123.5	3	0.70	0.88	0.78
rdNOUN_153	5	153	136	101	0	0.74	0.88	**0.80**
	10	153	141	100	0	0.71	0.92	**0.80**
	20	153	143	98	0	0.69	0.93	0.79

there was an increasing trend in the number of attempted questions with increasing the proximity window size, which consequently resulted in better coverage of questions. The accuracy on both rdCANADA_200 and rdNOUN_153 datasets decreased with increasing the window size, however on rd_100 dataset an opposite trend was observed between the accuracy and window size. On all three datasets, optimal H-measure values were achieved using a proximity window with $w = 10$.

Comparison with Existing Approaches

Table 7.2 compares the performance of APRM with multiple existing state-of-art approaches on the RDWP dataset. The results of all existing approaches (except Zesch [228]) on the word choice task are reported in [199]. All existing approaches other than Gurevych and Zesch are computed using Wikipedia as the underlying knowledge source. Gurevych is based on WordNet and Zesch is based on Wiktionary [228].

For comparison with existing approaches, APRM with three different window sizes is considered: APRM with $w = 5$, APRM with $w = 10$ and APRM with $w = 20$. The results are encouraging. APRM surpassed all

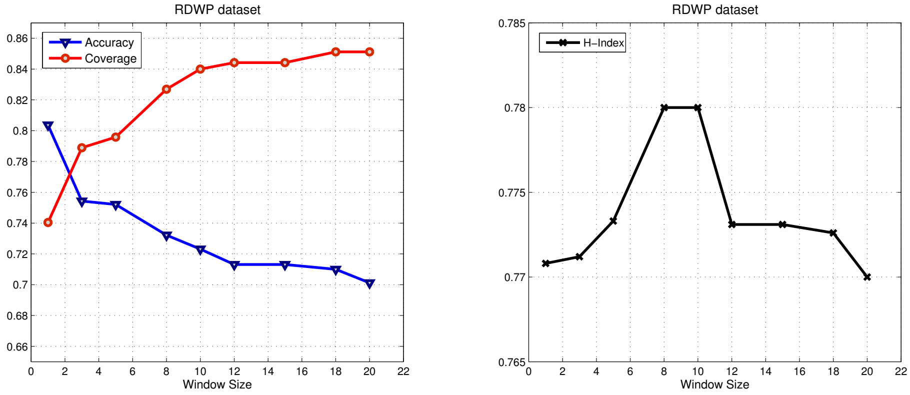
Table 7.2: Performance comparison of APRM with state-of-art approaches on the RDWP dataset. The best performance is shown in bold.

Method	Attempted	Score	# of Ties	Acc	Cov	H-measure
Lesk [111]	223	63	6	0.28	0.77	0.41
Rada [166]	222	79.7	70	0.36	0.77	0.49
WuP [222]	214	70.2	73	0.33	0.74	0.46
LC [106]	222	79.9	70	0.36	0.77	0.49
Resnik [171]	96	42.1	21	0.44	0.33	0.38
JC [97]	222	54.5	21	0.25	0.77	0.38
Lin [115]	222	52.5	201	0.24	0.77	0.37
ESA [58]	280	144	2	0.51	0.97	0.67
WLM [138]	141	69.8	3	0.50	0.49	0.49
Gurevych [69]	182	148.8	10	0.82	0.63	0.71
Zesch [228]	156	128.3	3	0.82	0.54	0.65
Mohammad [199]	253	196.66	14	0.77	0.87	**0.82**
APRM $(w = 5)$	232	174	2	0.75	0.80	0.78
APRM $(w = 10)$	244	176	2	0.72	0.84	0.78
APRM $(w = 20)$	248	173.5	3	0.70	0.86	0.77

other methods except that of Mohammad [199], who used a hybrid approach based on Wikipedia features mainly articles, categories, redirects and category graph. However, they reported their results using various thresholds ranging from $\theta = 0.30$ to $\theta = 0.40$, hence could not find an optimal threshold. The reason for the good performance of APRM on the RDWP dataset is mainly due to the nature of semantic associations of words included in the dataset. The RDWP dataset consists of similarity and relatedness based word choice questions. APRM, being a semantic association measure, is good at estimating these types of semantic associations, hence was able to perform well on this dataset.

Further Analysis

The performance of APRM on the RDWP dataset is further analyzed to understand the impact of window size on accuracy, coverage and H-measure.



(a) Accuracy Vs. Coverage (b) H measure

Figure 7.1: Correlation of the window size with performance evaluation metrics of the word choice task on the *Reader's Digest Word Power* (RDWP) dataset.

Figure 7.1(a) presents a comparison of accuracy and coverage of APRM on various window sizes. Again, note that the accuracy of the APRM-based system is inversely proportional to the window size. On the other hand, the coverage of the word choice questions increases with increasing window size. This analysis is useful because it empirically shows that for loosely-coupled synonyms and strongly related words, smaller windows result in higher accuracy by avoiding any possibility of having two terms in close proximity by chance. Hence, it demonstrates that a narrow proximity window leads to better system performance. Figure 7.1(b) analyses the impact of window size on the H-measure. A cursory look at the graph reveals that an optimal H-measure value (in other words an optimal combination of the accuracy and coverage) was achieved by using a proximity window size with $w = 8$ and $w = 10$. The graph shows that the performance of the system increases with the increase in window size till $w = 8$. Further experimentation revealed that by yielding the same performance (H-measure = 0.779) on $w = 9$, the system achieved an equilibrium from $w = 8$ to $w = 10$ after which it started decreasing again. hence, it can be deduced from this experiment that a good balance of the coverage and accuracy can be achieved on a size window that is neither too small nor too large ($w = \{8, 9, 10\}$).

In order to gain further insight into the behavior of the presented system for solving the word choice task, various questions were individually analyzed. Consider the following question as an example of the relatedness-based word choice task (correct answer is in bold).

`Therapy` (*Question # 197 — RDWP dataset*)

a) `Repeating Melody`
b) `Heat`
c) **`Healing`**
d) `Side Dish`

Here is the list of candidates ranked according to APRM-based association scores. The candidate word selected by APRM-based system as the answer is also shown in bold in the list of ranked-candidates.

a) **`Healing`**
b) `Heat`
c) `Side Dish`
d) `Repeating Melody`

It is clear from the ranked list of candidate words that the system was actually able to to identify the correct answer which is `Healing`. This shows that the APRM measure performed well on the relatedness-based word choice task. However, consider another question as an example of the synonymy-based word choice task from the same dataset.

Compassionate (*Question # 265 — RDWP dataset*)

a) Intense
b) **Sympathetic**
c) Indifferent
d) Friendly

The ranking of the candidates produced by the APRM-based system (with selected candidate shown in bold) is as follows.

a) **Friendly**
b) Sympathetic
c) Intense
d) Indifferent

The correct answer for this question is the candidate word Sympathetic whereas the system placed the word Friendly on the top due to its highest association score with the given word. The reason for this incorrect selection is the higher frequency of co-occurrence of the words Friendly and Compassionate as compared to the words Sympathetic and Compassionate (which being synonymous words do not co-occur in close proximity very often). Hence, due to the underlying proximity assumption of the APRM measure the system was unable to correctly solve the synonymy-based word choice question.

7.5.2 Synonymy-based Word Choice Problems

Another variant of the word choice task is the synonymy-based word choice task, in which the correct candidate word for a given problem word is always a tightly coupled synonymous word. To investigate the behavior of APRM on this synonymy task, two datasets consisting of synonymy-based word choice questions were used. The performance of APRM on both datasets are reported in Table 7.3.

On TOEFL dataset, students with English as their second language

Table 7.3: The performance of APRM on the synonymy-based word choice task using two datasets: ESL and TOEFL. The best performance is shown in bold.

Dataset	Window	Total	Attempted	Score	# of Ties	Acc	Cov	H-measure
ESL	5	50	45	30	0	0.66	0.90	**0.76**
	10	50	47	29	0	0.62	0.94	0.75
	20	50	48	27	0	0.56	0.96	0.71
TOEFL	5	80	72	52	0	0.73	0.90	**0.80**
	10	80	75	51.5	1	0.69	0.94	0.79
	20	80	75	50.5	1	0.67	0.94	0.78

achieved an accuracy of 67.5%. Many approaches tried to solve this task but recently, Bullinaria and Levy [24] have reported 100% accuracy on this task by solving all the questions correctly. Their approach is also uses word co-occurrences but is based on *topicality assumption* which considers two words to be similar if they share the same context. This approach favors the synonymous words which share maximum context. Table 7.3 shows that APRM performed reasonably well on H-measure though not the best on both synonymy-based datasets. It is worth mentioning that the smaller window size resulted in better predictions of synonyms for given problem words on both datasets. APRM relies mainly on the conditional probability of finding one term given another term, which logically makes it unsuitable for synonym detection as there is less chance of finding synonymous words in close proximity. However, due to huge knowledge coverage of Wikipedia corpus and the feature matching component of APRM, it was still able to find the synonyms reasonably well. For synonymy detection, distributional approaches based on weighted vectors of context terms (such as second order context vector based approaches) could be quite effective as these look for the neighboring context of both given words, which would have the highest overlap for synonymous words.

7.6 Chapter Summary

This chapter presented a task-driven evaluation of the Wikipedia-based APRM measure by investigating its performance in an application that critically relies on good estimates of semantic associations. For this purpose, APRM was used to solve the word choice task. The performance evaluation of the APRM-based system on word choice task demonstrated the effectiveness of the APRM measure. Overall, the APRM-based system surpassed most other systems on the relatedness-based word choice task. It was found that the increasing window size resulted in increasing coverage but decreasing accuracy on the word choice questions of the RDWP dataset. Empirical analysis of results demonstrates that APRM performed extremely well on relatedness-based word choice questions, but only reasonably well on synonymy-based word choice questions because synonymous words rarely occur in close proximity and hence are difficult to handle by approaches based on the proximity assumption.

Chapter 8

Common Sense Causality Detection

The previous chapter presented an application of the APRM measure in solving the word choice task and empirically demonstrated the effectiveness of the APRM measure on the task. This chapter investigates the performance of the APRM measure on the task of commonsense causality detection, which requires an exploitation of causal semantics. This is the first study (to the best of our knowledge) that exploits semantic capabilities of Wikipedia as a text corpus on the *Choice Of Plausible Alternatives* (COPA) task requiring commonsense causality detection [66]. The goal of COPA evaluation is to investigate an automated system's ability of finding causal connections between sentences.

Section 8.1 of the chapter introduces the idea of commonsense causal reasoning followed by an overview of the approaches that explored different aspects of commonsense reasoning. Section 8.2 discusses the COPA evaluation benchmarks and details the performance of existing approaches on these benchmarks. Section 8.3 presents a causality detection approach based on using APRM as the causal relatedness measure to solve the task. Section 8.4 presents the results and includes a detailed discussion of the effectiveness of the APRM measure in comparison with other Wikipedia-based measures on the task of commonsense causality detection. Further discussion identifies the potential factors that could lead to performance improvements and talks about future recommendations. Section 8.5 presents a brief summary of the chapter.

8.1 Common Sense Causal Reasoning

Linguistic discourse consists of a sequence of coherent clauses connected in a logical way by *discourse relations* such as temporal, causal and contrast [163]. These relations characterize the manner in which various parts of the text are connected. Depending on the usage of *discourse connectives*, also known as *triggers*, these relations are classified as either implicit or explicit relations. For instance, the trigger `however` signals a contradiction of something stated previously.

Commonsense reasoning involves the understanding of concepts and the relations that underlie everything we know and talk about. To make a computer capture and understand this knowledge is a difficult task because most of it is so obvious that it is rarely talked about in the knowledge sources. Exploring this implicit knowledge makes commonsense reasoning a challenging task in computational linguistics and artificial intelligence, hence it has been a research focus since its inception.

Common sense causality detection is the process of identifying the connection of two sentences based on the causal relation. Any causative argument involves two components: the *cause* and its consequent *effect*. For instance, consider the following causative argument.

> "The young campers felt scared **because** their camp counselor told them a ghost story."

Here, the *cause* is represented by the `ghost story telling` that resulted in the *effect* of `feeling scared`. Causative connections are fundamental to human language. English language involves a wide range of cause-effect relations. Girju [61] identified the following two broad classes of causative arguments in English.

- **Explicit causation** consists of a simple causative argument containing directly relevant words such as `cause`, `consequence`, `effect`. It also involves ambiguous causative arguments signaled by words such as `generate`, `create`, `trigger` and `produce`. Disambiguation of such words can be helpful to identify the causal relations.

- **Implicit causation** consists of more complex causative arguments involving inference based on lexical and semantic analysis of background knowledge.

Theoretical investigations of commonsense causal reasoning range from cognitive psychology to computational intelligence and from artificial intelligence to economics [41, 75, 32, 13, 86].

Fletcher *et al.* [55] analyzed the role of causal reasoning in the comprehension of simple narrative texts. Their research supported the claim that narrative comprehensions aim at finding the causal links that connect the opening of the text to its final outcome. Broek [208] proposed a process model of inference generation in text comprehension. According to him causal dependencies found in the text play a vital role in the construction of text representation in the short term memory. The criteria for causality guides the inferential process. He identified two types of causal inferences: *backward inferences*, which connects the current event to the prior events and *forward inferences*, which produce expectation about the upcoming events in the text.

Singer *et al.* [189] investigated the role of causal bridging in the discourse comprehension. Bridging inferences connect subsequent portions of the text to their antecedent portions. Their focus was on identifying the bridging inferences that connect two ideas by a causal relation. They identified various factors for detecting causal relations and emphasized on the importance of *change of state* in addition to causal conjunctions and causal verbs for detecting an *effect*.

Hobbs [86] introduced the notion of *causal complex* as a complete set of conditions necessary for a causal action to occur. He discussed the issue of differentiating the causal complex from what is outside its boundaries. He also analyzed the eventualities within the causal complex that can be referred to as the *cause* from those that do not.

Commonsense causal reasoning has a multidisciplinary scope in which various disciplines attempted to refine the notion of causality in agreement with human commonsense. This also involves an extensive discussion of what factors contribute to causal reasoning [86]. One factor that has profound impact on causal reasoning is the context of a causal relation, which is also referred to as *causal field* or *causal complex* [86]. Events that are involved in a causal relation always occur in some explicit or implicit context. This contextual information yields the inferences that bridge the two events into a causal relation.

Various research efforts to analyze and understand commonsense reasoning in general and commonsense causal reasoning in particular have been made. Speer *et al.* [191] proposed *AnalogySpace*, a technique that facilitates commonsense reasoning over a very large knowledge base of natural language assertions by forming an analogy closure over it. Their technique was able to make distinction between various analogous classes such as (`good,bad`) or (`easy,hard`). *AnalogySpace* used data from *Open Mind Common Sense* (OMCS) project and self organized it along dimensions that were able to distinguish between analogy classes. Girju [61] proposed an inductive learning approach for automatic detection and extraction of causal relations. His approach was able to automatically discover the lexical and semantic relations that are necessary constraints for disambiguating the causal relations used in the question answering task. Jung *et al.* [98] constructed a large-scale Situation ontology by mining the *how-to* instructions from the web.

Mihaila and Ananiadou [130] proposed a machine learning method for automatic recognition of discourse causality triggers in the biomedical scientific text. They concluded that the shallow approaches such as dictionary matching and lexical parsing performed very poorly due to highly ambiguous causal triggers. They showed that a classifier using a combination of lexical, syntactic and semantic features performed the best.

The ability to understand stories automatically involves filling in many missing pieces of information that is not explicitly stated in the story. The use of commonsense reasoning was investigated to comprehend script-based stories [145]. An information extraction module capable of handling the templates generated by commonsense reasoning module was used in comprehending script-based stories. Rajagopal *et al.* [168] proposed an approach for multi-word commonsense expression extraction from informal text. In their approach they used AffectNet [25] to derive commonsense knowledge based on syntactic and semantic relatedness. AffectNet is a huge matrix of commonsense in which the rows are concepts and columns are negative and positive assertions about that concept. For example, for the concept `penguin`, a positive assertion is `penguin is a bird` and a negative assertion is `penguin can not fly`. Gordon and Swanson [67] proposed an automated method for extracting millions of personal stories[1] from the Internet Weblog entries. Following this work, Gorden *et al.* [65] used personal stories as a source of information about the causal relations of everyday life to automatically reason about commonsense causality.

In contrast to expert knowledge that is explicit, commonsense knowledge is usually implicit. Hence, the goal of an automated commonsense reasoning system is to make this knowledge explicit. An attempt in similar direction was the development of Cyc [110] to formalize commonsense knowledge into a logical framework. It was developed by knowledge engineers who added 10^6 handcrafted assertions into the knowledge base.

[1] A personal story is defined as a narrative discourse that describes a specific series of causally related events.

Cyc is an expert system with commonsense knowledge spanning everyday life objects and actions such as causality, time, space, substances, planning, emotions, ambitions, uncertainty. Cyc sorts each of its assertions according to one or more explicit contexts. However, the process of using Cyc to reason about a text is quite complex as it requires the mapping of input text from natural language to a specific logical format described by its own language CycL. This involves mapping of inherently ambiguous natural language text to unambiguous logical formulation required by Cyc. This process hindered the full availability and usage of Cyc content for most natural language interpretation tasks.

Inspired by the range of commonsense concepts, knowledge of Cyc and the semantic structure of WordNet, Liu and Singh [117] constructed ConceptNet by combining the best of both worlds. While both Cyc and WordNet are handcrafted knowledge sources, ConceptNet is automatically generated from crowd sourced knowledge of *Open Mind Common Sense* (OMCS) Corpus [117]. ConceptNet is a huge semantic network of commonsense knowledge that has various kinds of inferential capabilities. It automatically mined 250,000 elements of commonsense knowledge from OMCS [44] using a set of commonsense extraction rules and constructed an ontology of predicate relations and arguments. The semantic structure of ConceptNet is similar to WordNet while being inferentially rich like Cyc. ConceptNet nodes consist of natural language fragments that conform to specific syntactic patterns and edges represent binary relations over the nodes. ConceptNet includes 19 semantic relations classified into 7 categories such as things, events and actions. The focus of ConceptNet is on contextual commonsense reasoning over real world objects and events.

There is a range of applications based on commonsense causal knowledge extraction and representation including information extraction, document categorization, opinion mining, sentiment analysis and social process modeling.

8.2 Choice Of Plausible Alternatives

Human actions in this world are based on exploiting knowledge of causality. Humans find it easy to connect a cause to the subsequent effect but formal reasoning about causality has proved to be a difficult task in automated NLP applications because it requires rich knowledge of all the relevant events and circumstances. Automated approaches to predicting causal connections attempt to partially capture this knowledge using commonsense reasoning based on lexical and semantic constraints. However, their performance is limited by the lack of sufficient breadth of commonsense knowledge to draw causal inferences.

The progress of research in commonsense reasoning has been slow. A major reason was the lack of a unified benchmark to compare the performance of automated commonsense reasoning systems. To address this issue, Roemmele *et al.* [175] devised a benchmark of binary choice questions for investigating an automated system's ability to make judgments on commonsense causality. This benchmark is referred to as Choice Of Plausible Alternatives (COPA) evaluation. The format of every question in this benchmark consists of two components: a text statement (the premise) and two choices (alternatives). Both alternatives can have a causal relation with the premise but the correct choice is the alternative that is more plausible given the premise. Depending on the nature of the premise, two types of question format were included in the dataset: *forward causal reasoning*, which views one of the alternatives as a *plausible effect* of the premise; and *backward causal reasoning*, which assumes the correct alternative to be a *plausible cause* of the premise. Examples of both types of questions are shown in Table 8.1.

Table 8.1: Examples of two types of COPA question format.

Question # 513 (find the `Cause`)
Premise: The check I wrote bounced.
Alternative 1: My bank account was empty.
Alternative 2: I earned a pay raise.

Question # 539 (find the `Effect`)
Premise: The woman bumped into the sofa
Alternative 1: The leg of the sofa came loose.
Alternative 2: She bruised her knee.

8.2.1 COPA Evaluation Setup

Choice of Plausible Alternatives (COPA) evaluation[2] consists of 1000 questions based on commonsense causal reasoning.

The authoring methodology ensured the breadth of the topics, clarity of the language and high inter-rater agreement. To ensure the breadth of the question set, the topics were drawn from two main sources:

- The first source was a corpus of one million personal stories written in the Weblogs in August and September 2008 [67]. The topics in this corpus are more focused on social and mental concepts and less on natural and physical concepts. Hence, the topics selected from this corpus were randomly selected accounts of personal stories.

- The second source was the Library of Congress thesaurus for Graphic Materials [156]. This corpus focused more on various kinds of people, places and things that appear in photographs and other images. Hence, the topics selected from this corpus were randomly selected subject terms from the Library of Congress thesaurus for Graphic Materials.

[2] COPA evaluation was included as task-7 in the 6th international workshop on Semantic Evaluation (SemEval-2012).

A set of terms was randomly selected from both sources as the question topics. From each topic, a causal argument was instantiated. Either the cause or effect in that causal argument was used as the premise depending on whether the question was about the forward or backward causal reasoning. To make this task more difficult for purely associative methods, the incorrect alternative was set closer in form (content) to the premise but with no obvious direct causal relation. The final set consisted of 1000 questions (COPA-all), each question validated by two raters with a high inter-rater agreement (Cohen's K = 0.965).

The COPA evaluation was designed so that a baseline choosing one of the alternatives randomly has exactly 50% accuracy. To develop the evaluation dataset, the set of 1000 questions was randomly split into two equally sized subsets of 500 questions to be used as: the development set (COPA-dev) and the test set (COPA-test). No training data was provided to give the automatic causal reasoning systems the freedom to use any text corpora or knowledge repositories in their construction. The order of correct alternative was also kept random in both subsets.

The evaluation data comes with gold standard answers that are used to evaluate the performance of any automatic system on this task using accuracy as the evaluation metric. The accuracy is defined as the number of correctly identified questions divided by the total number of questions.

8.2.2 COPA Vs. Textual Entailment

The COPA evaluation was originally inspired by the task of *Recognizing Textual Entailments* (RTE), which asks whether the interpretation of one sentence necessitates the truth of another sentence. RTE consists of two text sentences: a text `T` and a hypothesis `H`. The goal of the RTE challenge is to determine whether the truth of `H` is entailed (or can be inferred) from the text `T`. Although the RTE task requires some inferential capabilities, there exist some subtle differences between COPA and RTE tasks. The first difference lies in the structural construct of each task. Given a pair of text and hypothesis, RTE interprets the meaning of hypothesis by making inferences from the text, whereas the COPA task provides two alternatives, corresponding to a given premise, both of which are plausibly true but one of them is more plausible. Moreover, the RTE judgment on entailment of two text segments is strictly positive or negative, whereas COPA causal implications are judged in degrees of plausibility (deciding on which option is more plausible).

8.2.3 Existing Approaches to COPA Evaluation

This section discusses the performance of existing approaches that have used and reported results on the COPA evaluation.

PMI Gutenberg: This was the best performing baseline system on the COPA task proposed by Roemmele *et al.* [175]. They used averaged *Pointwise Mutual Information* (PMI) between unigram words of both premise and alternatives to pick up the most plausible alternative. The PMI statistics were calculated from the Gutenberg Project [3] using a proximity window of size $w = 5$. They computed the causality score of a premise-alternative pair by adding up the pairwise PMI-based correlation scores of their unigrams.

UTDHLT Bigram PMI: This was the unsupervised system of the only participating team in the SemEval-2012 task-7 that submitted their results successfully [64]. They processed the Gigaword Corpus[4] to collect PMI statistics of bigrams within a proximity window of size `w=100` and with a slop of 2 words in between the bigram tokens themselves.

[3] `http://www.gutenberg.org`

[4] `https://catalog.ldc.upenn.edu/`

UTDHLT SVM PMI: This was the supervised system of the participating team of SemEval-2012 task-7 [64]. It computed four features including the UTDHLT Bigram PMI as one of the features. The second feature was based on computing temporal links from the given cause and effect using PMI such that events that temporally occur more frequently were given higher PMI scores. For this purpose, they used the English Gigaword corpus [157] annotated with TimeML [165] annotations. The third feature was based on using twenty four manually crafted causal patterns to pick up the alternative having the maximum number of matched dependency structures with the premise. The fourth and final feature was based on using the Harvard General Inquirer [194] to extract the context-independent word polarities and the difference between the polarity scores of premise and alternatives was used as the feature. Then, an SVM-based classifier, using linear kernel, was trained on these features to select the more plausible alternative.

PMI Story 1M: This approach is quite similar to that of Roemmele *et al.* [175]. It was proposed by Gordeon *et al.* [65], who calculated PMI statistics using a corpus of one million personal stories written in the Internet Weblogs. However, they used a wider window of size $w = 25$ and suggested that the causal information is the strongest within the scope of adjacent sentences and clauses.

PMI Story 10M: This is another system presented by Gordon *et al.* [65], in which they explored the impact of using a larger corpus on computing PMI statistics for causal reasoning. They used 10 million personal stories and calculated PMI statistics using the same window size $w = 25$. Although this system yielded the best results so far, it produced only a very slight and statistically insignificant gain over their previous system—PMI Story 1M.

Table 8.2 presents the performance of existing approaches on the COPA evaluation using the accuracy metric. All methods shown in the table have used PMI as their semantic measure. However, each method has used a different window size and a different text corpus.

Table 8.2: Accuracy-based performance of existing approaches on the COPA-test benchmark dataset.

Approach	Window Size	Corpus	Accuracy
PMI Gutenberg	5	Gutenberg	58.8
UTDHLT Bigram PMI	100	GigaWord	61.8
UTDHLT SVM PMI	100	GigaWord	63.4
PMI Story 1M	25	Personal Stories	65.2
PMI Story 10M	25	Personal Stories	65.4

8.3 APRM-based Causality Detection System

This chapter presents a method to address the *Choice Of Plausible Alternatives* (COPA) task based on the assumption that causally related events occur closer in the written text. Consequently, the APRM measure will be able to capture some parts of this causal knowledge by considering the correlation statistics between words using their associative probabilities computed from the Wikipedia corpus.

8.3.1 From Sentence to Commonsense Concepts

The COPA task requires pre-processing of the premise and the two alternatives to convert informal text sentences to meaningful lexical chunks representing commonsense concepts before calculating their causality scores. Once the sentences are broken down to various chunks, their pairwise similarity scores are computed using the APRM measure. The pairwise scores are used to compute causality scores of each premise-alternative pair. An alternative that has higher causality score with the premise is selected as the correct answer. The preprocessing is performed online at the word/phrase level, after the content filtering of the premise and the two alternatives. Three different filtering models are used to generate the lexical chunks: *unigram model*, *bigram model* and *verbal-pattern model*.

Unigram Model

At unigram level, the content of premise and alternatives are preprocessed to get single word tokens. These unigrams are matched with a predefined list of stop words to get rid of common English words such as *it, of, as, on*. Each unigram is labeled with its corresponding part of speech (POS) tag and only unigrams with noun `_NN`, verb`_VB` and adjective `_JJ` tags are retained. Each unigram is matched with a huge repository of Wikipedia labels to pick the ones that exist in Wikipedia as labels. Since not all Wikipedia labels are equally useful, only those unigrams are selected that have a link probability value above a certain threshold $(\alpha > 0.001)$. Finally, the words are lemmatized before computing their APRM scores from the Wikipedia corpus.

Bigram Model

The bigram pattern extraction model converts a sentence to uni-gram and bi-gram tokens. After a number of filtering steps, all the meaningful unigrams and bigram chunks are added to the list of coherent concepts. The selected n-grams are lemmatized before computing causality scores.

In order to select the bigram chunks representing commonsense concept, various filtering steps are performed. After converting to n-grams, all stop word n-grams are filtered out. The remaining n-grams are marked up with their corresponding part of speech (POS) tags and only n-gram concepts with the following POS patterns are retained.

ADJECTIVE-NOUN: An adjective followed by a noun is a semantically useful pattern such as *bad day*. Hence, such bigram patterns are extracted and retained.

VERB-ADJECTIVE: A verb followed by an adjective is often a meaningful pattern for instance, *sounds good* or *tastes delicious*.

NOUN-NOUN: A noun can be followed by another noun as part of a single concept such as *credit card* or *ice cream*.

NOUN-VERB: A noun can be followed by a verb such as *Sea diving* or *day dreaming*.

VERB-NOUN: A verb followed by a noun is often a meaningful pattern such as *quit job* or *recycle papers*.

VERB/NOUN/ADJECTIVE: All unigrams with noun, verb and adjective POS tags are also considered useful, hence selected.

Multi-word expressions are identified and retained while getting rid of the pleonastic (providing redundant information) unigrams. For instance, the bigram `bubble wrap` was retained, while the unigrams `bubble` and `wrap` were filtered out. In order to avoid over-scoring the incorrect alternative due to the overlap of its content with that of the premise, concepts common to both premise and alternative are ignored while computing the pairwise scores between their commonsense concepts. Such concepts unnecessarily increase the overall scores without effectively contributing to the causal information, thus lead to the selection of the wrong alternative. After all these filtering steps, if no chunks are retrieved successfully then the filtering mode is switched to the unigram model with a lower link probability *LP* threshold (details of LP are included in Chapter 3).

Verbal-Pattern Model

The third model is a verbal-pattern extraction model. Based on the assumption that a verb can be a good indicator of causality, this model focuses on extracting the causal verbal patterns from the sentences and computing their associations. Like the previous two models, this model also converts a given sentence to unigram and bi-gram tokens. In order to pick the potential causal verbs, in the first filtering step, all bigrams that match an element of a predefined list of phrasal verbs [5] are selected. In the second filtering step, all stop words are filtered out of the original list of all unigrams and bigrams. After removing stop words, all n-grams are tagged with part of speech classes and only those that match with the following verbal patterns are retained.

Single VERBS: All Single verbs are good indicators of causality such as the verb *hitting* causes to *bruising*, hence all single verbs are retained.

VERB-ADJECTIVE: A pattern consisting of a verb followed by an adjective is selected such as *doing great* or *feeling bad*.

VERB-NOUN: Like the verb-object dependency, a verb followed by a noun can be useful such as *eating apple* or *playing football*.

NOUN-VERB: Like a subject-verb dependency, a noun can be followed by a verb such as *birds* followed by *flying*.

After filtering pleonastic unigrams, all multi-word expressions are retained. At the end of these filtering steps, if the final list of concepts is empty then the pre-processing mode shifts to the Bigram model.

[5] http://www.usingenglish.com/reference/phrasal-verbs/list.php

8.3.2 Causality Computation

Given a set of a premise p and two alternatives a_1 and a_2, the goal is to pick the most plausible alternative a' such that

$$a' = argmax_{a \in \{a_1, a_2\}} Causality(p, a) \tag{8.1}$$

The causality score between the premise p and alternative a is computed by averaging the pairwise APRM scores between commonsense concepts of the premise and each alternative as follows.

$$Causality(p, a) = \frac{\sum_{c_p \in p} \sum_{c_a \in a} APRM(c_p, c_a)}{N_p N_a} \tag{8.2}$$

where, N_p and N_a represent the number of commonsense concepts in p and a respectively. The APRM measure is calculated as a linear combination of directional association of c_p and c_a, given as follows:

$$APRM(c_p, c_a) = (1 - \lambda) \times Association(c_p \rightarrow c_a) + \lambda \times Association(c_a \rightarrow c_p)$$

where, c_p and c_q are commonsense concepts belonging to the premise and alternative respectively. The directional association of any two commonsense concepts c' and c'' of the premise and each alternative is given as

$$Association(c' \rightarrow c'') = \frac{|c' \; near \; c''|}{|c'|}$$

where, c' and c'' are searched in a proximity window of size $(2w + 1)$ with w representing the number of words on either side of a given word. For all bigram patterns in both these models, the slop between tokens of each bigram is set to 3 words while searching the patterns in the corpus.

8.4 Results and Discussion

Existing measures reported their results using the same PMI measure but different text corpora. The literature review shows that the PMI measure

Table 8.3: Performance comparison of Wikipedia-based semantic association measures using window Size with `w=10` on three benchmark datasets of COPA evaluation.

Model	Measure	Benchmarks		
		COPA Test	COPA Dev	COPA All
Unigram Model	APRM_avg	56.0	56.4	56.2
	APRM_max	55.4	55.6	55.5
	PMI	51.6	54.0	52.8
	Simpson	54.4	53.2	53.8
	NGD	52.4	53.4	52.9
	Dice	54.8	55.2	55.0
Bigram Model	APRM_avg	57.2	57.0	57.1
	APRM_max	56.8	56.2	56.5
	PMI	51.6	54.4	53.0
	Simpson	54.8	51.4	53.1
	NGD	56.0	55.6	55.8
	Dice	51.2	52.6	51.9
Verbal-pattern Model	APRM_avg	57.8	58.2	58.0
	APRM_max	58.8	59.6	59.2
	PMI	50.4	51.4	50.9
	Simpson	55.4	53.4	54.4
	NGD	55.8	54.2	55.0
	Dice	52.8	53.6	53.2

when using the Gutenberg corpus produced low accuracy (58.8) but the same PMI-based approach using the Personal Stories corpus led to improvement in the accuracy (65.4). Hence, we argue that the performance of a semantic association measure on causality detection is directly influenced by the underlying knowledge source. Changing the knowledge source or augmenting it with another complementary knowledge source could lead to improved results. Hence, the goal of this research is to compare different semantic association measures using the same text corpus under the same experimental settings. The experiment analyzed and compared the accuracy of the existing best performing semantic measure PMI with the asymmetric association based measure APRM using Wikipedia as the underlying corpus under the same experimental settings.

In order to investigate the behavior of Wikipedia-based semantic measures on the common sense causality detection task, three other proximity window-based measures were also adapted to the Wikipedia corpus: the Dice measure, the Simpson measure and the NGD measure. Table 8.3 presents the results on accuracy of the Wikipedia-based causality detection systems using three pre-processing models on three benchmark datasets of COPA evaluation: COPA-test, COPA-dev and COPA-all.

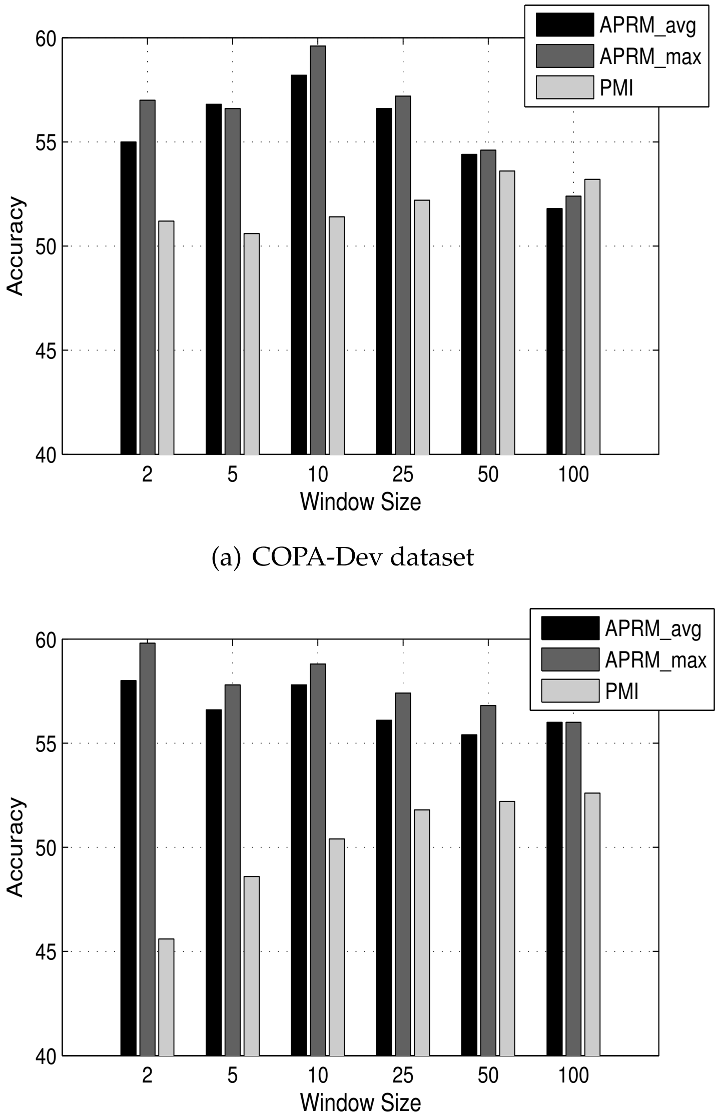
It is encouraging to note that on all benchmark datasets, the APRM measures performed much better than the PMI measure on the Wikipedia corpus. This experiment also encourages the use of the APRM measure on other augmented corpora to improve the performance of the causality detection system. It is also evident from the results that in case of verbal-pattern model APRM_max performed better than APRM_avg, which indicates the advantage of selecting the maximum score over averaged-score on causality computation.

In order to investigate the impact of the window size w on the performance of the semantic measures, the accuracy of both PMI and the APRM measures was compared on the COPA-test and COPA-dev benchmarks using six different window sizes ranging from 2 to 100. The results are shown in Figure 8.1.

On both benchmarks, the APRM measures surpassed the PMI measure by a good margin. The performance of both APRM measures was high in the lowest three windows sizes ($w = 2, w = 10, w = 25$) and degraded with the increase in the window size after $w = 10$. On the other hand, the PMI measure achieved high accuracy values on the highest three window sizes ($w = 25, w = 50, w = 100$). On the COPA-test benchmark, the accuracy of the PMI measure showed an increasing trend with increasing window size, achieving the highest accuracy on $w = 100$. On the same benchmark, the APRM measures achieved highest accuracy on the window size $w = 2$. On the COPA-test benchmark, however the best accuracy using the APRM measures was achieved on the window size $w = 10$. This resonates with the basic assumption of this research that the causally related events occur closer in the text. On the same dataset, PMI yielded best accuracy on $w = 50$. The PMI measure was able to outperform both the APRM measures only in one case—on the COPA-dev dataset—with a big window size ($w = 100$), which we believe is too wide to capture realistic causal semantics. In general, for each Wikipedia-based measure on both benchmarks, the difference of accuracy on different window sizes was not large. This experiment highlights the advantage of using asymmetric association based APRM measures over other standard measures. The experiment compares the performance of various semantic measures under the same experimental settings (with the same Wikipedia corpora and same size of proximity window). However, it will be useful to repeat this experiment on other kinds of text corpora.

8.4.1 Statistical Significance

In order to determine the statistical difference of APRM-based causality detection system with systems based on other measures, a statistical significance test based on approximate randomization [154] is used. This



(a) COPA-Dev dataset

(b) COPA-test dataset

Figure 8.1: Effect of changing window size on the performance of PMI and the APRM measures on COPA evaluation benchmarks using verbal-pattern model.

method uses a stratified shuffling approach to build a distribution of differences in the performance between any two methods. On the COPA-all dataset (which is a combination of COPA-test and COPA-dev), both causality detection systems based on the APRM_avg and APRM_max measures are statistically significantly different from the PMI-based causality detection system (with the p-value of 0.004 and 0.0004 respectively) as well as from the baseline system of COPA evaluation task (with the p-value of 0.001 and 0.0002 respectively). Also, on the COPA-all benchmark, the APRM_avg measure is statistically significantly different (with p-value of 0.04) from the APRM_max measure.

8.4.2 Further Analysis

The overall accuracy of the causality detection systems using Wikipedia-based APRM measures is much better than other Wikipedia-based measures and slightly better than the PMI measure using the Gutenberg corpus as the underlying knowledge source. However, its performance is not better than that of PMI on two other knowledge sources—the Gigaword corpus and the Personal Stories corpus. There are two important factors that critically affect the performance of any causality detection system on the COPA task: a sensible choice of the underlying knowledge source; and the nature of the COPA evaluation.

All existing approaches to COPA task focused their efforts on calculating the co-occurrence statistics between words and phrases of premise and alternatives using different knowledge sources. Each knowledge source has a different focus on the domain topics. While, the Personal Stories corpus focus more on everyday narrative discourse, Wikipedia talks primarily about the persons, places and things. Due to different nature of underlying knowledge sources, the same system using two different knowledge sources can lead to performance differences, as is evident from the performance of the PMI Gutenberg and PMI Story 10M systems.

The authors of the COPA evaluation used two sources to select question topics. While the focus of one of the sources (Library of Congress thesaurus for Graphic Materials) was on events, people, places and things, the other source (Personal Stories corpus) focused more on social and mental topics related to everyday life. The low performance of the APRM-based causality detection systems using Wikipedia is not surprising as the underlying knowledge source of APRM is Wikipedia (an independent knowledge source), which focuses on people, places and things more than social and mental behaviors of day-to-day life. We believe that combining Wikipedia with another complementary knowledge source on everyday life issues could lead to improvement of the system performance[6]. This may also explain why even the best performing existing system is still far behind the human performance–the Personal Stories corpus used by the current best approach focused primarily on the knowledge about everyday life. Performance of the existing approaches on the COPA evaluation is quite poor as compared to human judgments of causal connections. The best performing system's accuracy (65.4%) lies much below the humans accuracy (99%) on the COPA task. It is also worth mentioning that the Personal Stories corpus used by the best performing system is also used as one of the two sources for generating the COPA benchmarks, which may give an unfair bias to the best performing PMI Story system in comparison with other systems which did not used this corpus.

While devising the COPA evaluation datasets, the authors ensured that an incorrect alternative is also a plausible sentence in the same context. In COPA evaluation, most of the incorrect alternatives are found to be lexically closer to the premise than the correct alternative, so that the approaches based on shallow analysis of the causal semantics fail to find out the correct option. The analysis of the results also revealed that the co-occurrence-based approaches in general, fail to cope with the situation in which two clauses (connected through causal relation) consist of weakly

[6] Due to time constraints, this is set as a future goal.

related words. Consider for instance,

Premise: The man anticipated the team's victory.
Alternative 1: He met his friends to watch the game.
Alternative 2: He made a bet with his friends.

In this example, the frequency of co-occurrence of premise words with that of alternative 1 were much higher than that of alternative 2. The strongest associations were found between the words (*game, victory*), (*team, game*), (*team, friends*) and (*man, watch*). However, the only strong causal relation between premise and alternative 2 comes from the causal verbs *anticipate* and *make a bet*. Because the causality score of a premise-alternative pair is the sum of pairwise scores of all the commonsense concepts extracted from their content, the alternative 1 gets overall higher score than alternative 2. Consider another question.

Premise: The girl ran out of energy.
Alternative 1: She played checkers.
Alternative 2: He jumped rope.

For this question, the APRM-based causality detection system produced strongest associations for the word pairs *(girl, play)*, *(run out, play)* and *(energy, play)*. However, the APRM measure failed to find out the association of the word *run out* with the word *jump rope* which was central to the causal relation. Hence, selected alternative 1 as the answer to the question which was incorrect. Such cases are common in the benchmarks. Hence, any system that ignores causal indicators by focusing only on the general relatedness computation will necessarily perform poorly on many of the COPA questions.

Being a word relatedness approach, the performance of the APRM measures is also dependent on the preprocessing phase that converts the natural language text sentences into commonsense concepts, hence it is desirable to investigate the performance of different text parsers such as dependency parser and constituency parser in generating different classes of commonsense concepts during preprocessing phase. A related factor contributing to the low performance is the strong noun-to-noun associations. The reason for this strong association is the Wikipedia corpus, which being intrinsically an encyclopedia has a strong coverage of noun concepts. This leads to high noun-noun associative probabilities which have a pronounced effect on the causality scores. Consequently, the incorrect alternative having nouns is chosen over the correct answer leading to incorrect causal judgments.

The authoring procedure of COPA ensures that the competing method cannot easily solve this task without a fine-grained understanding of causality. Based on the experimental evaluation it is shown that while human raters performed extremely well on the causality reasoning, achieving near perfect inter-rater agreement, associative methods performed only moderately above the baseline. Certainly, the use of co-occurrence statistics is not sufficient to answer COPA questions. In order to reach the level of human inferential capability, future systems may need to look beyond simple corpus statistics to more advanced causality detection systems augmented with commonsense-based causal inferencing. Such a system should be able to focus on computing causal relatedness rather than general relatedness. Since COPA questions vary in their topics and subject matter with regards to the causal relation between the premise and alternatives, there is a need for a hybrid system capable of satisfying all kinds of causal scenarios. Such a system should facilitate more diverse as well as complementary features that are suitable for different classes of causal connections. However, the problem of combining multiple features best suited to different classes of causal relations is a non trivial task and requires deep explicit lexical and semantic reasoning over the implicit commonsense knowledge.

Even if a future system could perfectly extract the causal relations from a large text corpora, it is still difficult to imagine that it will be able to mirror human performance on commonsense causal reasoning. We believe that there are knowledge sources such as Cyc, OMCS and ConceptNet, which could be quite helpful in surmounting the obstacles faced by the existing approaches on the COPA evaluation. By using the resources that make commonsense knowledge explicit, the performance of the current system could be improved further on the COPA evaluation.

A range of challenges need to be overcome to reach the level of human commonsense on the COPA task. Current research efforts on commonsense causality detection may be of preliminary nature but are fundamental to give an insight into the task from the perspective of word associations. This could help future approaches to build more sophisticated approaches on these fundamentals. The application of commonsense reasoning for interpreting natural language semantics still remains an open challenge with a big margin of improvement. Despite many research efforts, the development of effective automated commonsense reasoning system backed up by powerful inferential capabilities approaching the level of humans remains an open research challenge.

8.5 Chapter Summary

This chapter presented a task-driven evaluation of the APRM measure for commonsense causality detection. The chapter demonstrated a way of exploiting semantic capabilities of Wikipedia in solving the choice of plausible alternatives task followed by a detailed empirical analysis and discussion of the performance indicators. Under the same experimental settings and the same causality detection system, the asymmetric association based semantic measure, APRM, outperformed the best existing measure, PMI. The experiments suggest that there is a significant advantage to recognizing the asymmetric nature of word associations and using that asymmetry in semantic association measures. The experiments only used Wikipedia as the text corpus; it will be important to repeat the same experiment using other text corpora as well to ensure that this result is not just a property of Wikipedia, and to compare APRM with the current best commonsense causality detection system that used a different text corpus (Personal Stories Corpus). In an application specific evaluation, it is difficult to demarcate the impact of different factors on the overall system performance. However, the choice of underlying corpus, the preprocessing phase and the need of commonsense-based causal inferencing were found to be among the most critical factors. Empirical results also highlighted the importance of carefully balancing and combing these three factors. Future approaches will require more powerful semantic expressions and tools for augmenting automated methods with rich causal semantics based on effective commonsense inferencing techniques.

Chapter 9

Conclusions

This thesis presented a detailed study investigating rich semantics mined from a collaboratively constructed knowledge source, Wikipedia, for solving the problem of semantic association computation. The overall goal was to make the process of semantic association computation more effective by exploiting the semantic capabilities of two different aspects of Wikipedia. In order to achieve this goal, multiple semantically rich elements of Wikipedia were used in various experiments: several semantic association measures were developed based on the hyperlink structure, informative-content and combinations of both elements. These measures can cope with various types of semantic relations. The effectiveness and reuseability of new semantic measures were evaluated by using well known benchmark datasets in both direct evaluation and task-oriented indirect evaluations. This chapter concludes the thesis by listing the main contributions and key findings of each chapter. Finally, it presents the related issues which are potential open research topics for future work.

9.1 Main Contributions

This section presents a summary of the main conclusions, corresponding to the three research objectives, drawn from the six major chapters (Chapter 3 to Chapter 8).

i) The thesis conducted an in-depth study on using a collaboratively constructed knowledge source, Wikipedia, for computing semantic associations of words. Two aspects of Wikipedia were investigated and compared in detail on the task of semantic association computation. The first aspect viewed Wikipedia as a well-structured and semantically-rich knowledge source. Using this aspect, various semantic elements of Wikipedia were explored such as hyperlink structure, informative-content, labels, redirects and disambiguation pages. The second aspect viewed Wikipedia as an unstructured text corpus with a huge coverage of knowledge. Based on this aspect, probabilistic word associations were used in semantic association computation. The thesis identified the key strengths and downsides of each aspect and demonstrated their potential applications. This contribution answers the first research question (posed in Section 1.3) of the thesis.

ii) Using Wikipedia as the source of background knowledge, new measures were developed based on various models of semantic associations. Three measures based on the knowledge source aspect of Wikipedia were presented in Chapter 3—WikiSim, OSR and CPRel. Both WikiSim and OSR are combinatorial measures using the feature-set based model of semantic associations. CPRel is a *Vector Space Model* based measure relying on the geometric model of associations. Chapter 4 presented a new measure of word association, DCRM, which is based on a combination of an asymmetric association based model and a feature set based model of associations. Chapter 5 presented a machine learning based hybrid model using the previously developed measures as features. Chapter 6 presented a semantic association measure APRM based on combining the asymmetric association based model of semantic associations with the feature set based model of synonymy. The experiments using asymmetric associations to develop new semantic measures systematically explored various characteristics of asymmetric word associations: first, the thesis exploited and compared the merits of symmetry and asymmetry in the performance of semantic measures; second, it demonstrated the benefits of using asymmetric association based semantic measures over symmetric semantic measures; third, it emphasized the need to evaluate semantic measures in a context-dependent way. Various factors contributing to the performance of each design model were detailed and compared with other models in different scenarios. The thesis demonstrated the trade-off between design complexity and performance of each model by testing it on various benchmarks datasets of semantic association computation. This contribution corresponds to the second research question (presented in Section 1.3) of the thesis.

iii) The APRM measure is applied in solving the word choice task, where given a target word and multiple candidate words, the goal is to find the most closely related candidate word. This task was studied at two levels: solving relatedness-based word choice problems and solving synonymy-based word choice problems. The asymmetric association based measure, APRM was found to surpass most other approaches in solving the relatedness-based word choice problems. Moreover, on synonymy-based word choice problems the APRM measure performed well on H-measure though not the best on accuracy due to its underlying proximity assumption. Various critical factors for performance enhancement on both levels of this task were also identified and discussed in detail. This contribution refers to the third research question (mentioned in Section 1.3) of the thesis.

iv) For the first time (to the best of our knowledge), a collaboratively constructed encyclopedia, Wikipedia, is used as an unstructured text corpus for estimating causal connections in the common sense causality detection task. This research employed proximity-based combinatorial semantic measures in solving the choice of plausible alternatives task using the COPA evaluation benchmarks [66]. The choice of plausible alternative is a subtask of common sense causality detection, where given a target sentence and two candidate clauses, the goal is to select the most plausible clause that is connected to the target sentence using either a cause or an effect relation. This semantic task was chosen to demonstrate the superiority of asymmetric association based measures over symmetric association based measures (including the best performing PMI measure) on the task of common sense causality detection. This study identified three important indicators that could positively contribute to the performance of a causality detection system: the choice of text corpus; the preprocessing phase; and the need of commonsense based causal inferencing. This research contribution answers the third research question (presented in Section 1.3) of the thesis.

9.2 Future Work

Exploiting a collaboratively constructed encyclopedia, Wikipedia, for semantic association computation in particular, and for supporting applications based on semantic associations in general, is still an open research area and there are many things yet to be done to come up with more powerful systems that would be capable of imitating humans' estimates of semantic associations. This section highlights some future research directions that can either be found on or motivate by the research work presented in this thesis.

i) The thesis developed new semantic association measures based on two different aspects of Wikipedia. However, it would be interesting to develop a semantic measure based on combining multiple features extracted from both aspects of Wikipedia. Such a measure may show performance improvement by combining the best of both views of Wikipedia—as an unstructured text corpora and as a well-structured knowledge source.

ii) The thesis has demonstrated that different knowledge sources are useful for estimating different kinds of semantic and lexical relations. It has empirically shown that the features with complementary nature resulted in better estimates of semantic associations than the individual features. We believe that the same assertion is applicable at the knowledge source level. Hence, it would be valuable to develop and examine the performance of a new semantic measure, based on combining a semantically rich multilingual online encyclopedia, Wikipedia, with a collaboratively constructed multilingual online dictionary, Wiktionary, on the task of semantic association computation. The complementary nature of the two collaboratively constructed resources could lead to better estimates of semantic associations of words. Literature review has demonstrated the effectiveness of each of these crowd sourced knowledge sources on the task of semantic association computation individually but using them together in developing a new measure of semantic associations is yet to be seen.

iii) Another aspect that needs more attention in semantic association computation is the nature of benchmark datasets available for evaluation of semantic association approaches. There are few datasets that are either too small or are generic in nature—there are too few datasets focusing on particular types of associations and POS-classes. Certain aspects of semantic associations should be explicitly investigated to understand the properties of any semantic association measure because every semantic measure cannot perform well on all kinds of semantic associations and relations. Hence, it will be useful to investigate the performance of semantic measures on various POS classes by constructing larger datasets having word pairs belonging to the same part-of-speech (such as noun, verb, adjectives, adverbs) as well as to different part-of-speech (such as noun-verb and adjective-noun term pairs) independently. Although a few research efforts on this perspective were discussed in Section 2.2 of the thesis, still there is a need for larger and more focused datasets targeting this aspect. Similarly, new datasets should be developed according to the type of semantic associations such as similarity, relatedness, collocations, lexical cohesion, co-occurrences and directional associations. One effort in this regard was made by Agirre *et al.* [1], who constructed similarity and relatedness based subsets of an existing dataset (WS-353), but this dataset was found to be culturally biased by other researchers. Nonetheless, it was useful for testing the behavior of semantic measures on similarity and relatedness based word pairs. There is a lot more to be done to generate new and focused datasets for extensively investigating various aspects of semantic measures in order to identify their semantic bias and limitations.

iv) Classical benchmark datasets on semantic association computation consist of word pairs along with the human judgment based scores. These datasets do not include any explicit context (except the ESL dataset). The research has already shown that the dataset with context resulted in higher inter-rater agreement (increased from 0.76 to 0.79), which means providing explicit context will also help the automatic approaches in computing more realistic word associations. The research in the thesis demonstrated that using different topical contexts for a word pair could lead to different association scores. Hence, it will be interesting to investigate the performance of semantic measures when each word pair is provided with an explicit topical context. For this purpose, no such dataset is available yet. So the future work includes constructing a new dataset for testing word associations in an explicitly provided topical context.

v) A large number of natural language semantic interpretation and processing approaches are essentially semantic association computation based tasks (see Chapter 2 of the thesis for details). Thus, many applications can benefit from the research in this thesis by using Wikipedia-based semantic measures. Each measure is suitable for different kinds of semantic association types and relations. Hence, the suitability of a semantic measure for a certain application must be carefully taken into account before employing it in an application. This research has barely scratched the surface by developing and using new measures of semantic associations (based on exploitation of semantic capabilities of Wikipedia) in standard natural language interpretation and processing tasks. The future work also involves exploring the impact of Wikipedia-based semantic measures on complex linguistic tasks requiring extraction of in-depth semantics, such as lexical chaining, propositional phrase attachment ambiguity resolution, conjunction scope identification, anaphora resolution, discourse structure analysis and knowledge representation. This will necessitate augmenting the word association semantics with syntactical information as well as linguistic dependencies in a more sophisticated manner.

It is hoped that future research will build on the experiments and discussions presented in this thesis to explore new avenues using Wikipedia for finding deeper and semantically more meaningful associations in a wide range of application areas.

List of Publications

Parts of research contributions have been published and presented in the following conferences and journals:

Jabeen, S., Gao, X. and Andreae, P. "Using Asymmetric Associations for Commonsense Causality Detection", In the *13th Pacific Rim International Conference on Artificial Intelligence-PRICAI'14*, Vol. 8862, pp. 877-883, 2014.

Jabeen, S., Gao, X. and Andreae, P. "A Hybrid Model for Learning Semantic Relatedness Using Wikipedia-Based Features". In *Web Information System Engineering-WISE'14*, Vol. 8786, pp. 523-533, 2014.

Jabeen, S., Gao, X. and Andreae, P. "Probabilistic Associations as a Proxy for Semantic Relatedness". In *Web Information System Engineering-WISE'14*, Vol. 8786, pp. 512-522, 2014.

Jabeen, S., Gao, X. and Andreae, P. "Directional Context Helps: Guiding Semantic Relatedness Computation by Asymmetric Word Associations". In *Web Information System Engineering-WISE'13*, Vol. 8080, pp. 92-101 2013.

Jabeen, S., Gao, X. and Andreae, P. "CPRel: Semantic Relatedness Computation Using Wikipedia based Context profiles", *Journal of Computing Science*, Vol. 70, pp. 55-66, 2013.

Jabeen, S., Gao, X. and Andreae, P. "Using Wikipedia as an External

Knowledge Source for Supporting Contextual Disambiguation", *Journal of Software Engineering and Applications*, Vol. 5, pp. 175-180, 2012.

Jabeen, S., Gao, X. and Andreae, P. "Harnessing Wikipedia for Computing Contextual Relatedness". In *PRICAI'12: 12th Pacific Rim International Conference on Artificial Intelligence*, Vol. 7458, pp. 861-865, 2012.

Jabeen, S., Gao, X. and Zhang, M.: "Leveraging Wikipedia Semantics for Computing Contextual Relatedness", In *NZCSRSC'12: New Zealand Computer Science Research Student Conference*, 2012.

Glossary

Semantics - The study of meanings at the level of words, phrases, sentences and larger units of text.

Syntactics - The study of the rules governing the sentence structure of a language.

Inter-rater Agreement - The degree of agreement among raters of an experiment.

Intra-rater Agreement - The degree of agreement among two groups of raters on the same set of data in two different experiment.

Zipf's law - Zipf's law states that for a given corpus, the frequency of any word is inversely proportional to its rank in the frequency table. Hence, the most frequent word occurs approximately twice as often as the second most frequent word, thrice the frequency of third most frequent word and so on.

Surface forms - A hypertext used to refer to a specific Wikipedia article, for instance the surface forms of the article `United states of America` are `USA`, `US`, `United States` and `America`.

Vector Space Model - An algebraic model, commonly used in information retrieval, for representing a text document as a vector in a multidimensional feature space. Each dimension is a keyword whose value is non-zero if it exists in the text document. Two text documents can be compared by generating their corresponding vectors and computing the similarity between their vectors using a similarity measure such as Cosine Similarity.

Cosine Similarity - A measure of similarity between two vectors using the cosine of angle between them. A value of cosine similarity equal to 1 means the two vectors are the same. Any value between 0 and 1 indicates the degree of similarity between any two vectors.

tf-idf - *Term Frequency Inverse Document Frequency* (*tf-idf*) is a weighting scheme that is used in vector Space Models to indicate the importance of a term to a document in a corpus or a collection. It is a common weighting factor used in information retrieval.

Stemming - The process of reducing a word to its base, stem or root form (which may or may not be a dictionary word). For instance, the words `stemming`, `stemmed` and `stemmer` are reduced to the root form `stem`.

Lemmatization - The process of determining the lemma or dictionary form of a word. For instance, the words `run`, `running` and `ran` are reduced to the root form `run`. Difference between lemmatization and stemming is that the stemming is context-independent and does not differentiate between different senses of a word based on different part of speech classes.

Taxonomy - A hierarchical structure to organize the data in the form of nodes connected by edges to represent some sort of relationships between them. The nodes are connected based on the generalization or specialization relation.

Parsing - Syntactic analysis of units of a language according to the rules of a formal grammar.

Lexicon - The word stock of a language. Grammar, is a set of rules that generates combination of these words into meaningful combinations.

n-gram - A contiguous sequence of *n* words extracted from a given piece of text. For instance, the bigrams of a sentence "*He ate apple*" will be *He ate* and *ate apple*.

Collocation - A sequence of adjacent words that occur together more often than by chance. For instance, `credit card`.

Pleonasm - In a broad sense, pleonasm is the use of words or parts of words than are necessary for expressing the meaning. One form of semantic pleonasm is when a word subsumes another word such as `Bear` is subsumed by `Black Bear`.

Synonym - Different words used to refer to the same concept. For instance, the synonym of the word `Rooster` is `cock` and that of the word `King` is `Monarch`.

Antonymy - Words that are opposite in meaning. For instance `day` and `night` are antonyms of each other.

Hyponymy and hypernymy - A relation in which a word (hyponym) has a *type-of* relation with another word (hypernym) such as `car` is a hyponym of `vehicle`. This is a *transitive superordinate-subordinate* relation. For instance, if `car` is a type of `vehicle` and `Ferrari` is a type of `car` then `Ferrari` is a type of `vehicle`.

Meronymy and holonymy - A relation in which a word (meronym) has the *part-of* or *member-of* relation with another word (holonym), for instance `wheel` is a meronym of `car` . Parts are inherited from their super-ordinates (if a `car` has wheels then `Ferrari` also has wheels) but are not inherited upwards (if `Ferrari` has a speed over 330 mph then not all `cars` have this same speed) as there might be some attributes of some kinds of specification rather than the class as a whole. These relations are *transitive* as well as *asymmetric* in nature.

Troponymy - Troponymy is a *manner-based relation* of verbs. It is similar to hyponymy. The difference lies in the fact that hyponymy is used for nouns whereas troponymy is used for verbs. For instance, `march` is a troponym of `walk` and `whisper` is a troponym of `speak`.

Entailment - These are unidirectional relations in which truth of one textual unit (`X`) requires the truth of another unit (`Y`). If `X` is false then `Y` must necessarily be false. For instance, you need to *have money* to *buy a computer*.

Cross-POS Relation - Cross-POS relations associate word belonging to different part of speech classes. For instance, the relation of `sorting` (verb) to the word `order` (noun).

Paradigmatic Relation - A relation between two words is paradigmatic if the two words can plausibly substitute one another in a sentence without affecting its structure and meanings. For instance, synonyms such as `slow` and `lazy` and antonyms such as `day` and `night`. Paradigmatic words generally have high semantic similarity (due to big overlap of their sets of neighboring words) and belong to same POS-class.

Syntagmatic Relation - A relation between two words which co-occur more frequently than by chance and belong to different part-of-speech classes. For instance, `bird` and `fly` or `drink` and `tea`.

Kolmogrov complexity - In information theory, Kolmogrov complexity is the amount of information contained by a finite object such as a piece of text.

Information Distance - Information distance, an extension of Kolmogrov complexity, is the minimum amount of information required to go from one object to another and vice versa. It was introduced in thermodynamics but later on adapted to normalized Google distance and normalized compression distance.

Process Flow Diagrams

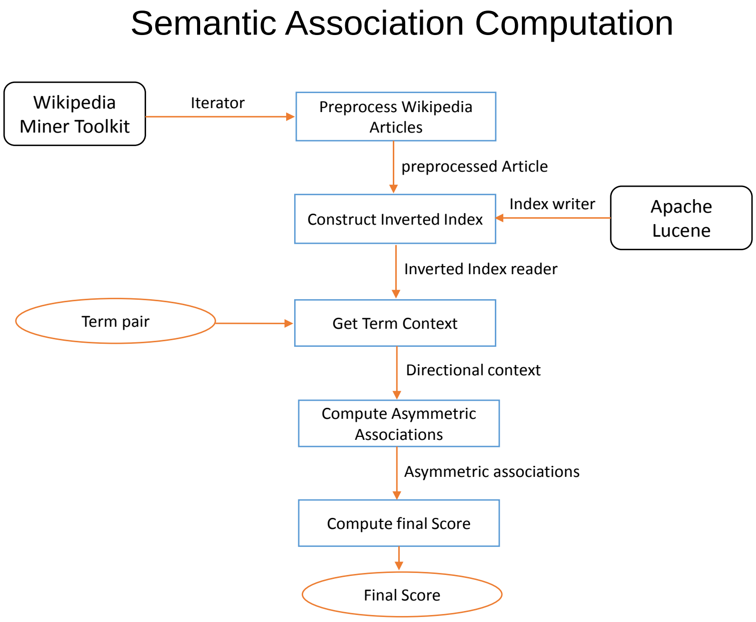


Figure 1: Process flow diagram of the APRM-based semantic association computation system.

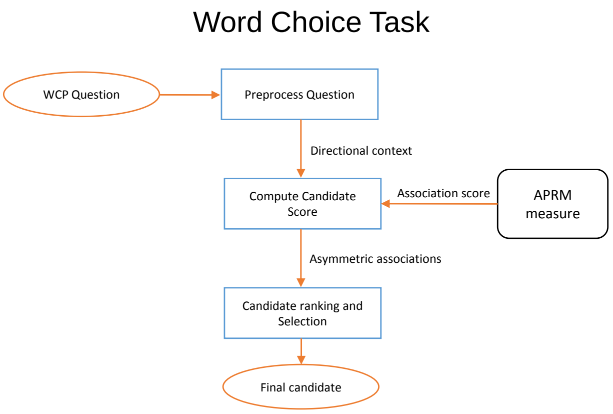


Figure 2: Process flow diagram of the APRM-based word choice system.

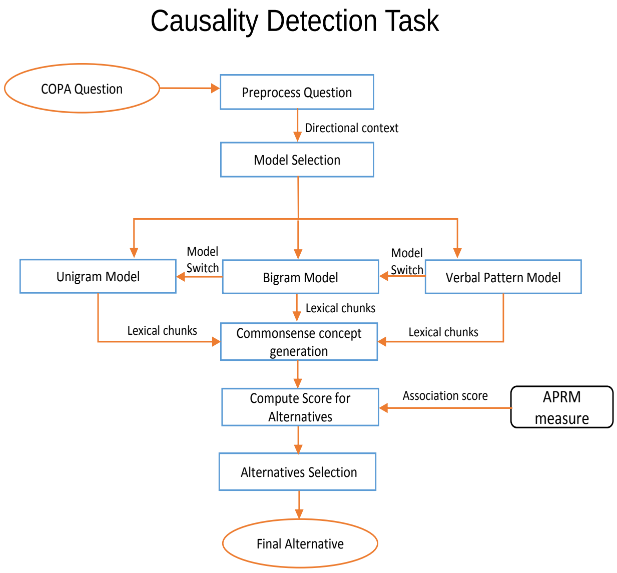


Figure 3: Process flow diagram of the APRM-based causality detection system.

Bibliography

[1] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)* (2009), pp. 19–27.

[2] Agirre, E., Unibertsitatea, E. H., and Rigau, G. Word sense disambiguation using conceptual distance. In *Proceedings of the 16th Conference on Computational Linguistics (COLING '96)* (1996), vol. 1, pp. 16–22.

[3] Agrawal, R., Imieliński, T., and Swami, A. Mining association rules between sets of items in large databases. vol. 22, pp. 207–216.

[4] Al-Mubaid, H., and Nguyen, H. A. Measuring semantic similarity between biomedical concepts within multiple ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C 39*, 4 (2009), pp. 389–398.

[5] Alonso, O., Strötgen, J., Baeza-Yates, R., and Gertz, M. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW '11)* (2011), pp. 1–8.

[6] ANDROUTSOPOULOS, I., AND MALAKASIOTIS, P. A survey of paraphrasing and textual entailment methods. *Journel of Artificial Intelligence Research 38*, 1 (2010), 135–187.

[7] BAEZA-YATES, R. A., AND RIBEIRO-NETO, B. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., 1999.

[8] BANERJEE, S., AND PEDERSEN, T. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence, (IJCAI '03)* (2003), pp. 805–810.

[9] BANERJEE, S., RAMANATHAN, K., AND GUPTA, A. Clustering short texts using wikipedia. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)* (2007), pp. 787–788.

[10] BARNBROOK, G., DANIELSSON, P., AND MAHLBERG, M. *Meaningful texts: the extraction of semantic information from monolingual and multilingual corpora*. Bloomsbury Publishing, 2005.

[11] BARZILAY, R., AND ELHADAD, M. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization* (1997), pp. 10–17.

[12] BATET, M., SÁNCHEZ, D., AND VALLS, A. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics 44*, 1 (2011), pp. 118–125.

[13] BERNDSEN, R., AND DANIELS, H. Causal reasoning and explanation in dynamic economic systems. *Journal of Economic Dynamics and Control 18*, 1 (1994), 251–271.

[14] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[15] BLEULER, E. *Dementia Praecox or the Group of Schizophrenias.* International Universities Press, 1911.

[16] BOLLEGALA, D., MATSUO, Y., AND ISHIZUKA, M. A web search engine-based approach to measure semantic similarity between words. *IEEE Transections on Knowledge and Data Engineering 23*, 7 (2011), 977–990.

[17] BOLSHAKOV, I., AND GELBUKH, A. Synonymous paraphrasing using wordnet and internet. In *Proceedings of 9th International Conference on Applications of Natural Language to Information Systems (NLDB '04* (2004), vol. 3136, pp. 312–323.

[18] BRANTS, T., CHEN, F., AND TSOCHANTARIDIS, I. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02)* (2002), pp. 211–218.

[19] BRIDGE, D. G. Defining and combining symmetric and asymmetric similarity measures. In *Advances in Case-Based Reasoning: Proceedings of 4th European Workshop on Case-Based Reasoning* (1998), vol. 1488, pp. 52–63.

[20] BRITANNICA, E. Fatally flawed: Refuting the recent study on encyclopedic accuracy by the journal nature, 2006.

[21] BRUSSEE, R., AND WARTENA, C. Automatic thesaurus generation using co-occurrence. In *Proceedings of the 20th Belgian Netherlands Conference on Artificial Intelligence (BNAIC 2008)* (2008), pp. 41–48.

[22] BUDANITSKY, A., AND HIRST, G. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the workshop on wordNet and other lexical resources, second meeting of the North American Chapter of the Association for Computational Linguistics* (2001).

[23] BUDANITSKY, A., AND HIRST, G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics 32* (2006), 13–47.

[24] BULLINARIA, J., AND LEVY, J. Extracting semantic representations from word co-occurrence statistics: A computational study. *Journal of Behavior Research Methods 39*, 3 (2007), 510–526.

[25] CAMBRIA, E., AND HUSSAIN, A. *Sentic Computing: Techniques, tools and applications*. Springer, Dordrecht, Netherlands, 2012.

[26] CARPINETO, C., OSIŃSKI, S., ROMANO, G., AND WEISS, D. A survey of web clustering engines. *ACM Computing Survey 41*, 3 (2009), 17:1–17:38.

[27] CARPINETO, C., AND ROMANO, G. A survey of automatic query expansion in information retrieval. *ACM Computing Survey 44*, 1 (2012), 1–50.

[28] CASSIDY, P. J. An investigation of the semantic relations in the roget's thesaurus: Preliminary results. In *Proceedings of the First International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '00)* (2000), pp. 181–204.

[29] CAVIEDES, J. E., AND CIMINO, J. J. Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics 37*, 2 (2004), 77–85.

[30] CHEN, H.-H., LIN, M.-S., AND WEI, Y.-C. Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL '06)* (2006), pp. 1009–1016.

[31] CHEN, Z., LIU, S., WENYIN, L., PU, G., AND MA, W.-Y. Building a web thesaurus from web link structure. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)* (2003), pp. 48–55.

[32] CHENG, P. W., HOLYOAK, K. J., NISBETT, R. E., AND OLIVER, L. M. Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive psychology 18*, 3 (1986), 293–328.

[33] CHESNEY, T. An empirical examination of wikipedia's credibility. *First Monday 11*, 11 (2006).

[34] CHURCH, K. W., AND HANKS, P. Word association norms, mutual information, and lexicography. *Computational Linguistics 16*, 1 (1990), 22–29.

[35] CILIBRASI, R. L., AND VITANYI, P. M. B. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng. 19*, 3 (2007), 370–383.

[36] CLIFTON, C., COOLEY, R., AND RENNIE, J. Topcat: Data mining for topic identification in a text corpus. *IEEE Transactions on Knowledge and Data Engineering 16*, 8 (2004), 949–964.

[37] COHEN, J. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 1988.

[38] COLLINS-THOMPSON, K., AND CALLAN, J. Query expansion using random walk models. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)* (2005), pp. 704–711.

[39] COURSEY, K., AND MIHALCEA, R. Topic identification using wikipedia graph centrality. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter*

of the Association for Computational Linguistics, (NAACL-Short '09) (2009), pp. 117–120.

[40] DAGAN, I., LEE, L., AND PEREIRA, F. C. N. Similarity-based models of word cooccurrence probabilities. *Journel of Machine Learning Research 34*, 1-3 (1999), 43–69.

[41] DAVIDSON, D. Actions, reasons, and causes. *Journal of Philosophy 60*, 23 (1963), 685–700.

[42] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science (JASIS) 41*, 6 (1990), 391–407.

[43] DICE, L. R. Measures of the amount of ecologic association between species. *Ecology 26* (1945), 297–302.

[44] EAGLE, N., SINGH, P., AND PENTLAND, A. Common sense conversations: Understanding casual conversation using a common sense database. In *Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003)* (2003).

[45] EDITORIAL. Britannica attacks... and we respond, 2006.

[46] EGLOFF, B., AND SCHMUKLE, S. C. Predictive validity of an implicit association test for assessing anxiety. *Journal of personality and social psychology 83*, 6 (2002), 1441.

[47] EMIGH, W., AND HERRING, S. C. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)* (2005), vol. 4, IEEE Computer Society, pp. 99a–99a.

[48] Erekhinskaya, T., and Moldovan, D. Lexical chains on wordnet and extensions. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference (FLAIRS '13)* (2013), pp. 52–57.

[49] Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. Open information extraction from the web. *Communications of the ACM 51*, 12 (2008), 68–74.

[50] Eysenck, H. J. Creativity and personality: Word association, origence, and psychoticism. *Creativity Research Journal 7*, 2 (1994), 209–216.

[51] Fernendo, S., and Stevenson, M. A semantic similarity approach to paraphrase detection. In *Proceedings of the Computational Linguistics UK (CLUK '04) 11th Annual Research Colloquium* (2008).

[52] Ferret, O. Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02)* (2002), vol. 1, pp. 1–7.

[53] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. Placing search in context: the concept revisited. *ACM Transactions on Informaiton Systems 20*, 1 (2002), 116–131.

[54] Firth, J. R. A synopsis of linguistic theory 1930-55. 1–32.

[55] Fletcher, C. R., and Bloom, C. P. Causal reasoning in the comprehension of simple narrative texts. *Journal of Memory and Language 27*, 3 (1988), 235–244.

[56] Fredembach, B., de Boisferon, A. H., and Gentaz, E. Learning of arbitrary association between visual and auditory novel stimuli in adults: The "bond effect" of haptic exploration. *Public Library of Science (PLoS) ONE 4*, 3 (2009), 44–48.

[57] GABRILOVICH, E., AND MARKOVITCH, S. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '06)* (2006), vol. 2, pp. 1301–1306.

[58] GABRILOVICH, E., AND MARKOVITCH, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, (IJCAI '07)* (2007), pp. 1606–1611.

[59] GALTON, F. Psychometric experiments. *Brain 2*, 2 (1879), 149–162.

[60] GANESAN, K., ZHAI, C., AND HAN, J. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)* (2010), pp. 340–348.

[61] GIRJU, R. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering (MultiSumQA '03)* (2003), vol. 12, pp. 76–83.

[62] GLEDSON, A., AND KEANE, J. Using web-search results to measure word-group similarity. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)* (2008), vol. 1, pp. 281–288.

[63] GONG, Y., AND LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)* (2001), pp. 19–25.

[64] GOODWIN, T., RINK, B., ROBERTS, K., AND HARABAGIU, S. M. Utdhlt: Copacetic system for choosing plausible alternatives. In *Proceedings of the First Joint Conference on Lexical and Computational Se-*

mantics (SemEval '12) - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (2012), pp. 461–466.

[65] GORDON, A. S., BEJAN, C. A., AND SAGAE, K. Commonsense causal reasoning using millions of personal stories. In *Proceedings of twenty Fifth Conference on Artificial Intelligence (AAAI '11)* (2011), pp. 7–11.

[66] GORDON, A. S., KOZAREVA, Z., AND ROEMMELE, M. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SemEval '12) - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (2012), pp. 394–398.

[67] GORDON, A. S., AND SWANSON, R. Identifying personal stories in millions of weblog entries. In *Proceedings of third International Conference on Weblogs and Social Media, Data Challenge Workshop* (2009).

[68] GRACIA, J., AND MENA, E. Web-based measure of semantic relatedness. In *Proceedings of 9th International Conference on Web Information Systems Engineering (WISE '08)* (2008), vol. 5175, pp. 136–150.

[69] GUREVYCH, I. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP '05)* (2005), pp. 767–778.

[70] HALAVAIS, A., AND LACKAFFB, D. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication 13*, 2 (2008), 429–440.

[71] HALAWI, G., DROR, G., GABRILOVICH, E., AND KOREN, Y. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)* (2012), pp. 1406–1414.

[72] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: An update. *The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD '09) 11*, 1 (2009), 10–18.

[73] HARISPE, S., RANWEZ, S., JANAQI, S., AND MONTMAIN, J. Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *CoRR* (2013).

[74] HARRIS, Z. S. Mathematical structures of language.

[75] HART, H. L. A., AND HONORÉ, T. *Causation in the Law.* Oxford University Press, 1985.

[76] HASSAN, S., AND MIHALCEA, R. Semantic relatedness using salient semantic analysis. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI '11)* (2011), pp. 884–889.

[77] HAUKE, J., AND TOMASZ, K. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae 30*, 2 (2011), 87–93.

[78] HE, XIAOFENG, C. H. Q. D. H. Z., AND SIMON, H. D. Automatic topic identification using webpage clustering. In *Proceedings of IEEE International Conference on Data Mining (ICDM '01)* (2001), pp. 195–202.

[79] HENNIG, L. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the Interna-*

tional Conference on Recent Advances in Natural language porocessing (RANLP '09) (2009), pp. 144–149.

[80] HESTILOW, T. J., AND HUANG, Y. Clustering of gene expression data based on shape similarity. *EURASIP Journal of Bioinformatics Systems and Biology 2009* (2009), 3:1–3:12.

[81] HIGGINS, D. Which statistics reflect semantics? rethinking synonymy and word similarity. In *Proceedings of International Conference on Linguistic Evidence* (2004), pp. 265–284.

[82] HIPP, J., GÜNTZER, U., AND NAKHAEIZADEH, G. Algorithms for association rule mining-a general survey and comparison. *The ACM SIGKDD Exploration Newsletter 2*, 1 (2000), 58–64.

[83] HIRST, G., AND BUDANITSKY, A. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering 11* (2005), 87–111.

[84] HIRST, G., AND ST ONGE, D. *Lexical Chains as representation of context for the detection and correction malapropisms*. The MIT Press, 1998.

[85] HLIAOUTAKIS, A., VARELAS, G., VOUTSAKIS, E., PETRAKIS, E. G., AND MILIOS, E. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems (IJSWIS '06) 2*, 3 (2006), 55–73.

[86] HOBBS, J. R. Toward a useful concept of causality for lexical semantics. *Journal of Semantics 22*, 2 (2005), 181–209.

[87] HOFMANN, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)* (1999), pp. 50–57.

[88] HU, X., ZHANG, X., LU, C., PARK, E. K., AND ZHOU, X. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)* (2009), pp. 389–396.

[89] HUANG, A., MILNE, D. N., FRANK, E., AND WITTEN, I. H. Clustering documents using a wikipedia-based concept representation. In *Proceedings of the 13th Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD '09)*.

[90] IOSIF, E., AND POTAMIANOS, A. Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering FirstView* (6 2014), 1–31.

[91] ISLAM, A., AND INKPEN, D. Second order co-occurrence pmi for determining the semantic similarity of words. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC '06)* (2006), pp. 1033–1038.

[92] ISLAM, A., AND INKPEN, D. Real-word spelling correction using google web it 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)* (2009), vol. 3, pp. 1241–1249.

[93] J., G. Internet encyclopedias go head to head, 2005.

[94] JACCARD, P. The distribution of the flora in the alpine zone. 1. *New phytologist 11*, 2 (1912), 37–50.

[95] JARMASZ, M., AND SZPAKOWICZ, S. Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural language porocessing (RANLP '03)* (2003), pp. 212–219.

[96] JARMASZ, M., AND SZPAKOWICZ, S. Roget's thesaurus: a lexical resource to treasure. *CoRR* (2012).

[97] JIANG, J., AND CONRATH, D. W. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING)* (1997), pp. 19–33.

[98] JUNG, Y., RYU, J., KIM, K.-M., AND MYAENG, S.-H. Automatic construction of a large-scale situation ontology by mining how-to instructions from the web. *Web Semant. 8*, 2-3 (2010), 110–124.

[99] KAMPS, J., AND KOOLEN, M. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)* (2009), pp. 232–241.

[100] KELLER, F., LAPATA, M., AND OURIOUPINA, O. Using the web to overcome data sparseness. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP '02)* (2002), vol. 10, pp. 230–237.

[101] KILGARRIFF, A., AND GREFENSTETTE, G. Introduction to the special issue on the web as corpus. *Journal of Computational Linguistics 29*, 3 (2003), 333–347.

[102] KLEBANOV, B. B., AND FLOR, M. Word association profiles and their use for automated scoring of essays. In *Proceedings of The Association for Computer Linguistics (ACL '13)* (2013), The Association for Computer Linguistics, pp. 1148–1158.

[103] KOEHN, P. *Statistical Machine Translation*, 1st ed. Cambridge University Press, New York, NY, USA, 2010.

[104] LANDAUER, T. K., AND DUMAIS, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* (1997), 211–240.

[105] LANDAUER, T. K., LAHAM, D., REHDER, B., AND SCHREINER, M. E. How well can passage meaning be derived without using word order:a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society, (CogSci '91)* (1991), pp. 412–417.

[106] LEACOCK, C., AND CHODOROW, M. *Combining Local Context and WordNet Similarity for Word Sense Identification*. The MIT Press, 1998, ch. 11, pp. 265–283.

[107] LEE, H., PEIRSMAN, Y., CHANG, A., CHAMBERS, N., SURDEANU, M., AND JURAFSKY, D. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (CONLL '11)* (2011), pp. 28–34.

[108] LEE, L. Measures of distributional similarity. In *Proceedings of The Association for Computer Linguistics (ACL '99)* (1999), pp. 25–32.

[109] LEECH, G. 100 million words of english. *English Today 9* (1993), 9–15.

[110] LENAT, D. B. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM 38*, 11 (1995), 33–38.

[111] LESK, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86)* (1986), pp. 24–26.

[112] LEVIN, E., SHARIFI, M., AND BALL, J. Evaluation of utility of lsa for word sense discrimination. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume (NAACL '06)* (2006), pp. 77–80.

[113] LIBERMAN, S., AND MARKOVITCH, S. Compact hierarchical explicit semantic representation. In *Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09)* (2009).

[114] LIH, A. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism* (2004), pp. 16–17.

[115] LIN, D. An information-theoretic definition of similarity. In *15th International Conference on Machine Learning (ICML '98)* (1998), pp. 296–304.

[116] LIU, H., BAO, H., AND XU, D. Concept vector for semantic similarity and relatedness based on wordnet structure. *Journal of Systems and Softwares 85* (2012), 370–381.

[117] LIU, H., AND SINGH, P. Commonsense reasoning in and over natural language. In *Proceeding of International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES '04)* (2004), vol. 3215, pp. 293–306.

[118] LOPEZ, A. Statistical machine translation. *ACM Computing Survey 40*, 3 (2008).

[119] LUND, K., AND BURGESS, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers 28* (1996), 203–208.

[120] MANNING, C. D., AND SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[121] MARCUS, M. P., SANTORINI, B., AND MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics 19*, 2 (1993), 313–330.

[122] MARTON, Y., CALLISON-BURCH, C., AND RESNIK, P. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)* (2009), vol. 1, pp. 381–390.

[123] MATSUO, Y., SAKAKI, T., UCHIYAMA, K., AND ISHIZUKA, M. Graph-based word clustering using a web search engine. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)* (2006), pp. 542–550.

[124] MATUSZEK, C., CABRAL, J., WITBROCK, M., AND DEOLIVEIRA, J. An introduction to the syntax and content of cyc. *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering 3864*, 1447 (2006), 44–49.

[125] MATUSZEK, C., WITBROCK, M., KAHLERT, R. C., CABRAL, J., SCHNEIDER, D., SHAH, P., AND LENAT, D. Searching for common sense: Populating cyc from the web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence* (2005), pp. 1430–1435.

[126] MEDELYAN, O., MILNE, D., LEGG, C., AND WITTEN, I. H. Mining meaning from wikipedia. *International Journal of Human Computer Studies 67* (2009), 716–754.

[127] MEDELYAN, O., WITTEN, I. H., AND MILNE, D. Topic indexing with wikipedia. In *Proceedings of of Association for the Advancement of Artificial Intelligence (AAAI '08), Wikipedia and Artificial Intelligence: An Evolving Synergy. Papers from the 2008 AAAI Workshop* (2008), pp. 19–24.

[128] MERTEN, T., AND FISCHER, I. Creativity, personality and word association responses: associative behaviour in forty supposedly cre-

ative persons. *Personality and Individual Differences 27*, 5 (1999), 933–942.

[129] MICHAEL, S., AND PONZETTO, S. P. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st national conference on Artificial intelligence (AAAI '06)* (2006), vol. 2, pp. 1419–1424.

[130] MIHAILA, C., AND ANANIADOU, S. What causes a causal relation? detecting causal triggers in biomedical scientific discourse. In *ACL (Student Research Workshop)* (2013), The Association for Computer Linguistics, pp. 38–45.

[131] MIHALCEA, R., CORLEY, C., AND STRAPPARAVA, C. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI '06)* (2006), pp. 775–780.

[132] MIHALCEA, R., AND CSOMAI, A. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on information and knowledge management, (CIKM '07)* (2007), pp. 233–242.

[133] MIHALCEA, R., AND MOLDOVAN, D. I. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)* (1999), pp. 152–158.

[134] MILLER, G. A. Wordnet: A lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[135] MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D., AND MILLER, K. Wordnet: An on-line lexical database. *International Journal of Lexicography 3* (1990), 235–244.

[136] MILLER, G. A., AND CHARLES, W. G. Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*, 1 (1991), 1–28.

[137] MILNE, D., MEDELYAN, O., AND WITTEN, I. H. Mining Domain-Specific thesauri from wikipedia: A case study. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI '06)* (2006), pp. 442–448.

[138] MILNE, D., AND WITTEN, I. H. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy* (2008), pp. 25–30.

[139] MILNE, D., AND WITTEN, I. H. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management, (CIKM '08)* (2008), pp. 509–518.

[140] MILNE, D., AND WITTEN, I. H. An open-source toolkit for mining wikipedia. *Artificial Intelligence 194* (2013), 222–239. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[141] MISCHEL, W. *Personality and assessment*. Psychology Press, 2013.

[142] MOHAMMAD, S., GUREVYCH, I., HIRST, G., AND ZESCH, T. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of EMNLP-CoNLL* (2007), pp. 571–580.

[143] MOHAMMAD, S., AND HIRST, G. Distributional measures of semantic distance: A survey. *CoRR abs/1203.1858* (2012).

[144] MORRIS, J., AND HIRST, G. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics (CLS '04)* (2004), pp. 46–51.

[145] MUELLER, E. T. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research 5*, 4 (2004), 307–340.

[146] NARAYANAN, S., AND HARABAGIU, S. Question answering based on semantic structures. In *Proceedings of the 20th international conference on Computational Linguistics (COLING '04)* (2004), Association for Computational Linguistics.

[147] NAVIGLI, R. Word Sense Disambiguation: a survey. *ACM Computing Surveys 41*, 2 (2009), 1–69.

[148] NAVIGLI, R., AND PONZETTO, S. P. Babelrelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI '12)* (2012).

[149] NELSON, D. L., MCEVOY, C. L., AND SCHREIBER, T. A. The university of south florida free association, rhyme, and word fragment norms. *Journal of Behavior Research Methods, Instruments, &amp; Computers* (2004), 402–407.

[150] NG, V. Shallow semantics for coreference resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (2007), pp. 1689–1694.

[151] NIELSEN, M. L. The word association test in the methodology of thesaurus construction. *Advances in Classification Research Online 8*, 1 (1997).

[152] NIELSEN, M. L., AND GJERLUF ESLAU, A. Corporate thesauri - how to ensure integration of knowledge and reflection of diversity. In *Proceedings of the Seventh International ISKO Conference* (2002), pp. 324–331.

[153] NIELSEN, M. L., AND INGWERSEN, P. The word association methodology: A gateway to work-task based retrieval. In *Proceedings of the Final Mira Conference on Information Retrieval Evaluation (MIRA '99)* (2006), pp. 17–27.

[154] NOREEN, E. W. *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience, 1989.

[155] PANG, B., AND LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2* (January 2008), 1–135.

[156] PARKER, E. B., AND OF CONGRESS., L. *LC thesaurus for graphic materials*. Cataloging Distribution Service, Library of Congress Washington, D.C, 1987.

[157] PARKER, R., Ed. *English gigaword fourth edition [electronic resource].* Linguistic Data Consortium, Philadelphia, PA, 2009.

[158] PATWARDHAN, S., BANERJEE, S., AND PEDERSEN, T. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, (CICLing '03)* (2003), pp. 241–257.

[159] PEDERSEN, T. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)* (Stroudsburg, PA, USA, 2010), Association for Computational Linguistics, pp. 329–332.

[160] PEDERSEN, T., PAKHOMOV, S. V. S., PATWARDHAN, S., AND CHUTE, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *Biomedical Informatics 40* (2007), 288–299.

[161] PILEHVAR, M. T., JURGENS, D., AND NAVIGLI, R. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of The Association for Computer Linguistics, ACL (1)* (2013), pp. 1341–1351.

[162] PIRRÓ, G., AND EUZENAT, J. A feature and information theoretic framework for semantic similarity and relatedness. In *Proceedings of the 9th international semantic web conference on The semantic web (ISWC '10)* (2010), vol. Part I, pp. 615–630.

[163] PITLER, E., RAGHUPATHY, M., MEHTA, H., NENKOVA, A., LEE, A., AND JOSHI, A. Easily Identifiable Discourse Relations. In *Coling 2008: Companion volume: Posters* (Manchester, UK, 2008), pp. 87–90.

[164] PONZETTO, S. P., AND STRUBE, M. Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research (JAIR) 30* (2007), 181–212.

[165] PUSTEJOVSKY, J., CASTAO, J., INGRIA, R., SAUR, R., GAIZAUSKAS, R., SETZER, A., AND KATZ, G. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of Fifth International Workshop on Computational Semantics (IWCS-5* (2003), vol. 3, pp. 28–34.

[166] RADA, R., MILI, H., BICKNELL, E., AND BLETTNER, M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics 19*, 1 (1989), 17–30.

[167] RADINSKY, K., AGICHTEIN, E., GABRILOVICH, E., AND MARKOVITCH, S. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World Wide Web (WWW '11)* (2011), pp. 337–346.

[168] RAJAGOPAL, D., CAMBRIA, E., OLSHER, D., AND KWOK, K. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW '13)* (2013), pp. 565–570.

[169] RAMAGE, D., RAFFERTY, A. N., AND MANNING, C. D. Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4 '09)* (2009), pp. 23–31.

[170] REED, S. L., AND LENAT, D. B. Mapping ontologies into cyc. In *Proceedings of the AAAI 2002 Conference Workshop on Ontologies For the Semantic Web* (2002), pp. 1–6.

[171] RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (AAAI '95)* (1995), pp. 448–453.

[172] RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR) 11* (1999), 95–130.

[173] RESNIK, P., AND DIAB, M. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, Philadelphia, PA* (2000), pp. 399–404.

[174] RICHMAN, A. E., AND SCHONE, P. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL '08)* (2008), pp. 1–9.

[175] ROEMMELE, M., BEJAN, C. A., AND GORDON, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (2011).

[176] ROGET, P. M. *Roget's Thesaurus of English words and phrases*. Project Gutemberg, Illinois Benedectine College, Lisle IL (USA), 1852.

[177] ROSCH, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General 104*, 3 (1975), 192–233.

[178] RUBENSTEIN, H., AND GOODENOUGH, J. B. Contextual correlates of synonymy. *Communications of the ACM 8* (1965), 627–633.

[179] RUIZ-CASADO, M., ALFONSECA, E., AND CASTELLS, P. Using context-window overlapping in synonym discovery and ontology extension. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '05)* (2005).

[180] SAHAMI, M., AND HEILMAN, T. D. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)* (2006), pp. 377–386.

[181] SAHLGREN, M. Vector-based semantic analysis: Representing word meanings based on random labels. In *Proceedings of ESSLI Workshop on Semantic Knowledge Acquistion and Categorization* (2001), Kluwer Academic Publishers.

[182] SALTON, G., WONG, A., AND YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (1975), 613–620.

[183] SÁNCHEZ, D., AND BATET, M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics 44*, 5 (2011), 749–759.

[184] SCHICKEL-ZUBER, V., AND FALTINGS, B. Oss: a semantic similarity function based on hierarchical ontologies. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI '07)* (2007), pp. 551–556.

[185] SCHONHOFEN, P. Identifying document topics using the wikipedia category network. In *Proceedings of the International Conference on Web Intelligence (WI '06)* (2006), IEEE Computer Society, pp. 456–462.

[186] SCHÜTZE, H. Automatic word sense discrimination. *Journal of Computational Linguistics 24*, 1 (1998), 97–123.

[187] SCHTZE, H. Dimensions of meaning. In *Proceedings of Supercomputing '92* (1992), pp. 787–796.

[188] SIMPSON, G. G. Mammals and the nature of continents. *American Journal of Science 241*, 1 (1943), 1–31.

[189] SINGER, M. Causal bridging inferences: Validating consistent and inconsistent sequences. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expÃ© rimentale 47*, 2 (1993), 340.

[190] SMITH, B., AND MARK, D. Ontology with human subjects testing: An empirical investigation of geographic categories. *American Journal of Economics and Sociology 58*, 2 (1999), 245–272.

[191] SPEER, R., HAVASI, C., AND LIEBERMAN, H. Analogyspace: Reducing the dimensionality of common sense knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI '08)* (2008), vol. 1, pp. 548–553.

[192] SPITERI, L. F. Word association testing and thesaurus construction: Defining inter-term relationships. In *Proceedings of the 30th Annual Conference of the Canadian Associaion for Information Science* (2002), pp. 24–33.

[193] STEINBACH, M., KARYPIS, G., AND KUMAR, V. A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining* (2000), vol. 400, pp. 525–526.

[194] STONE, P. J. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.

[195] STREHL, A., STREHL, E., GHOSH, J., AND MOONEY, R. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000* (2000), AAAI, pp. 58–64.

[196] STRÖTGEN, J., ALONSO, O., AND GERTZ, M. Identification of top relevant temporal expressions in documents. In *Proceedings of the 2nd Temporal Web Analytics Workshop (TempWeb '12)* (2012), pp. 33–40.

[197] SUCHANEK, F. M., KASNECI, G., AND WEIKUM, G. Yago: A large ontology from wikipedia and wordnet. *Web Semantics 6*, 3 (2008), 203–217.

[198] SUN, Q., LI, R., LUO, D., AND WU, X. Text segmentation with lda-based fisher kernel. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (2008), pp. 269–272.

[199] TAIEB, M. A. H., AOUICHA, M. B., AND HAMADOU, A. B. Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems 50* (2013), 260–278.

[200] TAMIR, R., AND SINGER, Y. On a confidence gain measure for association rule discovery and scoring. *The VLDB Journal 15*, 1 (2006), 40–52.

[201] TANG, J., YAO, L., AND CHEN, D. Multi-topic based query-oriented summarization. In *proceedings of the Siam International Conference on Data Mining (SDM '09)* (2009), pp. 1147–1158.

[202] TIMBERLAKE, W. Behavior systems, associationism, and Pavlovian conditioning. *Psychonomic Bulletin & Review 1*, 4 (1994), 405–420.

[203] TOIVONEN, H., GROSS, O., TOIVANEN, J. M., AND VALITUTTI, A. On creative uses of word associations. In *In Proceedings of Advances in Intelligent Systems and Computing (SMPS '12)* (2012), vol. 190, pp. 17–24.

[204] TSATSARONIS, G., VARLAMIS, I., AND VAZIRGIANNIS, M. Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research 37*, 1 (2010), 1–40.

[205] TURNEY, P. Learning algorithms for keyphrase extraction. *Information Retrieval 2* (2000), 303–336.

[206] TURNEY, P. D. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning (EMCL '01)* (2001), pp. 491–502.

[207] TVERSKY, B., AND HEMENWAY, K. Categories of environmental scenes. *Cognitive Psychology 15*, 1 (1983), 121–149.

[208] VAN DEN BROEK, P. The causal inference maker: Towards a process model of inference generation in text comprehension. *Comprehension processes in reading* (1990), 423–445.

[209] VARELAS, G., VOUTSAKIS, E., EURIPIDES, PETRAKIS, MILIOS, E., AND RAFTOPOULOU, P. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th ACM Internernational Workshop on Web Information and Data Management (WIDM '05* (2005), ACM Press, pp. 10–16.

[210] VASILESCU, F., LANGLAIS, P., AND LAPALME, G. Evaluating variants of the lesk approach for disambiguating words. In *Proceedings of Language Resources and Evaluation (LREC 2004)* (2004), pp. 633–636.

[211] VIÉGAS, F. B., WATTENBERG, M., AND DAVE, K. Studying cooperation and conflict between authors with history flow visualizations.

In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '04)* (2004), pp. 575–582.

[212] VOORHEES, E. *Query expansion using lexical-semantic relations*. Springer-Verlag New York, Inc., 1994, pp. 61–69.

[213] VOORHEES, E. M. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93)* (1993), ACM, pp. 171–180.

[214] WALES, J. http://en.wikipedia.org/.

[215] WANG, J. Z., DU, Z., PAYATTAKOOL, R., YU, P. S., AND CHEN, C.-F. A new method to measure the semantic similarity of go terms. *Bioinformatics 23*, 10 (2007), 1274–1281.

[216] WEALE, T., BREW, C., AND FOSLER-LUSSIER, E. Using the wiktionary graph structure for synonym detection. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web '09)* (2009), pp. 28–31.

[217] WEEDS, J., AND WEIR, D. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03)* (2003), pp. 81–88.

[218] WETTLER, M., AND RAPP, R. Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora* (1993), pp. 84–93.

[219] WILKINSON, D. M., AND HUBERMAN, B. A. Assessing the value of cooperation in wikipedia. *CoRR 12* (2007).

[220] WILKINSON, D. M., AND HUBERMAN, B. A. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis (WikiSym '07)* (2007), ACM, pp. 157–164.

[221] WU, F., AND WELD, D. S. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)* (2010), pp. 118–127.

[222] WU, Z., AND PALMER, M. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94)* (1994), pp. 133–138.

[223] YANG, D., AND POWERS, D. M. W. Measuring semantic similarity in the taxonomy of wordnet. In *Proceedings of the Twenty-eighth Australasian Conference on Computer Science (ACSC '05)* (2005), vol. 38, pp. 315–322.

[224] YAZDANI, M., AND POPESCU-BELIS, A. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence 194* (2013), 176–202.

[225] YEH, E., RAMAGE, D., MANNING, C. D., AGIRRE, E., AND SOROA, A. Wikiwalk: random walks on wikipedia for semantic relatedness. In *2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4 '09)* (2009), pp. 41–49.

[226] YIH, W., AND QAZVINIAN, V. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)* (2012), pp. 616–620.

[227] ZESCH, T., AND GUREVYCH, I. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances, (ACL '06)* (2006), pp. 16–24.

[228] ZESCH, T., AND GUREVYCH, I. Wisdom of crowds versus wisdom of linguists, measuring the semantic relatedness of words. *Natural Language Engineering 16*, 1 (2010), 25–59.

[229] ZESCH, T., GUREVYCH, I., AND MHLHUSER, M. Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '07)* (2007), pp. 205–208.

[230] ZESCH, T., MÜLLER, C., AND GUREVYCH, I. Using wiktionary for computing semantic relatedness. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI '08)* (2008), vol. 2, pp. 861–866.

[231] ZHAI, C. Statistical language models for information retrieval a critical review. *Foundation and Trends in Information Retrieval 2*, 3 (2008), 137–213.

[232] ZHANG, W., FENG, W., AND WANG, J. Integrating semantic relatedness and words' intrinsic features for keyword extraction. In *Proceedings of the twenty-third International Joint Conference on Artificial Intelligence (IJCAI '13)* (2013), pp. 2225–2231.

[233] ZHANG, Z., GENTILE, A., AND CIRAVEGNA, F. Recent advances in methods of lexical semantic relatedness-a survey. *Natural Language Engineering 1*, 1 (2012), 1–69.

[234] ZIPF, G. Human behaviour and the principle of least-effort. Addison-Wesley, 1949.

[235] ZOHAR, H., LIEBESKIND, C., SCHLER, J., AND DAGAN, I. Automatic thesaurus construction for cross generation corpus. *ACM Journal of Computing and Cultural Heritage 6*, 1 (2013), 1–19.