Measuring Semantic Distance using Distributional Profiles of Concepts

by

Saif Mohammad

A thesis submitted in conformity with the requirements
for the degree of
Graduate Department of Computer Science
University of Toronto

Abstract

Measuring Semantic Distance

using Distributional Profiles of Concepts

Saif Mohammad

Graduate Department of Computer Science

University of Toronto

2008

Semantic distance is a measure of how close or distant in meaning two units of language are. A large number of important natural language problems, including machine translation and word sense disambiguation, can be viewed as semantic distance problems. The two dominant approaches to estimating semantic distance are the **WordNet-based semantic measures** and the **corpus-based distributional measures**. In this thesis, I compare them, both qualitatively and quantitatively, and identify the limitations of each.

This thesis argues that estimating semantic distance is essentially a property of concepts (rather than words) and that two concepts are semantically close if they occur in similar contexts. Instead of identifying the co-occurrence (distributional) profiles of *words* (**distributional hypothesis**), I argue that **distributional profiles of concepts (DPCs)** can be used to infer the semantic properties of concepts and indeed to estimate semantic distance more accurately. I propose a new hybrid approach to calculating semantic distance that combines corpus statistics and a published thesaurus (*Macquarie Thesaurus*). The algorithm determines estimates of the DPCs using the categories in the thesaurus as very coarse concepts and, notably, without requiring any sense-annotated data. Even though the use of only about 1000 concepts to represent the vocabulary of a language seems drastic, I show that the method achieves results better than the state-of-the-art in a number of natural language tasks.

I show how **cross-lingual DPCs** can be created by combining text in one language with

a thesaurus from another. Using these cross-lingual DPCs, we can solve problems in one, possibly resource-poor, language using a knowledge source from another, possibly resource-rich, language. I show that the approach is also useful in tasks that inherently involve two or more languages, such as machine translation and multilingual text summarization.

The proposed approach is computationally inexpensive, it can estimate both semantic relatedness and semantic similarity, and it can be applied to all parts of speech. Extensive experiments on ranking word pairs as per semantic distance, real-word spelling correction, solving *Reader's Digest* word choice problems, determining word sense dominance, word sense disambiguation, and word translation show that the new approach is markedly superior to previous ones.

Dedication

To my brother, Zameer, and my parents, Shamim and Adil.

And to Toronto, for providing context to my thoughts, for brimming with inspiration, and for its extraordinary warmth.

Acknowledgements

I had a great time in Toronto and I had a great time working on this thesis. Both largely because of the unique and diverse group of people I had the opportunity of meeting while I was here. I am forever richer to have known them and deeply grateful for their indelible imprint on this thesis and my soul.

It was a joy and a privilege to work with Graeme Hirst. I thank him for having faith in me, for helping me develop ideas, for working patiently on my writing, and for helping me see the bigger picture. He is a model advisor, a brilliant mind, and a thorough gentleman. I will miss being able to walk across the corridor and talking to him, but I consider myself very fortunate to have had the opportunity to do exactly that for four years.

I thank Suzanne Stevenson, Gerald Penn, and Renée Miller for being on my thesis committee and giving valuable feedback. They were incredibly supportive and helped move this research in interesting directions. I thank Rada Mihalcea for agreeing to be the external examiner of this thesis and for finding time to come personally to Toronto.

I feel fortunate to have collaborated with Iryna Gurevych and Torsten Zesch (Darmstadt University of Technology) and Philip Resnik (University of Maryland, College Park). I gained valuable insights and good friends through these joint projects. I thank Alex Budanitsky (one of Graeme's earlier students) for allowing the use of his malapropism code. It was a pleasure to work with Yaroslav Riabinin (then a final year CS undergraduate at the University of Toronto, now a grad student) on a research project that intersected both our research goals.

I thank Diana McCarthy (University of Sussex) and Peter Turney (National Research Council of Canada) for their encouragement, especially early on when things seemed less clear.

Michael Demko deserves special praise. Whether wishing it or not an office mate can often find himself to be a bouncing board for the ideas of his colleague. Whether wishing it or not, Michael was *great* sounding board. The many conversations Michael, Robert Swier, and I had on numerous random issues (that oddly seemed very important at the time) were source of much joy.

In Afra Alishahi I found a dear friend, supportive colleague, and a wonderful companion to explore what this city had to offer. I thank Afsaneh Fazly for her sound advice and judgment. Much of the smooth sailing in the high seas of doctoral candidacy can be attributed to these two fond friends.

I thank my colleagues in the computational linguistics group: Faye Baron, Christopher Collins, Ol'ga Feiguina, Ulrich Germann, Cosmin Munteanu, Paul Cook, Diana Inkpen, Jane (Jianhua) Li, Meghana Marathe, Ryan North, Chris Parisien, Frank Rudzicz, Yun Niu, Vivian Tsang, Tony Wang, Amber Wilcox-O'Hearn, Xiaodan Zhu. They were a great support structure and they made grad school fun!

I will take this opportunity to express thanks and best wishes to Coco (Yan Ke) (currently pursuing her M.B.A. at Schulich School of Business, York University) and Tara Kahan (now an alumnus of the Department of Chemistry, University of Toronto). It was always fun hiking away summer weekends with good company and a pleasant change to sit-in on Environmental Chemistry seminars.

I thank the Toronto Outdoor Club and its founder Stephanie Amann, for letting me connect not just with nature but also with the people of Toronto. It is hard to forget: Cynthia Beernink with her camera, a great sense of social responsibility, and a ready smile; Marianna Danshina and her passion for culture; James R. Olchowy and his biking events; the always colorful Jason Mazariegos; and the endearing Joanna Reading.

I thank Michelle Chu, Jin Kang, Aravind Kumar, Tejaswi Popurri, Chaitanya Mishra, and Abhishek Ranjan for being interesting room-mates and good friends. Never a dull time at 5 Ross and one can always find some one to talk to—even if it is 2 in the morning. I thank Suraj Ramesh and Nikhil Ramesh for all the chess, Scrabble, cricket, and ping pong.

I will be remiss if I did not thank Ted Pedersen, my Master's advisor at the University of Minnesota, for introducing me to Computational Linguistics. I thank him and my fellow grad students from Duluth—Siddharth Patwardhan, Bridget McInnes, and Amruta Purandare—for their continued support.

Finally, I thank my family and friends back home in India for all their love and for letting me follow my dreams—even if it meant living, for extended periods of time, on the other side of the planet.

Contents

List of Tables

List of Figures

Chapter 1

Semantic Distance

1.1 Introduction

Semantic distance is a measure of how close or distant two units of language are, in terms of their meaning. The units of language may be words, phrases, sentences, paragraphs, or documents. The two nouns *dance* and *choreography*, for example, are closer in meaning than the two nouns *clown* and *bridge*, and so are said to be semantically closer. Units of language, especially words, may have more than one possible meaning. However, their context may be used to determine the intended senses. For example, *star* can mean both CELESTIAL BODY and CELEBRITY; however, *star* in the sentence below refers only to CELESTIAL BODY and is much closer to *sun* than to *famous*:

(1) *Stars are powered by nuclear fusion.*

Thus, semantic distance between words in context is in fact the distance between word senses or concepts. I use the terms *word senses* and *concepts* interchangeably here, although later on I will make a distinction. Figure 1.1 depicts that the concepts of DANCE and CHOREOGRAPHY are closer in meaning than the concepts of CLOWN and BRIDGE. Throughout the thesis, example words will be written in italics (as in the example sentence above), whereas example senses or concepts will be written in all capitals (as in Figure 1.1).

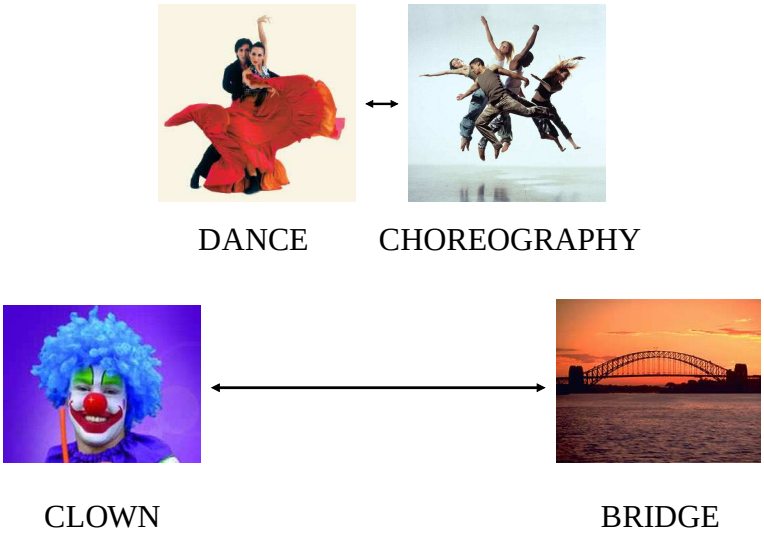


Figure 1.1: Semantic distance between example concepts.

Humans consider two concepts to be semantically close if there is a sharing of some meaning. Specifically, two concepts are semantically close if there is a **lexical semantic relation** between the concepts. Putting it differently, the reason why two concepts are considered semantically close can be attributed to a lexical semantic relation that binds them. According to Cruse (1986), a lexical semantic relation is the relation between **lexical units**—a surface form along with a sense. As he points out, the number of semantic relations that bind concepts is innumerable but certain relations, such as hyponymy, meronymy, antonymy, and troponymy, are more systematic and have enjoyed more attention in the linguistics community. However, as Morris and Hirst (2004) point out these relations are far out-numbered by others which they call **non-classical relations**. Here are a few of the kinds of non-classical relations they observed: positive qualities (BRILLIANT, KIND), concepts pertaining to a concept (KIND, CHIVALROUS, FORMAL pertaining to GENTLEMANLY), and commonly co-occurring words (locations such as HOMELESS, SHELTER; problem–solution pairs such as HOMELESS, DRUNK).

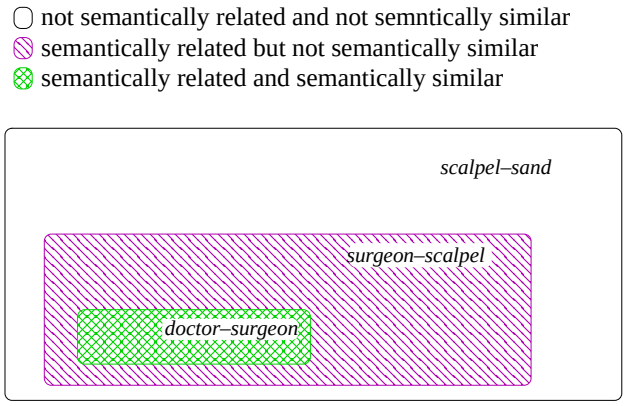


Figure 1.2: A Venn diagram of word pairs in semantic distance space.

1.1.1 Semantic relatedness and semantic similarity

Semantic distance is of two kinds: **semantic similarity** and **semantic relatedness**. The former is a subset of the latter (Figure 1.2), but the two may be used interchangeably in certain contexts, making it even more important to be aware of their distinction. Two concepts are considered to be semantically similar if there is a hyponymy (hypernymy), antonymy, or troponymy relation between them. Two concepts are considered to be semantically related if there is any lexical semantic relation between them—classical or non-classical.

Semantically similar concepts tend to share a number of common properties. For example, consider APPLES and BANANAS. They are both hyponyms of FRUIT. They are both edible, they grow on trees, they have seeds, etc. Therefore, APPLES and BANANAS are considered to be semantically similar. Another example of a semantically similar pair is DOCTOR and SURGEON. The concept of a DOCTOR is a hypernym of SURGEON. Therefore, they share the properties associated with a DOCTOR.

On the other hand, semantically related concepts may not have many properties in common, but have at least one classical or non-classical lexical relation between them which lends them the property of being semantically close. For example, DOOR and KNOB are semantically related as one is the meronym (is part) of another. The concept pairs, DOCTOR and SURGEON are semantically related (as well as being semantically similar) as one is the hyponym of the

other. Example pairs considered semantically related due to non-classical relations include SURGEON–SCALPEL and TREE–SHADE. Note that semantic similarity entails semantic relatedness (Figure 1.3 (a)) but the converse need not be true (Figure 1.3 (b)).

1.1.2 Can humans estimate semantic distance?

Many will agree that humans are adept at estimating semantic distance, but consider the following questions. How strongly will two people agree/disagree on distance estimates? Will the agreement vary over different sets of concepts? In our minds, is there a clear distinction between related and unrelated concepts or are concept-pairs spread across the whole range from synonymous to unrelated? Some of the earliest work that begins to answer these questions is by Rubenstein and Goodenough (1965a). They conducted quantitative experiments with human subjects (51 in all) who were asked to rate 65 English word pairs on a scale from 0.0 to 4.0 as per their semantic distance. The word pairs chosen ranged from almost synonymous to unrelated. However, they were all noun pairs and those that were semantically close were semantically similar; the dataset did not contain word pairs that are semantically related but not semantically similar (word pairs pertaining to the ▨ region of Figure 1.2). The subjects repeated the annotation after two weeks and the new distance values had a Pearson's correlation r of 0.85 with the old ones. Miller and Charles (1991) also conducted a similar study on 30 word pairs taken from the Rubenstein-Goodenough pairs. These annotations had a high correlation ($r = 0.97$) with the mean annotations of Rubenstein and Goodenough (1965a). Resnik (1999) repeated these experiments and found the inter-annotator agreement (r) to be 0.90.

Resnik and Diab (2000) conducted annotations of 48 verb pairs and found inter-annotator agreement (r) to be 0.76 (when the verbs were presented without context) and 0.79 (when presented in context). Gurevych (2005) and Zesch et al. (2007b) asked native German speakers to mark two different sets of German word pairs with distance values. Set 1 was a German translation of the Rubenstein and Goodenough (1965a) dataset. It had 65 noun–noun word pairs. Set 2 was a larger dataset containing 350 word pairs made up of nouns, verbs, and

DOCTOR SURGEON

semantic similarity

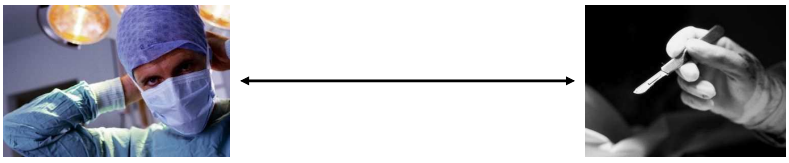
DOCTOR SURGEON

semantic relatedness

(a) Concept pair that is semantically related and semantically similar.

SURGEON SCALPEL

semantic relatedness



SURGEON SCALPEL

semantic similarity

(b) Concept pair that is semantically related but not semantically similar.

Figure 1.3: Examples

Table 1.1: Different datasets that are manually annotated with distance values. Pearson's correlation was used to determine inter-annotator agreement (last column).

Dataset	Year	Language	# pairs	PoS	# subjects	Agreement
Rubenstein and Goodenough	1965	English	65	N	51	-
Miller and Charles	1991	English	30	N	-	.90
Resnik and Diab	2000	English	27	V	-	.76 and .79
Gurevych	2005	German	65	N	24	.81
Zesch and Gurevych	2006	German	350	N, V, A	8	.69

adjectives. The semantically close word pairs in the 65-word set were mostly synonyms or hypernyms (hyponyms) of each other, whereas those in the 350-word set had both classical and non-classical relations with each other. Details of these **semantic distance benchmarks** are summarized in Table 1.1. Inter-subject agreements (last column in Table 1.1) are indicative of the degree of ease in annotating the datasets. The high agreement and correlation values suggest that humans are quite good and consistent at estimating semantic distance of noun-pairs; however, annotating verbs and adjectives and a combination of parts of speech is harder. This also means that estimating semantic relatedness is harder than estimating semantic similarity. It should be noted here that even though the annotators were presented with word-pairs and not concept-pairs, it is reasonable to assume that they were annotated as per their closest senses. For example, given the noun pair *bank* and *interest*, most if not all will identify it as semantically related even though both words have more than one sense and many of the sense–sense combinations are unrelated (for example, the RIVER BANK sense of *bank* and the SPECIAL ATTENTION sense of *interest*).

Apart from proving that humans can indeed estimate semantic distance, these datasets act as "gold standards" to evaluate automatic distance measures. However, lack of large amounts of data from human subject experimentation limits the reliability of this mode of evaluation. Therefore automatic distance measures are also evaluated by their usefulness in natural language tasks such as those described in the next section.

1.1.3 Pervasiveness of semantic distance in natural language processing

A large number of problems in natural language processing are in essence semantic-distance problems. Machine translation systems must choose a translation hypothesis in the target language that is semantically closest, if not identical, to the source language text. Paraphrases are pieces of text that can be used more or less interchangeably and can be identified by their property of being semantically close. The same is true, albeit to a lesser extent, for a phrase that entails another. Information retrieval involves the selection of documents closest in content to the query terms. Query-based summarization requires, among other things, choosing those sentences to be part of the summary that are closest to the query. Document clustering is the grouping of semantically close pieces of text. Discovering word senses from their usage involves grouping the usages so that those in the same group are semantically close to each other whereas those in different groups are distant—each such group represents a sense of the target. Word sense disambiguation is the identification of the sense closest to a particular instance of the target word. Identifying idioms and specific idiomatic usages of multiword expressions involves determining whether a usage (or a set of usages) of the expression is semantically distant from the usages of its components—if they are more distant, then the probability that the expression is used in a non-literal sense is higher. Real-word spelling errors can be detected by identifying words that are semantically distant from their context and the existence of a spelling variant that is close (Hirst and Budanitsky, 2005). Word completion and prediction algorithms rank those candidate words higher that are semantically close to the preceding context.

Thus, semantic distance plays a key role in natural language processing. As measures of semantic distance between concepts can be extended to calculate the distance between larger units of language, such as phrases and documents, understanding and improving these measures will have a significant and wide-ranging impact (see Table 1.2 for some recent applications). In this thesis, I will identify some of the key drawbacks and limitations of state-of-the-art distance measures, and propose a new class of measures that not only overcomes those problems but also lends itself for use in more tasks through its substantially new capabilities.

Table 1.2: Natural language tasks that have been attempted with measures of semantic distance.

Natural Language Task	**Approaches that use semantic distance**
Cognates (identifying)	Kondrak (2001)
Coreference resolution	Ponzetto and Strube (2006)
Document clustering	Wang and Hodges (2006)
Information extraction	Hassan et al. (2006); Stevenson and Greenwood (2005)
Information retrieval	Varelas et al. (2005)
Multiword expressions (identifying)	Baldwin et al. (2003); Cook et al. (2007)
Paraphrasing and textual entailment	Schilder and Thomson McInnnes (2006); Ferrández et al. (2006); Zanzotto and Moschitti (2006)
Question answering	Lamjiri et al. (2007)
Real-word spelling error detection	Hirst and Budanitsky (2005); Mohammad and Hirst (2006b)
Relation extraction	Chen et al. (2005)
Semantic similarity of texts	Corley and Mihalcea (2005)
Speech recognition	Inkpen and Desilets (2005)
Subjectivity (determining)	Wiebe and Mihalcea (2006)
Summarization	Gurevych and Strube (2004); Zhu and Penn (2005); Li et al. (2006)
Textual inference	Haghighi et al. (2005); Raina et al. (2005)
Word prediction	Pucher (2006)
Word sense disambiguation	Banerjee and Pedersen (2003); McCarthy (2006); Mohammad et al. (2007b); Patwardhan et al. (2007)
Word-sense discovery	Ferret (2004)
Word-sense dominance (determining)	McCarthy et al. (2004b); Mohammad and Hirst (2006a)
Word translation*	Mohammad et al. (2007b)

* Word translation refers to determining the translation of a word using its context.

1.1.4 Can machines estimate semantic distance?

Two classes of methods have been used in automatically determining semantic distance. **Knowledge-rich measures of concept-distance**, such as those of Jiang and Conrath (1997), Leacock and Chodorow (1998), and Resnik (1995), rely on the structure of a knowledge source, such as WordNet, to determine the distance between two concepts defined in it.[1] **Distributional measures of word-distance (knowledge-lean measures)**, such as cosine and α-skew divergence (Lee, 2001), rely on the **distributional hypothesis** which states that two words tend to be semantically close if they occur in similar contexts (Firth, 1957). These measures rely simply on text and can give the distance between any two words that occur at least a few times.

The various WordNet-based measures have been widely studied (Budanitsky and Hirst, 2006; Patwardhan et al., 2003). Even though individual distributional measures are being used more and more, the study of distributional measures on the whole, especially when work on this thesis commenced, received much less attention.[2] In Chapter 2, I summarize various knowledge-rich approaches to semantic distance and present a detailed analysis of the distributional measures.

1.2 Why the need for a better approach

Distributional word-distance and WordNet-based concept-distance measures each have certain uniquely attractive features: WordNet-based measures can capitalize on the manual encoding of lexical semantic relations, while distributional approaches are widely applicable because they need only raw raw text (and maybe some shallow syntactic processing). Unfortunately, these advantages come at a cost. I now flesh out the limitations of both kinds of measures.

[1] The nodes in WordNet (synsets) represent concepts and edges between nodes represent semantic relations such as hyponymy and meronymy.

[2] See Curran (2004) and Weeds et al. (2004) for other work that compares various distributional measures.

1.2.1 Limitations common to both WordNet-based concept-distance and corpus-based word-distance measures

1.2.1.1 Computational complexity and storage requirements

As applications for linguistic distance become more sophisticated and demanding, it becomes attractive to pre-compute and store the distance values between all possible pairs of words or senses. However both WordNet-based and distributional measures have large space requirements to do this, requiring matrices of size $N \times N$, where N is very large. In case of distributional measures, N is the size of the vocabulary (at least 100,000 for most languages). In case of WordNet-based measures, N is the number of senses (81,000 just for nouns). Given that the above matrices tend to be sparse[3] and that computational capabilities are continuing to improve, the above limitation may not seem hugely problematic, but as we see more and more natural language applications in embedded systems and hand-held devices, such as cell phones, iPods, and medical equipment, memory and computational power become serious constraints.

1.2.1.2 Reluctance to cross the language barrier

Both WordNet-based and distributional distance measures have largely been used in a monolingual framework. Even though semantic distance seems to hold promise in tasks, such as machine translation and multi-lingual text summarization, that inherently involve two or more languages, automatic measures of semantic distance have rarely been applied to these tasks. With the development of the EuroWordNet, involving interconnected networks of seven different languages, it is possible that we shall see more cross-lingual work using WordNet-based measures in the future. However, such an interconnected network will be very hard to create for more different language pairs such as English and Chinese or English and Arabic.

[3]Even though, WordNet-based and distributional measures give non-zero similarity and relatedness values to a large number of term pairs (concept pairs and word pairs), values below a suitable threshold can be reset to 0.

1.2.2 Further limitations of WordNet-based concept-distance measures

1.2.2.1 Lack of high-quality WordNet-like knowledge sources

Ontologies, WordNets, and semantic networks are available for a few languages such as English, German, and Hindi. Creating them requires human experts and it is time intensive. Thus, for most languages, we cannot use WordNet-based measures simply due to the lack of a WordNet in that language. Further, even if created, updating an ontology is again expensive and there is usually a lag between the current state of language usage/comprehension and the semantic network representing it. Further, the complexity of human languages makes creation of even a near-perfect semantic network of its concepts impossible. Thus in many ways the ontology-based measures are only as good as the networks on which they are based.

On the other hand, distributional measures require only text. Large corpora, billions of words in size, may now be collected by a simple web crawler. Large corpora of more-formal writing are also available (for example, the *Wall Street Journal* or the *American Printing House for the Blind (APHB)* corpus). This makes distributional measures very attractive.

1.2.2.2 Poor estimation of semantic relatedness

As Morris and Hirst (2004) pointed out, a large number of concept pairs, such as STRAWBERRY–CREAM and DOCTOR–SCALPEL, have a non-classical relation between them (STRAWBERRIES are usually eaten with CREAM and a DOCTOR uses a SCALPEL to make an incision). These words are not semantically similar, but rather semantically related. An ontology- or WordNet-based measure will correctly identify the amount of semantic relatedness only if such relations are explicitly coded into the knowledge source. Further, the most accurate WordNet-based measures rely only on its extensive is-a hierarchy. This is because networks of other lexical-relations such as meronymy are much less developed. Further, the networks for different parts of speech are not well connected. All this means that, while WordNet-based measures accurately estimate semantic similarity between nouns, their estimation of semantic relatedness

especially in pairs other than noun–noun is at best poor and at worse non-existent. On the other hand, distributional measures can be used to determine both semantic relatedness and semantic similarity (see Section 2.3.1 for more details).

1.2.2.3 Inability to cater to specific domains

Given a concept pair, measures that rely only on WordNet and no text, such as Rada et al. (1989), give just one distance value. However, two concepts may be very close in a certain domain but not so much in another. For example, SPACE and TIME are close in the domain of quantum mechanics but not so much in most others. Ontologies have been made for specific domains, which may be used to determine semantic similarity specific to these domains. However, the number of such ontologies is very limited. Some of the more successful WordNet-based measures, such as Jiang and Conrath (1997), that rely on text as well, do indeed capture domain-specificity to some extent, but the distance values are still largely shaped by the underlying network, which is not domain-specific. On the other hand, distributional measures rely primarily (if not completely) on text and large amounts of corpora specific to particular domains can easily be collected.

1.2.3 Further limitations of corpus-based word-distance measures

1.2.3.1 Conflation of word senses

The distributional hypothesis Firth (1957) states that words that occur in similar contexts tend to be semantically close. But when words have more than one sense, it is not at all clear what semantic distance between them actually means. Further, a word in each of its senses is likely to co-occur with different sets of words. For example, *bank* in the FINANCIAL INSTITUTION sense is likely to co-occur with *interest, money, accounts,* and so on, whereas the RIVER BANK sense might have words such as *river, erosion,* and *silt* around it. Since words that occur together in text tend to refer to senses that are closest in meaning to one another, in most natural

language applications, what is needed is the distance between the closest senses of the two target words. However, because distributional measures calculate distance from occurrences of the target word in all its occurrences and hence all its senses, they fail to get the desired result. Also note that the dimensionality reduction inherent to latent semantic analysis (LSA), a special kind of distributional measure, has the effect of making the predominant senses of the words more dominant while de-emphasizing the other senses. Therefore, an LSA-based approach will also conflate information from the different senses, and even more emphasis will be placed on the predominant senses. Given the semantically close target nouns *play* and *actor*, for example, a distributional measure will give a score that is some sort of a dominance-based average of the distances between their senses. The noun *play* has the predominant sense of CHILDREN'S RECREATION (and not DRAMA), so a distributional measure will tend to give the target pair a large (and thus erroneous) distance score. WordNet-based measures do not suffer from this problem as they give distance between concepts, not words.

1.2.3.2 Lack of explicitly-encoded world knowledge and data sparseness

It is becoming increasingly clear that more-accurate results can be achieved in a large number of natural language tasks, including the estimation of semantic distance, by combining corpus statistics with a knowledge source, such as a dictionary, published thesaurus, or WordNet. This is because such knowledge sources capture semantic information about concepts and, to some extent, world knowledge. For example, WordNet, as discussed earlier, has an extensive is-a hierarchy. If it lists one concept, say GERMAN SHEPHERD as a hyponym of another, say DOG, then we can be sure that the two are semantically close. On the other hand, distributional measures do not have access to such explicitly encoded information. Further, unless the corpus used by a distributional measure has sufficient instances of GERMAN SHEPHERD and DOG, it will be unable to deem them semantically close. Since Zipf's law seems to hold even for the largest of corpora, there will always be words that occur too few times to accurately determine their distributional distance from others.

1.3 A new approach: Distributional measures of concept-distance

In this thesis, I propose a new hybrid approach that combines a knowledge source with text to measure semantic distance. The new measures have the best features of both semantic and distributional measures and some additional advantages as well. They address, with varying degrees of success, the limitations of earlier approaches.

1.3.1 The argument: Distributional profiles of concepts for measuring semantic distance

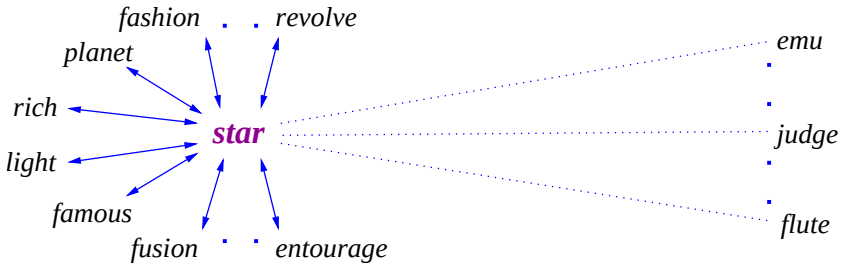
The central argument of this thesis is that semantic distance is essentially a property of concepts (rather than of words) and that two concepts are semantically close if they occur in similar contexts. This is similar to the distributional hypothesis except that the target is a word sense or concept (rather than a word). The set of contexts of a concept can be represented by what I will call the **distributional profile of the concept** (**DP of the concept** or simply **DPC**). The distributional profile of a concept is the set of words that co-occur with it in text, along with their strength of the co-occurrence association—a numeric value indicating how much more than random chance a word tends to co-occur with a concept (more details in Chapter 3). Thus, the semantic distance between two concepts can be determined by calculating the distance between the respective **DPCs**. The argument proposed here reduces to the distributional hypothesis when we consider words with just one sense or meaning. However, the words people use most tend to be highly ambiguous.

> It is a perverse feature of human languages that the words used most frequently tend to be the most polysemantic.
>
> — George A. Miller ("Ambiguous Words", *Impacts Magazine*, May 2001)

While the distributional hypothesis clumps all occurrences of a word into one bag (Figure 1.4), I propose profiling the different senses separately (Figure 1.5). The motivation is that a word when used in different senses tends to keep different company, that is, it co-occurs with a different sets of words. By profiling the contexts of different senses separately, we will be able to infer the semantic properties of the different senses and indeed estimate semantic distance more accurately.

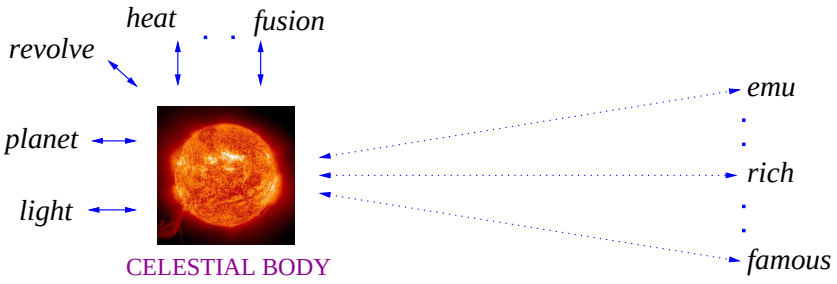
The creation of DPCs requires: (1) a concept inventory that lists all the concepts and words that refer to them, and (2) counts of how often a concept co-occurs with a word in text. We use the categories in the *Macquarie Thesaurus*, 812 in all, as very coarse-grained word senses or concepts (Figure 1.6). This is a departure from the norm in the computational linguistics community where the use of WordNet or another similarly fine-grained sense inventory is more common. However, this very aspect of fine-grainedness has been widely criticized for some time now (Agirre and Lopez de Lacalle Lekuona (2003) and citations therein), and is one of the reasons this work uses a published thesaurus; Section 1.3.2 presents further motivation.

Since words may be used in more than one sense and can refer to different concepts in different contexts, a direct approach to determining the concept–word co-occurrence counts requires sense-annotated text. However, manual annotation is tedious, expensive, and not easily scalable. This brings us to the following questions: (1) Can we determine accurate estimates of concept–word co-occurrence counts, and thereby determine DPCs, without the use of sense-annotated data? and (2) Can these estimates of DPCs be used to infer semantic properties of concepts, and indeed accurately measure semantic distance? This thesis claims that the answers to both of these questions are affirmative—an even stronger claim than the one made earlier in this section. In Chapter 3, I propose a bootstrapping and concept-disambiguation algorithm to create (estimates of) distributional profiles of concepts without the use of any human-annotated data. In Chapter 4, I show how DPCs can be created in a cross-lingual framework. Chapters 3 through 7 describe experiments in various natural language tasks that were attempted using

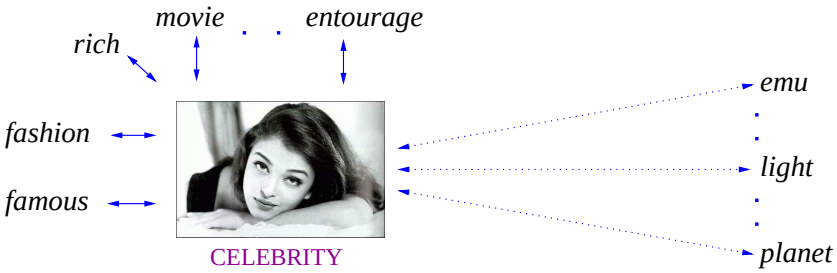


DP of *star*

Figure 1.4: Example distributional profile (DP) of the word *star*. A solid arrow indicates strong co-occurrence association whereas a dotted arrow indicates a weak co-occurrence association.

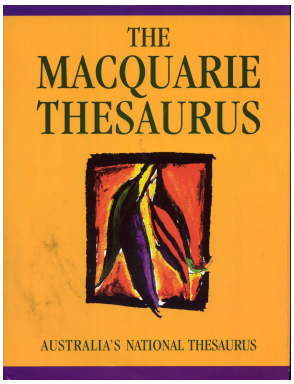


(a) DP of CELESTIAL BODY



(b) DP of CELEBRITY

Figure 1.5: Example distributional profiles of two senses of *star*.



302. FURNITURE	369. HONESTY
furniture	honesty
chair	incorruptness
stool	integrity
bench	plain dealing
couch	probity
cot	scrupulosity
table	sincerity
...	...
...	...
...	...

Figure 1.6: The *Macquarie Thesaurus* and fragments of its content.

these DPCs (Section 1.3.3 gives a brief outline) and whose results validate the claims made above.

1.3.2 A suitable knowledge source and concept inventory

Knowledge sources, such as dictionaries, thesauri, and wordnets, capture semantic information about concepts and, to some extent, world knowledge. The approach proposed here does not require a complex array of concepts interconnected by semantic relations as in WordNet. Nor does it require glosses that tend to be somewhat subjective and rigid. Instead, it requires only that the knowledge source provide a list of all the concepts in a language (or a subset of the language) and a set of words and/or multiword expressions representing each concept. I use the categories in the *Macquarie Thesaurus* as senses. Most published thesauri divide the vocabulary into about 1000 categories, which can be considered as the basic concepts represented by the language. The words listed under each category gloss the meaning of the concept. The concepts (categories) roughly correspond to very coarse-grained word senses (Yarowsky, 1992).

Published thesauri are available in a number of languages, although, admittedly many languages may not have comprehensive and high-quality ones. Resources that are not thesauri, per se, may also be used in place of a published thesaurus. (See Lapata and Keller (2007)

for a simplified version of our word sense dominance system (Mohammad and Hirst, 2006a) that uses WordNet instead of a thesaurus.) More importantly, as I will describe in the next subsection, my approach can determine semantic distance in one, possibly resource-poor, language using a thesaurus from another, possibly resource-rich language, thereby eliminating the knowledge-source bottleneck.

As applications for linguistic distance become more sophisticated and demanding, it becomes attractive to pre-compute and store the distance values between all possible pairs of words or senses. But both corpus-based word-distance and WordNet-based sense-distance measures have large space requirements, needing matrices of size $N \times N$, where N is the size of the vocabulary (perhaps 100,000 for most languages) in the case of distributional measures and the number of senses (75,000 just for nouns in WordNet) in the case of semantic measures. The use of categories in a thesaurus as concepts means that this approach requires a concept–concept distance matrix of size only about 1000×1000—much smaller than (about 0.01% the size of) the matrix required by traditional semantic and distributional measures. This makes the approach scalable to large amounts of text. Working in a relatively smaller number of dimensions (1000 concepts), as suggested above, means that on the one hand there will be a loss of information (in this case, a loss of distinction between near-synonyms) and yet on the other hand there is more information to accurately determine semantic distance between the coarse concepts. As I will show, through various experiments throughout this thesis, in a number of natural language applications, using semantic distance between these very coarse senses is just as useful if not more so. Further, I believe, this distance approach provides a powerful starting point to build on top of it a system that differentiates near-synonyms.

In this thesis, I go further and use the idea of a very coarse sense inventory to develop a framework for distributional measures of concept-distance that can more naturally and more accurately be used in place of semantic measures of word senses than distributional measures of word-distance.

1.3.3 Applications in measuring semantic distance

In Chapters 3 and 4, I will show how, using the DPCs, traditional distributional measures, such as cosine and α-skew divergence, can be used to measure the distance between concepts (rather than words). I will show that when applied even to domain-general noun concepts[4] in the tasks of **ranking word pairs in order of their semantic distance**, **correcting real-word spelling errors**, and **solving *Reader's Digest* word choice problems**, the newly proposed distributional concept-distance measures outperform traditional word-distance measures and are as accurate as, if not better than, the semantic measures.

Modeling co-occurrence distributions of concepts and words allows this approach (unlike the traditional semantic and distributional measures) to attempt in an unsupervised manner tasks that traditionally require sense-annotated data. In Chapter 5, I will show how distributional profiles of concepts can be used to **determine word sense dominance**—the proportion of the occurrences of a target word used in a particular sense—by both explicit and implicit word sense disambiguation. **Word sense disambiguation**, as mentioned earlier, is the identification of the sense closest to the context of a particular occurrence of the target word. Chapter 6 describes how the DPCs can be used to create an *unsupervised* naïve Bayes word sense classifier. This system participated in SemEval-07's English Lexical Sample Space coarse-grained word sense disambiguation task and was only about one percentage point below the best unsupervised system.[5]

Knowledge-rich measures of concept-distance and distributional measures of word-distance are largely monolingual, that is, they are used to quantify distance between concepts or words in the same language. Further, the use of semantic measures to estimate distance in one language requires a knowledge source in that (same) language. Unfortunately, most languages do not have knowledge sources such as WordNet. Even though many languages, as pointed

[4] As mentioned earlier, the performance of semantic measures is significantly worse for concept pairs other than noun–noun.

[5] SemEval-07 is a workshop of ACL-07, where systems compete in various semantic analysis tasks on newly compiled/created test data.

out in the previous section, have a published thesaurus, still many more do not. In Chapter 4, I will show how text in one language can be combined with a knowledge source in another, using a bilingual lexicon and the bootstrapping/concept-disambiguation algorithm, to create **cross-lingual distributional profiles of concepts**.

These cross-lingual DPCs model co-occurrence distributions of concepts, as per a knowledge source in one language, with words from another language. They can be used to obtain state-of-the-art accuracies in **estimating semantic distance in a resource-poor language using a knowledge source from a resource-rich one**. In Chapter 4, I will show how German–English DPCs can be created by combining a German corpus with an English thesaurus using a German–English bilingual lexicon. A comparison of this approach with strictly monolingual approaches that use GermaNet reveals that the cross-lingual approach performs just as well, if not better, thereby proving the worth of the approach to languages that lack a GermaNet, WordNet, or other such knowledge source.

Cross-lingual semantic distance and cross-lingual DPCs are also useful in tasks that inherently involve two or more languages. In Chapter 7, I will show how they can help **machine translation**—choosing a translation hypothesis in the target language that is semantically closest, if not identical, to the source language text. The implementation of a DPC-based unsupervised naïve Bayes classifier placed first among all unsupervised systems taking part in SemEval-07's Multilingual Chinese–English Lexical Sample Task, where suitable English translations of given target Chinese words in context were to be identified.

Together these results provide unequivocal and substantial evidence for the claim that estimates of distributional profiles of concepts, created without the use of any manually-annotated data, can be used to infer semantic properties of a concept, and indeed accurately measure semantic distance.

Chapter 2

State-of-the-art in estimating semantic distance

2.1 Knowledge sources

Automatic measures of semantic distance rely on one or more knowledge sources, such as text, dictionaries, thesauri, and WordNet. Those that rely simply on text and give distance between *words*, such as the distributional measures, are referred to as **knowledge-lean** whereas others, such as the WordNet-based measures that give distance between *concepts* are called **knowledge-rich**. Measures of concept-distance require both a concept inventory that lists all the concepts in a language and a lexicon that lists all the words that refer to them. WordNet acts as both the concept inventory and the lexicon for the WordNet-based measures, while the *Macquarie Thesaurus* plays those roles in the approach I propose.

2.1.1 Text

Words that occur within a certain window of a target word are called the **co-occurrences** of the word. The window size may be a few words on either side, the complete sentence, the paragraph or the entire document. Consider the sentence below:

*Nobody can cast a **spell** like Hermione*

If we consider the window size to be the complete sentence, then *spell* co-occurs with *nobody, can, cast, a, like,* and *Hermione*. The co-occurring words are also said to constitute the **context** of the target word.

The target word may have more than one meaning, but when used in a sentence it almost always refers to just one of these senses or concepts. Thus, the words that co-occur with the target word can also be said to co-occur with its intended sense. Although *spell* can mean A PERIOD OF TIME, in the example above it is used in the INCANTATION OR CHARM sense. We can therefore also say that *nobody, can, cast, a, like,* and *Hermione* co-occur with the concept of INCANTATION OR CHARM. Co-occurring words have long been used to determine semantic properties of the target word. In this thesis, the words that co-occur with a concept will be used to determine its semantic distance from other concepts.

Measures of Association

Some words co-occur with the target (word or concept) just by chance, whereas others tend to co-occur more often than chance. For example, *nobody* is expected to co-occur with *spell* (or INCANTATION OR CHARM) more or less by chance; however, *cast* is expected to co-occur with the same target much more often than chance. The stronger the association between the target and a co-occurring word, the more informative the co-occurring word is. The hypothesis is that the more two concepts are semantically related, the more they will be talked about together. Therefore, if inferences are to be made about the target from its co-occurring words, then more weight is given to information provided by stronger co-occurrences. The weight is proportional to the **strength of association**, which quantifies how strong the co-occurrence is. It can be calculated by applying a suitable statistic, such as pointwise mutual information

(PMI), to a **contingency table** of the target t (word or concept) and the co-occurring word w.

	t	$\neg t$
w	n_{wt}	$n_{w\neg}$
$\neg w$	$n_{\neg t}$	$n_{\neg\neg}$

A contingency table shows the number of times two events occur together (n_{wt}), the number of times one occurs while the other does not ($n_{\neg t}$ and $n_{w\neg}$), and the number of times neither occurs ($n_{\neg\neg}$). Strength of association values are calculated from observed frequencies ($n_{wt}, n_{\neg t}, n_{w\neg}$, and $n_{\neg\neg}$), marginal frequencies ($n_{w*} = n_{wt} + n_{w\neg}$; $n_{\neg *} = n_{\neg t} + n_{\neg\neg}$; $n_{*t} = n_{wt} + n_{\neg t}$; and $n_{*\neg} = n_{w\neg} + n_{\neg\neg}$), and the sample size ($N = n_{wt} + n_{\neg t} + n_{w\neg} + n_{\neg\neg}$). It should be noted here that when counting co-occurrence frequencies to populate the contingency table, one may choose whether or not to incorporate the order of the co-occurring terms. For example, n_{wt} may be chosen to be the number of times w co-occurs with t: (1) such that w is followed by t; or (2) irrespective of whether w follows t or the other way round. Both ways of determining the contingency table are defendable. For all the experiments conducted as part of this thesis, the order of co-occurrence is ignored.

Pointwise mutual information (PMI), is one of the most widely used measures of association. Its formula is given below:

$$pmi(w,t) = \log \frac{n_{wt} \times N}{n_{w*} \times n_{*t}}$$

PMI gives a score of 0 if the occurrence of one event is statistically independent of the other. Scores can reach positive infinity if the events are dependent and negative infinity if they are inversely dependent. Strictly speaking, the above formula does not truly represent PMI because while PMI calculations expect n_{wt} to be less than or equal to n_w, the way term co-occurrence in text is usually counted n_{wt} may be greater than n_w; for example, in a particular sentence, if there are two occurrences of t close to w, then n_{wt} is incremented by 2 whereas n_w is incremented by just 1. Church and Hanks (1990) pioneered the use of such a PMI-based measure of association and they called it **word association ratio** to differentiate it from PMI. Also, co-occurrence

counts for word association ratio respect the order of terms in text. Since the experiments in this thesis ignore order of co-occurrence and because the difference from PMI is only minor, rather than coining a new term and in accordance with the computational linguistics jargon, I will refer to the PMI-based measure of association simply as PMI.

The odds ratio (Tan et al., 2002) varies between 0 (inversely dependent) and positive infinity (dependent);

$$odds(w,t) = \frac{n_{wt} \times n_{\neg\neg}}{n_{w\neg} \times n_{\neg t}}$$

where a score of 1 indicates statistical independence. Yule's coefficient (Tan et al., 2002) transforms the odds ratio to a -1 to 1 scale with 0 representing independence.

$$Yule(w,t) = \frac{\sqrt{odds(w,t)} - 1}{\sqrt{odds(w,t)} + 1}$$

The cosine (van Rijsbergen, 1979) and Dice coefficient vary between 0 and 1, while the ϕ coefficient (Tan et al., 2002) gives values between 0 and infinity.

$$\cos(w,t) = \frac{n_{wt}}{\sqrt{n_{w*}} \times \sqrt{n_{*t}}}$$

$$Dice(w,t) = \frac{2 \times n_{wt}}{n_{w*} + n_{*t}}$$

$$\phi(w,t) = \frac{(n_{wt} \times n_{\neg\neg}) - (n_{w\neg} \times n_{\neg t})}{\sqrt{n_{w*} \times n_{\neg*} \times n_{*t} \times n_{*\neg}}}$$

There is no particular value signifying independence for these three measures. The higher the values, the stronger the association between the word and category.

2.1.2 WordNet

WordNet is described by its creators as a "large, electronically available, lexical database of English" (Fellbaum, 1998). It is a semantic network in which each node, called a synset, represents a fine-grained concept or word sense. Each synset is composed of a gloss and a set of near-synonymous words which refer to that concept. The synsets are connected by lexical relations such as hyponymy, meronymy, and so on.

WordNet 3.0, the current version as of this thesis, has more than 117,000 synsets and covers more than 155,000 word-types. It has more than 81,000 noun, 13,000 verb, 18,000 adjective, and 3,000 adverb synsets. It has a coverage of more than 117,000 noun, 11,000 verb, 22,000 adjective, and 4,000 adverb word-types.

Since its creation, WordNet has been used by the computational linguistics research community for a wide range of tasks from machine translation to text categorization to identifying cognates. Its remarkable success has propelled creation of wordnets for numerous other languages too. For example, GermaNet is a wordnet that connects German nouns, verbs, and adjectives. It has more than 60,000 synsets.

However, the fine-grainedness of WordNet remains one of its key drawbacks. WordNet-based measures of semantic distance require matrices of size $N \times N$, where N is the number of senses—81,000 just for nouns. The approach proposed in this thesis relies on a published thesaurus, but for the sake of comparison I also conducted experiments using state-of-the art approaches that rely on WordNet and GermaNet.

2.1.3 Thesauri

Published thesauri, such as *Roget's* and *Macquarie*, divide the English vocabulary into around a thousand **categories** of near-synonymous and semantically related words. Words with more than one meaning are listed in more than one category. For every word-type in the vocabulary of the thesaurus, the index specifies the categories that list it. Categories roughly correspond to very coarse word senses or concepts (Yarowsky, 1992), and the terms will be used interchangeably. For example, in the *Macquarie Thesaurus*, *bark* is listed in the categories ANIMAL NOISES and MEMBRANE. These categories represent the coarse senses of *bark*. A published thesaurus thus provides us with a very coarse human-developed set or inventory of word senses or concepts that are more intuitive and discernible than the "concepts" generated by dimensionality-reduction methods such as latent semantic analysis. Using coarse senses from a known inventory means that the senses can be represented unambiguously by a large

number of possibly ambiguous words (conveniently available in the thesaurus)—a feature I will exploit to determine useful estimates of the strength of association between a concept and co-occurring words.

We use the *Macquarie Thesaurus* (Bernard, 1986) categories as very coarse word senses. It has 812 categories with around 176,000 word-tokens and 98,000 word-types. This allows us to have a much smaller **concept–concept distance matrix** of size just 812×812 (roughly .01% the size of matrices required by existing measures).

Note that in published thesauri, such as *Roget's* and *Macquarie*, categories are further divided into paragraphs and paragraphs into semicolon groups. Words within a semicolon group tend to be semantically closer to each other than those in different semicolon groups of the same paragraph. Likewise, words within a paragraph tend to be semantically closer than those in different paragraphs. The experiments described in this thesis do not take advantage of this information, except those detailed in Chapter 4. are structurally quite different from the so called "distributional thesaurus" automatically generated by Lin (1998b), wherein a word has exactly one entry, and its neighbors may be semantically related to it in any of its senses. All future mentions of *thesaurus* in this thesis will refer to a published thesaurus.

2.2 Knowledge-rich approaches to semantic distance

Creation of electronically available ontologies and semantic networks like WordNet has allowed their use to help solve numerous natural language problems including the measurement of semantic distance. Budanitsky and Hirst (2006), Hirst and Budanitsky (2005), and Patwardhan et al. (2003) have done an extensive survey of the various WordNet-based measures, their comparisons with human judgment on selected word pairs, and their usefulness in applications such as real-word spelling correction and word sense disambiguation. Hence, this section provides only a brief summary of the major knowledge-rich measures of semantic distance.

2.2.1 Measures that exploit WordNet's semantic network

A number of WordNet-based measures consider two concepts to be close if they are close to each other in WordNet. One of the earliest and simplest measures is the Rada et al. (1989) **edge-counting** method. The shortest path in the network between the two target concepts (**target path**) is determined. The more edges there are between two words, the more distant they are. Elegant as it may be, the measure hinges on the largely incorrect assumption that all the network edges correspond to identical semantic distance.

Nodes in a network may be connected by different kinds of lexical relations such as hyponymy, meronymy, and so on. Edge counts apart, the Hirst and St-Onge (1998) measure takes into account the fact that if the target path consists of edges that belong to many different relations, then the target concepts are likely more distant. The idea is that if we start from a particular node c_1 and take a path via a particular relation (say, hyponymy), to a certain extent the concepts reached will be semantically related to c_1. However, if during the way we take edges belonging to different relations (other than hyponymy), very soon we may reach words that are unrelated. Hirst and St-Onge's measure of semantic relatedness is listed below:

$$HS(c_1, c_2) = C - path\ length - k \times d \tag{2.1}$$

where c_1 and c_2 are the target concepts, d is the number of times an edge pertaining to a relation different from that of the preceding edge is taken, and C and k are empirically determined constants. More recently, Yang and Powers (2005) propose a weighted edge-counting method to determine semantic relatedness using the hypernymy/hyponymy, holonymy/meronymy, and antonymy links in WordNet.

Leacock and Chodorow (1998) used just one relation (hyponymy) and modified the path length formula to reflect the fact that edges lower down in the *is-a* hierarchy correspond to smaller semantic distance than the ones higher up. For example, synsets pertaining to *sports car* and *car* (low in the hierarchy) are much more similar than those pertaining to *transport* and *instrumentation* (higher up in the hierarchy) even though both pairs of nodes are separated by

exactly one edge in WordNet's *is-a* hierarchy.

$$LC(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2D} \tag{2.2}$$

where D is the depth in the taxonomy.

Resnik (1995) suggested a measure that combines corpus statistics with WordNet. He proposed that since the **lowest common subsumer** or **lowest super-ordinate (lso)** of the target nodes represents what is similar between them, the semantic similarity between the two concepts is directly proportional to how specific the lso is. The more general the lso is, the larger the semantic distance between the target nodes. This specificity is measured by the formula for information content (IC):

$$Res(c_1, c_2) = IC(lso(c_1, c_2)) = -\log p(lso(c_1, c_2)) \tag{2.3}$$

Observe that using information content has the effect of inherently scaling the semantic similarity measure by depth of the taxonomy. Usually, the lower the lowest super-ordinate, the lower the probability of occurrence of the lso and the concepts subsumed by it, and hence, the higher its information content.

As per Resnik's formula, given a particular lowest super-ordinate, the exact positions of the target nodes below it in the hierarchy do not have any effect on the semantic similarity. Intuitively, we would expect that word pairs closer to the lso are more semantically similar than those that are distant. Jiang and Conrath (1997) and Lin (1997) incorporate this notion into their measures which are arithmetic variations of the same terms. The Jiang and Conrath (1997) measure (*JC*) determines how dissimilar each target concept is from the lso ($IC(c_1) - IC(lso(c_1, c_2))$ and $IC(c_2) - IC(lso(c_1, c_2))$). The final semantic distance between the two concepts is then taken to be the sum of these differences. Lin (1997) (like Resnik) points out that the lso is what is common between the two target concepts and that its information content is the common information between the two concepts. His formula (*Lin*) can be thought of as

taking the Dice coefficient of the information in the two target concepts.

$$JC(c_1,c_2) = 2\log p(lso(c_1,c_2)) - (\log(p(c_1)) + (\log(p(c_2))) \quad (2.4)$$

$$Lin(c_1,c_2) = \frac{2 \times \log p(lso(c_1,c_2))}{\log(p(c_1)) + (\log(p(c_2))} \quad (2.5)$$

Budanitsky and Hirst (2006) show that the Jiang-Conrath measure has the highest correlation (0.850) with the Miller and Charles noun pairs and performs better than all other measures considered in a spelling correction task. Patwardhan et al. (2003) get similar results using the measure for word sense disambiguation.

All of the approaches described above rely heavily (if not solely) on the hypernymy/hyponymy network in WordNet; they are designed for, and evaluated on, noun–noun pairs. However, more recently, Resnik and Diab (2000) and Yang and Powers (2006a) developed measures aimed at verb–verb pairs. Resnik and Diab (2000) ported several measures which are traditionally applied on the noun hypernymy/hyponymy network (edge counting, Resnik (1995), and Lin (1997)) to the relatively shallow verb troponymy network. The two information content-based measures best ranked a carefully chosen set of 48 verbs in order of their semantic distance.[1] Yang and Powers (2006a) ported their earlier work on nouns (Yang and Powers, 2005) to verbs. In order to compensate for the relatively shallow verb troponymy hierarchy and the lack of a corresponding holonymy/meronymy hierarchy, they proposed several back-off models—the most useful one being the distance between a noun pair that has the same lexical form as the verb pair. However, the approach has too many tuned parameters (9 in all) and performed poorly on a set of 36 TOEFL word choice questions involving verb targets and alternatives.

2.2.2 Measures that rely on dictionaries and thesauri

Lesk (1986) introduced a method to perform word sense disambiguation using word glosses (definitions). The glosses of the senses of a target word are compared with those of its context

[1] Only those verbs were selected which require a theme and the sub-categorization frames of verb pairs had to match.

and the number of word overlaps is determined. The sense with the most number of overlaps is chosen as the intended sense of the target. Inspired by this approach, Banerjee and Pedersen (2003) proposed a semantic relatedness measure that deems to concepts to be more semantically related if there is more overlap in their glosses. Notably, they overcome the problem of short glosses by considering the glosses of concepts related to the target concepts through the WordNet lexical semantic relations such as hyponymy/hypernymy. They also give more weight to larger overlap sequences. Patwardhan and Pedersen (2006) proposed another gloss-based semantic relatedness measure which performed slightly worse than the extended gloss overlap measure in a word sense disambiguation task, but markedly better at ranking the Miller and Charles (1991) word pairs. Their approach has certain similarities to the one proposed in this thesis and so will be discussed in more detail in the Section 3.6 (*Related work*) of the next chapter.

Jarmasz and Szpakowicz (2003) use the taxonomic structure of the *Roget's Thesaurus* to determine semantic similarity. Two words are considered maximally similar if they occur in the same semicolon group in the thesaurus. Then on, decreasing in similarity are word pairs in the same paragraph, words pairs in different paragraphs belonging to the same part of speech and within the same category, word pairs in the category, and so on until word pairs which have nothing in common except that they are in the thesaurus (maximally distant). They show that this simple approach performs remarkably well at ranking word pairs and determining the correct answer in sets of TOEFL, ESL, and *Reader's Digest* word choice problems.

2.3 Knowledge-lean approaches to semantic distance

2.3.1 The distributional hypotheses: the original and the new

Distributional measures are inspired by the maxim "You shall know a word by the company it keeps" (Firth, 1957). These measures rely simply on raw text and possibly some shallow syntactic processing. They are much less resource-hungry than the semantic measures, but

they measure the distance between words rather than word-senses or concepts. Two words are considered close if they occur in similar contexts. Statistics acquired from large text corpora are used to determine how similar the contexts of the two words are. This distance between sets of contexts can be used as a proxy for semantic distance as words found in similar contexts tend to be semantically similar—the **distributional hypothesis** (Firth, 1957; Harris, 1968).

The hypothesis makes intuitive sense, as Budanitsky and Hirst (2006) point out: If two words have many co-occurring words in common, then similar things are being said about both of them and so they are likely to be semantically similar. Conversely, if two words are semantically similar, then they are likely to be used in a similar fashion in text and thus end up with many common co-occurrences. For example, the semantically similar *bug* and *insect* are expected to have a number of common co-occurring words such as *crawl, squash, small, woods*, and so on, in a large enough text corpus.

The distributional hypothesis only mentions semantic similarity and not semantic relatedness. This coupled with the fact that the difference between semantic relatedness and semantic similarity is somewhat nuanced, and can be missed, meant that almost all work employing the distributional hypothesis was labeled as estimating semantic similarity. However, it should be noted that distributional measures can be used to estimate both semantic similarity and semantic relatedness. Even though Schütze and Pedersen (1997) and Landauer et al. (1998), for example, use the term *similarity* and not *relatedness*, their LSA-based distance measures in fact estimate semantic relatedness and not semantic similarity. I propose more specific distributional hypotheses that make clear how distributional measures can be used to estimate semantic similarity and how they can be used to measure semantic relatedness:

> **Hypothesis of the distributionally close and semantically related:**
> Two target words are distributionally close and semantically related if they have many common strongly co-occurring words.
> (For example, *doctor–surgeon* and *doctor–scalpel*. See example co-occurring words in Table 2.1.)

Table 2.1: Example: Common syntactic relations of target words with co-occurring words.

	Co-occurring words		
	cut (v)	*hardworking* (adj)	*patient* (n)
Semantically similar target pair			
doctor (n)	subject–verb	noun–qualifier	subject–object
surgeon (n)	subject–verb	noun–qualifier	subject–object
Semantically related target pair			
doctor (n)	subject–verb	noun–qualifier	subject–object
scalpel (n)	prepositional object–verb	–	prepositional object–object

Hypothesis of the distributionally close and semantically similar:

Two target words are distributionally close and semantically similar if they have many common strongly co-occurring words that each have the same syntactic relation with the two targets.

(For example, *doctor–surgeon*, but not *doctor–scalpel*. See syntactic relations with example co-occurring words in Table 2.1.)

The idea is that both semantically similar and semantically related word pairs will have many common co-occurring words. However, words that are semantically similar belong to the same broad part of speech (noun, verb, etc.), but the same need not be true for words that are semantically related. Therefore, words that are semantically similar will tend to have the same syntactic relation, such as verb–object or subject–verb, with most common co-occurring words. Thus, the two words are considered semantically related simply if they have many common co-occurring words. But to be semantically similar as well, the words must have the same syntactic relation with co-occurring words. Consider the word pair *doctor–operate*. In a large enough body of text, the two words are likely to have the following common co-occurring

words: *patient, scalpel, surgery, recuperate,* and so on. All these words will contribute to a high score of relatedness. However, they do not have the same syntactic relation with the two targets. (The word *doctor* is almost always used as a noun while *operate* is a verb.) Thus, as per the two newly proposed distributional hypotheses, *doctor* and *operate* will correctly be identified as semantically related but not semantically similar. The word pair *doctor–nurse,* on the other hand, will be identified as both semantically related and semantically similar.

In order to clearly differentiate from the distance as calculated by a WordNet-based semantic measure (described earlier in Section 2.2.1), the distance calculated by a corpus-based distributional measure will be referred to as **distributional distance**.

2.3.2 Corpus-based measures of distributional distance

I now describe specific distributional measures that rely on the distributional hypotheses; depending on which specific hypothesis they use, they mimic either semantic similarity or semantic relatedness.

2.3.2.1 Spatial Metrics: Cos, L_1, L_2

Consider a multidimensional space where the number of dimensions is equal to the size of the vocabulary. A word *w* can be represented by a point in this space such that the component of $\vec{w}$ in a dimension (corresponding to word *x*, say) is equal to the strength of association (SoA) of *w* with *x* ($SoA(w,x)$) (Figure 2.1 (a)). Thus, the vectors corresponding to two words are *close* together, and thereby get a low distributional distance score, if they share many co-occurring words and the co-occurring words have more or less the same strength of association with the two target words (Figure 2.1 (b)). The distance between two vectors can be calculated in different ways as described below.

Cosine

The **cosine** method (denoted by **Cos**) is one of the earliest and most widely used distribu-

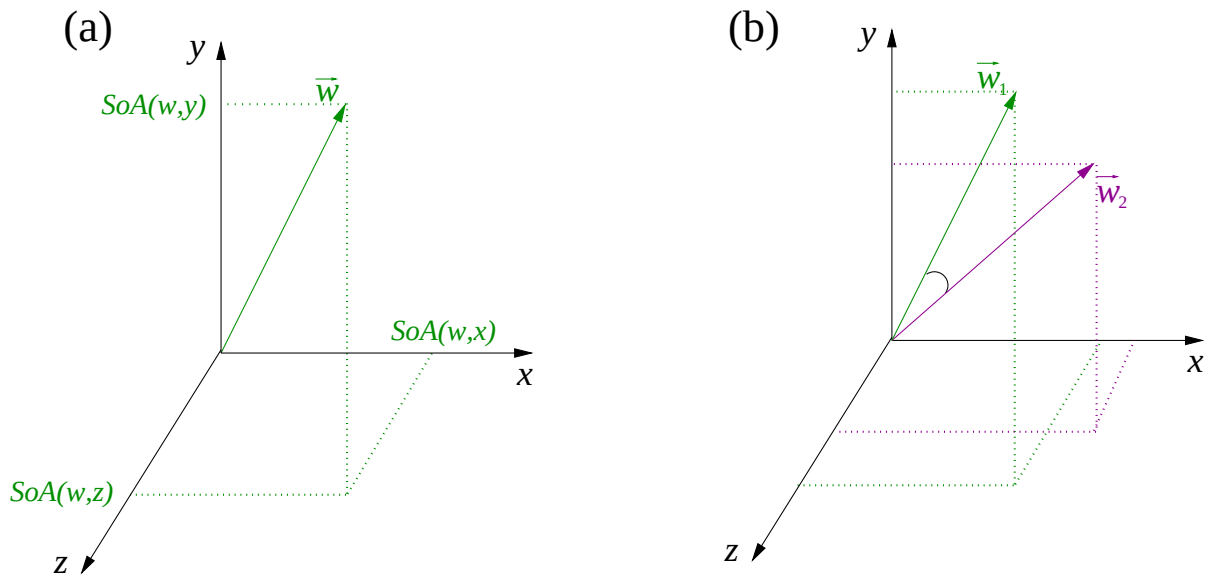


Figure 2.1: (a) A representation of word *w* in co-occurrence vector space. Values *wx*, *wy*, and *wz* are its strengths of association with *x*, *y*, and *z*, respectively. (b) Spatial distributional distance between target words w_1 and w_2.

tional measures. Given two words w_1 and w_2, the cosine measure calculates the cosine of the angle between $\vec{w}_1$ and $\vec{w}_2$. If a large number of words co-occur with both w_1 and w_2, then $\vec{w}_1$ and $\vec{w}_2$ will have a small angle between them and the cosine will be large; signifying a large relatedness/similarity between them. The cosine measure gives scores in the range from 0 (unrelated) to 1 (synonymous).

$$Cos(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) \times P(w|w_2))}{\sqrt{\sum_{w \in C(w_1)} P(w|w_1)^2} \times \sqrt{\sum_{w \in C(w_2)} P(w|w_2)^2}} \tag{2.6}$$

where $C(t)$ is the set of words that co-occur (within a certain window) with the word t in a corpus. In this example, conditional probability of the co-occurring words given the target words is used as the strength of association.

The cosine was used, among others, by Schütze and Pedersen (1997) and Yoshida et al. (2003), who suggest methods of automatically generating distributional thesauri from text corpora. Schütze and Pedersen (1997) use the Tipster category B corpus (Harman, 1993) (450,000 unique terms) and the *Wall Street Journal* to create a large but sparse co-occurrence matrix of

3,000 medium-frequency words (frequency rank between 2,000 and 5,000). Latent semantic indexing (singular value decomposition) (Schütze and Pedersen, 1997) is used to reduce the dimensionality of the matrix and get for each term a word vector of its 20 strongest co-occurrences. The cosine of a target word's vector with each of the other word vectors is calculated and the words that give the highest scores comprise the thesaurus entry for the target word.

Yoshida et al. (2003) believe that words that are closely related for one person may be distant for another. They use around 40,000 HTML documents to generate personalized thesauri for six different people. Documents used to create the thesaurus for a person are retrieved from the subject's home page and a web crawler which accesses linked documents. The authors also suggest a root-mean-squared method to determine the similarity of two different thesaurus entries for the same word.

Manhattan and Euclidean Distances

Distance between two points (words) in vector space can also be calculated using the formulae for **Manhattan distance** a.k.a. the $\mathbf{L_1}$ **norm** (denoted by $\mathbf{L_1}$) or **Euclidean distance** a.k.a. the $\mathbf{L_2}$ **norm** (denoted by L_2). In the Manhattan distance (2.7) (Dagan et al. (1997), Dagan et al. (1999), and Lee (1999)), the difference in strength of association of w_1 and w_2 with each word that they co-occur with is summed. The greater the difference, the greater is the distributional distance between the two words. Euclidean distance (2.8) (Lee, 1999) employs the root mean square of the difference in association to get the final distributional distance. Both the L_1 and L_2 norms give scores in the range between 0 (zero distance or synonymous) and infinity (maximally distant or unrelated).

$$L_1(w_1,w_2) = \sum_{w \in C(w_1) \cup C(w_2)} | P(w|w_1) - P(w|w_2) | \tag{2.7}$$

$$L_2(w_1,w_2) = \sqrt{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) - P(w|w_2))^2} \tag{2.8}$$

The above formulae use conditional probability of the co-occurring words given a target word as the strength of association.

Lee (1999) compared the ability of all three spatial metrics to determine the probability of an unseen (not found in training data) word pair. The measures in order of their performance (from better to worse) were: L_1 norm, cosine, and L_2 norm. Weeds (2003) determined the correlation of word pair ranking as per a handful of distributional measures with human rankings (Miller and Charles (1991) word pairs). She used verb-object pairs from the *British National Corpus (BNC)* and found the correlation of L_1 norm with human rankings to be 0.39.

2.3.2.2 Mutual Information–Based Measures: *Hindle, Lin*

Hindle (1990) was one of the first to factor the strength of association of co-occurring words into a distributional similarity measure.[2] Consider the nouns n_j and n_k that exist as objects of verb v_i in different instances within a text corpus. Hindle used the following formula to determine the distributional similarity of n_j and n_k solely from their occurrences as object of v_i:

$$Hin_{obj}(v_i, n_j, n_k) = \begin{cases} \min(I(v_i, n_j), I(v_i, n_k)), \\ \qquad \text{if } I(v_i, n_j) > 0 \text{ and } I(v_i, n_k) > 0 \\ |\max(I(v_i, n_j), I(v_i, n_k))|, \\ \qquad \text{if } I(v_i, n_j) < 0 \text{ and } I(v_i, n_k) < 0 \\ 0, \qquad \text{otherwise} \end{cases} \tag{2.9}$$

$I(n, v)$ stands for the PMI between the noun n and verb v (Note that in case of negative PMI values, the maximum function captures the PMI, which is lower in absolute value). The measure follows from the distributional hypothesis—the more similar the associations of co-occurring words with the two target words, the more semantically similar they are. Hindle used pointwise mutual information (PMI)[3] as the strength of association. The minimum of the two PMIs

[2] See Grefenstette (1992) for an approach that does NOT incorporate strength of association of co-occurring words. He, like Hindle (1990), uses syntactic dependencies to create distributional profiles of words. The Jaccard coefficient is applied to a pair of such distributional profiles to determine their similarity.

[3] In their respective papers, Donald Hindle and Dekang Lin refer to pointwise mutual information as mutual

captures the similarity in the strength of association of v_i with each of the two nouns.

Hindle used an analogous formula to calculate distributional similarity (Hin_{subj}) using the subject-verb relation. The overall distributional similarity between any two nouns is calculated by the formula:

$$Hin(n_1,n_2) = \sum_{i=0}^{N} \big(Hin_{obj}(v_i,n_1,n_2) + Hin_{subj}(v_i,n_1,n_2)\big) \tag{2.10}$$

The measure gives similarity scores from 0 (maximally dissimilar) to infinity (maximally similar or synonymous). Note that in Hindle's measure, the set of co-occurring words used is restricted to include only those words that have the same syntactic relation with both target words (either verb–object or verb–subject). This is therefore a measure that mimics semantic similarity and not semantic relatedness. A form of Hindle's measure where all co-occurring words are used, making it a measure that mimics semantic relatedness, is shown below:

$$Hin_{rel}(w_1,w_2) = \sum_{w \in C(w)} \begin{cases} \min(I(w,w_1),I(w,w_2)), \\ \qquad \text{if } I(w,w_1) > 0 \text{ and } I(w,w_2) > 0 \\ |\max(I(w,w_1),I(w,w_2))|, \\ \qquad \text{if } I(w,w_1) < 0 \text{ and } I(w,w_2) < 0 \\ 0, \qquad \text{otherwise} \end{cases} \tag{2.11}$$

where $C(t)$ is the set of words that co-occur with word t.

Lin (1998b) suggests a different measure derived from his information-theoretic definition of similarity (Lin, 1998a). Further, he uses a broad set of syntactic relations apart from just subject–verb and verb–object relations and shows that using multiple relations is beneficial even by Hindle's measure. He first extracts triples of the form (x,r,y) from the partially parsed text, where the word x is related to y by the syntactic relation r. If a particular triple (x',r',y') occurs c times in text, then the pointwise mutual information $I(x',r',y')$ is the information contained in the proposition: the triple (x,r,y) occurred a constant c times. Lin defines the

information.

distributional similarity between two words, w_1 and w_2, as follows:

$$Lin(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w') \in T(w_1)} I(w_1, r, w') + \sum_{(r,w'') \in T(w_2)} I(w_2, r, w'')} \tag{2.12}$$

where $T(x)$ is the set of all word pairs (r, y) such that the pointwise mutual information $I(x, r, y)$, is positive. Note that this is different from Hindle (1990) where even the cases of negative PMI were considered. As mentioned earlier, Church and Hanks (1990) show that it is hard to accurately predict negative word association ratios with confidence. Thus, co-occurrence pairs with negative PMI are ignored. The measure gives similarity scores from 0 (maximally dissimilar) to 1 (maximally similar).

Like Hindle's measure, Lin's is a measure of distributional *similarity*. However, it distinguishes itself from that of Hindle in two respects. First, Lin normalizes the similarity score between two words (numerator of (2.12)) by their cumulative strengths of association with the rest of the co-occurring words (denominator of (2.12)). This is a significant improvement as now high PMI of the target words with shared co-occurring words alone does not guarantee a high distributional similarity score. As an additional requirement, the target words must have low PMI with words they do not both co-occur with. Second, Hindle uses the minimum of the PMI between each of the target words and the shared co-occurring word, while Lin uses the sum. Taking the sum has the drawback of not penalizing for a mismatch in strength of co-occurrence, as long as w_1 and w_2 both co-occur with a word.

Hindle (1990) used a portion of the *Associated Press* news stories (6 million words) to classify the nouns into semantically related classes. Lin (1998b) used his measure to generate a distributional thesaurus from a 64-million-word corpus of the *Wall Street Journal, San Jose Mercury*, and *AP Newswire*. He also provides a framework for evaluating such automatically generated thesauri by comparing them with WordNet-based and Roget-based thesauri. He shows that the distributional thesaurus created with his measure is closer to the WordNet and Roget-based thesauri than that created using Hindle's measure.

2.3.2.3 Relative Entropy–Based Measures: KLD, ASD, JSD

Kullback-Leibler divergence

Given two probability mass functions $p(x)$ and $q(x)$, their **relative entropy** $D(p\|q)$ is:

$$D(p\|q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \qquad \text{for } q(x)\neq 0 \tag{2.13}$$

Intuitively, if $p(x)$ is the accurate probability mass function corresponding to a random variable X, then $D(p\|q)$ is the information lost when approximating $p(x)$ by $q(x)$. In other words, $D(p\|q)$ is indicative of how different the two distributions are. Relative entropy is also called the **Kullback-Leibler divergence** or the **Kullback-Leibler distance** (denoted by **KLD**).

Pereira et al. (1993) and Dagan et al. (1994) point out that words have probabilistic distributions with respect to neighboring syntactically related words. For example, there exists a certain probabilistic distribution ($d_1(P(v|n_1))$, say) of a particular noun n_1 being the object of any verb. This distribution can be estimated by corpus counts of parsed or chunked text. Let d_2 ($P(v|n_2)$) be the corresponding distribution for noun n_2. These distributions (d_1 and d_2) define the contexts of the two nouns (n_1 and n_2, respectively). As per the distributional hypothesis, the more these contexts are similar, the more n_1 and n_2 are semantically similar. Thus the Kullback-Leibler distance between the two distributions is indicative of the semantic distance between the nouns n_1 and n_2.

$$\begin{aligned} KLD(n_1,n_2) &= D(d_1\|d_2) \\ &= \textstyle\sum_{v\in \mathrm{Vb}} P(v|n_1)\log\frac{P(v|n_1)}{P(v|n_2)} && \text{for } P(v|n_2)\neq 0 \\ &= \textstyle\sum_{v\in \mathrm{Vb}\prime(n_1)\cap \mathrm{Vb}\prime(n_2)} P(v|n_1)\log\frac{P(v|n_1)}{P(v|n_2)} && \text{for } P(v|n_2)\neq 0 \end{aligned} \tag{2.14}$$

where Vb is the set of all verbs and $Vb\prime(x)$ is the set of verbs that have x as the object. Note again that the set of co-occurring words used is restricted to include only verbs that each have the same syntactic relation (verb-object) with both target nouns. This too is therefore a measure that mimics semantic similarity and not semantic relatedness.

It should be noted that the verb-object relationship is not inherent to the measure and that one or more of any other syntactic relations may be used. One may also estimate semantic relat-

edness by using all words co-occurring with the target words. Thus a more generic expression of the Kullback-Leibler divergence is as follows:

$$\begin{aligned} KLD(w_1, w_2) &= D(d_1 \| d_2) \\ &= \sum_{w \in V} P(w|w_1) \log \frac{P(w|w_1)}{P(w|w_2)} \quad \text{for } P(w|w_2) \neq 0 \\ &= \sum_{w \in C(w_1) \cup C(w_2)} P(w|w_1) \log \frac{P(w|w_1)}{P(w|w_2)} \quad \text{for } P(w|w_2) \neq 0 \end{aligned} \tag{2.15}$$

where V is the vocabulary (all the words found in a corpus). $C(t)$, as mentioned earlier, is the set of words occurring (within a certain window) with word t.

It should be noted that the Kullback-Leibler distance is not symmetric, that is, the distance from w_1 to w_2 is not necessarily, and even not likely, the same as the distance from w_2 to w_1. This asymmetry is counter-intuitive to the general notion of semantic similarity of words, although Weeds (2003) has argued in favor of asymmetric measures. Further, it is very likely that there are instances such that $P(w_1|v)$ is greater than 0 for a particular verb v, while due to data sparseness or grammatical and semantic constraints, the training data has no sentence where v has the object w_2. This makes $P(w_2|v)$ equal to 0 and the ratio of the two probabilities infinite. Kullback-Leibler divergence is not defined in such cases but approximations may be made by considering smoothed values for the denominator.

Pereira et al. (1993) used KLD to create clusters of nouns from verb-object pairs corresponding to a thousand most frequent nouns in the *Grolier's Encyclopedia*, June 1991 version (10 million words). Dagan et al. (1994) used KLD to estimate the probabilities of bigrams that were not seen in a text corpus. They point out that a significant number of possible bigrams are not seen in any given text corpus. The probabilities of such bigrams may be determined by taking a weighted average of the probabilities of bigrams composed of distributionally similar words. Use of Kullback-Leibler distance as the semantic distance metric yielded a 20% improvement in perplexity on the *Wall Street Journal* and dictation corpora provided by ARPA's HLT program Paul (1991).

It should be noted here that the use of distributionally similar words to estimate unseen bigram probabilities will likely lead to erroneous results in case of less-preferred and strongly-

preferred collocations (word pairs). Inkpen and Hirst (2002) point out that even though words like *task* and *job* are semantically very similar, the collocations they form with other words may have varying degrees of usage. While *daunting task* is a strongly-preferred collocation, *daunting job* is rarely used. Thus using the probability of one bigram to estimate that of another will not be beneficial in such cases.

α-skew divergence

The **α-skew divergence** (*ASD*) is a slight modification of the Kullback-Leibler divergence that obviates the need for smoothed probabilities. It has the following formula:

$$ASD(w_1, w_2) = \sum_{w \in C(w_1) \cup C(w_2)} P(w|w_1) \log \frac{P(w|w_1)}{\alpha P(w|w_2) + (1-\alpha) P(w|w_1)} \tag{2.16}$$

where α is a parameter that may be varied but is usually set to 0.99. Note that the denominator within the logarithm is never zero with a non-zero numerator. Also, the measure retains the asymmetric nature of the Kullback-Leibler divergence. Lee (2001) shows that α-skew divergence performs better than Kullback-Leibler divergence in estimating word co-occurrence probabilities. Weeds (2003) achieves a correlation of 0.48 and 0.26 with human judgment on the Miller and Charles word pairs using $ASD(w_1, w_2)$ and $ASD(w_2, w_1)$, respectively.

Jensen-Shannon divergence

A relative entropy–based measure that overcomes the problem of asymmetry in Kullback-Leibler divergence is the **Jensen-Shannon divergence** a.k.a. **total divergence to the average**

a.k.a. **information radius**. It is denoted by **JSD** and has the following formula:

$$JSD(w_1,w_2) = D\left(d_1\|\frac{1}{2}(d_1+d_2)\right)+D\left(d_2\|\frac{1}{2}(d_1+d_2)\right) \quad (2.17)$$

$$= \sum_{w\in C(w_1)\cup C(w_2)}\left(P(w|w_1)\log\frac{P(w|w_1)}{\frac{1}{2}(P(w|w_1)+P(w|w_2))}+ P(w|w_2)\log\frac{P(w|w_2)}{\frac{1}{2}(P(w|w_1)+P(w|w_2))}\right) \quad (2.18)$$

The Jensen-Shannon divergence is the sum of the Kullback-Leibler divergence between each of the individual co-occurrence distributions d_1 and d_2 of the target words with the average distribution ($\frac{d_1+d_2}{2}$). Further, it can be shown that the Jensen-Shannon divergence avoids the problem of zero denominator. The Jensen-Shannon divergence is therefore always well defined and, like α-skew divergence, obviates the need for smoothed estimates.

The Kullback-Leibler divergence, α-skew divergence, and Jensen-Shannon divergence all give distributional distance scores from 0 (synonymous) to infinity (unrelated).

2.3.2.4 Latent Semantic Analysis

Latent semantic analysis (LSA) (Landauer et al., 1998) can be used to determine distributional distance between words or between sets of words.[4] Unlike the various approaches described earlier where a word–word co-occurrence matrix is created, the first step of LSA involves the creation of a word–paragraph, word–document, or similar such word-passage matrix, where a *passage* is some grouping of words. A cell for word *w* and passage *p* is populated with the number of times *w* occurs in *p* or, for even better results, a function of this frequency that captures how much information the occurrence of the word in a text passage carries.

Next, the dimensionality of this matrix is reduced by applying **singular value decomposition (SVD)**, a standard matrix decomposition technique. This smaller set of dimensions represent abstract (unknown) concepts. Then the original word–passage matrix is recreated,

[4]Landauer et al. (1998) describe it as a measure of *similarity*, but in fact it is a distributional measure that mimics semantic relatedness.

but this time from the reduced dimensions. Landauer et al. (1998) point out that this results in new matrix cell values that are different from what they were before. More specifically, words that are expected to occur more often in a passage than what the original cell values reflect, are incremented. Then a standard vector distance measure, such as cosine, that captures the distance between distributions of the two target words is applied.

LSA was used by Schütze and Pedersen (1997) and Rapp (2003) to measure distributional distance, with encouraging results. However, there is no non-heuristic way to determine when the dimension reduction should stop. Further, the generic concepts represented by the reduced dimensions are not interpretable; that is, one cannot determine which concepts they represent in a given sense inventory. This means that LSA cannot directly be used for tasks such as unsupervised sense disambiguation or estimating semantic similarity of known concepts. Finally, it has two of the biggest problems that plague all distributional word-distance measures—conflation of word senses and computational complexity. More about these and other limitations of distributional and WordNet-based measures is given in Section 1.2 ahead.

2.3.3 The anatomy of a distributional measure

Even though there are numerous distributional measures, many of which may seem dramatically different from each other, all distributional measures perform two functions: (1) create **distributional profiles (DPs)**, and (2) calculate the distance between two DPs.

The distributional profile of a word is the strength of association between it and each of the lexical, syntactic, and/or semantic units that co-occur with it. Commonly used **measures of strength of association** are conditional probability (0 to 1) and pointwise mutual information ($-\infty$ to ∞). Commonly used units of co-occurrence with the target are other *words*, and so we speak of the **lexical distributional profile of a word (lexical DPW)**. The co-occurring words may be all those in a predetermined window around the target, or may be restricted to those that have a certain syntactic (*e.g.,* verb–object) or semantic (*e.g.,* agent–theme) relation with the target word. We will refer to the former kind of DPs as **relation-free**. Usually in the latter

case, separate association values are calculated for each of the different relations between the target and the co-occurring units. We will refer to such DPs as **relation-constrained**. Typical relation-free DPs are those of Schütze and Pedersen (1997) and Yoshida et al. (2003). Typical relation-constrained DPs are those of Lin (1998a) and Lee (2001). Below are contrived, but plausible, examples of each for the word *pulse*; the numbers are conditional probabilities:

relation-free DP

pulse: *beat* .28, *racing* .2, *grow* .13, *beans* .09, *heart* .04, ...

relation-constrained DP

pulse: <*beat*, subject–verb> .34, <*racing*, noun–qualifying adjective> .22, <*grow*, subject–verb> .14, ...

Since the DPs represent the contexts of the two target words, the distance between the DPs is the distributional distance and, as per the distributional hypothesis, a proxy for semantic distance. A **measure of DP distance**, such as cosine, calculates the distance between two distributional profiles. While any of the measures of DP distance may be used with any of the measures of strength of association, in practice only certain combinations are used (see Table 2.2) and certain other combinations may not be meaningful, for example, Kullback-Leibler divergence with ϕ coefficient. Observe from Table 2.2 that all standard-combination distributional measures (or at least those that are described in this chapter) use either conditional probability or PMI as the measure of association.[5]

In this thesis, I show how distributional word-distance measures can be used to estimate *concept-distance*. All experiments will use standard combinations of measure of DP distance and measure of association. Therefore, to avoid clutter, instead of referring to a distributional measure by its measure of DP distance and measure of association (for example, α-skew divergence—conditional probability), I will refer to it simply by the measure of DP distance (in this case, α-skew divergence).

[5]Sense dominance experiments in Chapter 5 use all measures of strength of association listed in table 2.2.

Table 2.2: Measures of DP distance, measures of strength of association, and standard combinations. Measures of DP distance that are part of experiments in this thesis as well as the measures of strength of association that they are traditionally used in combination with, are marked in bold.

Measures of DP distance
α-skew divergence (ASD)
cosine (Cos)
Dice coefficient (Dice)
Euclidean distance (L_2 norm)
Hindle's measure (Hin)
Kullback-Leibler divergence (KLD)
Manhattan distance (L_1 norm)
Jensen–Shannon divergence (JSD)
Lin's measure (Lin)

Measures of strength of association
ϕ coefficient (Phi)
conditional probability (CP)
cosine (Cos)
Dice coefficient (Dice)
odds ratio (Odds)
pointwise mutual information (PMI)
Yule's coefficient (Yule)

Standard combinations
α-skew divergence—ϕ coefficient (ASD–CP)
cosine—conditional probability (Cos–CP)
Dice coefficient—conditional probability (Dice–CP)
Euclidean distance—conditional probability (L_2 norm–CP)
Hindle's measure—pointwise mutual information (Hin–PMI)
Kullback-Leibler divergence—conditional probability (KLD–CP)
Manhattan distance—conditional probability (L_1 norm–CP)
Jensen–Shannon divergence—conditional probability (JSD–CP)
Lin's measure—pointwise mutual information (Lin–PMI)

2.4 Other semantic distance work

Apart from the work described so far, which aims at estimating semantic distance between pairs of concepts and pairs of words, there is a large amount of work that focusses on estimating semantic distance between larger units of language. Tsang and Stevenson (2004, 2006) propose ways to determine fine-grained semantic distance between two texts, each of which is represented by a set of words weighted by their frequency of occurrence in text. They map the words to WordNet's is-a hierarchy and use graph-theoretic approaches to find the distance between two concept distributions. In paraphrasing (Barzilay and Lee, 2003; Schilder and Thomson McInnnes, 2006), machine translation (see Lopez (2007) and Knight and Marcu (2005) for surveys), text summarization (Gurevych and Strube, 2004; Zhu and Penn, 2005), and others, the aim is to estimate the distance between two phrases or sentences; in information retrieval (Varelas et al., 2005), to estimate the distance between a word (or a few words) and a document; in text clustering (see Steinbach et al. (2000) for survey), authorship attribution (Feiguina and Hirst, 2007), and others, to estimate the distance between two documents; and in determining selectional preferences (Resnik, 1996), detecting verb argument alternations (McCarthy, 2000; Tsang and Stevenson, 2004), and others, the goal is to estimate the distance between two word–frequency pair sets. Some of the above algorithms explicitly use the semantic distance between a pair of words (or concepts) as the starting point (see Table 1.2 in Chapter 1 for examples), while others implicitly do so by utilizing networks of semantically related concepts and/or co-occurrence information from text.

There is also work on estimating the strength of specific semantic relations between concept pairs—recall that semantic relatedness is a function of closeness as per each of the semantic relations between the target pair and that semantic similarity is a function of closeness as per synonymy, hypernymy/hyponymy, and antonymy. See Mirkin et al. (2007) for work on lexical entailment, Lucero et al. (2004) for detecting antonyms, and Lin et al. (2003) for detecting synonyms.

The vastness of literature pertaining to the tasks mentioned in this sub-section precludes

their discussion in this thesis. However, a lion's share of the future work (see Section 8.5) will be the use of ideas proposed in this thesis to both determine semantic distance between larger units of language and to estimate specific lexical semantic relations such as antonymy.

Chapter 3

Distributional Measures of Concept-Distance

3.1 A very coarse concept inventory

In this chapter, I will propose a new distributional concept-distance approach that combines corpus statistics with a published thesaurus to overcome, with varying degrees of success, many of the limitations of earlier approaches. The categories in the thesaurus are used as very coarse senses or concepts; most published thesauri have around a thousand categories. This allows investigating the impact of choosing a coarse concept inventory—an area not explored by other approaches, which tend to use the relatively much more fine-grained WordNet (with more than 100,000 senses). Further, it means that pre-computing a complete concept–concept distance matrix now involves the creation of a matrix approximately only 1000×1000 in size (much smaller and roughly .01% the size of matrices required by existing measures). This makes the new approach computationally less expensive and the storage requirements easy to meet.

3.2 The distributional hypothesis for concepts

As discussed earlier in Sections 1.3.1 and 1.2.3.1, the central limitation of using the distributional hypothesis[1] to estimate semantic distance is the conflation of word senses. While in most cases, the semantic distance between two concepts (or between the closest senses of two words) is required, distributional measures of word-distance give some sort of a dominance-based average of relevant sense-pairs. Further, words when used in different senses tend to keep different "company" (co-occurring words). For example, consider the contrived but plausible distributional profile of *star*:

> ***star***: *space* 0.21, *movie* 0.16, *famous* 0.15, *light* 0.12, *constellation* 0.11, *heat* 0.08, *rich* 0.07, *hydrogen* 0.07, ...

Observe that it has words that co-occur both with *star*'s CELESTIAL BODY sense and *star*'s CELEBRITY sense. Thus, it is clear that different senses of a word will probably have very different distributional profiles. Using a single DP for the word will mean the union of those profiles. While this might be useful for certain applications, this thesis will argue that in a number of tasks (including estimating semantic distance), acquiring different DPs for the different senses is not only more intuitive, but also, as I will show through numerous experiments, more useful. In other words, **distributional profiles of senses or concepts (DPCs)** can be used to infer semantic properties of the senses:

> You know a *concept* by the company it keeps.

Therefore, I propose profiling the co-occurrence distributions of word senses or concepts, rather than those of words, to determine distributional distance between concepts, rather than the distance between words. The closer the distributional profiles of two concepts, the smaller is their semantic distance. Below are example distributional profiles of two senses of STAR:

> CELESTIAL BODY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...

[1] Recall that the distributional hypothesis states, "You know a word by the company it keeps".

CELEBRITY: *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, ...

The values are the strength of association (usually pointwise mutual information or conditional probability) of the target concept with co-occurring words. It should be noted that creating such distributional profiles of concepts is much more challenging than creating distributional profiles of words which involve simple word–word co-occurrence counts. (In the next section, I show how these profiles may be estimated without the use of sense-annotated data). However, once created, any of the many distributional measures can be used to estimate the distance between the DPs of two target concepts (just as in the case of traditional word-distance measures, distributional measures are used to estimate the distance between the DPs of two target words). For example, here is how cosine is traditionally used to estimate distributional distance between two words (as described in Section 2.3.2.1 earlier):

$$Cos(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) \times P(w|w_2))}{\sqrt{\sum_{w \in C(w_1)} P(w|w_1)^2} \times \sqrt{\sum_{w \in C(w_2)} P(w|w_2)^2}} \tag{3.1}$$

$C(t)$ is the set of words that co-occur (within a certain window) with the word t in a corpus. The conditional probabilities in the formula are taken from the distributional profiles of words. We can adapt the formula to estimate distributional distance between two concepts as shown below:

$$Cos_{cp}(c_1, c_2) = \frac{\sum_{w \in C(c_1) \cup C(c_2)} (P(w|c_1) \times P(w|c_2))}{\sqrt{\sum_{w \in C(c_1)} P(w|c_1)^2} \times \sqrt{\sum_{w \in C(c_2)} P(w|c_2)^2}} \tag{3.2}$$

$C(x)$ is now the set of words that co-occur with *concept x* within a pre-determined window. The conditional probabilities in the formula are taken from the distributional profiles of concepts.

With the new approach, if the distance between two words is required, then the distance between all relevant sense pairs is determined and the minimum is chosen. For example, if *star* has the two senses mentioned above and *fusion* has one (let's call it FUSION), then the distance between them is determined by first applying cosine (or any distributional measure) to the DPs of CELESTIAL BODY and FUSION:

CELESTIAL BODY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...

FUSION: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

then applying cosine to the DPs of CELEBRITY and FUSION:

CELEBRITY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...

FUSION: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

and finally choosing the one with minimum semantic distance, that is, maximum similarity/relatedness:

$$\textit{distance}(\textit{star}, \textit{fusion}) = \max(\textit{Cos}(\text{CELEBRITY}, \text{FUSION}), \textit{Cos}(\text{CELESTIAL BODY}, \text{FUSION})) \tag{3.3}$$

Note that the maximum value is chosen above because cosine is a similarity/relatedness measure. In case of distance measures, such as α-skew divergence, the lower of the two values will be chosen.

3.3 Estimating distributional profiles of concepts

Determining distributional profiles of *words* simply involves making word–word co-occurrence counts in a corpus. Determining distributional profiles of *concepts*, on the other hand, requires information about which words co-occur with which concepts. This means that a direct approach requires the text, from which counts are made, to be sense annotated. Since existing labeled data is minimal and manual annotation is far too expensive, indirect means must be used. I now present a way to estimate distributional profiles of concepts from raw text, using a published thesaurus (the concept inventory) and a bootstrapping algorithm.

3.3.1 Creating a word–category co-occurrence matrix

A **word–category co-occurrence matrix (WCCM)** is created having word types w as one dimension and thesaurus categories c as another.

	c_1	c_2	...	c_j	...
w_1	m_{11}	m_{12}	...	m_{1j}	...
w_2	m_{21}	m_{22}	...	m_{2j}	...
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
w_i	m_{i1}	m_{i2}	...	m_{ij}	...
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\ddots$

The matrix is populated with co-occurrence counts from a large corpus. A particular cell m_{ij}, corresponding to word w_i and category or concept c_j, is populated with the number of times w_i co-occurs (in a window of ± 5 words) with any word that has c_j as one of its senses (i.e., w_i co-occurs with any word listed under concept c_j in the thesaurus). For example, assume that the concept of CELESTIAL BODY is represented by four words in the thesaurus: *constellation, planet*, *star* and *sun*. If the word *space* co-occurs with *constellation* (15 times), *planet* (50 times), *star* (40 times), and *sun* (65 times) in the given text corpus, then the cell for *space* and CELESTIAL BODY in the WCCM is populated with 170 ($15+50+40+65$). This matrix, created after a first pass of the corpus, is called the **base word–category co-occurrence matrix (base WCCM)**.

The choice of ± 5 words as window size is somewhat arbitrary and hinges on the intuition that, in text and speech, words close to a target word are more indicative of its semantic properties than those more distant. Church and Hanks (1990), in their seminal work on word–word co-occurrence association, also use a window size of ± 5 words and argue that this size is large enough to capture many verb–argument dependencies and yet small enough so that adjacency information is not diluted too much. In the word sense dominance experiments (described ahead in Chapter 5 and through which the WCCM was first evaluated), using the whole sentence as context resulted in a lower accuracy than when using the ± 5 word window. While, it is

reasonable to determine optimal window sizes for different applications from held out datasets, we decided to use a fixed window size for all our experiments because for most of the tasks there were only limited amounts of gold standard evaluation data.

A contingency table for any particular word *w* and category *c* can be easily generated from the WCCM by collapsing cells for all other words and categories into one and summing up their frequencies.

	c	$\neg c$
w	n_{wc}	$n_{w\neg}$
$\neg w$	$n_{\neg c}$	$n_{\neg\neg}$

The application of a suitable statistic, such as pointwise mutual information or conditional probability, will then yield the strength of association between the word and the category.

As the base WCCM is created from unannotated text, it will be noisy. For example, out of the 40 times *star* co-occurs with *space*, 25 times it may have been used in the CELESTIAL BODY sense and 15 times in the CELEBRITY sense. However, since this information was not known to the system, the cell for *space*—CELESTIAL BODY in the base WCCM was incremented by 40 rather than 25. Similarly, the cell for *space*—CELEBRITY was also incremented by 40 rather than 15. That said, the base WCCM does capture strong word–category co-occurrence associations reasonably accurately. This is because the errors in determining the true category that a word co-occurs with will be distributed thinly across a number of other categories. For example, even though we increment counts for both *space*–CELESTIAL BODY and *space*–CELEBRITY for a particular instance where *space* co-occurs with *star*, *space* will co-occur with a number of words such as *planet*, *sun*, and *constellation* that each have the sense of *celestial body* in common (Figure 3.1), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count, and thereby strength of association, of *space* and CELESTIAL BODY will be relatively higher than that of *space* and CELEBRITY (Figure 3.2). For more details, see discussion of the general principle in Resnik (1998).

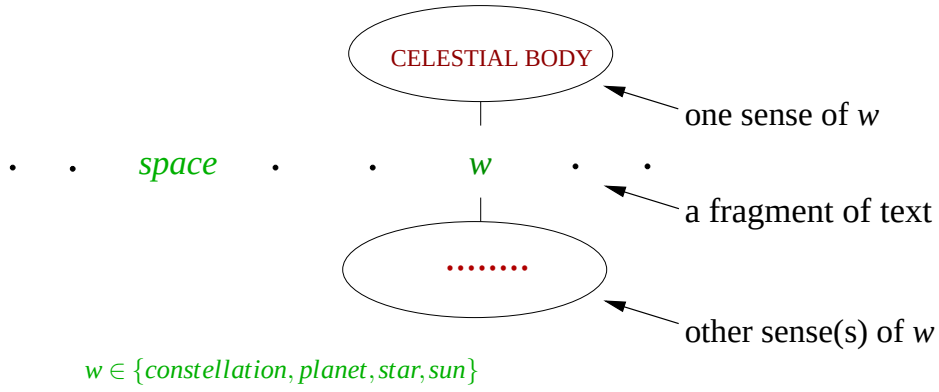


Figure 3.1: The word *space* will co-occur with a number of words X that each have one sense of CELESTIAL BODY in common.

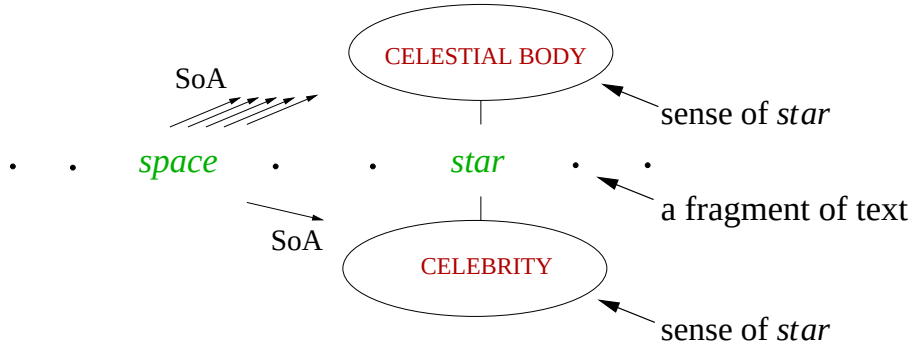


Figure 3.2: The base WCCM captures strong word–category co-occurrence associations.

3.3.2 Bootstrapping

I now describe a bootstrapping procedure which can be used to reduce, even more, the errors in the WCCM due to word sense ambiguity. Words that occur close to a target word tend to be good indicators of its intended sense. Therefore, a second pass of the corpus is made and the base WCCM is used to roughly disambiguate the words in it. Each word in the corpus is considered as the target one at a time. For each sense of the target, its strength of association with each of the words in its context (± 5 words) is summed. The sense that has the highest cumulative association with co-occurring words is chosen as the intended sense of the target word. In this second pass, a new **bootstrapped WCCM** is created such that each cell m_{ij}, cor-

responding to word w_i and concept c_j, is populated with the number of times w_i co-occurs with any word *used in sense* c_j. For example, consider again the 40 times *star* co-occurs with *space*. If the contexts of 25 of these instances have higher cumulative strength of association with CELESTIAL BODY than CELEBRITY, suggesting that in only these 25 of those 40 occurrences *star* was used in CELESTIAL BODY sense, then the cell for *space*–CELESTIAL BODY is incremented by 25 rather than 40 (as was the case in the base WCCM). This bootstrapped WCCM, created after simple and fast[2] word sense disambiguation, is expected to better capture word–concept co-occurrence values, and hence strengths of association values, than the base WCCM. The base and bootstrapped WCCMs were first evaluated through word sense dominance experiments (described ahead in Chapter 5); the bootstrapped WCCM gave markedly better results. Further iterations of the bootstrapping procedure did not, however, improve results. This is not surprising because the base WCCM was created without any word sense disambiguation and so the first bootstrapping iteration with word sense disambiguation is expected to markedly improve the matrix. The same is not true for subsequent iterations. Therefore, all other experiments that use a word–concept co-occurrence matrix, including the ones described ahead in this chapter, use the bootstrapped matrix (created after one bootstrapping iteration over the base WCCM).

3.3.3 Mimicking semantic relatedness and semantic similarity

The distributional profiles created by the above methodology are relation-free. This is because (1) all co-occurring words (not just those that are related to the target by certain syntactic relations) are used, and (2) the WCCM, as described above in Sections 3.3.1 and 3.3.2, does not maintain separate counts for the different syntactic relations between the target and co-occurring words. Thus, distributional measures that use these WCCMs will estimate semantic *relatedness* between concepts. Distributional measures that mimic semantic *similarity*, which

[2]Speed of disambiguation is important here as all words in the corpus are to be disambiguated. After determining co-occurrence counts from the BNC (a 100 million word corpus), creating the bootstrapped WCCM from the base WCCM took only about 4 hours on a 1.3GHz machine with 16GB memory.

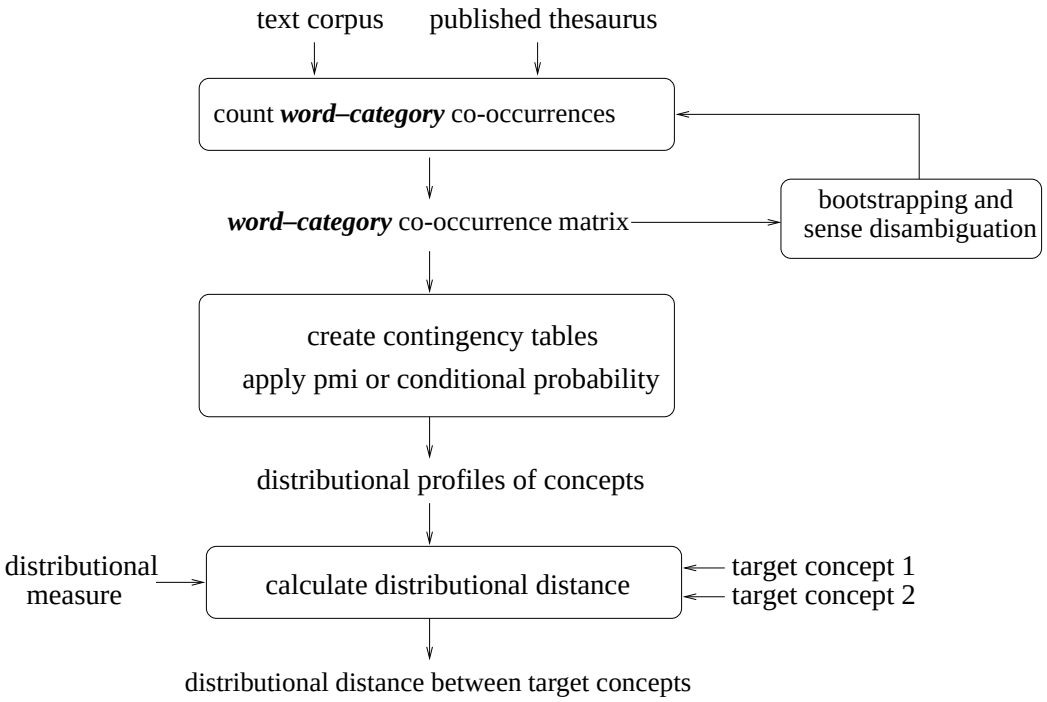


Figure 3.3: An overview of the new distributional concept-distance approach.

require relation-constrained DPCs, can easily be created from WCCMs that have rows for each word–syntactic relation pair (rather than just words). (See earlier Sections 2.3.1 and 2.3.3 for more discussion of the motivating principles.) However, in this thesis, all experiments are conducted using distributional measures that estimate semantic relatedness mainly because of time constraints, because relatedness subsumes similarity, and because there is a need for good relatedness measures (WordNet-based measures are especially poor at estimating semantic relatedness).

3.4 An overview of the evaluation

I evaluate the newly proposed distributional concept-distance approach (Figure 3.3) through its usefulness in various natural language tasks. Tasks or applications that use distance values can be classified as what I will call concept-distance tasks and word-distance tasks. **Concept-distance tasks** require distance between explicit senses or concepts (from a sense inventory)

and other units of text. For example, unsupervised word sense disambiguation (Lin, 1997) requires the semantic distance between a word's sense and its context. These tasks must be solved with the sense inventory associated with the task. For example, if the task is to do unsupervised sense disambiguation on data annotated with senses from WordNet, then either a WordNet-based approach must be used or the senses from another knowledge source (for example, the thesaurus) must be mapped to WordNet. **Word-distance tasks**, *seemingly* at least, require distances between words. For example, malapropism correction (Hirst and Budanitsky, 2005) requires the semantic distance between the spelling variants of the target word and its context. However, semantic distance is essentially a property of word senses or concepts and not words (see discussion in the introduction of this thesis—Section 1.1); even though it seems as if the word-distance task involves only words, what is really needed is the distance between the intended senses of those words, which tends to be the distance between their closest senses. These tasks may be independently solved with different sense inventories or even without using any sense inventory.[3] Concept-distance tasks can be attempted with WordNet-based measures of concept-distance or distributional measures of concept-distance, but not the traditional distributional measures of word-distance. Word-distance tasks can be attempted with any of the three types of measures.

In the following section, I will describe experiments using all three kinds of distance measures to solve two word-distance tasks. In the next chapter, I will show how distributional concept-distance measures can be used to estimate semantic distance in one language using a knowledge source from another. I will evaluate this cross-lingual approach on another pair of word-distance tasks. In Chapters 5 and 6, I will show the newly proposed approach can be used in the concept-distance tasks of word sense dominance and word sense disambiguation. In Chapter 7, I will show how my cross-lingual semantic distance approach not only overcomes the knowledge source bottleneck (Chapter 4) but is also useful in tasks that inherently involve

[3]Traditional distributional word-distance measures do not require any sense inventory; however, as a consequence, they conflate the many senses of a word and give a dominance-based average semantic distance of the relevant sense pairs.

two or more languages, such as machine translation.

3.5 Evaluation: monolingual, word-distance tasks

In this section, I describe experiments that evaluated the distributional concept-distance measures on two monolingual word-distance tasks: ranking word pairs in order of their semantic distance and correcting real-word spelling errors. Each task will be described in the subsections below. I will compare the new approach with state-of-the-art distributional word-distance measures.

The distributional profiles of concepts were created from the *British National Corpus (BNC)* and the *Macquarie Thesaurus*. 22.85% of the $98,000 \times 812$ cells in the base WCCM had non-zero values whereas the statistic in the bootstrapped WCCM was 9.1%.[4] The word-distance measures used a word–word co-occurrence matrix created from the *BNC* alone. The *BNC* was not lemmatized, part-of-speech tagged, nor chunked. The vocabulary was restricted to the words present in the thesaurus (about 98,000 word types) both to provide a level evaluation platform and to filter out named entities and tokens that are not actually words (for example, the BNC has *Hahahahahahahahaaaaa, perampam,* and *Owzeeeyaaaah*). Also, in order to overcome large computation times of distributional word-distance measures, co-occurrence counts less than five were reset to zero, and words that co-occurred with more than 2000 other words were stoplisted (543 in all). This resulted in a word–word co-occurrence matrix having non-zero values in 0.02% of its $98,000 \times 98,000$ cells.

I used α-skew divergence (ASD) ($\alpha = 0.99$), cosine (Cos), Jensen–Shannon divergence (JSD), and Lin's distributional measure (Lin_{dist})[5] to populate corresponding concept–concept distance matrices and word–word distance matrices. While it is easy to completely pre-compute the concept–concept distance matrix (due to its small size), completely populating the word–

[4]The *Macquarie Thesaurus* has 98,000 word types and 812 categories.

[5]Although Lin (1998a) used relation-constrained DPs, in these experiments all DPs are relation-free.

word distance matrix is non-trivial because of memory and time constraints. Therefore, the word–word distance matrix was populated on the fly and only to the extent necessary.

The same distributional measures will be used to solve the word-pair ranking and spelling correction tasks in two different ways: first by calculating word-distance, and then by calculating concept-distance. This allows for an even-keeled comparison of the two approaches. However, comparison with WordNet-based measures is not so straightforward. Both of the above-mentioned semantic distance tasks have traditionally been performed using WordNet-based measures—which are good at estimating semantic similarity between nouns but particularly poor at estimating semantic relatedness between concept pairs other than noun–noun. This has resulted in the creation of "gold-standard" data only for nouns. As creating new gold-standard data is arduous, we perform experiments on existing noun data. Of course, even though it is a given that WordNet-based measures are significantly less applicable than the proposed new approach, it will be interesting to determine how competitive the new approach is on concept-pairs for which WordNet-based measures can be used and perform best on.

3.5.1 Ranking word pairs

A direct approach to evaluate semantic distance measures is to determine how close they are to human judgment and intuition. Given a set of word-pairs, humans can rank them in order of their distance—placing near-synonyms on one end of the ranking and unrelated pairs on the other. Rubenstein and Goodenough (1965a) provide a "gold-standard" list of 65 human-ranked word-pairs (based on the responses of 51 subjects). An automatic distance measure is deemed to be more accurate than another if its ranking of word-pairs correlates more closely with the human ranking. Measures of concept-distance can determine distance between each word-pair by first finding the concept-distance between all pairs of senses of the two words, and then choosing the shortest distance. This is based on the assumption that when humans are asked to judge the semantic distance between a pair of words, they implicitly consider its closest senses. For example, most people will agree that *bank* and *interest* are semantically

related, even though both have multiple senses—most of which are unrelated. Alternatively, a concept-distance method can take the average of the distance between each of the relevant pairs of senses.

Table 3.1 lists correlations of human rankings with those created using the word–word co-occurrence matrix–based traditional distributional word-distance measures and the correlations using the newly proposed word–concept co-occurrence matrix–based distributional concept-distance measures. Observe that the distributional concept-distance measures give markedly higher correlation values than distributional word-distance measures. (Figure 3.4 depicts the results in a graph.) Also, using the distance of the closest sense pair (for Cos and Lin_{dist}) gives much better results than using the average distance of all relevant sense pairs. (We do not report average distance for ASD and JSD because they give very large distance values when sense-pairs are unrelated—values that dominate the averages, overwhelming the others, and making the results meaningless.) These correlations are, however, notably lower than those obtained by the best WordNet-based measures (not shown in the table), which fall in the range .78 to .84 (Budanitsky and Hirst, 2006).

3.5.2 Correcting real-word spelling errors

The set of Rubenstein and Goodenough word pairs is much too small to safely assume that measures that work well on them do so for the entire English vocabulary. Consequently, semantic measures have traditionally been evaluated through more extensive applications such as the work by Hirst and Budanitsky (2005) on correcting **real-word spelling errors** (or **malapropisms**). If a word in a text is not semantically close to any other word in its context, then it is considered a **suspect**. If the suspect has a spelling-variant that *is* semantically close to a word in its context, then the suspect is declared a probable real-word spelling error and an **alarm** is raised; the semantically close spelling-variant is considered its **correction**. Hirst and Budanitsky tested the method on 500 articles from the 1987–89 *Wall Street Journal* corpus for their experiments, replacing one noun in every 200th word by a spelling-variant

Table 3.1: Correlations with human ranking of Rubenstein and Goodenough word pairs of automatic rankings using traditional word–word co-occurrence–based distributional word-distance measures and the newly proposed word–concept co-occurrence–based distributional concept-distance measures. Best results for each measure-type are shown in boldface.

	Measure-type		
	Word-distance	Concept-distance	
Distributional measure		closest	average
α-skew divergence	0.45	0.60	–
cosine	**0.54**	0.69	0.42
Jensen–Shannon divergence	0.48	0.61	–
Lin's distributional measure	0.52	**0.71**	**0.59**

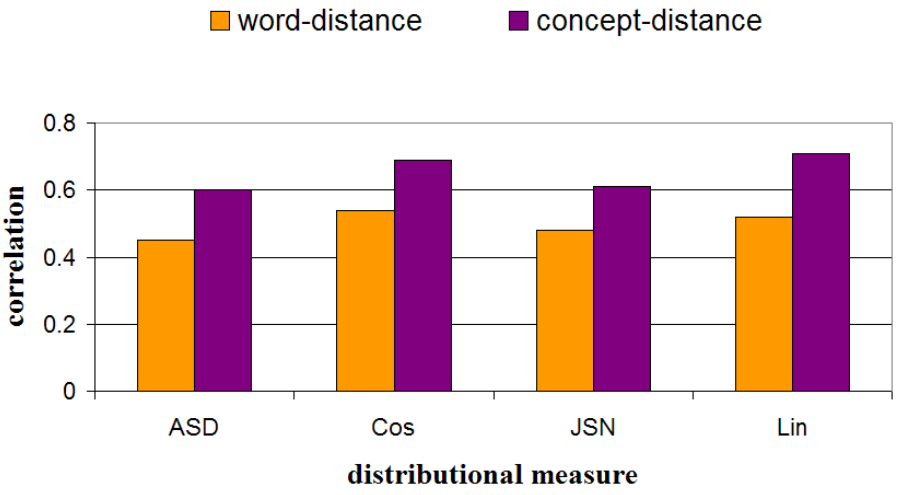


Figure 3.4: Correlations with human ranking of Rubenstein and Goodenough word pairs of automatic rankings using traditional word–word co-occurrence–based distributional word-distance measures and the newly proposed word–concept co-occurrence–based distributional concept-distance measures. Best results for each measure-type are shown in boldface.

and looking at whether the method could restore the original word. This resulted in text with 1408 real-word spelling errors out of a total of 107,233 noun tokens. I adopt this method and this test data, but whereas Hirst and Budanitsky used WordNet-based semantic measures, I use distributional concept- and word-distance measures.

In order to determine whether two words are "semantically close" or not as per any measure of distance, a **threshold** must be set. If the distance between two words is less than the threshold, then they will be considered **semantically close**. Hirst and Budanitsky (2005) pointed out that there is a notably wide band in the human ratings of the Rubenstein and Goodenough word pairs such that no word-pair was assigned a distance value between 1.83 and 2.36 (on a scale of 0–4). They argue that somewhere within this band is a suitable threshold between semantically close and semantically distant, and therefore set thresholds for the WordNet-based measures such that there was maximum overlap in what the automatic measures and human judgments considered semantically close and distant. Following this idea, I use an automatic method to determine thresholds for the various distributional concept- and word-distance measures. Given a list of Rubenstein and Goodenough word pairs ordered according to a distance measure, I repeatedly consider the mean of all adjacent distance values as **candidate thresholds**. Then I determine the number of word-pairs correctly classified as semantically close or semantically distant for each candidate threshold, considering which side of the band they lie as per human judgments. The candidate threshold with highest accuracy is chosen as the threshold.

I follow the Hirst and St-Onge (1998) metrics to evaluate real-word spelling correction. **Suspect ratio** and **alarm ratio** evaluate the processes of identifying suspects and raising alarms, respectively.

$$\textit{suspect ratio} = \frac{\frac{\text{number of true-suspects}}{\text{number of malapropisms}}}{\frac{\text{number of false-suspects}}{\text{number of non-malapropisms}}} \tag{3.4}$$

$$\textit{alarm ratio} = \frac{\frac{\text{number of true-alarms}}{\text{number of true-suspects}}}{\frac{\text{number of false-alarms}}{\text{number of false-suspects}}} \tag{3.5}$$

Detection ratio is the product of the two, and measures overall performance in detecting the

errors.

$$\textit{detection ratio} = \frac{\frac{\text{number of true-alarms}}{\text{number of malapropisms}}}{\frac{\text{number of false-alarms}}{\text{number of non-malapropisms}}} \tag{3.6}$$

Correction ratio indicates overall correction performance, and is the "bottom-line" statistic.

$$\textit{correction ratio} = \frac{\frac{\text{number of corrected malapropisms}}{\text{number of malapropisms}}}{\frac{\text{number of false-alarms}}{\text{number of non-malapropisms}}} \tag{3.7}$$

Values greater than 1 for each of these ratios indicate results better than random guessing. The ability of the system to determine the intended word, given that it has correctly detected an error, is indicated by the **correction accuracy** (0 to 1).

$$\textit{correction accuracy} = \frac{\text{number of corrected malapropisms}}{\text{number of true-alarms}} \tag{3.8}$$

Notice that the correction ratio is the product of the detection ratio and correction accuracy. The overall (single-point) precision (P), recall(R), and F-score (F) of detection are also computed.

$$P = \frac{\text{number of true-alarms}}{\text{number of alarms}} \tag{3.9}$$

$$R = \frac{\text{number of true-alarms}}{\text{number of malapropisms}} \tag{3.10}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3.11}$$

The product of detection F-score and correction accuracy, which we will call **correction performance**, can also be used as a bottom-line performance metric.

Table 3.2 details the performance of distributional word- and concept-distance measures. For comparison, the table also lists results obtained by Hirst and Budanitsky (2005) using WordNet-based concept-distance measures: Hirst and St-Onge (1998), Jiang and Conrath (1997), Leacock and Chodorow (1998), Lin (1997), and Resnik (1995). These information content measures rely on finding the lowest common subsumer (lcs) of the target synsets in WordNet's hypernym hierarchy and use corpus counts to determine how specific or general this concept is. The more specific the lcs is and the smaller the difference of its specificity with that of the target concepts, the closer the target concepts are considered. (See Section 2.2.1 for more details.)

Table 3.2: Real-word spelling error correction. The best results as per the two bottom-line statistics are shown in boldface.

Measure	*suspect ratio*	*alarm ratio*	*detection ratio*	*correction accuracy*	***correction ratio***	*detection*			***correction performance***
						P	*R*	*F*	
Distributional$_{word}$									
α-skew divergence	3.36	1.78	5.98	0.84	5.03	7.37	45.53	12.69	10.66
cosine	2.91	1.64	4.77	0.85	4.06	5.97	37.15	10.28	8.74
Jensen–Shannon divergence	3.29	1.77	5.82	0.83	4.88	7.19	44.32	12.37	10.27
Lin's distributional measure	3.63	2.15	7.78	0.84	**6.52**	9.38	58.38	16.16	**13.57**
Distributional$_{concept}$									
α-skew divergence	4.11	2.54	10.43	0.91	**9.49**	12.19	25.28	16.44	**14.96**
cosine	4.00	2.51	10.03	0.90	9.05	11.77	26.99	16.38	14.74
Jensen–Shannon divergence	3.58	2.46	8.79	0.90	7.87	10.47	34.66	16.08	14.47
Lin's distributional measure	3.02	2.60	7.84	0.88	6.87	9.45	36.86	15.04	13.24
WNet$_{concept}$									
Hirst–St-Onge	4.24	1.95	8.27	0.93	7.70	9.67	26.33	14.15	13.16
Jiang–Conrath	4.73	2.97	14.02	0.92	**12.91**	14.33	46.22	21.88	**20.13**
Leacock–Chodrow	3.23	2.72	8.80	0.83	7.30	11.56	60.33	19.40	16.10
Lin's WordNet-based measure	3.57	2.71	9.70	0.87	8.48	9.56	51.56	16.13	14.03
Resnik	2.58	2.75	7.10	0.78	5.55	9.00	55.00	15.47	12.07

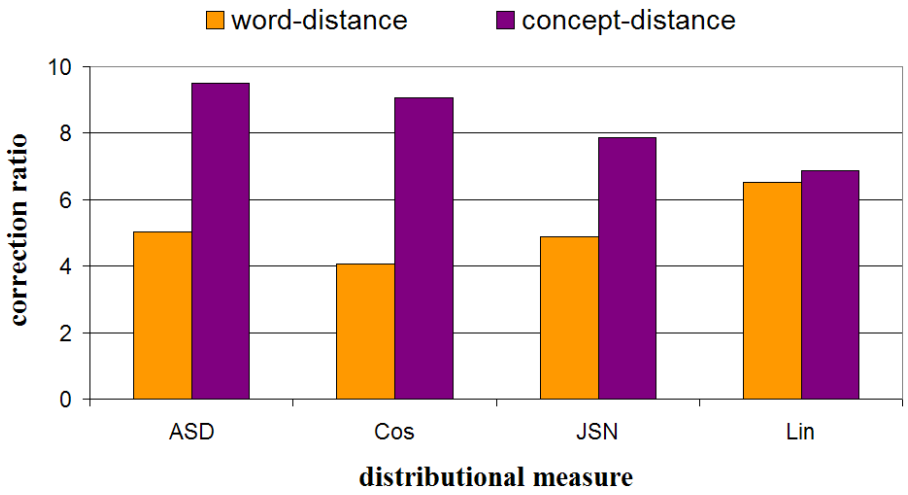


Figure 3.5: Correcting real-word spelling errors

Observe that the correction ratio results for the distributional word-distance measures are poor compared to distributional concept-distance measures; the concept-distance measures are clearly superior, in particular α-skew divergence and cosine. (Figure 3.5 depicts the results in a graph.) Moreover, if we consider correction ratio to be the bottom-line statistic, then three of the four distributional concept-distance measures outperform all WordNet-based measures except the Jiang–Conrath measure. If we consider correction performance to be the bottom-line statistic, then again we see that the distributional concept-distance measures outperform the word-distance measures, except in the case of Lin's distributional measure, which gives slightly poorer results with concept-distance. Also, in contrast to correction ratio values, using the Leacock–Chodorow measure results in relatively higher correction performance values than the best distributional concept-distance measures. While it is clear that the Leacock–Chodorow measure is relatively less accurate in choosing the right spelling-variant for an alarm (correction accuracy), detection ratio and detection F-score present contrary pictures of relative performance in detection. As the correction ratio is determined by the product of a number of ratios, each evaluating the various stages of malapropism correction (identifying suspects, raising alarms, and applying the correction), I believe it is a better indicator of overall performance than correction performance, which is a not-so-elegant product of an F-score and accu-

racy. However, no matter which of the two is chosen as the bottom-line performance statistic, the results show that the newly proposed distributional concept-distance measures are clearly superior to word-distance measures. Further, of all the WordNet-based measures, only that proposed by Jiang and Conrath outperforms the best distributional concept-distance measures consistently with respect to both bottom-line statistics.

3.6 Related work

Apart from the vast array of work on WordNet-based and distributional word-distance measures (summarized in Chapter 2), below is a brief description of work related specifically to that described in this chapter.

Yarowsky (1992) proposed a model for unsupervised word sense disambiguation using *Roget's Thesaurus*. A mutual information–like measure was used to identify words that best represent each category in the thesaurus, which he calls the **salient words**. The presence of a salient word in the context of a target word is evidence that the word is used in a sense corresponding to the salient word. The evidence is incorporated in a Bayesian model. The word-category co-occurrence matrix I created can be seen as a means of determining the degree of salience of any word co-occurring with a concept. I further improved the accuracy of the WCCM using simple bootstrapping techniques.

Pantel (2005) also provides a way to create co-occurrence vectors for WordNet senses. The lexical co-occurrence vectors of words in a leaf node are propagated up the WordNet hierarchy. A parent node inherits those co-occurrences that are shared by its children. Lastly, co-occurrences not pertaining to the leaf nodes are removed from its vector. Even though the methodology attempts at associating a WordNet node or sense with only those co-occurrences that pertain to it, no attempt is made at correcting the frequency counts. After all, *word1–word2* co-occurrence frequency (or association) is likely not the same as SENSE1–*word2* co-occurrence frequency (or association), simply because *word1* may have senses other than

SENSE1, as well. Further, in the Pantel (2005) system, the co-occurrence frequency associated with a parent node is the weighted sum of co-occurrence frequencies of its children. The frequencies of the child nodes are used as weights. Sense ambiguity issues apart, this is still problematic because a parent concept (say, BIRD) may co-occur much more frequently (or infrequently) with a word than its children. In contrast, the bootstrapped WCCM not only identifies which words co-occur with which concepts, but also has more accurate estimates of the co-occurrence frequencies.

Patwardhan and Pedersen (2006) create **aggregate co-occurrence vectors** for a WordNet sense by adding the co-occurrence vectors of the words in its WordNet gloss. The distance between two senses is then determined by the cosine of the angle between their aggregate vectors. However, such aggregate co-occurrence vectors are expected to be noisy because they are created from data that is not sense-annotated. The bootstrapping procedure introduced in Section 3.3.2 minimizes such errors and as I will show in Chapter 5 markedly improves accuracies of natural language tasks that use these co-occurrence vectors.

Véronis (2004) presents a graph theory–based approach to identify the various senses of a word in a text corpus without the use of a dictionary. For each target word, a graph of inter-connected nodes is created. Every word that co-occurs with the target word is a node. Two nodes are connected with an edge if they are found to co-occur with each other. Highly interconnected components of the graph represent the different senses of the target word. The node (word) with the most connections in a component is representative of that sense and its associations with words that occur in a test instance are used to quantify evidence that the target word is used in the corresponding sense. However, these strengths of association are at best only rough estimates of the associations between the sense and co-occurring words, since a sense in his system is represented by a single (possibly ambiguous) word.

3.7 Conclusion

I have proposed a framework that allows distributional measures to estimate concept-distance using a published thesaurus and raw text. These distributional concept-distance measures are more intuitive proxies for semantic measures than distributional word-distance measures. I evaluated them in comparison with traditional distributional word-distance measures and WordNet-based measures through their ability to rank word-pairs in order of their human-judged linguistic distance, and their ability to correct real-word spelling errors.

I showed that distributional concept-distance measures outperformed word-distance measures in both tasks. They do not perform as well as the best WordNet-based measures in ranking a small set of word pairs, but in the task of correcting real-word spelling errors, they beat all WordNet-based measures except for Jiang–Conrath (which is markedly better) and Leacock–Chodorow (which is slightly better if we consider correction performance as the bottom-line statistic, but slightly worse if we rely on correction ratio). It should be noted that the Rubenstein and Goodenough word-pairs used in the ranking task, as well as all the real-word spelling errors in the correction task, are nouns. We expect that the WordNet-based measures will perform poorly when other parts of speech are involved, as those hierarchies of WordNet are not as extensively developed. Further, the various hierarchies are not well connected, nor is it clear how to use these interconnections across parts of speech for calculating semantic distance. On the other hand, our DPC-based measures do not rely on any hierarchies (even if they exist in a thesaurus) but on sets of words that unambiguously represent each sense. Further, because our measures are tied closely to the corpus from which co-occurrence counts are made, we expect the use of domain-specific corpora to give even better results.

Both DPW- and WordNet-based measures have large space and time requirements for pre-computing and storing all possible distance values for a language. However, by using the categories of a thesaurus as very coarse concepts, pre-computing and storing all possible distance values for our DPC-based measures requires a matrix of size only 812×812. This level of concept-coarseness might seem drastic at first glance, but results show that distributional mea-

sures of distance between these coarse concepts are surprisingly accurate in natural language tasks. Part of future work is to try an intermediate degree of coarseness (still much coarser than WordNet) by using the paragraph subdivisions of the thesaurus instead of its categories to see if that gives even better results (see Future Directions Section 8.5) for more discussion.

This newly proposed distributional approach of concept-distance has all the attractive features of a distributional measure, and yet avoids problems of sense-conflation (limitation 1.2.3.1) and computationally complexity (limitation 1.2.1.1). As it calculates distance between coarse senses, each represented by many words, even if some words are not seen often in a text corpus, all concepts have sufficient representation even in small corpora, thereby avoiding the data sparseness problem (limitation 1.2.3.2). However, because this method uses a published thesaurus, the lack of high-quality knowledge sources in most languages (limitation 1.2.2.1) remains a problem. Also, the approach as proposed is still monolingual (limitation 1.2.1.2). The next chapter addresses both these issues by making the approach cross-lingual.

Chapter 4

Cross-lingual Semantic Distance[1]

4.1 The knowledge-source bottleneck

Accurately estimating semantic distance, as discussed earlier in Section 1.1 of Chapter 1, has pervasive applications in computational linguistics, including machine translation, information retrieval, speech recognition, spelling correction, and text categorization. However, applying algorithms for semantic distance to most languages is hindered by the lack of high-quality linguistic resources. WordNet-based measures of semantic distance, such as those of Jiang and Conrath (1997) and Resnik (1995), require a WordNet which does not exist for most languages. Distributional measures of word-distance, such as cosine and α-skew divergence, rely simply on raw text, but as I showed in the previous chapter, are much less accurate because they conflate the many senses of a word. Distributional measures of concept-distance combine written text with a published thesaurus to measure distance between *concepts* (or *word senses*) using distributional measures, such as cosine and α-skew divergence. They avoid sense conflation and achieve results better than the traditional word-distance measures and indeed also most

[1]This chapter describes work done in collaboration with Torsten Zesch and Iryna Gurevych of Darmstadt University of Technology. They played a pivotal role in the evaluation of the ideas presented here. They compiled the ”gold-standard” data for the *Reader's Digest* word choice task and the ranking of German word pairs in order of their semantic distance. They also provided baseline semantic distance values as per state-of-the-art GermaNet measures. I am grateful for their contributions and an enriching collaboration.

of the WordNet-based semantic measures (also shown in the previous chapter). Further, the distributional concept-distance measures are much more applicable than the WordNet-based measures, which are only good at estimating semantic similarity between noun pairs. However, the high-quality thesauri and (to a much greater extent) WordNet-like resources that these concept-distance methods require do not exist for most of the 3000–6000 languages in existence today and they are costly to create. While such linguistic resources are being created for English, Chinese, Spanish, Bengali, Hindi, and German—languages that enjoy a large number of speakers—others such as Pashto (Afghanistan), Kannada (South Indian), Greek, and Kazakh are largely ignored, let alone Swahili (African), Cherokee (native American), Guarani (indigenous South American), and such. This chapter proposes a way to overcome this knowledge bottleneck.

I introduce **cross-lingual distributional measures of concept-distance**, or simply **cross-lingual measures**, that determine the distance between a word pair in resource-poor language L_1 using a knowledge source in a resource-rich language L_2. An L_1–L_2 bilingual lexicon[2] will be used to map words in the resource-poor language to words in the resource-rich one. I will compare this approach with the best monolingual approaches, which usually require high-quality knowledge sources in the same language (L_1); the smaller the loss in performance, the more capable the cross-lingual algorithm is of overcoming ambiguities in word translation. An evaluation, therefore, requires an L_1 that in actuality has adequate knowledge sources. Therefore I chose German to stand in as the resource-poor language L_1; the monolingual evaluation in German will use GermaNet. I chose English as the resource-rich L_2; the cross-lingual evaluation will use the *Macquarie Thesaurus*. The evaluation tasks will involve estimating the semantic distance between German words. Both monolingual and cross-lingual approaches will use the same German corpus, but while the monolingual approach will use a knowledge source in the same language, the German GermaNet, the cross-lingual approach (which I will

[2]For most languages that have been the subject of academic study, there exists at least a bilingual lexicon mapping the core vocabulary of that language to a major world language and a corpus of at least a modest size.

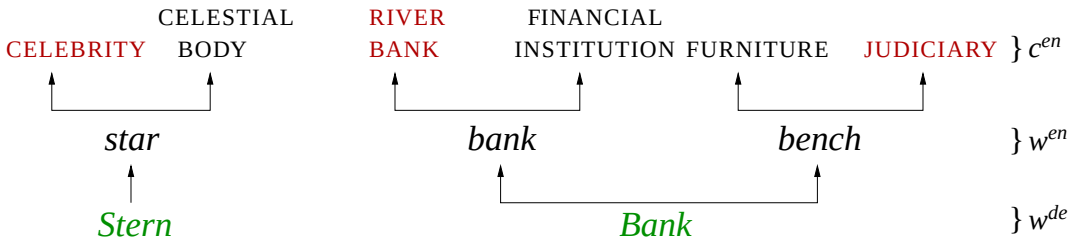


Figure 4.1: The cross-lingual candidate senses of German words *Stern* and *Bank*. In red are concepts not really senses of the German words, but simply artifacts of the translation step.

describe ahead) will use a knowledge source from another language, the English *Macquarie Thesaurus*. The remainder of the chapter describes our approach in terms of German and English, but the algorithm itself is language independent.

4.2 Cross-lingual senses, cross-lingual distributional profiles, and cross-lingual distributional distance

Given a German word w^{de} in context, we use a German–English bilingual lexicon to determine its different possible English translations. Each English translation w^{en} may have one or more possible coarse senses, as listed in an English thesaurus. These English thesaurus concepts (c^{en}) will be referred to as the **cross-lingual candidate senses** of the German word w^{de}. Figure 4.1 depicts examples. They are called "candidate" because some of the senses of w^{en} might not really be senses of w^{de}. For example, CELESTIAL BODY and CELEBRITY are both senses of the English word *star*, but the German word *Stern* can only mean CELESTIAL BODY and not CELEBRITY. Similarly, the German *Bank* can mean FINANCIAL INSTITUTION or FURNITURE, but not RIVER BANK or JUDICIARY. An automated system has no straightforward method of teasing out the actual cross-lingual senses of w^{de} from those that are an artifact of the translation step. So we treat them all as its senses. Now, I proceed to determine semantic distance just as in the monolingual case, except that the words are German and their senses are English thesaurus categories. Table 4.1 presents a mini vocabulary of German words needed to understand the

Table 4.1: Vocabulary of German words needed to understand this discussion.

German word	Meaning(s)	German word	Meaning(s)
Bank	1. financial institution	*Licht*	light
	2. bench (furniture)	*Morgensonne*	morning sun
berühmt	famous	*Raum*	space
Bombe	bomb	*reich*	rich
Erwärmung	heat	*Sonne*	sun
Film	movie (motion picture)	*Star*	star (celebrity)
Himmelskörper	heavenly body	*Stern*	star (celestial body)
Konstellation	constellation	*Verschmelzung*	fusion

discussion in this chapter.

As in the monolingual estimation of distributional concept-distance, the distance between two concepts is calculated by first determining their DPs. Recall the example monolingual DPs of the two senses of *star*:

CELESTIAL BODY (*celestial body, sun, . . .*): *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, . . .

CELEBRITY (*celebrity, hero, . . .*): *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, . . .

In the cross-lingual approach, a concept is now glossed by near-synonymous words in an *English* thesaurus, whereas its profile is made up of the strengths of association with co-occurring *German* words. I will call them **cross-lingual distributional profiles of concepts** or just **cross-lingual DPCs**. Here are constructed examples for the two cross-lingual candidate senses of the German word *Stern*:

CELESTIAL BODY (*celestial body, sun, . . .*): *Raum* 0.36, *Licht* 0.27, *Konstellation* 0.11, . . .

CELEBRITY (*celebrity, hero, . . .*): *berühmt* 0.24, *Film* 0.14, *reich* 0.14, . . .

The values are the strength of association (usually pointwise mutual information or conditional probability) of the target concept with co-occurring words. In order to calculate the strength of association, we must first determine individual word and concept counts, as well as their co-occurrence counts. The next section describes how these can be estimated without the use of any word-aligned parallel corpora and without any sense-annotated data. The closer the cross-lingual DPs of two concepts, the smaller is their semantic distance. Just as in the case of monolingual distributional concept-distance measures (described in the previous chapter), distributional measures can be used to estimate the distance between the cross-lingual DPs of two target concepts. For example, recall how cosine is used in a monolingual framework to estimate distributional distance between two concepts (described in Section 3.2 earlier):

$$Cos_{cp}(c_1, c_2) = \frac{\sum_{w \in C(c_1) \cup C(c_2)} (P(w|c_1) \times P(w|c_2))}{\sqrt{\sum_{w \in C(c_1)} P(w|c_1)^2} \times \sqrt{\sum_{w \in C(c_2)} P(w|c_2)^2}} \tag{4.1}$$

$C(x)$ is the set of English words that co-occur with English *concept x* within a pre-determined window. The conditional probabilities in the formula are taken from the monolingual distributional profiles of concepts. We can adapt the formula to estimate cross-lingual distributional distance between two concepts as shown below:

$$Cos(c_1^{en}, c_2^{en}) = \frac{\sum_{w^{de} \in C(c_1^{en}) \cup C(c_2^{en})} \left(P(w^{de}|c_1^{en}) \times P(w^{de}|c_2^{en})\right)}{\sqrt{\sum_{w^{de} \in C(c_1^{en})} P(w^{de}|c_1^{en})^2} \times \sqrt{\sum_{w^{de} \in C(c_2^{en})} P(w^{de}|c_2^{en})^2}} \tag{4.2}$$

$C(x)$ is now the set of German words that co-occur with English concept x within a pre-determined window. The conditional probabilities in the formula are taken from the cross-lingual DPCs.

If the distance between two German words is required, then the distance between all relevant English cross-lingual candidate sense pairs is determined and the minimum is chosen. For example, if *Stern* has the two cross-lingual candidate senses mentioned above and *Verschmelzung* has one (FUSION), then the distance between them is determined by first applying Cosine (or any distributional measure) to the cross-lingual DPs of CELESTIAL BODY and FUSION:

CELESTIAL BODY (*celestial body, sun, ...*): *Raum* 0.36, *Licht* 0.27, *Konstellation* 0.11, ...

FUSION (*thermonuclear reaction, atomic reaction, ...*): *Erwäermung* 0.16, *Bombe* 0.09, *Licht* 0.09, *Raum* 0.04, ...

Then applying cosine to the cross-lingual DPs of CELEBRITY and FUSION:

CELEBRITY (*celebrity, hero, ...*): *berühmt* 0.24, *Film* 0.14, *reich* 0.14, ...

FUSION (*thermonuclear reaction, atomic reaction, ...*): *Erwärmung* 0.16, *Bombe* 0.09, *Licht* 0.09, *Raum* 0.04, ...

And finally choosing the one with minimum semantic distance, that is, maximum similarity/relatedness:

$$distance(Stern, Verschmelzung) = \max(Cos(\text{CELEBRITY}, \text{FUSION}), Cos(\text{CELESTIAL BODY}, \text{FUSION})) \tag{4.3}$$

Maximum is chosen because cosine is a similarity/relatedness measure. In case of distance measures, such as α Skew Divergence, the minimum will be chosen.

4.3 Estimating cross-lingual DPCs

Determining cross-lingual distributional profiles of concepts requires information about which words in one language L_1 co-occur with which concepts as defined in another language L_2. This means that a direct approach requires the text in L_1, from which counts are made, to have a word-aligned parallel corpus in L_2. Further, the L_2 text must be sense annotated. Such data exists rarely, if at all, and it is expensive to create. Thus, another way to obtain these counts must be devised. I now present a way to estimate cross-lingual distributional profiles of concepts from raw-text (in one language, L_1) and a published thesaurus (in another language, L_2) using an L_1–L_2 bilingual lexicon and a bootstrapping algorithm.

4.3.1 Creating cross-lingual word–category co-occurrence matrix

I create a cross-lingual word–category co-occurrence matrix with German word types w^{de} as one dimension and English thesaurus concepts c^{en} as another.

	c_1^{en}	c_2^{en}	...	c_j^{en}	...
w_1^{de}	m_{11}	m_{12}	...	m_{1j}	...
w_2^{de}	m_{21}	m_{22}	...	m_{2j}	...
⋮	⋮	⋮	⋱	⋮	⋮
w_i^{de}	m_{i1}	m_{i2}	...	m_{ij}	...
⋮	⋮	⋮	...	⋮	⋱

The matrix is populated with co-occurrence counts from a large German corpus. A particular cell m_{ij}, corresponding to word w_i^{de} and concept c_j^{en}, is populated with the number of times the German word w_i^{de} co-occurs (in a window of ± 5 words) with any German word having c_j^{en} as one of its *cross-lingual candidate senses*. For example, the *Raum*–CELESTIAL BODY cell will have the sum of the number of times *Raum* co-occurs with *Himmelskörper, Sonne, Morgensonne, Star, Stern*, and so on (see Figure 4.2). This matrix, created after a first pass of the corpus, is called the **cross-lingual base WCCM**. A contingency table for any particular German word w^{de} and English category c^{en} can be easily generated from the WCCM by collapsing cells for all other words and categories into one and summing up their frequencies.

	c^{en}	$\neg c^{en}$
w^{de}	$n_{w^{de}c^{en}}$	$n_{w^{de}\neg}$
$\neg w^{de}$	$n_{\neg c^{en}}$	$n_{\neg\neg}$

The application of a suitable statistic, such as PMI or conditional probability, will then yield the strength of association between the German word and the English category.

As the cross-lingual base WCCM is created from unannotated text, it is expected to be noisy (for the same word-sense-ambiguity reasons as to why the monolingual base WCCM is noisy—explained in Section 3.3.1 earlier). Yet, again, the cross-lingual base WCCM does capture strong associations between a category (concept) and co-occurring words (just like the

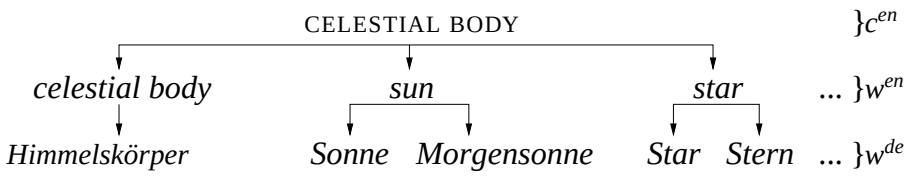


Figure 4.2: Words having CELESTIAL BODY as one of their cross-lingual candidate senses.

monolingual base WCCM). For example, even though we increment counts for both *Raum*–CELESTIAL BODY and *Raum*–CELEBRITY for a particular instance where *Raum* co-occurs with *Star*, *Raum* will co-occur with a number of words such as *Himmelskörper, Sonne*, and *Morgensonne* that each have the sense of CELESTIAL BODY in common (see Figures 4.2 and 4.3), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count of *Raum* and CELESTIAL BODY, and thereby their strength of association, will be relatively higher than those of *Raum* and CELEBRITY (Figure 4.4).

4.3.2 Bootstrapping

As in the monolingual case, a second pass of the corpus is made to disambiguate the (German) words in it. Each word in the corpus is considered as the target one at a time. For each cross-lingual candidate sense of the target, its strength of association with each of the words in its context ($\pm$5 words) is summed. The sense that has the highest cumulative association with co-occurring words is chosen as the intended sense of the target word. A new bootstrapped WCCM is created by populating each cell m_{ij}, corresponding to word w_i^{de} and concept c_j^{en}, with the number of times the German word w_i^{de} co-occurs with any German word *used in cross-lingual sense* c_j^{en}. (Again, this is just like the monolingual bootstrapping—explained earlier in Section 3.3.2.) A statistic such as PMI is then applied to these counts to determine the strengths of association between a target concept and co-occurring words, giving the cross-lingual distributional profile of the concept.

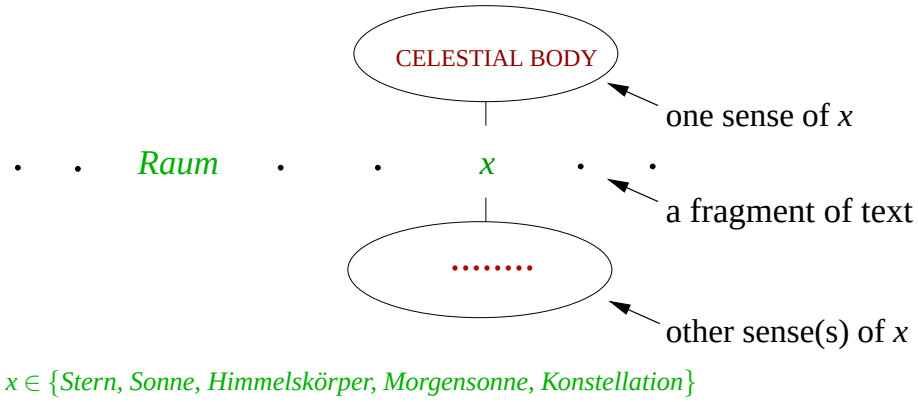


Figure 4.3: The word *Raum* will also co-occur with a number of other words *x* that each have one sense of CELESTIAL BODY in common.

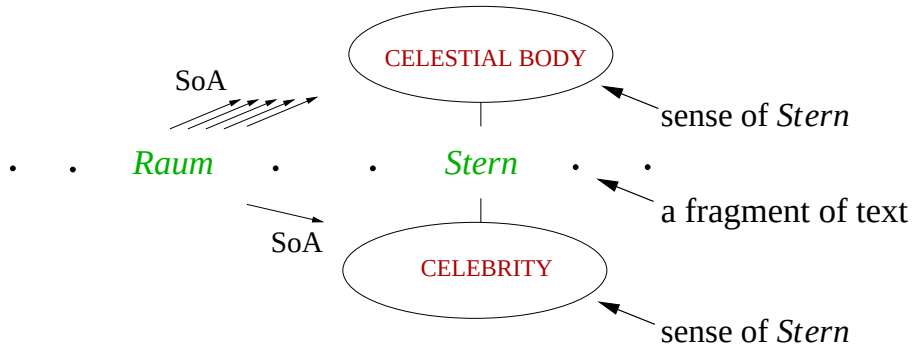


Figure 4.4: The base WCCM captures strong word–category co-occurrence associations.

4.4 Evaluation

We evaluate the newly proposed cross-lingual distributional measures of concept-distance on the tasks of (1) measuring semantic distance between German words and ranking German word pairs according to semantic distance, and (2) solving German 'Word Power' questions from *Reader's Digest*. The cross-lingual approach uses the following resources: the German newspaper corpus *taz*[3] (Sep 1986 to May 1999; 240 million words), the English *Macquarie Thesaurus* (Bernard, 1986) (about 98,000 words), and the German–English bilingual lexicon BEOLINGUS[4] (about 265,000 entries). Multi-word expressions in the thesaurus and the bilin-

[3]http://www.taz.de
[4]http://dict.tu-chemnitz.de

Table 4.2: Distance measures used in the experiments.

(Cross-lingual) Distributional Measures	**(Monolingual) GermaNet Measures**	
	Information Content–based	**Lesk-like**
α-skew divergence (Lee, 2001)	Jiang and Conrath (1997)	hypernym pseudo-gloss (Gurevych, 2005)
cosine (Schütze and Pedersen, 1997)	Lin (1998c)	radial pseudo-gloss (Gurevych, 2005)
Jensen-Shannon divergence (Dagan et al., 1994)	Resnik (1995)	
Lin (1998a)		

gual lexicon were ignored. We used a context of ± 5 words on either side of the target word for creating the base and bootstrapped WCCMs. No syntactic pre-processing was done, nor were the words stemmed, lemmatized, or part-of-speech tagged.

In order to compare results with state-of-the-art monolingual approaches we conducted experiments using GermaNet measures as well. The specific distributional measures and GermaNet-based measures used are listed in Table 4.2. Jensen–Shannon divergence and α-skew divergence calculate the difference in distributions of words that co-occur with the targets. Lin's distributional measure and Lin's GermaNet measure follow from his information-theoretic definition of similarity (Lin, 1998c). The GermaNet measures used are of two kinds: (1) information content measures, and (2) Lesk-like measures that rely on *n*-gram overlaps in the glosses of the target senses, proposed by Gurevych (2005). As GermaNet does not have glosses for synsets, Gurevych (2005) proposed a way of creating a bag-of-words-type pseudo-gloss for a synset by including the words in the synset and in synsets close to it in the network. The information content measures rely on finding the lowest common subsumer (lcs) of the target synsets in a hypernym hierarchy and using corpus counts to determine how specific or general this concept is. The more specific the lcs is and the smaller the difference of its specificity with that of the target concepts, the closer the target concepts are.

Table 4.3: Comparison of datasets used for evaluating semantic distance in German.

Dataset	Year	Language	# pairs	PoS	Scores	# subjects	Correlation
Gur65	2005	German	65	N	discrete {0,1,2,3,4}	24	.810
Gur350	2006	German	350	N, V, A	discrete {0,1,2,3,4}	8	.690

4.4.1 Ranking word pairs

4.4.1.1 Data

A direct approach to evaluate distance measures is to compare them with human judgments. Gurevych (2005) and Zesch et al. (2007b) asked native German speakers to mark two different sets of German word pairs with distance values. Set 1 (**Gur65**) is the German translation of the English Rubenstein and Goodenough (1965b) dataset. It has 65 noun–noun word pairs. Set 2 (**Gur350**) is a larger dataset containing 350 word pairs made up of nouns, verbs, and adjectives. The semantically close word pairs in Gur65 are mostly synonyms or hypernyms (hyponyms) of each other, whereas those in Gur350 have both classical and non-classical relations (Morris and Hirst, 2004) with each other. Details of these **semantic distance benchmarks**[5] are summarized in Table 4.3. Inter-subject correlations are indicative of the degree of ease in annotating the datasets.

4.4.1.2 Results and Discussion

Word-pair distances determined using different distance measures are compared in two ways with the two human-created benchmarks. The rank ordering of the pairs from closest to most distant is evaluated with Spearman's rank order correlation ρ; the distance judgments themselves are evaluated with Pearson's correlation coefficient r. The higher the correlation, the more accurate the measure is. Spearman's correlation ignores actual distance values after a list is ranked—only the ranks of the two sets of word pairs are compared to determine correlation.

[5]The datasets are publicly available at `http://www.ukp.tu-darmstadt.de/data/semRelDatasets`.

On the other hand, Pearson's coefficient takes into account actual distance values. So even if two lists are ranked the same, but one has distances between consecutively-ranked word-pairs more in line with human-annotations of distance than the other, then Pearson's coefficient will capture this difference. However, this makes Pearson's coefficient sensitive to outlier data points, and so one must interpret it with caution.

Table 4.4 shows the results. Observe that on both datasets and by both measures of correlation, cross-lingual measures of concept-distance perform not just as well as the best monolingual measures, but in fact better. (Figure 4.5 depicts the results in a graph.) In general, the correlations are lower for Gur350 as it contains cross-PoS word pairs and non-classical relations, making it harder to judge even by humans (as shown by the inter-annotator correlations for the datasets in Table 4.3).[6] Considering Spearman's rank correlation, α-skew divergence and Jensen-Shannon divergence perform best on both datasets. The correlations of cosine and Lin's distributional measure are not far behind. Amongst the monolingual GermaNet measures, radial pseudo-gloss performs best. Considering Pearson's correlation, Lin's distributional measure performs best overall and radial pseudo-gloss does best amongst the monolingual measures.

4.4.2 Solving word choice problems from *Reader's Digest*

4.4.2.1 Data

Our approach to evaluating distance measures follows that of Jarmasz and Szpakowicz (2003), who evaluated semantic similarity measures through their ability to solve synonym problems (80 TOEFL (Landauer and Dumais, 1997), 50 ESL (Turney, 2001), and 300 (English) *Reader's Digest* Word Power questions). Turney (2006) used a similar approach to evaluate the identification of semantic relations, with 374 college-level multiple-choice word analogy questions.

[6] One can also note that the drop in correlation when moving from classical to non-classical relations is somewhat higher for the automatic measures than for humans. However, it is unclear what we can conclude from this.

Table 4.4: Correlations of distance measures with human judgments. The best results obtained using monolingual and cross-lingual measures are marked in bold.

	Gur65		Gur350	
Measure	**Spearman's rank correlation**	**Pearson's correlation**	**Spearman's rank correlation**	**Pearson's correlation**
Monolingual				
hypernym pseudo-gloss	0.672	0.702	0.346	0.331
radial pseudo-gloss	**0.764**	0.565	**0.492**	0.420
Jiang and Conrath measure	0.665	**0.748**	0.417	0.410
Lin's GermaNet measure	0.607	0.739	0.475	**0.495**
Resnik's measure	0.623	0.722	0.454	0.466
Cross-lingual				
α-skew divergence	**0.794**	0.597	**0.520**	0.413
cosine	0.778	0.569	0.500	*0.212*
Jensen-Shannon divergence	**0.793**	0.633	**0.522**	0.422
Lin's distributional measure	0.775	**0.816**	0.498	**0.514**

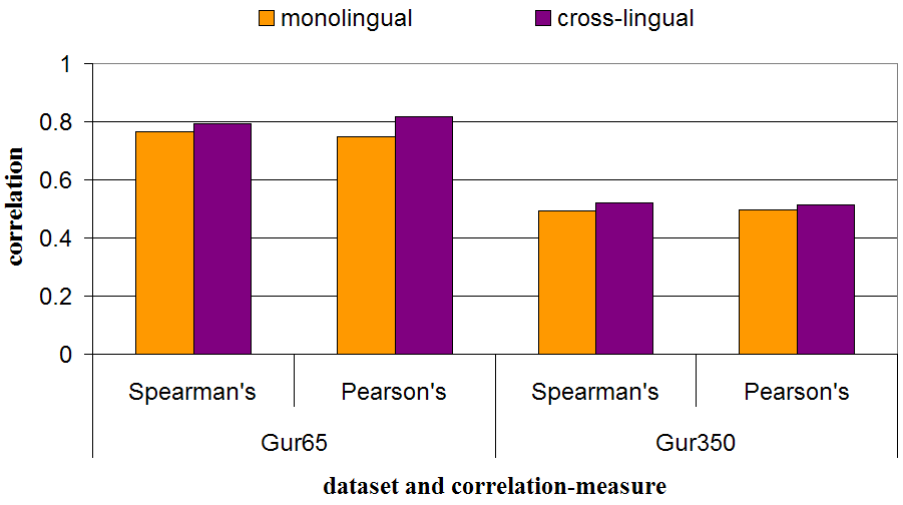


Figure 4.5: Ranking German word pairs

Issues of the German edition of *Reader's Digest* include a word choice quiz called 'Word Power'. Each question has one target word and four alternative words or phrases; the objective is to pick the alternative that is most closely related to the target. The correct answer may be a near-synonym of the target or it may be related to the target by some other classical or non-classical relation (usually the former). For example:[7]

Duplikat (duplicate)

a. *Einzelstück* (single copy) b. *Doppelkinn* (double chin)

c. *Nachbildung* (replica) d. *Zweitschrift* (copy)

As part our collaboration, Torsten Zesch compiled the **Reader's Digest Word Power (RDWP) benchmark** for German, which consists of 1072 of these word-choice problems collected from the January 2001 to December 2005 issues of the German-language edition (Wallace and Wallace, 2005). Forty-four problems that had more than one correct answer and twenty problems that used a phrase instead of a single term as the target were discarded. The remaining 1008 problems form our evaluation dataset, which is significantly larger than any of the previous datasets employed in a similar evaluation.

We evaluate the various cross-lingual and monolingual distance measures by their ability to choose the correct answer. The distance between the target and each of the alternatives is computed by a measure, and the alternative that is closest is chosen. If two or more alternatives are equally close to the target, then the alternatives are said to be **tied**. If one of the tied alternatives is the correct answer, then the problem is counted as correctly solved, but the corresponding score is reduced. The system assigns a score of 0.5, 0.33, and 0.25 for 2, 3, and 4 tied alternatives, respectively (in effect approximating the score obtained by randomly guessing one of the tied alternatives). If more than one alternative has a sense in common with the target, then the thesaurus-based cross-lingual measures will mark them each as the closest sense. However, if one or more of these tied alternatives is in the same semicolon group of the thesaurus as the target, then only these are chosen as the closest senses. Recall that words

[7] English translations are in parentheses.

in a thesaurus category are further partitioned into different paragraphs and each paragraph into semicolon groups. Words within a semicolon group are more closely related than those in semicolon groups of the same paragraph or category.

Even though we discard questions from the German RDWP dataset that contained a phrasal target, we did not discard questions that had phrasal alternatives simply because of the large number of such questions. Many of these phrases cannot be found in the knowledge sources (GermaNet or *Macquarie Thesaurus* via translation list). In these cases, we remove stopwords (prepositions, articles, etc.) and split the phrase into component words. As German words in a phrase can be highly inflected, all components are lemmatized. For example, the target *imaginär* (*imaginary*) has *nur in der Vorstellung vorhanden* (*exists only in the imagination*) as one of its alternatives. The phrase is split into its component words *nur, Vorstellung,* and *vorhanden*. The system computes semantic distance between the target and each phrasal component and selects the minimum value as the distance between target and potential answer.

4.4.2.2 Results and Discussion

Table 4.5 presents the results obtained on the German RDWP benchmark for both monolingual and cross-lingual measures. Only those questions for which the measures have some distance information are attempted; the column '# attempted' shows the number of questions attempted by each measure, which is the maximum score that the measure can hope to get. Observe that the thesaurus-based cross-lingual measures have a much larger coverage than the GermaNet-based monolingual measures. The cross-lingual measures have a much larger number of correct answers too (column '# correct'), but this number is bloated due to the large number of ties. We see more ties when using the cross-lingual measures because they rely on the *Macquarie Thesaurus*, a very coarse-grained sense inventory (around 800 categories), whereas the monolingual measures operate on the fine-grained GermaNet. 'Score' is the score each measure gets after it is penalized for the ties. The cross-lingual measures cosine, Jensen-Shannon divergence, and Lin's distributional measure obtain the highest scores. But 'Score'

by itself does not present the complete picture either as, given the scoring scheme, a measure that attempts more questions may get a higher score just from random guessing. We therefore present precision (P), recall (R), and F measure (F):

$$P = \frac{\text{Score}}{\text{\# attempted}} \tag{4.4}$$

$$R = \frac{\text{Score}}{1008} \tag{4.5}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{4.6}$$

Figure 4.6 depicts the results in a graph. Observe that the cross-lingual measures have a higher coverage (recall) than the monolingual measures but lower precision. The F measures show that the best cross-lingual measures do slightly better than the best monolingual ones, despite the large number of ties. The measures of cosine, Jensen-Shannon divergence, and Lin's distributional measure remain the best cross-lingual measures, whereas hypernym pseudo-gloss and radial pseudo-gloss are the best monolingual ones.

4.5 Conclusion

I have proposed a new method to determine semantic distance in a possibly resource-poor language by combining its text with a knowledge source in a different, preferably resource-rich, language. Specifically, I combined German text with an English thesaurus to create cross-lingual distributional profiles of concepts—the strengths of association between English thesaurus senses (concepts) of German words and co-occurring German words—using a German–English bilingual lexicon and a bootstrapping algorithm designed to overcome ambiguities of word-senses and translations. Notably, I do so without the use of sense-annotated text or word-aligned parallel corpora. I did not parse or chunk the text, nor did I stem, lemmatize, or part-of-speech-tag the words.

I used the cross-lingual DPCs to estimate semantic distance by developing new cross-lingual distributional measures of concept-distance. These measures are like the distributional

Table 4.5: Performance of distance measures on word choice problems. The best results obtained using monolingual and cross-lingual measures are marked in bold.

	Reader's Digest Word Power benchmark						
Measure	**# attempted**	**# correct**	**# ties**	**Score**	**P**	**R**	**F**
Monolingual							
hypernym pseudo-gloss	222	174	11	**171.5**	.77	.17	**.28**
radial pseudo-gloss	266	188	15	**184.7**	.69	.18	**.29**
Jiang and Conrath	357	157	1	156.0	.44	.16	.23
Lin's GermaNet measure	298	153	1	152.5	.51	.15	.23
Resnik's measure	299	154	33	148.3	.50	.15	.23
Cross-lingual							
α-skew divergence	438	185	81	151.6	.35	.15	.21
cosine	438	276	90	**223.1**	.51	.22	**.31**
Jensen-Shannon divergence	438	276	90	**229.6**	.52	.23	**.32**
Lin's distributional measure	438	274	90	**228.7**	.52	.23	**.32**

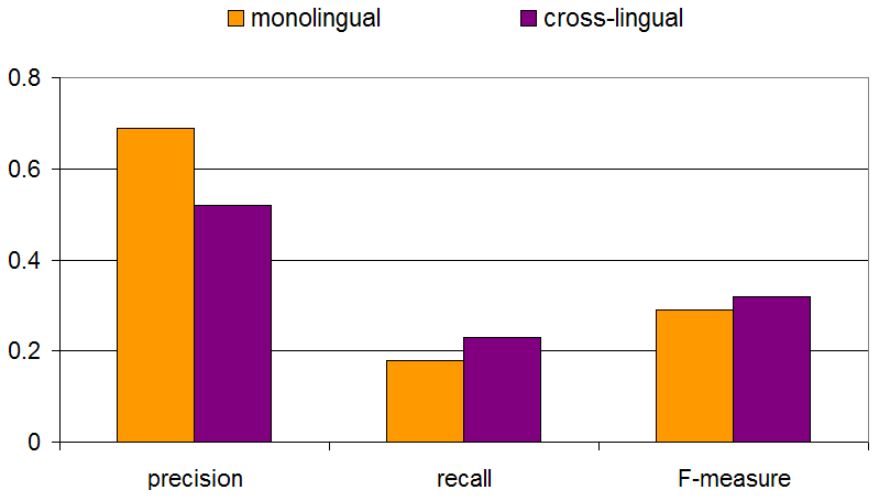


Figure 4.6: Solving word choice problems.

measures of concept-distance (Mohammad and Hirst, 2006a, 2006b), except they can determine distance between words in one language using a thesaurus in a different language. I evaluated the cross-lingual measures against the best monolingual ones operating on a WordNet-like resource, GermaNet, through an extensive set of experiments on two different German semantic distance benchmarks. In the process, my collaborators (Iryna Gurevych and Torsten Zesch) compiled a large German benchmark of *Reader's Digest* word choice problems suitable for evaluating semantic-relatedness measures. Most previous semantic distance benchmarks are either much smaller or cater primarily to semantic similarity measures.

Even with the added ambiguity of translating words from one language to another, the cross-lingual measures performed better than the best monolingual measures on both the word-pair task and the *Reader's Digest* word-choice task. Further, in the word-choice task, the cross-lingual measures achieved a significantly higher coverage than the monolingual measure. The richness of English resources seems to have a major impact, even though German, with GermaNet, a well-established resource, is in a better position than most other languages. This is indeed promising, because achieving broad coverage for resource-poor languages remains an important goal as we integrate state-of-the-art approaches in natural language processing into real-life applications. These results show that the proposed algorithm can successfully combine German text with an English thesaurus using a bilingual German–English lexicon to obtain state-of-the-art results in measuring semantic distance.

These results also support the broader and far-reaching claim that natural language problems in a resource-poor language can be solved using a knowledge source in a resource-rich language (for example the cross-lingual PoS tagger of Cucerzan and Yarowsky (2002)). Cross-lingual DPCs also have tremendous potential in tasks inherently involving more than one language. In Chapter 7 ahead, I investigate the use of cross-lingual DPCs in word translation. This work will act as a launching pad for other multilingual efforts on machine translation (Section 8.5.1), multi-language multi-document summarization (Section 8.5.2), multilingual information retrieval (Section 8.5.3), and multilingual document clustering (Section 8.5.4). I believe

that the future of natural language processing lies not in standalone monolingual systems but in those that are powered by automatically created multilingual networks of information.

Chapter 5

Determining Word Sense Dominance

5.1 Introduction

In the last two chapters, I showed how corpus statistics can be combined with a published thesaurus to estimate semantic distance. I evaluated the new approach, in both monolingual and cross-lingual frameworks, on certain word-distance tasks (tasks that do not explicitly require distance between a concept and another unit of language, but rather, seemingly at least, require the distance between words). Those were tasks where distributional profiles of words (DPWs) can and have been used, but, as I have shown, using distributional profiles of concepts (DPCs) gives as much better results. In this chapter, I describe the use of DPCs in a task where DPWs alone cannot help. This chapter describes the evaluation of the new approach on a *concept-distance* task—determining word sense dominance.

In text, the occurrences of the senses of a word usually have a skewed distribution (Gale et al., 1992; Ng and Lee, 1996; Sanderson and van Rijsbergen, 1999). For example, in a set of randomly acquired sentences containing the word *dam*, it is probable that most of the instances correspond to the BODY OF WATER sense as opposed to the rather infrequent UNIT OF LENGTH or FEMALE PARENT OF AN ANIMAL senses. Further, the distribution varies in accordance with the domain or topic of discussion. For example, the ASSERTION OF ILLEGALITY sense

of *charge* is more frequent in the judicial domain, whereas in the domain of economics, the EXPENSE/COST sense occurs more often. Formally, the **degree of dominance of a** particular **sense** of a word (**target word**) in a given text (**target text**) may be defined as the proportion of occurrences of the sense to the total occurrences of the target word. The sense with the highest dominance in the target text is called the **predominant sense** of the target word.

Determination of word sense dominance has many uses. An unsupervised system will benefit by backing off to the predominant sense in case of insufficient evidence (Hoste et al., 2001). The dominance values may be used as prior probabilities for the different senses, obviating the need for labeled training data in a sense disambiguation task. Natural language systems can choose to ignore infrequent senses of words (McCarthy et al., 2004a) or consider only the most dominant senses (McCarthy et al., 2004b). An unsupervised algorithm that discriminates instances into different usages can use word sense dominance to assign labels to the different clusters generated.

Word sense dominance may be determined by simple counting in sense-tagged data. However, as mentioned earlier, dominance varies with domain and existing sense-tagged data is largely insufficient to meet these needs. I propose four new measures to accurately determine word sense dominance using raw text and a published thesaurus. Unlike the McCarthy et al. (2004b) system, these measures can be used on relatively small target texts, without the need for a *similarly-sense-distributed* auxiliary text. Further, given a new target text, the measures are much faster and they can be employed not just for nouns but for any part of speech. I perform an extensive evaluation using artificially generated thesaurus-sense-tagged data.

5.2 Related work

McCarthy et al. (2004b) automatically determine domain-specific predominant senses of words by using both a measure of distributional similarity (Lin, 1998b) and a measure of semantic similarity (Jiang and Conrath, 1997). The system (Figure 5.1) automatically generates a distri-

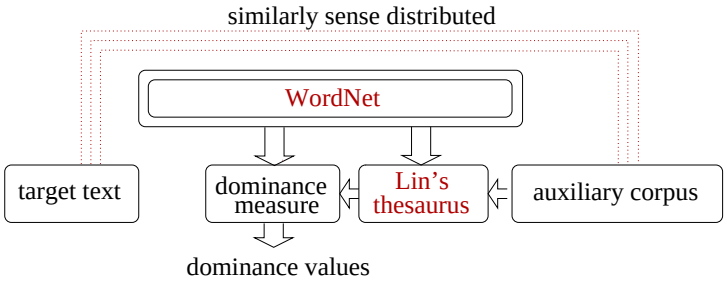


Figure 5.1: The McCarthy et al. system. Its limitations include: (1) requirement of a large corpus similarly sense distributed as the target text, (2) its reliance on WordNet-based semantic distance measures which are good only for noun pairs, and (3) need to re-create Lin's distributional thesaurus for each new text with a different sense distribution.

butional thesaurus from a large corpus. The target text is used for this purpose, provided it is large enough. Otherwise a large corpus with sense distribution similar to the target text (text pertaining to the specified domain) must be used.

The thesaurus has an entry for each word type, which lists a limited number of words (**neighbors**) that are distributionally most similar to it. Since Lin's distributional measure overestimates the distributional similarity of more-frequent word pairs (Mohammad and Hirst, 2005), the neighbors of a word corresponding to the predominant sense are distributionally closer to it than those corresponding to any other sense. For each sense s of a target word t, the distributional similarity scores of t with all it neighbors are summed using the semantic similarity of s with the closest sense of the neighbor as weight. The sense that gets the highest score is chosen as the predominant sense.

The McCarthy et al. system needs to re-train (create a new thesaurus) every time it is to determine predominant senses in data from a different domain. This requires large amounts of part-of-speech-tagged and chunked data from that domain. Further, the target text must be large enough to learn a thesaurus from (Lin (1998b) used a 64-million-word corpus), or a large auxiliary text with a sense distribution similar to the target text must be provided (McCarthy et al. (2004b) separately used 90-, 9.1-, and 32.5-million-word corpora). As the McCarthy et al.

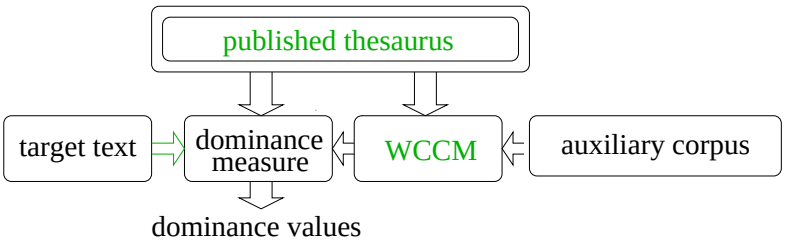


Figure 5.2: My word-sense-dominance system. Notably: (1) it too uses an auxiliary corpus, but it does not need to have a sense distribution similar to the target text, (2) the word–category co-occurrence matrix is created just once, and (3) it relies on a published thesaurus and can be applied to content words of any part of speech.

system relies on a WordNet-based measure of semantic distance as well, and as WordNet-based measures are particularly poor at estimating semantic relatedness, the approach, in practice, is applicable only to nouns and it is unable to exploit information from semantically related, albeit semantically dissimilar, co-occurring words.

5.3 My word-sense-dominance system

I present a method (Figure 5.2) that, in contrast to the McCarthy et al. (2004b) system, determines word sense dominance even in relatively small amounts of target text (a few hundred sentences); although it does use a corpus, it does not require a *similarly-sense-distributed* corpus. Nor does my system need any part-of-speech-tagged data (although that may improve results further), and it does not need to generate a thesaurus or execute any such time-intensive operation at run time. The approach stands on the hypothesis that words surrounding the target word are indicative of its intended sense, and that the dominance of a particular sense is proportional to the relative strength of association between it and co-occurring words in the target text. As shown in the previous two chapters, this strength of association can be determined not just for a noun sense and co-occurring words but also for any other part of speech. Therefore, this sense dominance approach can be applied not just to nouns, but to any part of speech.

5.3.1 Small target texts and a domain-free auxiliary corpus

As before, I use the *Macquarie Thesaurus*, with its 812 categories, as a coarse-grained sense inventory. I create a word–category co-occurrence matrix (WCCM) and distributional profiles of these concepts or categories (DPCs) using the bootstrapping algorithm described earlier in Section 3.3 and a subset of the *British National Corpus (BNC)* (Burnard, 2000); I use all except every twelfth sentence of the BNC and keep the remaining for evaluation purposes.[1] This corpus, used in addition to the target text, will be called the **auxiliary corpus**. If the target text belongs to a particular domain, then the creation of the WCCM from an auxiliary text of the same domain is expected to give better results than the use of a domain-free text. The target text itself may be used as the auxiliary corpus if it is large enough. However, the key feature of my approach is that the target text does not have to be large and even a domain-free auxiliary corpus can help obtain accurate results.

5.3.2 Dominance measures

I examine each occurrence of the target word in a given untagged target text to determine dominance of any of its senses. For each occurrence t' of a target word t, let T' be the set of words (tokens) co-occurring within a predetermined window around t'; let T be the union of all such T' and let $\mathscr{X}_t$ be the set of all such T'. (Thus $|\mathscr{X}_t|$ is equal to the number of occurrences of t, and $|T|$ is equal to the total number of words (tokens) in the windows around occurrences of t.) I propose four methods (Figure 5.3) to determine dominance ($D_{I,W}, D_{I,U}, D_{E,W}$, and $D_{E,U}$) and the underlying assumptions of each.

$D_{I,W}$ is based on the assumption that the more dominant a particular sense is, the greater the strength of its association with words that co-occur with it. For example, if most occurrences of *bank* in the target text correspond to RIVER BANK, then the strength of association of RIVER

[1] Note that even though we use a subset of the the BNC for evaluation, as described ahead in Section 5.5 ahead, we create different test sets pertaining to different sense distributions from this subset. Thus, we are not assuming that the corpus used to create the WCCM and the test sets have the same sense distribution. In fact, they do not.

	Weighted voting	Unweighted voting
Implicit sense disambiguation	$D_{I,W}$	$D_{I,U}$
Explicit sense disambiguation	$D_{E,W}$	$D_{E,U}$

Figure 5.3: The four dominance methods.

BANK with all of *bank*'s co-occurring words will be larger than the sum for any other sense. Dominance $D_{I,W}$ of a sense or category (c) of the target word (t) is:

$$D_{I,W}(t,c) = \frac{\sum_{w \in T} A(w,c)}{\sum_{c' \in senses(t)} \sum_{w \in T} A(w,c')} \tag{5.1}$$

where A is any one of the measures of association, such as pointwise mutual information, described earlier in Section 2.1.1—cosine (Cos), Dice coefficient (Dice), odds ratio (Odds), pointwise mutual information (PMI), Yule's coefficient (Yule), and ϕ coefficient. Metaphorically, words that co-occur with the target word give a weighted vote to each of its senses. The weight is proportional to the strength of association between the sense and the co-occurring word. The dominance of a sense is the ratio of the total votes it gets to the sum of votes received by all the senses.

A slightly different assumption is that the more dominant a particular sense is, the greater the number of co-occurring words having highest strength of association with that sense (as opposed to any other). This leads to the following methodology. Each co-occurring word casts an equal, unweighted vote. It votes for that sense (and no other) of the target word with which it has the highest strength of association. The dominance $D_{I,U}$ of the sense is the ratio of the votes it gets to the total votes cast for the word (number of co-occurring words).

$$D_{I,U}(t,c) = \frac{|\{w \in T : Sns_1(w,t) = c\}|}{|T|} \tag{5.2}$$

$$Sns_1(w,t) = \underset{c' \in senses(t)}{\operatorname{argmax}} A(w,c') \tag{5.3}$$

Observe that in order to determine $D_{I,W}$ or $D_{I,U}$, we do not need to explicitly disambiguate the senses of the target word's occurrences. I now describe alternative approaches that may be

used for explicit sense disambiguation of the target word's occurrences and thereby determine sense dominance (the proportion of occurrences of that sense). $D_{E,W}$ relies on the hypothesis that the intended sense of any occurrence of the target word has highest strength of association with its co-occurring words.

$$D_{E,W}(t,c) = \frac{|\{T' \in \mathscr{X}_t : Sns_2(T',t) = c\}|}{|\mathscr{X}_t|} \tag{5.4}$$

$$Sns_2(T',t) = \underset{c' \in senses(t)}{\operatorname{argmax}} \sum_{w \in T'} A(w,c') \tag{5.5}$$

Metaphorically, words that co-occur with the target word give a weighted vote to each of its senses just as in $D_{I,W}$. However, votes from co-occurring words in an occurrence are summed to determine the intended sense (sense with the most votes) of the target word. The process is repeated for all occurrences that have the target word. If each word that co-occurs with the target word votes as described for $D_{I,U}$, then the following hypothesis forms the basis of $D_{E,U}$: in a particular occurrence, the sense that gets the maximum votes from its neighbors is the intended sense.

$$D_{E,U}(t,c) = \frac{|\{T' \in \mathscr{X}_t : Sns_3(T',t) = c\}|}{|\mathscr{X}_t|} \tag{5.6}$$

$$Sns_3(T',t) = \underset{c' \in senses(t)}{\operatorname{argmax}} |\{w \in T' : Sns_1(w,t) = c'\}| \tag{5.7}$$

In methods $D_{E,W}$ and $D_{E,U}$, the dominance of a sense is the proportion of occurrences of that sense.

The degree of dominance provided by all four methods has the following properties: (i) The dominance values are in the range 0 to 1—a score of 0 implies lowest possible dominance, while a score of 1 means that the dominance is highest. (ii) The dominance values for all the senses of a word sum to 1.

5.4 Pseudo-thesaurus-sense-tagged data

To evaluate the four dominance methods we would ideally like sentences that have target words annotated with senses from the thesaurus (the concept inventory). However, human annotation is both expensive and time intensive. So I present an alternative approach of artificially generating thesaurus-sense-tagged data following the ideas of Leacock et al. (1998) and Mihalcea and Moldovan (1999). Around 63,700 of the 98,000 word types in the *Macquarie Thesaurus* are **monosemous**—listed under just one of the 812 categories. This means that on average around 77 words per category are monosemous. **Pseudo-thesaurus-sense-tagged (PTST) data** for a non-monosemous target word *t* (for example, *brilliant*) used in a particular sense or category *c* of the thesaurus (for example, INTELLIGENCE) may be generated as follows. Identify monosemous words (for example, *clever*) belonging to the same category as *c*. Pick sentences containing the monosemous words from an untagged auxiliary text corpus.

*Hermione had a **clever** plan.*

In each such sentence, replace the monosemous word with the target word *t*. In theory the words in a thesaurus category are near-synonyms or at least strongly related words, making the replacement of one by another acceptable. For the sentence above, we replace *clever* with *brilliant*. This results in (artificial) sentences with the target word used in a sense corresponding to the desired category. Figure 5.4 summarizes the process.

Clearly, many of these sentences will not be linguistically well formed, but the non-monosemous word used in a particular sense is likely to have similar co-occurring words as the monosemous word of the same category.[2] This justifies the use of these pseudo-thesaurus-sense-tagged data for the purpose of evaluation.

[2]Strong collocations are an exception to this, and their effect must be countered by considering larger window sizes. Therefore, we do not use a window size of just one or two words on either side of the target word, but rather windows of ± 5 words in our experiments.

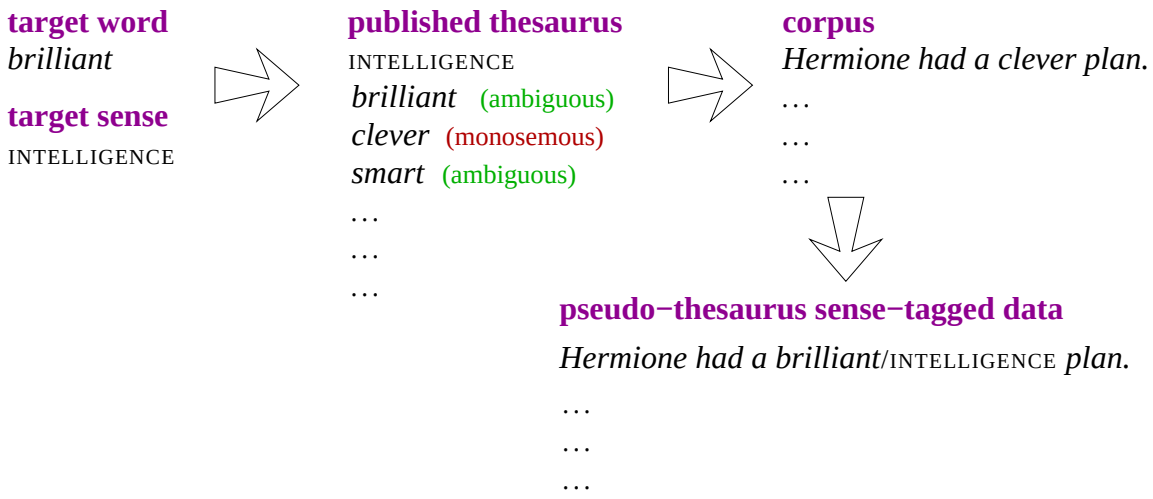


Figure 5.4: An overview of how pseudo-thesaurus-sense-tagged data was created.

I generated PTST test data for the head words in SENSEVAL-1 English lexical sample space[3] using the *Macquarie Thesaurus* and the held out subset of the *BNC* (every twelfth sentence).

5.5 Evaluation

I evaluated the four dominance methods, like McCarthy et al. (2004b), through the accuracy of a naïve word sense disambiguation system that always gives out the predominant sense of the target word. The predominant sense is determined by each of the four dominance methods, individually. The more accurately a measure determines the predominant sense of the target words, the higher will be the accuracy of the word sense disambiguation system. I used the following setup to study the effect of sense distribution on performance.

[3]SENSEVAL-1 head words have a wide range of possible senses, and availability of alternative sense-tagged data may be exploited in the future.

5.5.1 Setup

For each target word for which we have PTST data, the two most dominant senses are identified, say s_1 and s_2. If the number of sentences annotated with s_1 and s_2 is x and y, respectively, where $x > y$, then all y sentences of s_2 and the first y sentences of s_1 are placed in a **data bin**. Eventually the bin contains an equal number of PTST sentences for the two most dominant senses of each target word. Our data bin contained 17,446 sentences for 27 nouns, verbs, and adjectives. We then generate different test data sets d_α from the bin, where α takes values $0, 0.1, 0.2, \ldots, 1$, such that the fraction of sentences annotated with s_1 is α and those with s_2 is $1-\alpha$. Thus the data sets have different dominance values even though they have the same number of sentences. (Note that because of the way the datasets are compiled, each has half as many sentences as there are in the bin.)

Each data set d_α is given as input to the naïve word sense disambiguation system. If the predominant sense is correctly identified for all target words, then the system will achieve highest accuracy, whereas if it is falsely determined for all target words, then the system achieves the lowest accuracy. The value of α determines this **upper bound** and **lower bound**. If α is close to 0.5, then even if the system correctly identifies the predominant sense, the naive disambiguation system cannot achieve accuracies much higher than 50%. On the other hand, if α is close to 0 or 1, then the system may achieve accuracies close to 100%. A disambiguation system that randomly chooses one of the two possible senses for each occurrence of the target word will act as the baseline. Note that no matter what the distribution of the two senses (α), this system will get an accuracy of 50%.

5.5.2 Results

Highest accuracies achieved using the four dominance methods and the measures of association that worked best with each are shown in Figure 5.5. The table below the figure shows **mean distance below upper bound (MDUB)** for all α values considered. Measures that perform

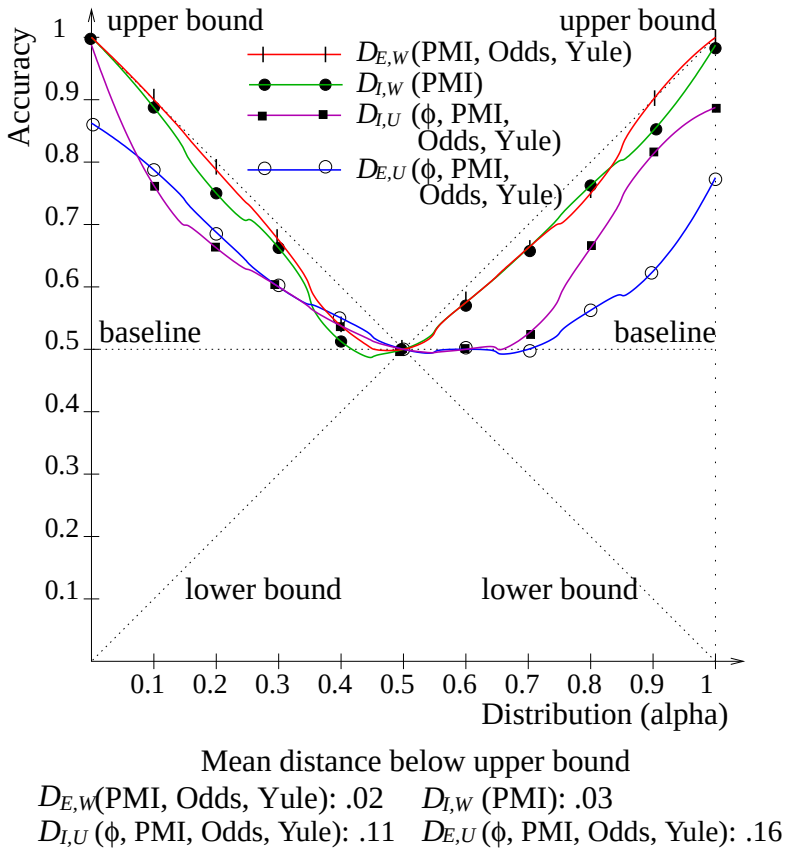


Figure 5.5: Best results: four dominance methods

almost identically are grouped together and the MDUB values listed are averages. The window size used was ± 5 words around the target word. Each dataset d_α, which corresponds to a different target text in Figure 2, was processed in less than 1 second on a 1.3GHz machine with 16GB memory. Weighted voting methods, $D_{E,W}$ and $D_{I,W}$, perform best with MDUBs of just 0.02 and 0.03, respectively. Yule's coefficient, odds ratio, and PMI give near-identical, maximal accuracies for all four methods with a slightly greater divergence in $D_{I,W}$, where PMI does best. The ϕ coefficient performs best for unweighted methods. Dice and cosine do only slightly better than the baseline. In general, results from the method–measure combinations are symmetric across $\alpha = 0.5$, as they should be.

Marked improvements in accuracy were achieved as a result of bootstrapping the WCCM (Figure 5.6). Most of the gain was provided by the first bootstrapping iteration itself whereas

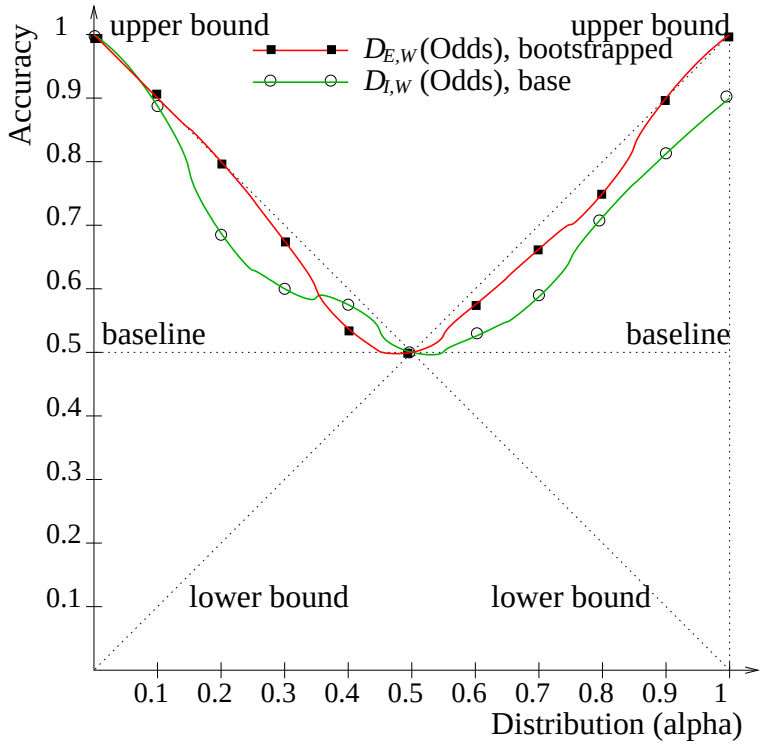


Figure 5.6: Best results: base vs. bootstrapped

further iterations did not improve accuracy ($D_{E,W}$ and $D_{I,W}$ still had MDUBs of 0.02 and 0.03, respectively). This is not surprising because it is in the first bootstrapping iteration that word sense disambiguation is first done (not while creating the base WCCM). The marginal improvements to the WCCM by subsequent iterations do not have as much effect. All bootstrapped results reported in this thesis pertain to just one iteration.[4] The bootstrapped WCCM is 72% smaller, and 5 times faster at processing the data sets, than the base WCCM.

[4]Even though the DPCs were conceived with the intention of estimating semantic distance, this work on word sense dominance (Mohammad and Hirst, 2006a) preceded the experiments on the word-distance tasks described in the last two chapters (Mohammad and Hirst, 2006b; Mohammad et al., 2007a). Once it was determined through these sense dominance experiments that bootstrapping once was optimal, the same was done for all other experiments described in this thesis, with occasional sanity checks.

5.5.3 Discussion

Considering that this is a completely unsupervised approach, not only are the accuracies achieved using the weighted methods well above the baseline, but also remarkably close to the upper bound. This is especially true for α values close to 0 and 1. The lower accuracies for α near 0.5 are understandable as the amount of evidence towards both senses of the target word are nearly equal.

Odds ratio, pointwise mutual information, and Yule's coefficient perform almost equally well for all methods. Since the number of times two words co-occur is usually much less than the number of times they occur individually, pointwise mutual information tends to approximate the logarithm of odds ratio. Also, Yule's coefficient is a derivative of odds ratio. Thus all three measures will perform similarly in case the co-occurring words give an unweighted vote for the most appropriate sense of the target as in $D_{I,U}$ and $D_{E,U}$. For the weighted voting schemes, $D_{I,W}$ and $D_{E,W}$, the effect of scale change is slightly higher in $D_{I,W}$ as the weighted votes are summed over the complete text to determine dominance. In $D_{E,W}$ the small number of weighted votes summed to determine the sense of the target word may be the reason why performances using pointwise mutual information, Yule's coefficient, and odds ratio do not differ markedly. Dice coefficient and cosine gave below-baseline accuracies for a number of sense distributions. This suggests that the normalization[5] to take into account the frequency of individual events inherent in the Dice and cosine measures may not be suitable for this task.

The accuracies of the dominance methods remain the same if the target text is partitioned as per the target word, and each of the pieces is given individually to the disambiguation system. The average number of sentences per target word in each dataset d_α is 323. Thus the results shown above correspond to an average target text size of only 323 sentences.

I repeated the experiments on the base WCCM after filtering out (setting to 0) cells with frequency less than 5 to investigate the effect on accuracies and gain in computation time (pro-

[5]If two events occur individually a large number of times, then they must occur together much more often to get substantial association scores through PMI or odds ratio, as compared to cosine or the Dice coefficient.

portional to size of WCCM). There were no marked changes in accuracy but a 75% reduction in size of the WCCM. Using a window equal to the complete sentence as opposed to ± 5 words on either side of the target resulted in a drop of accuracies.

5.6 Conclusions

I proposed four methods to determine the degree of dominance of a sense of a word using distributional profiles of concepts. I used the *Macquarie Thesaurus* as a very coarse concept inventory. I automatically generated sentences that have a target word annotated with senses from the published thesaurus, which were used to perform an extensive evaluation of the dominance methods. The system achieved near-upper-bound results using all combinations of the the weighted dominance methods ($D_{I,W}$ and $D_{E,W}$) and three different measures of association (Odds, PMI, and Yule).

We cannot compare accuracies with McCarthy et al. (2004b) because use of a thesaurus instead of WordNet means that knowledge of exactly how the thesaurus senses map to WordNet is required. However, I showed that, unlike the McCarthy et al. system, this new system gives accurate results without the need for a large *similarly-sense-distributed* text or retraining. The target texts used were much smaller (a few hundred sentences) than those needed for automatic creation of a distributional thesaurus (a few million words). My system does not perform any time-intensive operation, such as the creation of Lin's thesaurus, at run time; and it can be applied to all parts of speech—not just nouns.

Chapter 6

Unsupervised Word Sense Disambiguation

6.1 Introduction

Word sense disambiguation or **WSD** is the task of determining the intended sense or meaning of an ambiguous **target word** from its context. The context may be a few words on either side of the target, the complete sentence, or it could include a few sentences around it as well. Humans are skilled at word sense disambiguation. For example, even though *weakness* can mean either AN INSTANCE OR PERIOD OF LACKING IN STRENGTH, FAULT, or A SPECIAL FONDNESS, in the sentence below:

The Dark Lord has a weakness for ice cream.

we very quickly home into the SPECIAL FONDNESS sense, and often without conscious effort. However, automatic word sense disambiguation has proved to be much harder. There are many reasons for this including the difficulties of encoding comprehensive world knowledge, determining what the senses of a word must be, how coarse or fine this sense-inventory must be, and so on.

That said, determining the intended sense of a word is potentially useful in many natural language tasks including machine translation and information retrieval. The more-accurate approaches for word sense disambiguation are supervised (Pradhan et al., 2007; Pedersen, 2001;

Ng and Lee, 1996; McRoy, 1992). These systems rely on sense-annotated data to identify co-occurring words that are indicative of the use of the target word in each of its senses.

However, only limited amounts of sense-annotated data exist and it is expensive to create. Thus, a number unsupervised but knowledge-rich approaches have been proposed that do not require sense-annotated data but make use of one or more of the lexical semantic networks in WordNet (Sussna, 1993; Banerjee and Pedersen, 2003; Yang and Powers, 2006b). In this thesis, I have proposed an unsupervised approach to determine the strength of association between a sense or concept and its co-occurring words—the distributional profile of a concept (DPC)—relying simply on raw text and a published thesaurus. I now show how these distributional profiles of concepts can be used to create an *unsupervised* naïve Bayes word-sense classifier (determining both the prior probability and the likelihood in an unsupervised manner). I will compare it with a baseline classifier that also uses the strength of association between the senses of the target and co-occurring words, but relies only on contextual evidence. Since I use pointwise mutual information (PMI) to measure the strength of association, I will refer to the baseline classier as the PMI-based classifier. Both the naïve Bayes and the PMI-based classifiers participated in SemEval-07's[1] English Lexical Sample Task (task #17). Most other unsupervised word sense disambiguation (as opposed to *discrimination*) systems, such as those mentioned above, rely on a language-specific knowledge source such as WordNet and as a consequence are monolingual. The approach proposed here uses raw text and a published thesaurus and it can be used both monolingually (as shown ahead in this chapter) and cross-lingually (as I will show in the next chapter in the guise of a word-translation task). Notably, when used cross-lingually the system can perform word sense disambiguation even in a resource-poor language by combining its text with a published thesaurus from a resource-rich one.

[1]SemEval-07 is a workshop of ACL-07, where systems compete in various semantic analysis tasks on newly compiled/created test data.

6.2 The English Lexical Sample Task

The English Lexical Sample Task (Pradhan et al., 2007) was a traditional word sense disambiguation task wherein the intended (WordNet) sense of a target word was to be determined from its context. The training and test data had 22,281 and 4,851 instances respectively for 100 target words (50 nouns and 50 verbs). They are in Senseval-2 data format. WordNet 2.1 was used as the sense inventory for most of the target words, but certain words were assigned one or more senses from OntoNotes (Hovy et al., 2006). Many of the fine-grained senses were grouped into coarser ones. Here is an example training instance:

<instance id="29:0@5@wsj12wsj_1253@wsj@en@on" docsrc="wsj">
<answer instance="29:0@5@wsj12wsj_1253@wsj@en@on" senseid="1" wn="1,2,3" wn-version="2.1">
<context>
This is just not so . The reality is that Bank finances are rock solid . As of June 30 , 1989 – the day our past fiscal year came to a close – only 4.1 % of the Bank 's portfolio was <head> affected <head> by arrears of over six months . This is an enviably low level . Moreover , the Bank follows a prudent provisioning policy and has set aside $ 800 million against possible loan losses .
</context>
</instance>

The target word *affect* is enclosed with the <head> and </head> tags. From the answer instance line (second line) we know that the above instance is annotated with sense identifier 1 (senseid="1"), and that the intended sense of the target is a concept defined as the grouping of three wordnet synsets (wn="1,2,3"). Below is an example test instance:

<instance id="15:0@35@browncfcf11@brown@en@on" docsrc="brown">
<context>
every orthodontist sees children who are embarrassed by their malformed teeth .

Some such youngsters rarely smile , or they try to speak with the mouth closed . In certain cases , as in Dick Stewart 's , a child 's personality is <head> affected <head> . Yet from the dentist 's point of view , bad-fitting teeth should be corrected for physical reasons . Bad alignment may result in early loss of teeth through a breakdown of the bony structure that supports their roots .

</context>

</instance>

Note that the test instance does not have an answer-instance tag.

6.3 Coping with sense-inventory mismatch

As described earlier, I create distributional profiles of concepts or senses by first representing the sense with a number of near-synonymous words. A thesaurus is a natural source of such synonyms. Even though the approach can be ported to WordNet,[2] there was no easy way of representing OntoNotes senses with near-synonymous words. Therefore, I asked four native speakers of English to map the WordNet and OntoNotes senses of the 100 target words to the *Macquarie Thesaurus* and continue to use it as sense inventory. I also wanted to examine the effect of using a very coarse sense inventory, such as the categories in a published thesaurus, (only 812 in all) on word sense disambiguation.

The annotators were presented with a target word, its WordNet/OntoNotes senses, and the Macquarie senses. WordNet senses were represented by synonyms, glosses, and example usages. The OntoNotes senses were described through syntactic patterns and example usages (provided by the task organizers). The Macquarie senses (categories) were described by the category head (a representative word for the category) and five other words in the category. Specifically, words in the same semicolon group as the target were chosen, as words within

[2]The synonyms within a synset, along with its one-hop neighbors and all its hyponyms, can represent that sense.

a semicolon group of a thesaurus tend to be more closely related than words across groups. Annotators 1 and 2 labeled each WordNet/OntoNotes sense of the first 50 target words with one or more appropriate *Macquarie Thesaurus* categories. Annotators 3 and 4 labeled the senses of the other 50 words. I combined all four annotations into a **WordNet–Macquarie mapping file** by taking, for each target word, the union of categories chosen by the two annotators.

6.4 The DPC-based classifiers

I will now describe the two classifiers that took part in the English Lexical Sample Task. Both use words in the context of the target word as features and both rely on a word–category co-occurrence matrix to determine the evidence towards each sense of the target. The general structure of the WCCM is shown below again for ease of reference.

	c_1	c_2	...	c_j	...
w_1	m_{11}	m_{12}	...	m_{1j}	...
w_2	m_{21}	m_{22}	...	m_{2j}	...
⋮	⋮	⋮	⋱	⋮	⋮
w_i	m_{i1}	m_{i2}	...	m_{ij}	...
⋮	⋮	⋮	...	⋮	⋱

It has words in one dimension and categories in the other. A particular cell m_{ij} for word w_i and sense or category c_j contains the number of times they co-occur in text. As described earlier in Section 3.3, I created the word–category co-occurrence matrix (WCCM) using a bootstrapping algorithm and the *British National Corpus (BNC)* (Burnard, 2000).

6.4.1 Unsupervised naïve Bayes classifier

The naïve Bayes classifier uses the following formula to determine the intended sense c_{nb}:

$$c_{nb} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{w_i \in W} P(w_i|c_j) \tag{6.1}$$

where C is the set of possible senses of the target word (as listed in the *Macquarie Thesaurus*) and W is the set of words that co-occur with the target (I used a window of ± 5 words).[3]

Traditionally, prior probabilities of the senses ($P(c_j)$) and the conditional probabilities in the likelihood ($\prod_{w_i \in W} P(w_i|c_j)$) are determined by simple counts in sense-annotated data. I approximated these probabilities using counts from the word–category co-occurrence matrix, thereby obviating the need for manually-annotated data.

$$P(c_j) = \frac{\sum_i m_{ij}}{\sum_{i,j} m_{ij}} \tag{6.2}$$

$$P(w_i|c_j) = \frac{m_{ij}}{\sum_i m_{ij}} \tag{6.3}$$

Here, m_{ij} is the number of times the word w_i co-occurs with the category c_j—as listed in the word–category co-occurrence matrix (WCCM).

6.4.2 PMI-based classifier

Pointwise mutual information (PMI) between a sense of the target word and a co-occurring word is calculated using the following formula:

$$PMI(w_i, c_j) = \log \frac{P(w_i, c_j)}{P(w_i) \times P(c_j)} \tag{6.4}$$

$$\text{where} \quad P(w_i, c_j) = \frac{m_{ij}}{\sum_{i,j} m_{ij}} \tag{6.5}$$

$$\text{and} \quad P(w_i) = \frac{\sum_j m_{ij}}{\sum_{i,j} m_{ij}} \tag{6.6}$$

Here, m_{ij} is the count in the WCCM and $P(c_j)$ is as in equation 6.2. For each sense of the target word, the sum of the strength of association (PMI) between it and each of the co-occurring words (in a window of ± 5 words) is calculated. The sense with the highest sum is chosen as

[3]Note that while it is reasonable to filter out non-content stopwords, it is not necessary in the case of the two classifiers described in this and the next subsection. These words will have a small and more-or-less identical co-occurrence strength of association with all concepts and so will not play a significant role in determining the intended sense. Of course, in certain cases the exact position of non-content words (for example, the target word being immediately preceded by *on*) is indicative of the intended sense, but these classifiers do not make use of such exact positional information.

the intended sense.

$$c_{pmi} = \underset{c_j \in C}{\operatorname{argmax}} \sum_{w_i \in W} PMI(w_i, c_j) \tag{6.7}$$

Note that even though the PMI-based classifier uses prior probabilities of the categories $P(c_j)$ (determined from the WCCM) to determine the strength of association of c_j with co-occurring words, the classifier does not bias (multiply) this contextual evidence with $P(c_j)$. Since it uses only contextual evidence, I call the PMI-based classifier a baseline to the classifier described in the previous sub-section.

6.5 Evaluation

Both the naïve Bayes classifier and the PMI-based classifier were applied to the training data of English Lexical Sample Task. For each instance, the *Macquarie Thesaurus* category *c* that best captures the intended sense of the target was determined. The system then labels an instance with all the WordNet senses that are mapped to *c* in the WordNet–Macquarie mapping file (described earlier in Section 4.1). Multiple answers for an instance are given partial credit as per SemEval's scoring program.

6.5.1 Results

Table 6.1 shows the performances of the two classifiers on the training data. The system attempted to label all instances and so we report accuracy values instead of precision and recall. The naïve Bayes classifier performed markedly better in training than the PMI-based one and so was applied to the test data. (Figure 6.1 depicts the results in a graph.) The table also lists baseline results obtained when a system randomly guesses one of the possible senses for each target word. Note that since this is a completely unsupervised system, it is not privy to the dominant sense of the target words. We do not rely on the ranking of senses in WordNet as that would be an implicit use of the sense-tagged SemCor corpus. Therefore, the most-frequent-sense baseline does not apply.

Table 6.1: English Lexical Sample Task: Results obtained using the PMI-based classifier and the naïve Bayes classifier on the **training data**.

WORDS	**BASELINE**	PMI-BASED	NAÏVE BAYES
all	27.8	41.4	50.8
nouns only	25.6	43.4	53.6
verbs only	29.2	38.4	44.5

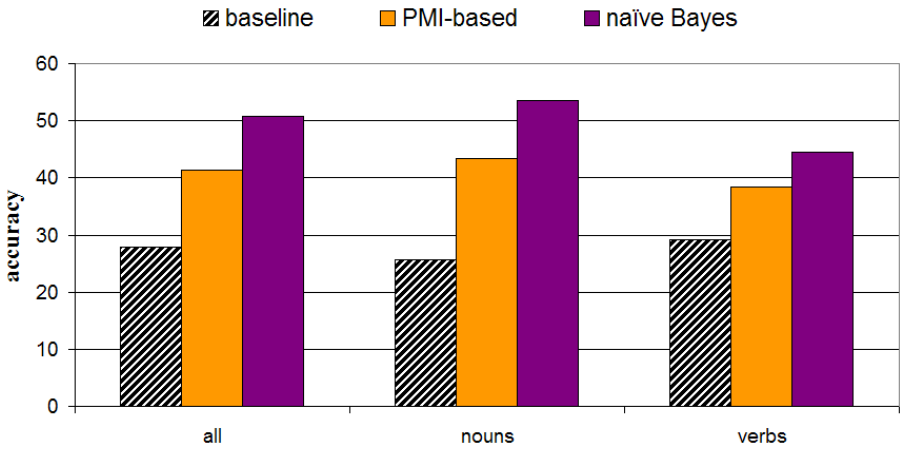


Figure 6.1: English Lexical Sample Task: Results obtained using the PMI-based classifier and the naïve Bayes classifier on the **training data**.

Table 6.2 shows results obtained by the naïve Bayes classifier on the test data. It also shows results obtained using just the prior probability and likelihood components of the naïve Bayes formula. Note that the prior probability alone gives much lower accuracies than likelihood for nouns whereas in case of verbs, prior probability does better. Overall, for all target words, the accuracy of the naïve Bayes classifier is better than that of individual components. (Figure 6.2 depicts the results in a graph.)

6.5.2 Discussion

The naïve Bayes classifier's accuracy is only about one percentage point lower than that of the best unsupervised system taking part in the task (Pradhan et al., 2007). One reason that it does better than the PMI-based one is that it takes into account prior probabilities of the categories. Further, PMI is not very accurate when dealing with low frequencies (Manning and Schütze, 1999). In case of verbs, lower combined accuracies compared to when using just prior probabilities suggests that the bag-of-words type of features are not very useful. It is expected that more syntactically oriented features will give better results. Using window sizes of $\pm 1, \pm 2,$ and ± 10 on the training data resulted in lower accuracies (exact values not shown here) than that obtained using a window of ± 5 words. A smaller window size is probably missing useful co-occurring words, whereas a larger window size is adding words that are not indicative of the target's intended sense.

The use of a sense inventory (*Macquarie Thesaurus*) different from that used to label the data (WordNet) clearly will have a negative impact on the results. The mapping from WordNet/OntoNotes to the *Macquarie Thesaurus* is likely to have some errors. Further, for 19 WordNet/OntoNotes senses, none of the annotators found a thesaurus category close enough in meaning. This meant that the system had no way of correctly disambiguating instances with these senses. Also impacting accuracy is the significantly fine-grained nature of WordNet compared to the thesaurus. For example, following are the three so-called coarse senses for the noun *president* in WordNet: (1) executive officer of a firm or college, (2) the chief executive

Table 6.2: English Lexical Sample Task: Results obtained using the naïve Bayes classifier on the **test data**.

WORDS	**BASELINE**	PRIOR	LIKELIHOOD	NAÏVE BAYES
all	27.8	37.4	49.4	52.1
nouns only	25.6	18.1	49.6	49.7
verbs only	29.2	58.9	49.1	54.7

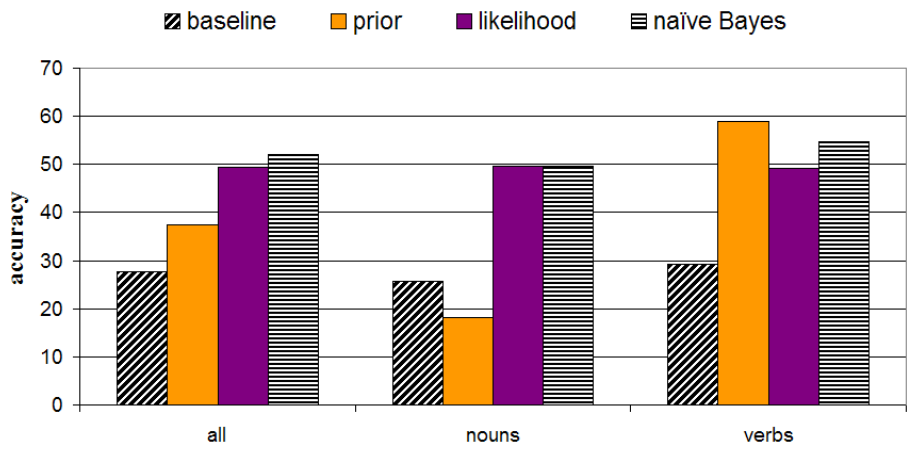


Figure 6.2: English Lexical Sample Task: Results obtained using the naïve Bayes classifier on the **test data**.

of a republic, and (3) President of the United States. The last two senses will fall into just one category for most, if not all, thesauri.

6.6 Conclusions

In this chapter, I showed how distributional profiles of concepts can be used in place of sense-annotated data. I implemented an unsupervised naïve Bayes word-sense classifier estimating its probabilities from a word–category co-occurrence matrix. It estimates the semantic distance between the senses of the target word and its context. I compared it with a baseline PMI-based classifier. Both classifiers took part in SemEval-07's English Lexical Sample task. On the training data, the naïve Bayes classifier achieved markedly better results than the PMI-based classifier and so was applied to the respective test data. On both test and training data, the classifiers achieved accuracies well above the random baseline. Further, the naïve Bayes classifier placed close to one percentage point from the best unsupervised system.

Chapter 7

Machine Translation[1]

7.1 Introduction

Cross-lingual distributional profiles of concepts (introduced in Chapter 4), are useful not only to solve natural language problems in a resource-poor language using knowledge sources from a resource-rich one (as shown in Chapter 4), but are also useful in tasks that inherently involve two or more languages. This is because the cross-lingual DPCs provide a seamless transition from words in one language to concepts in another. In this chapter, I will explore the use of cross-lingual DPCs in one such task—**machine translation (MT)**.

Machine translation is the task of automatically translating text in one language (**source**) into another (**target**). In other words, given a sentence in the source language, machine translation is the task of constructing/determining that sentence in the target language which is closest in meaning to it. For example, the following is a good input–output pair of a machine translation system:

Source sentence (in English): *You know a person by the company they keep*.

Target sentence (in German): *Das Wesen eines Menschen erkennt man an der*

[1]This chapter describes work done in collaboration with Philip Resnik, University of Maryland. Philip played a crucial role in identifying the potential of distributional profiles of concepts in a cross-lingual framework. He provided access to Chinese text used in experiments, as well. I am grateful for the insights and helpful discussions.

Gesellschaft, mitder er sich umgibt.

The need for accurate machine translation is simple and compelling: it allows understanding of foreign-language text. By *foreign language*, I mean a language that the reader does not understand. Machine translation is vital towards eliminating the language divide and allowing speakers of all languages easy access to information.

Given its significance, it is not surprising that machine translation has enjoyed the attention of a large number of researchers, giving rise to a rich diversity of approaches and ideas. In the last fifteen years, statistical machine translation has evolved as the dominant methodology. To learn more about machine translation and popular approaches, see the machine translation chapter in "*Foundations of Statistical Natural Language Processing*" (Manning and Schütze, 1999) and the "*Statistical Machine Translation*" textbook (Koehn, 2007). See Dorr et al. (1999) for a survey of approaches in machine translation and Lopez (2007) and Knight and Marcu (2005) for recent surveys of statistical machine translation.

Statistical machine translation involves learning from example translations inherent in parallel corpora—corpora that are translations of each other. However, parallel corpora are a limited resource. Like sense-annotated data, not many parallel corpora exist, and none for most language pairs. In this chapter, I show how cross-lingual distributional profiles of concepts can be useful in machine translation. Notably, my approach does not require any parallel corpora or sense-annotated data. An implementation of such a system participated in the Multilingual Chinese–English Lexical Sample Task and placed first among the unsupervised systems. It should be noted that the experiments presented here are only an initial exploration of the abilities of cross-lingual DPCs in machine translation and multilingual tasks in general. Experiments on a full-scale machine translation system are planned for the near future in collaboration with Philip Resnik (see Section 8.5.1 on future work).

7.2 The Multilingual Chinese–English Lexical Sample Task

The objective of the Multilingual Chinese–English Lexical Sample Task (Jin et al., 2007) was to select from a given list a suitable English translation of a Chinese target word in context. The training and test data had 2686 and 935 instances respectively for 40 target words (19 nouns and 21 verbs). The instances were taken from the January, February, and March 2000 editions of the *People's Daily*—a popular Chinese newspaper. The organizers used the *Chinese Semantic Dictionary (CSD)*, developed by the Institute of Computational Linguistics, Peking University, both as a sense inventory and as a bilingual lexicon (to extract a suitable English translation of the target word once the intended Chinese sense was determined). A CSD-based system can use the bilingual lexicon to determine which senses of the Chinese target word correspond to its given English translations. The system then analyzes an occurrence of the target word to determine which of these Chinese senses is intended. The instance is then labeled with the corresponding English translation.

However, one of the motivations for this task was that traditional word sense disambiguation tasks force all competing systems to work with the same sense-inventory. By presenting the sense disambiguation task in the guise of a word-translation task, such a restriction is no longer obligatory. In that spirit, my system does not use the CSD, but rather the English *Macquarie Thesaurus*. In order to determine the English translations of Chinese words in context, our system first determines the intended cross-lingual candidate sense. Recall from Section 4.2 that cross-lingual candidate senses of a target word in one language L_1 are those categories in the thesaurus of another language L_2 that are reachable by looking up the target in an L_1–L_2 bilingual lexicon and the translations in the L_2 thesaurus. See Figure 7.2 for Chinese–English examples. Also recall that they are called "candidate" senses because some of the senses of an L_2 word might not really be senses of the L_1 word—for example, CELEBRITY, PRACTICAL LESSON, and STATE OF ATMOSPHERE in the example of the figure. Using the English thesaurus instead of CSD also means that the system requires a mapping of the given English translations to the English thesaurus categories (rather than a mapping from the En-

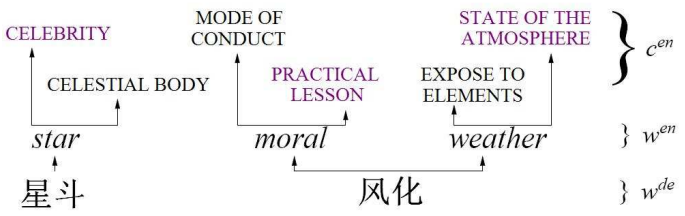


Figure 7.1: The cross-lingual candidate senses of example Chinese words. In red are concepts not really senses of the Chinese words, but simply artifacts of the translation step.

glish translations to Chinese senses—as provided by CSD). A native speaker of Chinese and I mapped the English translations of the Chinese target words to appropriate *Macquarie Thesaurus* categories—referred to as the **English translations–Macquarie category mapping**. We used three examples (from the training data) per English translation for this purpose. Once the intended sense (thesaurus category) of the Chinese word is determined, the system uses this English category–English word mapping to assign the appropriate English translation to the Chinese target word.

7.3 The cross-lingual DPC–based classifiers

In the subsections below, I will describe the two word sense classifiers that took part in the Multilingual Chinese–English Lexical Sample Task. Both use Chinese words in the context of the Chinese target word as features to determine its intended cross-lingual sense (*Macquarie Thesaurus* category). Both classifiers rely on the cross-lingual (Chinese–English) word–category co-occurrence matrix to determine the evidence towards each English cross-lingual candidate sense of the target. In chapter 4, I described how German text can be combined with an English thesaurus using a German–English bilingual lexicon to create German–English word–category co-occurrence matrix. Using the same algorithm, I now create a cross-lingual (Chinese–English) word–category co-occurrence matrix with Chinese word types w^{ch} as one

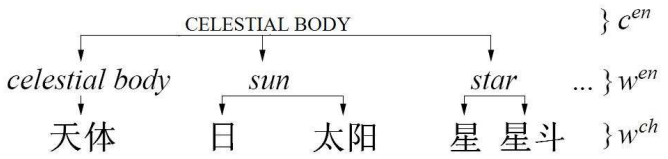


Figure 7.2: Chinese words having CELESTIAL BODY as cross-lingual candidate senses.

dimension and English thesaurus concepts c^{en} as another.

	c_1^{en}	c_2^{en}	...	c_j^{en}	...
w_1^{ch}	m_{11}	m_{12}	...	m_{1j}	...
w_2^{ch}	m_{21}	m_{22}	...	m_{2j}	...
⋮	⋮	⋮	⋱	⋮	⋮
w_i^{ch}	m_{i1}	m_{i2}	...	m_{ij}	...
⋮	⋮	⋮	...	⋮	⋱

The matrix is populated with co-occurrence counts from a large Chinese corpus; we used a collection of LDC-distributed corpora[2]—*Chinese Treebank English Parallel Corpus*, *FBIS data, Xinhua Chinese–English Parallel News Text Version 1.0 beta 2, Chinese English News Magazine Parallel Text, Chinese News Translation Text Part 1,* and *Hong Kong Parallel Text.* A particular cell m_{ij}, corresponding to word w_i^{ch} and concept c_j^{en}, is populated with the number of times the Chinese word w_i^{ch} co-occurs with any Chinese word having c_j^{en} as one of its *cross-lingual candidate senses*. For example, the cell for 太空 (SPACE) and CELESTIAL BODY will have the sum of the number of times 太空 co-occurs with 天体, 日, 太阳, 星, 星斗, and so on (see Figure 7.2). As before, we used the *Macquarie Thesaurus* (Bernard, 1986) (about 98,000 words). The possible Chinese translations of an English word were taken from the Chinese–English Translation Lexicon version 3.0 (Huang and Graff, 2002) (about 54,000 entries).

As described earlier too, this base word–category co-occurrence matrix (base WCCM), created after a first pass of the corpus, captures strong associations between a category (concept) and co-occurring words. For example, even though we increment counts for both 太空–

[2]http://www.ldc.upenn.edu

CELESTIAL BODY and 太空–CELEBRITY for a particular instance where 太空 co-occurs with 星斗, 太空 will co-occur with a number of words such as 天体, 太阳, and 日 that each have the sense of CELESTIAL BODY in common (see Figure 7.2), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count, and thereby the strength of association, of 太空 and CELESTIAL BODY will be relatively higher than that of 太空 and CELEBRITY.

Again as described for the German–English case in Chapter 4, a second pass of the corpus is made to disambiguate the (Chinese) words in it. A new bootstrapped WCCM is created by populating each cell m_{ij}, corresponding to word w_i^{ch} and concept c_j^{en}, with the number of times the Chinese word w_i^{ch} co-occurs with any Chinese word *used in cross-lingual sense* c_j^{en}.

7.3.1 Cross-lingual naïve Bayes classifier

The cross-lingual naïve Bayes classifier has the following formula to determine the intended English sense c_{nb}^{en} of the Chinese target word w_{target}^{en}:

$$c_{nb}^{en} = \underset{c_j^{en} \in C^{en}}{\operatorname{argmax}} P(c_j^{en}) \prod_{w_i^{ch} \in W^{ch}} P(w_i^{ch} | c_j^{en}) \tag{7.1}$$

where C^{en} is the set of possible senses (as listed in the *Macquarie Thesaurus*) and W^{ch} is the set of Chinese words that co-occur with the target w_{target}^{en} (we used a window of ± 5 words).

A direct approach to determine these probabilities, prior probabilities of the senses ($P(c_j^{en})$) and the conditional probabilities in the likelihood ($\prod_{w_i^{ch} \in W^{ch}} P(w_i^{ch} | c_j^{en})$), require word-aligned parallel corpora and sense-annotated corpora. Both are expensive and hard-to-find resources. I approximate these probabilities using counts from the cross-lingual word–category co-occurrence matrix, thereby obviating the need for manually-annotated data.

$$P(c_j^{en}) = \frac{\sum_i m_{ij}}{\sum_{i,j} m_{ij}} \tag{7.2}$$

$$P(w_i^{ch} | c_j^{en}) = \frac{m_{ij}}{\sum_i m_{ij}} \tag{7.3}$$

Here, m_{ij} is the number of times the Chinese word w_i^{ch} co-occurs with the English category c_j^{en}—as listed in the Chinese–English word–category co-occurrence matrix.

7.3.2 PMI-based classifier

Pointwise mutual information (PMI) between a cross-lingual candidate sense of a Chinese target word and a co-occurring Chinese word is calculated using the following formula:

$$PMI(w_i^{ch}, c_j^{en}) = \log \frac{P(w_i^{ch}, c_j^{en})}{P(w_i^{ch}) \times P(c_j^{en})} \tag{7.4}$$

$$\text{where} \quad P(w_i^{ch}, c_j^{en}) = \frac{m_{ij}}{\sum_{i,j} m_{ij}} \tag{7.5}$$

$$\text{and} \quad P(w_i^{ch}) = \frac{\sum_j m_{ij}}{\sum_{i,j} m_{ij}} \tag{7.6}$$

Here, m_{ij} is the count in the WCCM and $P(c_j)$ is as in equation 7.2. For each cross-lingual candidate sense of the Chinese target, the sum of the strength of association (PMI) between it and each of the co-occurring Chinese words (in a window of ± 5 words) is calculated. The sense with the highest sum is chosen as the intended sense.

$$c_{pmi}^{en} = \underset{c_j^{en} \in C^{en}}{\text{argmax}} \sum_{w_i^{ch} \in W^{ch}} PMI(w_i^{ch}, c_j^{en}) \tag{7.7}$$

Note that even though the PMI-based classifier uses prior probabilities of the categories $P(c_j)$ (determined from the cross-lingual WCCM) to determine the strength of association of c_j with co-occurring words, the classifier does not bias (multiply) this contextual evidence with $P(c_j)$. Since it uses only contextual evidence, I call the PMI-based classifier a baseline to the classifier described in the previous sub-section.

7.4 Evaluation

Both the naïve Bayes classifier and the PMI-based classifier were applied to the SemEval training data. For each instance, the Macquarie category, say c^{en}, that best captures the intended

Table 7.1: Multilingual Chinese–English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on the **training data**.

	BASELINE		PMI-BASED		NAÏVE BAYES	
WORDS	micro	macro	micro	macro	micro	macro
all	33.1	38.3	33.9	40.0	38.5	44.7
nouns only	41.9	43.5	43.6	45.0	49.4	50.5
verbs only	28.0	34.1	28.0	35.6	31.9	39.6

sense of the target Chinese word w_i^{ch} was determined. The system then labels an instance with all the English translations that are mapped to c in the English translations–Macquarie category mapping (described earlier in Section 7.2). Multiple answers for an instance are given partial credit as per SemEval's scoring program. However, the multilingual Chinese–English lexical sample task evaluation script did not give partial credit in case of multiple answers, and so an answer is chosen at random from the tied alternatives.

7.4.1 Results

Table 7.1 shows accuracies of the two classifiers. **Macro average** is the ratio of the number of instances correctly disambiguated to the total, whereas **micro average** is the average of the accuracies achieved on each target word. As in the English Lexical Sample Task, both classifiers, especially the naïve Bayes classifier, perform well above the baseline classifier which chooses one of the possible English translations at random. (Figure 7.3 depicts the results in a graph.) Since the naïve Bayes classifier performed markedly better than the PMI-based one too, it was applied to the test data.

Table 7.2 shows results obtained on the test data. Again the results are well above baseline. The table also presents results obtained using the individual components of the naïve Bayes classifier, likelihood and prior probability. In general, prior probabilities are less useful than

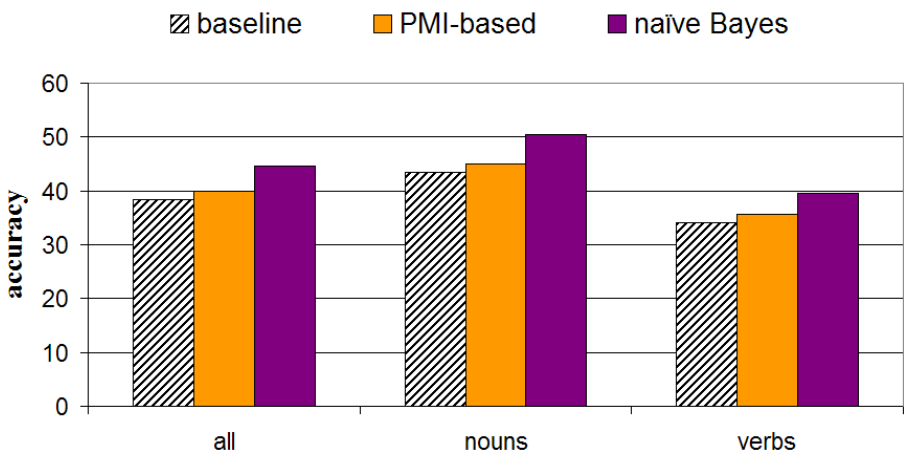


Figure 7.3: Multilingual Chinese–English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on the **training data**.

the likelihood, so much so that they are negatively impacting the overall performance in some cases. (Figure 7.4 depicts the results in a graph.)

7.4.2 Discussion

Our naïve Bayes classifier was a clear first among the two unsupervised systems taking part in the task (Jin et al., 2007). The use of a sense inventory different from that used to label the data (Macquarie as opposed to CSD) again will have a negative impact on the results as the mapping may have a few errors. The annotators believed none of the given Macquarie categories could be mapped to two Chinese Semantic Dictionary senses. This meant that our system had no way of correctly disambiguating instances with these senses.

There were also a number of cases where more than one CSD sense of a word was mapped to the same Macquarie category. This occurred for two reasons: First, the categories of the *Macquarie Thesaurus* act as very coarse senses. Second, for certain target words, the two CSD senses may be different in terms of their syntactic behavior, yet semantically very close (for example, the BE SHOCKED and SHOCKED senses of 震惊). This many-to-one mapping meant that for a number of instances more than one English translation was chosen. Since the task

Table 7.2: Multilingual Chinese–English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on **test data**.

	BASELINE		PRIOR		LIKELIHOOD		NAÏVE BAYES	
WORDS	micro	macro	micro	macro	micro	macro	micro	macro
all	33.1	38.3	35.4	41.7	38.8	44.6	37.5	43.1
nouns only	41.9	43.5	45.3	47.1	48.1	50.8	50.0	51.6
verbs only	28.0	34.1	29.1	36.8	32.9	39.0	29.6	35.5

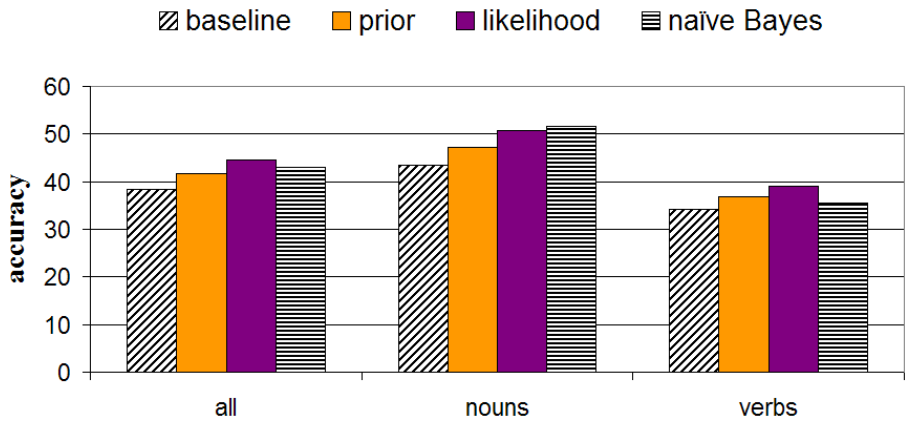


Figure 7.4: Multilingual Chinese–English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on **test data**.

required us to provide exactly one answer (and there was no partial credit in case of multiple answers), a category was chosen at random.

7.5 Conclusions

In this chapter, I showed how cross–lingual distributional profiles of concepts can be useful in tasks that inherently involve two or more languages. Specifically, I show that they are useful in machine translation. I implemented an unsupervised naïve Bayes word-sense classifier that uses cross-lingual (Chinese–English) distributional profiles of concepts to determine the intended English sense of a given Chinese word from its context. Notably, I do so without using any manually sense-annotated data or parallel corpora. Once the intended English sense was determined, the exact English translation was determined from a manually mapped file. Automated approaches for this step can also be used.

I compared the cross-lingual naïve Bayes classifier with a baseline cross-lingual PMI-based classifier. Both classifiers took part in SemEval-07's Multilingual Chinese–English Lexical Sample Task. Just as in the English Lexical Sample Task (described in the previous chapter), on the training data, the naïve Bayes classifier achieved markedly better results than the PMI-based classifier and so was applied to the test data. On both test and training data, the classifiers achieved accuracies well above the random baseline. Further, the cross-lingual naïve Bayes classifier placed first among the unsupervised systems.

Applying cross-lingual DPCs to other multilingual tasks such as multilingual document clustering, summarization, and information retrieval is an especially exciting aspect of future work (see Section 8.5 for more details).

Chapter 8

Conclusions

8.1 Distributional concept-distance

In this thesis, I have proposed a new hybrid approach that combines a published thesaurus with text to measure semantic distance. The central argument is that the semantic distance between two concepts can be accurately determined by calculating the distance between their distributional profiles. The distributional profile of a concept is the strength of association between it and each of the words that co-occur with it. The argument is similar to the distributional hypothesis—"you know a word by the company it keeps". However, there the targets are words whereas here the targets are word senses or concepts.

Determining distributional profiles of concepts is more difficult than determining distributional profiles of words, which require only simple word–word co-occurrence counts. A direct approach for estimating concept–word co-occurrence counts (needed to create DPCs) requires sense-annotated data, which is rare and expensive to create. I proposed a way to estimate these counts, and thereby the DPCs, using a bootstrapping algorithm. Notably, I do so without the use of any sense-annotated data. I use the categories in a published thesaurus (812 in all) as concepts or coarse senses. This newly proposed approach fills a void created by the many limitations of existing approaches described next.

8.2 Problems with earlier approaches

One of the contributions of this thesis was to do a comparative study of previous distance approaches and to flesh out the problems associated with them. The various WordNet-based measures have been widely studied (Budanitsky and Hirst, 2006; Patwardhan et al., 2003), and even though individual distributional measures are being used more and more, the study of distributional measures on the whole has received much less attention. In Chapter 2, I presented a detailed analysis of distributional measures and a qualitative comparison with WordNet-based measures. I summarize the key limitations of both WordNet-based and distributional word-distance measures here.

The best WordNet-based measures of concept-distance rely on an extensive hierarchy of hyponymy relationships for nouns and are therefore only good at estimating semantic similarity between nouns. They are particularly poor at estimating semantic relatedness between all other part-of-speech pairs and cross-part-of-speech pairs such as a noun–verb and adjective–noun. Further, WordNet, with more than 117,000 synsets, is a very fine-grained sense inventory (Agirre and Lopez de Lacalle Lekuona (2003) and citations therein). This itself leads to several problems: (1) Creating such an extensively connected network of concepts for another language is an arduous task. Even if there are WordNet projects in the pipeline for a handful of languages, most languages will have to make do without one. (2) Fine-grained senses may have the effect of erroneously splitting the semantic relatedness/similarity score. (3) It is computationally expensive and memory-intensive to pre-compute all sense–sense distance values—a pre-requisite for use in real-time applications.

Distributional measures of word-distance conflate all possible senses of a word, giving a dominance-based average of the distances between the senses of the target words. Therefore distributional word-distance measures perform poorly when compared to concept-distance measures, because in most natural language tasks, when given a target word pair, we usually need the distance between their closest senses. Distributional measures of word-distance, like the WordNet-based measures, also require the computation of large distance matrices ($V \times V$),

where V is the size of the vocabulary (usually at least 100,000). And finally, both distributional word-distance and WordNet-based concept-distance measures have largely been used monolingually. They do not lend themselves easily to tasks that involve multiple languages.

8.3 Features of the new approach

In contrast, distributional measures of concept-distance determine profiles of concepts (word senses) and do not conflate the distances between the many senses of a word. I have shown that they are markedly more accurate than distributional word-distance measures through experiments on a number of natural language tasks: ranking word pairs in order of their semantic distance, correcting real-word spelling errors, and solving word-choice problems. The newly proposed approach can accurately measure both semantic relatedness and semantic similarity. Further, it can do so for all part-of-speech pairs and not just for noun pairs (as in case of the best WordNet-based measures). When measuring semantic similarity between English noun pairs, the distributional concept-distance measures perform competitively, but the Jiang Conrath measure which uses WordNet does better. When measuring semantic similarity between German noun pairs, the cross-lingual distributional concept-distance measures perform better than the best monolingual GermaNet-based measures, including the Jiang Conrath measure.

I use the *Macquarie Thesaurus* categories (812 in all) as concepts. Drastic as this may seem, I have shown through experiments in a number of natural language tasks that accurate results can be obtained even with such a coarse sense-inventory. The use of thesaurus categories as concepts means that to pre-compute all distance values we now require a concept–concept distance matrix of size only 812×812—much smaller than (and about 0.01% the size of) the matrix required by traditional semantic and distributional measures. This also means that the distance between two concepts (categories) is calculated from the occurrences of *all* the words listed under those categories and so the approach largely avoids the data-sparseness problems of distributional word-distance measures (poor word–word distance estimation due to insufficient

number of occurrences of the target word(s) in a corpus).

As mentioned earlier, distributional measures of concept-distance combine text and a published thesaurus. I have shown how this can be done in a monolingual framework, with both text and thesaurus belonging to the same language, and in a cross-lingual framework, where they belong to different languages. Cross-lingual distributional profiles created in the latter case provide a seamless transition from words in one language to concepts in another. This allows the use of these cross-lingual DPCs to attempt tasks in a resource-poor language using a knowledge source from a resource-rich one. It also allows us to attempt tasks that inherently involve two or more languages, such as machine translation and multilingual information retrieval.

8.4 How the new approach helps

A large number of natural language tasks are essentially semantic-distance tasks (see Section 1.1.3 for more discussion). Thus, potentially, they can all benefit from the new distributional concept-distance approach. Certain kinds of tasks are especially well suited to take advantage of the unique features of this new approach and I describe them below along with specific conclusions from my experiments.

8.4.1 Moving from profiles of words to profiles of concepts

Firth's (1957) distributional hypothesis states that words occurring in similar contexts tend to be semantically similar. Distributional word-distance measures estimate how similar two words are by quantifying the similarity between their profiles. However, most words have more than one meaning and semantic similarity of two words can vary greatly depending on their intended senses. In Chapter 3, I showed that with distributional profiles of concepts, the same distributional measures can now be used to estimate semantic distance between word senses or concepts. Further, I showed that the newly proposed concept-distance measures are

more in line with human notions of semantic distance and attain markedly higher accuracies in (1) ranking word pairs as per their semantic distance and in (2) correcting real-word spelling errors.

8.4.2 Obviating the need for sense-annotated data

A large number of problems, such as word sense disambiguation (WSD), are traditionally approached with sense-annotated data. Other problems, such as determining word sense dominance, become trivial given such data. However, manually annotated data is expensive to create and not much exists—a problem further exacerbated by the practical need for domain-specific data for many different domains. The distributional profiles of concepts proposed in this thesis are created in an unsupervised manner and can be used in place of sense-annotated data.

In Chapter 5, I proposed methods to determine the degree of dominance of a sense of a word using distributional profiles of concepts. They achieved near-upper-bound results even when the target text was relatively small (a few hundred sentences as opposed to thousands used by other approaches). Unlike the McCarthy (2006) system, I showed that these new methods do not require large *similarly-sense-distributed* text or retraining. The methods do not perform any time-intensive operation, such as the creation of Lin's thesaurus, at run time; and they can be applied to all parts of speech—not just nouns. In the process of evaluation, I automatically generated sentences that have a target word annotated with senses from the published thesaurus. One of the future directions is to automatically generate sense-annotated data using various sense-inventories, and in different domains.

In Chapter 6, I described an unsupervised naïve Bayes word-sense classifier that estimates its prior and likelihood probabilities from the word–category co-occurrence matrix. The classifier obtained close to one percentage point from the top unsupervised system that took part in SemEval-07's English Lexical Sample task.

8.4.3 Overcoming the resource-bottleneck

In a majority of languages, estimating semantic distance, and indeed many other natural language problems, is hindered by a lack of manually created knowledge sources such as WordNet for English and GermaNet for German. In Chapter 4, I presented the idea of estimating semantic distance in one, possibly resource-poor, language using a knowledge source from another, possibly resource-rich, language. (This is work done in collaboration with Torsten Zesch and Iryna Gurevych of the Darmstadt University of Technology.) I did so by creating cross-lingual distributional profiles of concepts, using a bilingual lexicon and a bootstrapping algorithm, but without the use of any sense-annotated data or word-aligned corpora. The cross-lingual measures of semantic distance were evaluated on two tasks: (1) estimating semantic distance between words and ranking the word pairs according to semantic distance, and (2) solving *Reader's Digest* 'Word Power' problems. "Gold standard" evaluation data for these tasks were compiled by Zesch and Gurevych. We compared results with those obtained by conventional state-of-the-art monolingual approaches to determine the amount of loss-in-accuracy due to the translation step. Apart from traditional information-content measures proposed by Resnik (1995) and Jiang and Conrath (1997), we also compared the cross-lingual distributional measures with Lesk-like measures proposed specifically for GermaNet (Gurevych, 2005).

The thesaurus-based cross-lingual approach gave much better results than monolingual approaches that do not use a knowledge source. Further, in both tasks and all the experiments, the cross-lingual measures performed as well if not slightly better than the GermaNet-based monolingual approaches. This shows that the proposed cross-lingual approach, while allowing the use of a superior knowledge source from another language, is able to keep at a minimum losses due to the translation step. We show that in languages that lack a knowledge source, large gains in accuracy can be obtained by using the proposed cross-lingual approach. Further, even if the language has a semi-developed knowledge source, better results can be obtained by using the cross-lingual approach and a superior knowledge source from another language.

8.4.4 Crossing the language barrier

Cross-lingual distributional profiles of concepts (introduced in Chapter 4), are useful not only to solve natural language problems in a resource-poor language with knowledge sources from a resource-rich one (as shown in Chapter 4), but are also useful in tasks that inherently involve two or more languages. In Chapter 7, I showed how cross–lingual distributional profiles of concepts can be useful in machine translation. I implemented an unsupervised naïve Bayes word-sense classifier that uses cross-lingual (Chinese–English) distributional profiles of concepts to determine the intended English sense of a given Chinese word from its context. Notably, I did so without using any manually sense-annotated data or parallel corpora. The classifier placed first among the unsupervised systems that took part in SemEval-07's Multilingual Chinese–English Lexical Sample Task.

Applying cross-lingual DPCs to other multilingual tasks such as multilingual document clustering, summarization, and information retrieval is an especially exciting aspect of future work (see sub-sections 8.5.4, 8.5.2, and 8.5.3 ahead).

8.5 Future directions

Future work on this topic can be divided into two kinds: (1) improving the estimation of DPCs by using better algorithms, task-suited sense-inventories, and syntactic information; and (2) using the DPC-based approach in a variety of other natural language tasks that can take advantage of its features.

I have used categories in the thesaurus as concepts. However, most published thesauri divide categories into paragraphs, and paragraphs into semicolon groups. On certain tasks it may be more beneficial to use these as less-coarse sense-inventories. Also, I have used all words in a category, irrespective of their part of speech. It will be interesting to determine the role of different parts of speech in different tasks. It is also worth comparing performance of thesaurus-based DPCs with those created from other knowledge sources, especially Wikipedia.

Wikipedia is appealing, not just because it is created by the community (75,000 active contributors) and so in many ways reflects language and concepts as they are used and understood, but also because of its very high coverage (5.3 million articles in 100 languages). Of course, it is challenging to organize Wikipedia concepts as in a published thesaurus (see Zesch et al. (2007a) and Milne et al. (2006) for some exploration in this area); the DPC-based approach of using a thesaurus to estimate semantic distance can be used to evaluate different thesaurus-representations of Wikipedia.

The distributional profiles of concepts I used were calculated from simple word–concept co-occurrences without incorporating any syntactic information. Yet they have achieved competitive results in various natural language tasks. The next step will be to use only those co-occurring words that stand in certain syntactic relations, such as verb–object and subject–verb, with each other and determine if that leads to significant improvements in accuracies of DPC-based applications.

Finally, the ideas presented in this thesis can be applied to a number of natural language tasks. Below are some of the applications that I am especially interested in.

8.5.1 Machine translation

The experiments described in Chapter 7 constitute only the first stage of using cross-lingual DPCs for machine translation (MT). I intend to determine (domain-specific) probabilities of possible English translations of Chinese words and use them as prior probabilities in a full-fledged MT system. The next step will be to do the same for phrases. I am also interested in determining how useful cross-lingual DPCs are in choosing the correct target hypothesis from the top k that an MT system picks. It is worth determining whether combining a traditional word-based language model with a concept-based language model will improve results.

8.5.2 Multilingual multi-document summarization

Consider the task of summarizing several news articles about the same event that are written in several different languages. Conventional approaches may, at best, find concepts pertinent only within the scope of each document and include them in the summary. Further, identifying sentences that convey more or less the same information across articles in different languages is a problem. Using my algorithm we can create different DPCs corresponding to words in each language and concepts as per *one common* inventory of senses. Therefore, we can determine concepts that are deemed significant by the document set as a whole and also identify sentences that convey more or less the same information to create a more pertinent and non-redundant summary.

8.5.3 Multilingual information retrieval

State-of-the-art models in information retrieval (IR) have traditionally made certain independence assumptions so that their models remain relatively simple. They assume that different word-types in the document and query are independent. However, if the query has a term *scalpel*, then we would want the system to score a document higher if it has *surgeon* as opposed to another completely unrelated word. Recently, there has been some encouraging work incorporating such dependencies and semantic relations into IR systems (Cao et al., 2005; Gao et al., 2004). However these methods are computationally expensive. As my approach uses a very coarse sense-inventory (only 812 concepts), it can easily pre-compute semantic relatedness values between all concepts pairs and use it to estimate term dependencies. Even so, I believe the crucial benefit of my approach will be in cross-lingual IR, where the documents and queries belong to different languages; using cross-lingual DPCs not only can we place all queries and documents in the same concept-space (as described in the previous section), we can also incorporate term dependencies between terms that belong to different languages.

8.5.4 Multilingual document clustering

A large number of tasks such as estimating semantic distance and document clustering involve vector representations of linguistic units such as words, concepts, documents and so on in word-space. This has largely worked out nicely; however, in certain tasks the approach can fall on its face, a case in point being multilingual document clustering. If words in a document are used as features, then no two documents from different languages will be grouped together, even though they may have similar content. Part of my future work is to represent documents in the same concept-space by using cross-lingual distributional profiles of concepts. For example, if the document pool has articles in English, Spanish, and Russian, then I can use English–English, Spanish–English, and Russian–English distributional profiles to represent each document in an English thesaurus's concept-space. Then a standard clustering algorithm can be used to group them according to content.

As part of a monolingual baseline for this, I have already conducted some experiments in collaboration with Yaroslav Riabinin[1]. We used the *Reuters-21578* corpus for our experiments. It had 21,578 documents; 3,746 of these were labeled with more than one topic/class and were discarded. The baseline system simply replaced every word in a document with its coarse senses (thesaurus categories) and applied the *k*-means clustering algorithm. It obtained a purity of 0.79.[2]

The state-of-the-art bag-of-words model, which clusters the documents using words as features, obtained a purity of 0.86 (which is comparable to published results on this corpus). However, as pointed out, that approach will not work on multilingual data. The next step will be to cluster a multi-lingual dataset using a cross-lingual baseline—replace each word with its cross-lingual candidate senses. Then more sophisticated systems can be developed that make

[1] Yaroslav Riabinin was a fourth-year undergraduate student at the University of Toronto when we collaborated. He is now a graduate student in the same university.

[2] Purity is one of the standard metrics used to evaluate automatic document clustering. It is the proportion of documents clustered correctly. A document is considered to be clustered correctly if it is placed in a cluster where documents similar to it form the majority.

use of cross-lingual distributional profiles.

8.5.5 Enriching ontologies

Human-created knowledge sources, such as WordNet and Roget's Thesaurus, are widely used to solve natural language problems. However, as language evolves and new words are coined in different domains, the lack of coverage becomes an issue. I am presently developing an algorithm to supplement a published thesaurus with new and previously unknown words. This is essentially a classification task where a category of the thesaurus that best represents the usages of the unknown target is to be chosen. One way of doing this is to represent each category by a vector in some multi-dimensional feature-space. A traditional word-distribution approach to this will require sense-annotated data and the categories will be represented in (high-dimensional) word-space. My algorithm uses DPCs to estimate these vectors, therefore does *not* require sense-annotated data, and places the vectors in low-dimensional category-space. Initial experiments show a large gain over the baseline. I am also interested in using the DPCs to automatically enrich an ontology with more information, such as identifying lexical entailment (Mirkin et al., 2007) and antonymy (Muehleisen, 1997; Lucero et al., 2004).

8.5.6 Word prediction/completion

Word prediction or completion is the task of predicting or completing a partially typed word using the preceding context as cue. Unigram and bigram models (Nantais et al., 2001) and those combined with some part-of-speech information (Fazly, 2002) have been shown to perform reasonably well; yet there is plenty of room for improvement. It will be interesting to determine if their performance can be improved on by using them in combination with measures of semantic distance. The hypothesis is that given a list of possible words, the intended one is that which is closely related to the preceding context (Li, 2006; Li and Hirst, 2005).

8.5.7 Text segmentation

A document may be about a certain broad topic, but different portions of the document tend to be about different sub-topics. Text segmentation is the task of partitioning a document into (possibly multi-paragraph) units or passages such that each passage is about a different sub-topic than the ones adjacent to it. These passages are characterized by the presence of more-than-random number of semantically related words (Morris and Hirst, 1991; Halliday and Hasan, 1976). Further, these semantically related words may belong to different parts of speech. Identifying such links between words (possibly to form lexical chains) is crucial to automatic segmentation. However, WordNet-based measures of semantic distance are good only at estimating semantic similarity between nouns and distributional word-distance measures are much less accurate. The distributional concept-distance approach proposed in this thesis has neither limitation. Further, as I use a very coarse sense-inventory (thesaurus categories), the method is expected to yield more, longer, and accurate lexical chains. Hearst (1997) uses word–word co-occurrence distributions for text segmentation with encouraging results. However, those distributions suffer from problems due to word sense ambiguity. Distributional profiles of concepts provide a natural way to determine associations between co-occurring word senses. All these factors suggest that the ideas proposed in this thesis hold promise in text segmentation as well.

Bibliography

Eneko Agirre and Oier Lopez de Lacalle Lekuona. 2003. Clustering WordNet word senses. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, Borovets, Bulgaria.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the Association for Computational Linguistics Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 805–810, Acapulco, Mexico.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL,HLT-2003)*, pages 16–23, Edmonton, Canada.

John R. L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services, Oxford, England.

Guihong Cao, Jian-Yun Nie, and Jing Bai. 2005. Integrating word relationships into language models. In *Proceedings of the 28th Annual International Association for Computing Ma-*

chinery Special Interest Group on Information Retrieval (SIGIR-2005), pages 298–305, Salvador, Brazil.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Unsupervised feature selection for relation extraction. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-2005)*, Jeju Island, Republic of Korea.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Association for Computational Linguistics Workshop on A Broader Perspective on Multiword Expressions (Poster Session)*, pages 41–48, Prague, Czech Republic.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the Association for Computational Linguistics Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan.

D. Allen Cruse. 1986. *Lexical semantics*. Cambridge University Press, Cambridge, UK.

Silviu Cucerzan and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th Conference on Computational Natural Language Learning*, pages 132–138, Taipei, Taiwan.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics (ACL-1994)*, pages 272–278, Las Cruces, New Mexico.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL-1997)*, pages 56–63, Madrid, Spain.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69.

Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. 1999. A survey of current research in machine translation. *Advances in Computers*, 49:1–68.

Afsaneh Fazly. 2002. The use of syntax in word completion utilities. Master's thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Ol'ga Feiguina and Graeme Hirst. 2007. Authorship attribution for small texts: Literary and forensic experiments. In *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 30th Annual International ACM SIGIR Conference (SIGIR-2007)*, Amsterdam, Netherlands.

Christianne Fellbaum, editor. 1998. *WordNet. An electronic lexical database*. MIT Press, Cambridge, MA.

Óscar Ferrández, Rafael M. Terol, Rafael Muniõz, Patricio Martínez-Barco, and Manuel Palomar. 2006. An approach based on logic forms and WordNet relationships to textual entailment performance. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 22–26, Venice, Italy.

Olivier Ferret. 2004. Discovering word senses from a network of lexical cooccurrences. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 1326–1332, Geneva, Switzerland.

John R. Firth. 1957. A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32, Oxford, England. The Philological Society.

William Gale, Kenneth W. Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-1992)*, pages 249–256, Newark, Delaware.

Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence language model for information retrieval. In *Proceedings of the 28th Annual International Association for Computing Machinery Special Interest Group on Information Retrieval (SIGIR-2004)*, pages 25–29, Sheffield, UK.

Gregory Grefenstette. 1992. Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-1992)*, pages 324–326, Newark, Delaware.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 767–778, Jeju Island, Republic of Korea.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770, Geneva, Switzerland.

Aria Haghighi, Andrew Ng, and Christopher Manning. 2005. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT,EMNLP-2005)*, pages 387–394, Vancouver, Canada.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. Cohesion in english. *Longman*.

Donna Harman. 1993. Overview of the first text retrieval conference. In *Proceedings of the 16th Annual International Association for Computing Machinery Special Interest Group on Information Retrieval (ACM-SIGIR) conference on Research and Development in information retrieval*, pages 36–47, Pittsburgh, Pennsylvania.

Zellig Harris. 1968. *Mathematical Structures of Language*. Interscience Publishers, New York, NY.

Hany Hassan, Ahmed Hassan, and Ossama Emam. 2006. Unsupervised information extraction approach using graph mutual reinforcement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 501–508, Sydney, Australia.

Marti Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association of Computational Linguistics (ACL-1990)*, pages 268–275, Pittsburgh, Pennsylvania.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 13, pages 305–332. The MIT Press, Cambridge, MA.

Veronique Hoste, Anne Kool, and Walter Daelemans. 2001. Classifier optimization and combination in the English all-words task. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 83–86, Toulouse, France.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 57–60, New York, NY.

Shudong Huang and David Graff. 2002. Chinese–english translation lexicon version 3.0. *Linguistic Data Consortium*.

Diana Inkpen and Alain Desilets. 2005. Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT,EMNLP-2005)*, pages 49–56, Vancouver, British Columbia.

Diana Inkpen and Graeme Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Linguistics*, Philadelphia, PA.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, Borovets, Bulgaria.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics (ROCLING X)*, Taipei, Taiwan.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample task. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Prague, Czech Republic.

Kevin Knight and Daniel Marcu. 2005. Machine translation in year 2004. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, pages 45–48, Philadelphia, PA.

Philip Koehn. 2007. *Statistical Machine Translation*. Cambridge University Press, Cambridge, UK.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pages 103–110, Pittsburgh, Pennsylvania.

Abolfazl K. Lamjiri, Leila Kosseim, and Thiruvengadam Radhakrishnan. 2007. A hybrid unification method for question answering in closed domains. In *Proceedings of the 3rd International Workshop on Knowledge and Reasoning for Answering Questions (KRAQ-2007) – A Workshop of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 36–42, Hyderabad, India.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.

Mirella Lapata and Frank Keller. 2007. An information retrieval approach to sense ranking. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT,NAACL-2007)*, pages 348–355, Rochester.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 11, pages 265–283. The MIT Press, Cambridge, MA.

Claudia Leacock, Martin Chodorow, and George Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th conference on Association for Computational Linguistics (ACL-1999)*, pages 25–32, College Park, Maryland.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of the Eigth International Workshop on Artificial Intelligence and Statistics (AISTATS-2001)*, pages 65–72, Key West, Florida.

Michael Lesk. 1986. 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*, pages 24–26, Toronto, Canada.

Jianhua Li. 2006. Modelling semantic knowledge for a word completion task. Master's thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Jianhua Li and Graeme Hirst. 2005. Semantic knowledge in a word completion task. In *Proceedings of the 7th International Association for Computing Machinery Special Interest Group on Accessible Computing Conference on Computers and Accessibility (SIGACCESS-2005)*, pages 18–23, Baltimore, MD.

Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. 2006. Extractive summarization using inter- and intra event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 369–376, Sydney, Australia.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL,EACL-1997)*, pages 64–71, Madrid, Spain.

Dekang Lin. 1998a. Automatic retreival and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, pages 768–773, Montreal, Canada.

Dekang Lin. 1998b. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, pages 768–773, Montreal, Canada.

Dekang Lin. 1998c. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco, CA. Morgan Kaufmann.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1492–1493, Acapulco, Mexico.

Adam Lopez. 2007. A survey of statistical machine translation. In *Technical report 2006-47*, University of Maryland, Insitute for Advanced Computer Studies, College Park, Maryland.

Cupertino Lucero, David Pinto, and Hctor Jimnez-Salazar. 2004. A tool for automatic detection of antonymy relations. In *Proceedings of the Workshop on Lexical Resources and the Web for Word Sense Disambiguation (IBERAMIA-2004)*, Peubla, Mexico.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, pages 256–263, Seattle, WA.

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 17–24, Trento, Italy.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004a. Automatic identification of infrequent word senses. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1220–1226, Geneva, Switzerland.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 280–267, Barcelona, Spain.

Susan McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

Rada Mihalcea and Dan I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-1999)*, pages 461–466, Orlando, Florida.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

David Milne, Olena Medelyan, and Ian H. Witten. 2006. Mining domain-specific thesauri from wikipedia: A case study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence table of contents*, pages 442–448, Hong Kong, China.

Shachar Mirkin, Ido Dagan, and Maayan Geffet. 2007. Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics Main Conference Poster Sessions*, pages 579–586, Sydney, Australia.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007a. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, pages 571–580, Prague, Czech Republic.

Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. *In submission*, http://www.cs.toronto.edu/compling/Publications.

Saif Mohammad and Graeme Hirst. 2006a. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 121–128, Trento, Italy.

Saif Mohammad and Graeme Hirst. 2006b. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 35–43, Sydney, Australia.

Saif Mohammad, Graeme Hirst, and Philip Resnik. 2007b. Tor, tormd: Distributional profiles of concepts for unsupervised word sense disambigution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-07)*, pages 326–333, Prague, Czech Republic.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–51, Boston, Massachusetts.

Victoria Muehleisen. 1997. *Antonymy and Semantic Range in English*. Ph.D. thesis, Department of Linguistics, Northwestern University, Evanston, IL.

Tom Nantais, Fraser Shein, and Mattias Johansson. 2001. Efficacy of the word prediction algorithm in wordq. In *Proceedings of the Annual Conference of Rehabilitation Engineering and Assistive Technology Society of North America (RESNA-2001)*, pages 77–79, Washington, D.C.

Hwee Tou Ng and Hian B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-1996)*, pages 40–47, Santa Cruz, California.

Patrick Pantel. 2005. Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 125–132, Ann Arbor, Michigan.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. 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)*, pages 17–21, Mexico City, Mexico.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2007. Umnd1: Unsupervised word sense disambiguation using contextual semantic relatedness. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 390–393, Prague, Czech Republic.

Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop Making Sense of Sense—Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.

Douglas B. Paul. 1991. Experience with a stack decoder-based hmm csr and back-off n-gram language models. In *Proceedings of the Speech and Natural Language Workshop*, pages 284–288, Palo Alto, California.

Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, pages 79–86, Pittsburgh, Pennsylvania.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics (ACL-1993)*, pages 183–190, Columbus, Ohio.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2006)*, pages 192–199, New York, NY.

Sameer Pradhan, Martha Palmer, and Edward Loper. 2007. SemEval-2007 task 17: English lexical sample, English SRL and English all-words tasks. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SemEval-2007)*, Prague, Czech Republic.

Michael Pucher. 2006. Performance evaluation of WordNet-based semantic relatedness measures for word prediction in coversational speech. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)*, pages 332–342, Tilburg, Netherlands.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

Rajit Raina, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via learning and abductive reasoning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-2005)*, pages 1099–1105, Pittsburgh, Pennsylvania.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Machine Translation Summit IX*, pages 315–322, New Orleans, Louisiana.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Philip Resnik. 1998. Wordnet and class-based probabilities. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 239–263. The MIT Press, Cambridge, Massachusetts.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Communications of the Association of Computing Machinery*, 11:95–130.

Philip Resnik and Mona Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 399–404, Philadelphia, Pennsylvania.

Herbert Rubenstein and John B. Goodenough. 1965a. Contextual correlates of synonymy. *Communications of the Association of Computing Machinery*, 8(10):627–633.

Herbert Rubenstein and John B. Goodenough. 1965b. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Mark Sanderson and Cornelius J. van Rijsbergen. 1999. The impact on retrieval effectiveness of skewed frequency distributions. *Association of Computing Machinery Transactions on Information Systems (TOIS)*, 17(4):440–465.

Frank Schilder and Bridget Thomson McInnnes. 2006. Word and tree-based similarities for textual entailment. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 140–145, Venice, Italy.

Hinrich Schütze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retreival. *Information Processing and Management*, 33(3):307–318.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *Proceedings of the Knowledge Discovery and Data Mining Workshop on Text Mining (KDD-2000)*, Boston, MA.

Mark Stevenson and Mark Greenwood. 2005. A semantic approach to ie pattern induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 379–386, Ann Arbor, Michigan.

Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on Information and knowledge management (IKM-1993)*, pages 67–74, Washington, D.C.

Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. 2002. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th Association of Computing Machinery SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 32–41, Edmonton, Canada.

Vivian Tsang and Suzanne Stevenson. 2004. Calculating semantic distance between word sense probability distributions. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 81–88, Boston, MA.

Vivian Tsang and Suzanne Stevenson. 2006. Context comparison as a minimum cost flow problem. In *Proceedings of the HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*, New York, NY.

Peter Turney. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freiburg, Germany.

Peter Turney. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 313–320, Sydney, Australia.

Cornelius J. van Rijsbergen. 1979. *Information Retrieval*. London: Butterworths.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, and Evangelos E. Milios. 2005. Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proceedings of the 7th Annual Association of Computing Machinery International Workshop on Web Information and Data Management*, pages 10–16, Bremen, Germany.

Jean Véronis. 2004. Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language. Special Issue on Word Sense Disambiguation*, 18(3):223–252.

DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Jan 2001–Dec 2005. Verlag Das Beste, Stuttgart.

Yong Wang and Julia Hodges. 2006. Document clustering with semantic analysis. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, Washington, D.C.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1015–1021, Geneva, Switzerland.

Julie E. Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Department of Informatics, University of Sussex, Brighton, UK.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia.

Dongqiang Yang and David Powers. 2005. Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the Twenty-eighth Australasian Computer Science Conference (ACSC-2005)*, pages 315–322, Newcastle, Australia.

Dongqiang Yang and David Powers. 2006a. Verb similarity on the taxonomy of WordNet. In *Proceedings of the Third International WordNet Conference (GWC-2006)*, pages 121–128, Jeju Island, Republic of Korea.

Dongqiang Yang and David Powers. 2006b. Word sense disambiguation using lexical cohesion in the context. In *Proceedings of the Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING,ACL-2006)*, pages 929–936, Sydney Australia.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, pages 454–460, Nantes, France.

Sen Yoshida, Takashi Yukawa, and Kazuhiro Kuwabara. 2003. Constructing and examining personalized cooccurrence-based thesauri on web pages. In *Proceedings of the 12th International World Wide Web Conference*, pages 20–24, Budapest, Hungary.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 401–408, Sydney, Austrailia.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007a. Analyzing and accessing Wikipedia as a lexical semantic resource. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen, Tübingen, Germany.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007b. 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-2007)*, pages 205–208, Rochester, New York.

Xiadon Zhu and Gerald Penn. 2005. Evaluation of sentence selection for speech summarization. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-2005), Workshop on Crossing Barriers in Text Summarization Research*, pages 39–45, Borovets, Bulgaria.