UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Discovery of Noun Semantic Relations based on Sentential Context Analysis

Rumen Valentinov Moraliyski

Tese para a obtenção do Grau de Doutor em

Engenharia Informática

(3º ciclo de estudos)

Orientador: Prof. Doutor Gaël Harry Dias

Covilhã, Fevereiro de 2013

The work on this dissertation was supported by the Programa Operacional Potencial Humano of the Quadro de Referência Estratégico Nacional, by the PhD grant SFRH/BD/19909/2004 of the Fundação para a Ciência e a Tecnologia (FCT) and a 9 months grant of the project MEDON PTDC/EIA/80772/2006 funded by FCT.

Acknowledgements

With pleasure I begin to write this section about the people who helped me to get to this stage.

I would like to thank first my PhD advisor, Gaël Dias for his moral support, for being incessant source of ideas and for the positive work atmosphere he created for me and my colleagues. Thank you for all the help and motivation and for the challenges you made me to face!

To my colleague João Cordeiro who provided his data. That resource gave ground for the most interesting part of my work. Thanks go to him, as well, for the help with LaTeX. He also contributed to parts of Chapter 5.

I would also like to thank my colleagues Isabel and Sebastião for the company throughout these years, for making me not miss home, for the support and for the delicious meals that we prepared together.

The main culprits to take this direction in my life is my family and my professors from the University of Plovdiv "Paisii Hilendarski". Thus I owe my curiosity and critical view about the world to my grandfather Kiril. Then, my parents and environment nourished it. Later, came my professors from the Faculty of Mathematics and Informatics of the University of Plovdiv, especially Professor Petko Proinov and Professor Todor Mollov who were impressive with their enthusiasm about mathematics and teaching. For me and my colleagues at the Faculty, the greatest and the most persistent motivator was and continues to be Professor Dimitar Mekerov. I also received invaluable support from Professor Asen Rahnev, who is completely dedicated to the faculty and educational initiatives of all kinds.

The connection between both sides, the University of Plovdiv and the environment where I executed the work presented here, the University of Beira Interior are Professor Veska Noncheva and my colleague and friend Zornitsa Kozareva who always helped me to improve my texts and subjected my work to a constructive criticism.

At the end, I would especially like to thank my dear cousin Dari who read and edited parts of the text.

Abstract

The last years saw a surge in the statistical processing of natural language and in particular in corpus based methods oriented to language acquisition. Polysemy is pointed at as the main obstacle to many tasks in the area and to thesaurus construction in particular. This dissertation summarizes the current results of a work on automatic synonymy discovery. The accent is focused on the difficulties that spring from polysemy and on linguistically and empirically motivated means to deal with it. In particular, we propose an unsupervised method to identify word usage profiles pertinent to specific word meanings.

Further, we show that the routine to verify every possibility in search of semantic relations is not only computationally expensive but is rather counterproductive. As a consequence, we propose an application of a recently developed system for paraphrases extraction and alignment so that the exhaustive search is avoided in an unsupervised manner. This led to a method, that creates short lists of pairs of words that are highly probable to be in synonymy relation.

The results show that the negative impact of polysemy is significantly reduced for part of the polysemy specter that covers about two thirds of the vocabulary. Besides the increased probability to discover frequently manifested synonymy relations, paraphrase alignment proved to highlight infrequent word meanings, and to reliably identify a set of very specific semantic relations.

Keywords

synonymy, lexical analysis, semantic relations, natural language processing

Resumo

Nos últimos anos surgiu um aumento no tratamento estatístico da linguagem natural, em particular nos métodos baseados em corpus orientados para a compreensão da linguagem. A polissemia foi apontada como o principal obstáculo para muitas tarefas nesta área, onde se destaca a construção de dicionários de sinónimos.

Esta dissertação resume os resultados atuais de um trabalho que tem como objetivo a descoberta de sinónimos de modo automático. A ênfase recai sobre as dificuldades que advêm da polissemia onde as mesmas são superadas através de métodos linguísticos e empíricos. Propomos um método não supervisionado para fazer a comparação entre os diversos perfis de uso de palavras polissémicas. Esta é a nossa principal contribuição.

Além disso, mostramos que as formas habituais de verificar todas as possibilidades na busca de relações semânticas, têm um custo computacional elevado e não apresentam resultados satisfatórios. São contraproducente.

Como consequência, propomos a utilização de um sistema recentemente desenvolvido para a extração e alinhamento de paráfrases. Assim, conseguimos evitar de forma não supervisionada a busca exaustiva e criar listas curtas de pares de palavras que são altamente prováveis de estarem em relação de sinonímia.

Os resultados mostram que o impacto negativo da polissemia é significativamente reduzido para uma fração do espectro da polissemia que abrange cerca de dois terços do vocabulário. Obtivemos probabilidades elevadas para descobrir relações de sinonímia que se manifestam frequentemente. Conseguimos também provar que, a partir do alinhamento de paráfrases, se destaca o significado de palavras não frequentes e é possível identificar com segurança um conjunto de relações semânticas específicas.

Palavras-chave

sinonímia, análise lexical, relações semânticas, processamento da linguagem natural

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of variance.
AP	Associated Press.
API	Application Programming Interface.
BNC	British National Corpus.
CR	Co-occurrence Retrieval.
EI	Equivalence Index.
ESL	English as a Second Language.
HAL	Hyperspace Analogue to Language.
IE	Information Extraction.
IR	Information Retrieval.
LSA	Latent Semantic Analysis.
LSI	Latent Semantic Indexing.
MFS	Maximal Frequent Sequence.
MRD	Machine Readable Dictionary.
NANC	North American News Corpus.
NLP	Natural Language Processing.
PMI	Pointwise Mutual Information.
POS	Part-Of-Speech.
QT	Quality Threshold.
RCV1	Reuters Corpus Volume 1.
RD	Reader's Digest.
RI	Random Indexing.
SVD	Singular Value Decomposition.
TOEFL	Test of English as a Foreign Language.
USFFAN	University of South Florida Free Association Norms.
WCSD	Web Corpus for Synonym Detection.
WSD	Word Sense Disambiguation.
WSJ	Wall Street Journal.

Chapter 1

Introduction

Thesauri, that list the most salient semantic relations between words have mostly been compiled manually. Therefore, the inclusion of an entry depends on the subjective decision of the lexicographer. Unfortunately, those resources are incomplete. Indeed, thesauri unlikely include syntagmatic semantic relations[1]. Levin (1993) is certainly the most comprehensive effort, to date, to categorize the verb part of the vocabulary with respect to the kind of constructions a word can participate in. Consider the following simple sentence: *The words of a phrase relate in many ways to each other.* Probably only the pair $<$*word , phrase*$>$ would be listed in a manual resource with its semantic relation, but interpretation clues for a polysemous word like *way* would be more difficult to code in a thesaurus.

In text understanding, humans are capable, up to a variable extent, of uncovering these relations (Morris & Hirst, 2004). Natural Language Processing (NLP) systems, however, need either a complete inventory of the semantic relations or a module capable to infer them from the text in order to perform human like interpretation. There exist attempts towards a number of scientific and practical directions. The early endeavors were limited to find empirical evidences, based on few manually examined examples, that when words share contexts they are somehow semantically and associatively related. The consequent developments were to test how far one can reach exploiting this kind of information. It turned out that there were many different ways to define what is context - some took the context to be the entire discourse, while others focused on more specific syntactic relations. The results and the possible applications varied from test solving through relatively successful attempts at thesaurus construction. However, most methods were unaware of what kinds of semantic relations they do register, thus the entries they produced were rather heterogeneous. The necessity to introduce concrete knowledge was met by pattern based techniques, that when combined with the distributional analysis, were able, for example, to distinguish, quite accurately, between synonyms and antonyms or between words of different levels of specificity.

[1] Syntagmatic relation are the relations that hold between the elements that co-occur in order to form a unit. In this sense, syntagmatic relations are the various ways in which words within the same sentence may be related to each other.

1.1 Synonymy

The main focus below will fall on the synonymy semantic relationship. In the literature, it is discussed the paradox that the notion of synonymy is intuitively clear by, usually, some very informal definition. For example, in WordNet (Fellbaum, 1998) reads:

> synonym, equivalent word - (two words that can be interchanged in a context are said to be synonymous relative to that context).

This notion is incomplete as it does not mention anything about similarity in meanings. Probably a better definition is as in the Merriam-Webster On-Line[1] where we can read:

> synonym - one of two or more words or expressions of the same language that have the same or nearly the same meaning in some or all senses.

A discussion in Charles (2000) is based on a formulation of Gottfried Leibniz, that states "*two words are synonymous if they are interchangeable in statements without a change in the truth value of the statements where the substitution occurs*". Some authors prefer to avoid the term synonymy in favor of the term near-synonymy since *there are very few absolute synonyms, if they exist at all. So-called dictionaries of synonyms actually contain near-synonyms* (Inkpen & Hirst, 2006). However, a more common view of synonymy is as a continuous scale where the judged sense similarity is proportional to the amount of the occasions when both words can be used in order to express the same idea. This observation and existing empirical data supports the assumption that synonyms can be discovered by means of lexical analysis.

Throughout the exposition below we will discuss three more semantic relations that are close siblings of synonymy. They all derive from the *hyponym / hypernym* relation. A definition given by Miller *et al.* (1990) goes as follows:

> a concept expressed by the words $\{x, x', \dots\}$ is said to be a hyponym of the concept expressed by the words $\{y, y', \dots\}$ if native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y*, for example *A crane is a (kind of) bird*.

The hypernym of a concept is frequently called subsumer in order to emphasize the hierarchical relation. In the literature, this relation is more commonly known as *Is a* relation. We will prefer to use *Is a* when it is not necessary to specify the exact direction of the relation.

[1] `http://www.merriam-webster.com/` [*13th July, 2011*]

The third relation is referred to as *Instance of*. As its name suggests, it relates proper names with more general concepts. For example, *Ronaldo* can be defined as an *instance of* a *football player*.

The last relation we need to define here is *Co-hyponymy*. Two words are said to be co-hyponyms when they have common direct subsumer. For example, according to WordNet, the words *car* and *motorcycle* are in a *co-hyponymy* relation as they share the same direct hypernym, which is *motor vehicle*.

It is frequently difficult to distinguish synonymy from the other relations if only the quality of substitutability is considered. Thus, words may have contexts in common in spite of dissimilarity in meanings. Further, words with disparate meanings still can have contexts in common and Charles (2000) found that occasionally they can be substituted even within a sentence frame.

1.2 Objectives of Synonym Discovery

Synonymy being common among all the parts of speech is considered to be the main organizing principle of the human semantic memory (see Miller *et al.* (1990) for comprehensive reference list). The objective pursued here is to develop a synonymy discovery procedure. The definition states clearly the objective of synonymy discovery, it suggests the method also. In order to contribute with new qualities to the works developed before, our method needs to be more robust with respect to words' polysemy and to avoid any use of manually compiled resources except a morphosyntactic analyzer. However, its scope must be limited in certain ways.

The verbs provide most of the semantic frame of sentences and are considered the most important syntactic category. Although each syntactically correct sentence must have a verb but not necessarily a noun, in language there are more nouns than verbs, e.g., WordNet[1], lists $11,488$ verbs versus $117,097$ nouns. Verbs rarely have true synonyms and when synonyms exist, frequently, one of them is from Anglo-Saxon origin and the other has Greco-Latinate root. The verbs from Greco-Latinate origin are used in more formal contexts (Fellbaum, 1990). Due to the incompatible stylistic usage they rarely co-occur in the same discourse and the substitution of one for the other in context is usually unacceptable. Although the verb category is less populous it still needs to cover the entire diversity of communication necessities. This is compensated for by higher polysemy and more flexible interpretation of the verb's meaning. Because of this property of verbs, they are probably the category that is most difficult to study. In Miller *et al.* (1990), we can read:

> "*[…] some words in the same syntactic category (particularly verbs) express very similar concepts, yet can not be interchanged without making the sentence ungrammatical.*"

[1] Hereafter, each reference to WordNet signifies WordNet 2.1, unless otherwise specified.

This means that for verbs, simple substitutability, wherever it occurs, is most probably a misleading indicator of synonymy.

In the domain of adjectives, synonymy and antonymy are equally important and play complementary roles in their organization. The relation of similarity between adjectives orders them in gradable scales rather than in sets of synonyms and although pairs of synonymous adjectives do exist, they rarely have the same antonyms. Further, only a fraction of the adjectives have true antonyms. This means, for example, that although the pair *weighty* and *light* are conceptually opposite, they are not true antonyms. Empirical evidences (Justeson & Katz, 1991; Mohammad *et al.*, 2008) show that antonymous adjectives tend to co-occur in close proximity. At the same time they are interchangeable in context, which is, as well, the behavior of synonymous words. This means that for adjectives, attributional similarity would discover a range of semantic relations, from synonymy, through adjectives expressing variable levels of the same quality, e.g., names of colors, to antonymy, e.g., *poor*, *mediocre*, *average*, *fair*, *good*, *great*, *superb*. The mentioned idiosyncrasies of the adjectives and verbs render any synonymy discovery procedure based on in-context substitutability evidence inadequate.

Nouns are organized mainly on the basis of substitutability. Though not for all, the main semantic component of a noun is its functional features and they are one of the two means to define a noun concept in a dictionary. The second direction of organization of nouns is vertical, i.e., hierarchical. We perceive the hierarchical structure by the adjectives of which a noun can be an argument. At the same time, when the context does not allow ambiguity, then superordinate terms can serve as anaphors referring back to their subordinates. Thus synonymy and hypernymy are not always clearly distinct. We also refer to events and entities by naming them with nouns. Thus, while verbs are fundamental for the syntactic structure of the sentence, nouns represent the subject matter of the message. Due to the necessity to repeat ideas and to accentuate different aspects, there exists a great diversity of manners to say the same thing. This is achieved by synonymy on lexical level, or by paraphrasing on sentence level when more complex notions are concerned. Further, diversity in expression is created by the fact that words with similar yet not equivalent meanings can be used in place of synonyms, for example hypernymy. In order to understand a message, to perceive its redundancies and important details, one has to be capable to map the variable forms to their conceptual space. Thus, the use and the comprehension of synonyms is fundamental for communication.

A second problem faced by any natural language processing system is polysemy. It is the quality of the words to express different meanings depending on the context. This means that *line* and *cable* can refer to the same object in the electrical domain while in general they signify different concepts. Most efforts to discover synonyms ignore polysemy entirely although approximately 25% of the words have more than one meaning.

The objective we set in this work is to develop a noun synonymy discovery method, based solely on lexical analysis that is robust against polysemy. We will pursue this objective in two stages. First, we plan to identify the mechanism by which polysemy impedes lexical analysis. Then, we intend to propose a conjunction of techniques to avoid the negative influence of polysemy.

1.3 Applications of Synonym Discovery

The primary reason to make attempt at automatic discovery of any semantic property of the words is the intention to build a system, that is capable to mimic human cognition and reasoning. Thus, the organizing principles of the human semantic memory, revealed by psycho-linguistic research, give the ground on which machine language acquisition should be implemented.

The most important relation for the organization of the human semantic memory is similarity of meaning. Its manifestation in the lexical dimension of the language is synonymy. Thus, while we are in search of manners to represent meaning, the synonymy relation between words is the available proxy to the similarity of meaning.

Meanwhile, multiple applications of NLP benefit from lexical semantic similarity information. Query expansion (Ruge, 1997) in Information Retrieval (IR) is an application area, which requires such knowledge at any phase. It is known that people use many different ways to express the same idea. For example, Furnas *et al.* (1983) note that people choose the same key word for a single well-known object less than 20% of the time. Therefore, it is important that the system is capable to retrieve many documents, which do not contain the terms in the query. At the same time, a quarter of the vocabulary words are polysemous, and when such word is placed in the query its intended meaning has to be determined and only afterwards the relevant documents can be properly identified. Each of these steps requires semantic similarity in particular and semantic relation information in general in order to be accomplished.

Language models, developed for the necessities of speech recognition and statistical parsing often need to deal with extremely sparse data. Even in very large corpora many feasible combinations are not observed. One can expect 14.7% of the word trigrams in any new English text to be unseen in a corpus of 366 million words (Brown *et al.*, 1992). Similarity based smoothing is a way to take advantage of linguistic knowledge in order to achieve better efficiency in comparison to the formal statistical methods. It is based on the intuition that words that have many contexts in common are semantically similar as well. Then, we can restore some incomplete data about the contexts of a word knowing the words to which it is contextually similar (Brown *et al.*, 1992; Hindle, 1990; Pereira *et al.*, 1993).

The hypothesis that "*the meaning of entities, and the meaning of grammatical relations among them,*

is related to the restriction of combinations of these entities relative to other entities", formulated in Harris (1968), encouraged a great number of attempts in automatic thesaurus construction (Curran & Moens, 2002; Grefenstette, 1994; Hindle, 1990; Lin, 1998a). However, those resources can only be evaluated against existing, manually built gold standard. A more sensible application would have been to augment an existing general purpose dictionaries with domain specific semantic information, as proposed for example in McCarthy *et al.* (2004). This direction continues to another important application, i.e., dictionary maintenance. Since the language is a dynamic system, it changes all the time adopting new words, abolishing old ones and shifting word meanings. This dynamics requires constant effort to maintain the existing resources up to date. Moreover, dictionaries do not usually list collocations that have compositional meaning, e.g., *broken bone*. However, a computer system requires some clue as how to interpret them. This application is suggested in Hearst (1992) where an algorithm to discover pairs in hypernymy-hyponymy relation is developed. An extension of this idea could take advantage of any feasible semantic relation that signifies semantic similarity, synonymy, functional features, or specific attributes in common with another lexical unit with known meaning.

Since synonyms appear naturally in speech and text, a natural language generation system requires knowledge about felicitous usage and substitution patterns in pursue of authentic sounding. Adequate semantic information have strong impact on intermediate tasks such as natural language generation (Inkpen & Hirst, 2006) and word sense disambiguation (Dagan & Itai, 1994) that are enabling steps for more applied technologies as for example machine translation, message compression, and summarization.

Finally, knowledge about synonymous pairs is frequently used in order to extend some seed sets. For example, Esuli & Sebastiani (2007) and Kozareva *et al.* (2008) classify WordNet synsets with respect to a set of predefined classes beginning with few instances per class, relying on the assumption that similarity in meaning is indicator of equality with respect to some specific semantic aspect, positive or negative connotation in the former case, and membership to a more specific semantic domain in the later.

1.4 Different Approaches for Synonymy Discovery

Probably, the first method to be mentioned here is the manual way. The usual disadvantages of this method are common characteristics to the manual work, i.e., it is tedious and as a consequence expensive to compile dictionaries manually. For many domains, it is unjustified to spend the resource required. Even when the endeavor is undertaken, at the moment of completion the dictionary is already outdated due to language dynamics.

Some authors, e.g., Baroni & Bisi (2004), stress the subjective factor as a flaw. This sounds as an invitation to leave the machine to decide instead of a lexicographer what is synonymy. At the current

level of understanding of the linguistic phenomena, it seems more reasonable to take advantage of the computational power of the machines to propose a list of feasible candidates plus concordance information to facilitate the decision of the lexicographer. This is the viewpoint we focus on.

WordNet is a notable example of a manually built thesaurus. WordNet "*[…] is an attempt to organize lexical information in terms of word meanings, rather than word forms. In that respect, WordNet resembles a thesaurus more than a dictionary*" (Miller *et al.*, 1990). It has great value not only for the community concerned with automatic language processing, but the process of its compilation brings to light vast amount of otherwise unrealized linguistic knowledge.

We overview a number of automatic approaches to synonymy discovery in Chapter 2. What is common between all of the works is that semantic information is sought to be extracted from structured or raw textual corpora with the use of as little as possible language specific knowledge (Firth, 1957; Lyons, 1968). Since information about the meaning of the lexical units is practically unavailable, the only basis for comparison is the knowledge of what company the words keep.

Some naïve methods look for semantic relations among words that co-occur side by side. In the literature, this event is called *first order co-occurrence*. It is very simple to acquire first order co-occurrence statistics and calculate the respective association strength value. This simplicity was valued feature in the early attempts when computational power was limited factor. With the advance of the computing technology, it became possible to couple first order co-occurrence information with more complex mathematical methods and to scale up the coverage. Hypothetically, words that are similar in meaning should occur together more frequently compared to less similar words (Turney, 2001).

More linguistically motivated methods resorted to so called *second order association*, which is a similarity estimation calculated between two lexical units based on the amount of first order co-occurrences that they have in common. Words that have a common first order co-occurrence are said to be in *second order co-occurrence*. The name is misleading since second order associations have radically different properties and applications. While the methods based on *first order co-occurrence* tend to discover syntagmatic relations and are useful in resolving prepositional attachment or noun bracketing, the methods based on *second order association* tend to discover paradigmatic relations and can be used to factorize the vocabulary in Part-Of-Speech sets or to find out synonyms. It seems appealing to conclude that since synonymous words can take the same contexts, then words that have strong second order association must be semantically similar.

1.5 Problems Encountered by Previous Methodologies

Major difficulty in front of NLP stems from the fact that the nature of meaning is virtually unknown. Thus, any pursue of semantics is indeed an underdetermined task and the lack of conceptualized knowledge about word meaning hinders the studies and makes impossible the formal analysis. Consequently, the existing methods to find words' semantics make use of the manifestations of the investigated aspect of meaning as an indirect evidence of its hidden regularities. The automatic processing of natural language falls back on lexical analysis as a machine accessible registration of the manifestations of human conceptual space and reasoning. Although attempts in multimodal dialogue systems exist, their focus is not the language itself.

Different approaches aim to list a finite number of representative semantic axes and eventually linear combinations between them that catch the exact meaning of vocabulary words. Those approaches are not principally different from manual construction of dictionaries as the semantic axes are determined manually. The most prevalent current paradigm takes the surrounding context as a factor defining and describing semantics. However, it has significant limitations (Freitag *et al.*, 2005; Terra & Clarke, 2003).

In human interaction, there are two contrary forces. At one side is the producer of the utterance, whose minimization of processing requirement imposes limitations on the size of the possible vocabulary. This causes that some symbols have to express several, sometimes disparate, ideas. At the other side is the receiver of the message, that is exposed to a stream of symbols, which meaning has to be looked up in the conceptual space. Thus, message interpretation imposes some processing. This processing is minimal, when the message is complete and does not require further communication in order to be clarified and is proportional to the number of variants of interpretation. Those two factors govern a property of lexis called polysemy (Köhler, 1987).

Although in normal interaction the disambiguation process remains virtually unnoticed, the polysemy is so common, that no system for automatic language processing can hope to ignore it. This led to a great number of attempts at Word Sense Disambiguation (WSD). However, most of the approaches require some training, but annotated data is scarce and expensive to obtain. This circumstance coupled with the complexity of the problem explains why the most frequent sense is that powerful a heuristics that any more sophisticated WSD system achieves only marginal improvement (Navigli *et al.*, 2007; Pradhan *et al.*, 2007).

It was found that the performance of a synonymy discovery system is proportional to the volume of statistical data only to a certain point (Terra & Clarke, 2003). Further growth of the statistics, however, is not beneficial. Excessive statistical data only strengthens the representation of the most-frequent sense of polysemous words and obscures less frequent senses. Thus the maximal

performance attainable has an upper limit, that depends on the bias towards one sense or another.

Because the proper meanings of the words are not accessible and in place of proxies are used vague symbols, formal inferences about the semantics are rarely precise. The specificity of the semantic relations discovered varies according to the definition of context. When more loose context, as for example in Landauer & Dumais (1997) where the context spans the entire paragraph, is used, relations in the order of similarity and relatedness are discovered, e.g., *doctor* and *disease*. A tighter context definition, one that spans a phrase or just a specific syntactic relation, e.g., *verb - direct object*, has usually more discriminative power and thus discovers more specific relations, e.g., synonyms, hyponyms and antonyms, with greater probability. Usually, a successive manual or automatic step is required in order to determine what is the exact semantic relation that holds between the listed pairs (Church & Hanks, 1990; Sahlgren, 2006b).

The exhaustive search is the obvious way to verify all the possible connections between words of the vocabulary. However, comparison based on word usage can only highlight those terms that are *highly* similar in meaning. This method of representation is usually unable to distinguish between *middle strength* and *weak* semantic relations (Rubenstein & Goodenough, 1965). Thus, the relative success of the Vector Space Model paradigm on synonymy tests (Ehlert, 2003; Freitag *et al.*, 2005; Jarmasz & Szpakowicz, 2004; Landauer & Dumais, 1997; Rapp, 2003; Sahlgren, 2006b; Terra & Clarke, 2003; Turney *et al.*, 2003) is due to the tests' structure. As a matter of fact, the results on synonymy tests depend very much on the number of the candidates among which choice has to be made. We conducted a simple experiment with a set of 1000 random test cases, created in the manner described in Freitag *et al.* (2005), with up to 10 decoy words and 1 synonym. We then solved the test cases using a contextual similarity measure, i.e., a second order association. In particular, we used the Cosine similarity measure (see Equation 3.4) with features weighted by the Pointwise Mutual Information (PMI) (see Equation 3.2). The increase of the number of decoys caused a rapid drop of the probability to rank the synonym at the first position as shown in Figure 1.1.

Thus, the exhaustive search is only capable of finding the most salient semantic relations, the ones that are established in the language and are frequent enough to be well represented, the ones that are usually included in the manually built thesauri. At the same time, neologisms, recently adopted foreign words and names, which consist that part of the current vocabulary that needs constant update, elude characterization since they are not always well established and represented by written media.

1.6 Our Proposal

We propose two possible ways, motivated by previous works, to alleviate the problems mentioned in the previous section. The problems are closely related, i.e., the substitution of concepts for

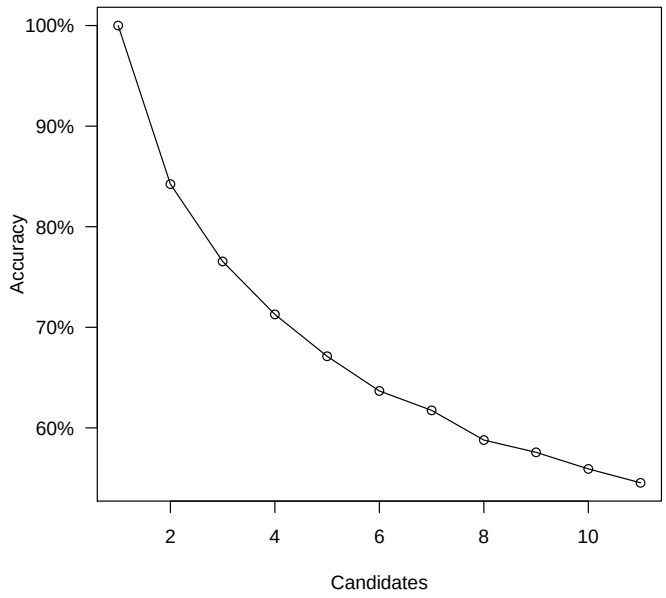

Figure 1.1: Accuracy by candidates count.

polysemous symbols introduces uncertainty, and the uncertainty grows uncontrollably with the size of the problem. This situation imposes significant limitations of lexical analysis applicability in real world. Thus, we first focus our attention on the problem related to words, i.e., polysemy, and afterwards on the difficulty to relate pairs of words by means of exhaustive search.

The polysemy problem can be, at least, partially mitigated. Most of the works conducted so far ignore entirely word meanings as reality, e.g., (Landauer & Dumais, 1997). This comes to draw our attention to the fact that most of the time we process some text in an automated manner, we do not even care about the meanings of the words, and we can do well anyway. Knowing the exact word meaning is not always that important.

It was observed that words have strong tendency to express the same, or closely related ideas every time when they reoccur within the limits of discourse. This means that a partial description of a word meaning can be learned from within a coherent piece of text. This property, combined with the nature of the synonymy, which presumes that synonyms are used in order to avoid repetitions affords for a pair of words to be compared within the limit of a discourse as if they were monosemous.

We evaluate this first step of the proposal against a set of standard Test of English as a Foreign Language (TOEFL)[1] tests and a larger set of automatically created TOEFL-like tests. This evaluation proves that the polysemy problem might be alleviated up to such a level that a significant part of

[1] Educational Testing Service, `http://www.ets.org/` [*13th July, 2011*]

the vocabulary can be safely accounted for. This is our first contribution, which was published in (Moraliyski & Dias, 2007) and (Dias & Moraliyski, 2009).

Next, we address the second problem of the lexical analysis, the rapid increase of error rate with the increase of the number of possible alternatives. In order to discover pairs of semantically related words that may be used in figurative or rare sense, we need to have them highlighted by their environment as in the information extraction strategy. This particular environment is the sentence. It was shown in Charles (2000) that a sentence as a context limits the possible substitutions of words almost exclusively to close synonyms. The concrete implementation of this idea is a recently published method, which uses paraphrase extraction and alignment algorithm (Dias *et al.*, 2010). This is our second contribution.

In summary, we take advantage of the fact that words similar in meaning are strongly associated in a first-order fashion. They are, also, in strong first-order association with other related but more distant in meaning words. At the same time, semantically similar words are strongly associated in a second-order manner. This is in contrast to the words that are only related, but not similar in meaning. In order to limit further the number of possible alternatives we require that the synonym candidates are interchangeable at least in one sentential context and consequently we check whether they have strong second-order association.

The contributions of the current work are three-fold: we identify polysemy and exhaustive search as main obstacles to successful language acquisition; we propose two distinct techniques to sidestep the outlined difficulties; and perform an empirical evaluation and comparison of the proposed model against state-of-the-art word-meaning models.

1.7 Plan of The Thesis

In the following chapter, we give an overview of the most important works, that consider the same matter, published so far. It is intended to serve as a detailed introduction to the problem under consideration and to give support and motivation to the methodologies that we later propose.

Based on the analysis and discussion of related work, we propose our first contribution in Chapter 3. Here, we give thorough definitions of context and contextual similarity as common concepts. This is the place where we introduce context weighting schemas and measures of contextual similarity. In this chapter, we develop the theoretical motivation of our proposal. It is based on the *One sense per discourse* hypothesis as of Gale *et al*. (1992), which is used in a way to reduce the negative effect of polysemy on a concrete practical application.

Experiments and empirical evaluation are given in Chapter 4, where we make comparison between a

standard off-line corpus for text categorization and IR and a Web derived corpus, specifically tailored to show representative statistics. In this chapter, we criticize various characteristics of the common approach to contextual similarity and contrast it to the newly proposed similarity measures.

In Chapter 5, we elaborate on a real-world application for synonymy detection based on paraphrase extraction and alignment. We describe an algorithm capable to automatically create TOEFL-like test cases, which purpose is to substitute the exhaustive search. Chapter 6 then examines various issues in current paraphrase detection methodology and we provide further evidences of the viability of our new measure.

Finally, Chapter 7 summarizes the main contributions, and gives possible future directions of development.

Chapter 2

Related Work

The work surveyed here could probably be classified by various criteria, i.e., objectives (test solving, thesaurus construction, ontology construction, specific application), by evaluation, which more or less overlaps with the classification by objectives, by resources required (raw textual corpus, annotated training data, knowledge rich resources such as dictionaries, thesaurus, ontologies etc.). However, we are interested in the various methods to discover close semantic relations. This is why we classify them under the following categories: pattern-based methods (Section 2.1), document as a context (Section 2.2), attributional similarity (Section 2.3) and few other works that employ more unusual techniques and sources of evidences (Sections 2.5, 2.6 and 2.7).

2.1 Pattern-based Synonymy Extraction

Due to the heterogeneity and the lack of structure, automated discovery of targeted or unexpected information from raw text presents many challenging research problems. Within this context, patterns can only be helpful to learn knowledge that can possibly be expressed by constructions known in advance. In this section, we give a number of references to works on semantic relations detection based on pattern recognition in order to lay down base for comparison to the method proposed in Chapter 5.

In NLP, patterns are applied to a wide variety of problems, including, but not limited to Information Extraction (IE) (Huffman, 1996; Muslea, 1999; Tang *et al.*, 2006), subcategorization frames extraction (Brent, 1991), shallow parsing (Liu, 2004) and parsing (Noord, 2006).

It is possible to manually build comprehensive grammars for Part-Of-Speech (POS) tagging or even for parsing as proposed in Lin (1994). Further, methods that learn by set of unambiguous seed patterns and consequently apply the acquired knowledge in order to learn new patterns have been developed for the sake of subcategorization frame extraction (Brent, 1991) and WSD (Yarowsky, 1995). Similar to other problems in NLP, those methods are only successful for the most frequent events and require large corpora, since they rely on very simple schemes as opposed to deeper understanding.

The pattern-based approach to semantics discovery is hindered by the extreme variability in the way

semantic relations can be activated in text. If the pattern-based approaches were the only possible way to discover the relations between words, it would have reduced to manual compilation of the corresponding dictionary, a prohibitively time consuming and tedious task.

One way in which this variety can be reduced is to use controlled sources of textual data. For example, Markowitz *et al*. (1986) and Nakamura & Nagao (1988) propose the use of Machine Readable Dictionary (MRD), since dictionaries are designed to explicitly convey semantic information through a restricted vocabulary and a set of unambiguous constructions. Those works aim to process structured corpora, which allow to extract detailed information (Ahlswede & Evens, 1988).

However, it is apparent that although highly accurate semantic information can be extracted, its coverage is limited by the content of the dictionary. In order to overcome this limitation, a number of works, starting with Hearst (1992), proposed to use unrestricted text as a corpus.

The problem considered in Hearst (1992) is the automatic discovery of pairs of words that are in *hyponym* / *hypernym* relation (see Section 1.1). The process goes as follows: First, pairs of words for which *hyponym* / *hypernym* relation is known to hold, e.g., *bicycle - vehicle*, are sought in text in syntactic proximity. Then, those contexts that reoccur are scrutinized in order to find appropriate patterns. However, this phase is underdetermined and requires human intervention.

Appropriate pattern can simply be a definition, e.g., *Cranes are large birds*, which is very productive and for this reason very noisy. However, in the habitual communication, there are other highly probable constructions that have different communication intents and the *hyponym* / *hypernym* relation is understood only as a secondary aspect, e.g., *Coarse-grained rocks, such as granite, consist of interlocking mineral crystals*. Through this procedure a number of patterns are identified, that can be summarized as follows:

- such NP as $(NP,)*$ {or | and} NP
- NP $(,NP)*$ $(,)$? {or | and} {other | another} NP
- NP $(,)$? {including | especially} $(NP,)*$ {or | and} NP

When applied to a MRD, e.g., *Academic American Encyclopedia*[1], those patterns are unambiguous and produce *Is a* pairs with high accuracy. However, again, the coverage of the discovered relations is limited to the coverage of the dictionary. In a more general corpus, 20 million words of the New York Times, the patterns and their meaning as *hyponym* / *hypernym* indicator were much more rare events and worse accuracy was noted.

[1] Grolier Electronic Publishing, 1990

The difficulties encountered by Hearst (1992), Caraballo (1999) and Maynard *et al.* (2009), evidence that the lexico-syntactic patterns tend to be quite ambiguous with respect to which relations they indicate. At the same time, only a subset of possible instances of hyponymy relation will appear in a particular form. This imposes to be used as many patterns as possible.

The success can be measured by the property of the method to, at least partially, explain phrases with composite semantics, that would not be normally included in regular dictionary, e.g., *hyponym (broken bone, injury)*, which needs a definition when it is processed in automated manner. On the other side, the same patterns catch metonymic pairs such as *metonym (king, institution)* or such that make sense only in certain contexts, e.g., *hyponym (aircraft, target)*, the semantics of which is not usually thought as a thesaurus relation.

Cederberg & Widdows (2003) developed further on the idea by improving both precision and recall of the extraction process. The rationale is that correct *hyponym* / *hypernym* pair of words tend to have very similar usage while other candidates that fit the patterns but do not share meaning are unlikely to have many contexts in common. Cederberg & Widdows (2003) illustrate this with a pair of sentences from the British National Corpus (BNC). Both sentences

- Illnesses, including chronic muscle debility, herpes, tremors and eye infections, have come and gone.
- Often entire families including young children need practical home care […]

fit the same pattern

- NP $(,)$? {including | especially} $(NP,)*$ {or | and}? NP

However, between *families* and *young children* does not hold an *Is a* relation. After the inappropriate cases are filtered out, coordination information is introduced as it is a reliable clue of semantic similarity and thus indicates that when a pair of the coordinate terms is known to be in *Is a* relation, it holds for the other members of the coordinate expression as well.

Hearst (1992) noted that it was more difficult to succeed on the *Part of* relation, known to lexical semanticists as *meronymy* / *holonymy*, relation[1]. However, Berland & Charniak (1999) used two frequent patterns

[1] A concept represented by the synset $\{x, x', \dots\}$ is a meronym of a concept represented by the synset $\{y, y', \dots\}$ if native speakers of English accept sentences constructed from such frames as *A y has an x (as a part)* or *An x is a part of y* (Miller *et al.*, 1990).

- $wholeNP$'s $partNP$
  …building's basement …
- $partNP$ of {the | a} $wholeNP$
  …basement of a building …

and applied a statistics that takes under consideration the statistical significance of the observed evidences. When the most confidential candidates extracted from the North American News Corpus (NANC), a 100 million words corpus, were taken and manually evaluated by a number of human subjects, it was found that about 55% of the relations were perceived as correct. This evaluation procedure was adopted due to the lack of a clear definition and agreement as to what is *Part of* relation.

A common difficulty of both works was to deal with idiomatic expressions, noun phrase modifiers and wrong POS tagging. Indeed, the patterns are very ambiguous and while they might express *Part of* relation, they can be used to denote qualities, e.g., *value of a car*, events, e.g., *collapse of a regime* or even idiomatic ideas, e.g., *a jalopy of a car* (Berland & Charniak, 1999). In order to exclude many candidates that denote qualities through constructions that match both of the patterns, they discarded those nouns that end with suffixes *-ing*, *-ness* or *-ity*. Idiomatic expressions can be filtered out by recognizing the fact that they appear in one only form and thus match either of the patterns, but not both. However, the adoption of this heuristic would have required a much larger corpus.

Another work, that builds on the findings of Hearst (1992), elaborated in Caraballo (1999) makes indeed more complete use of the same set of patterns. They extract the conjunctive and appositional data from parsed text[1]. Following the discussion in Riloff & Shepherd (1997) and Roark & Charniak (1998), nouns in conjunctive and appositive constructions tend to be semantically related. This information is integrated in a clustering process, which results in an unlabeled hierarchy similar to the *Is a* hierarchy of WordNet. At the successive steps, the internal nodes are given labels with respect to the votes for the various possible hypernyms of the words at leaf levels, as caught by Hearst's patterns.

This algorithm has the tendency to annotate several internal levels with the same hypernym, usually of some intermediate level of generality. Further, Caraballo & Charniak (1999) noted that the pattern-based methods have difficulty in getting beyond the middle level of generality to more abstract categories.

Lin *et al*. (2003) consider a hybrid application of patterns and search engine results (see Section 2.2.2) in order to distinguish between synonyms, antonyms and other kinds of semantic relations that con-

[1] 1 million words of the 1989 Wall Street Journal (WSJ) material annotated in the Penn Treebank (Marcus *et al*., 1993) plus automatically parsed 1987 WSJ texts.

dition high contextual similarity. They achieve the distinction with the help of patterns that indicate semantic incompatibility:

- from X to Y
- either X or Y.

If two words X and Y appear in one of these patterns, they are very likely to be semantically incompatible. A compatibility score is defined in Equation 2.1.

$$score(x,y) = \frac{hits(xNEARy)}{\sum_{pat \in P} hits(pat(x,y)) + \epsilon} \tag{2.1}$$

where $hits(query)$ is the number of hits returned by AltaVista[1] for the *query*, P is the set of patterns and ϵ is a small constant to prevent division by zero. They set a threshold of $\theta = 2000$ above which a pair of words is classified as synonymous. The measure is applied to an automatically compiled thesaurus (Lin, 1998a) and achieves 86.4% precision in partitioning of a set of 80 synonymous and 80 antonymous pairs. To summarize, the method applies second order association measure in order to find feasible thesaurus candidates, among which synonyms are present. Afterwards it applies first order association information in order to filter out the non-synonyms. In this respect, it is similar to Giovannetti *et al.* (2008), where they apply two kinds of lexico-syntactic patterns with an intermediate phase of contextual similarity filtering.

Snow *et al.* (2006) describe a statistically sound approach to the construction of hierarchy around the *Is a* relation. They combine evidences from classifiers of a number of lexical relations within a single framework by defining a set of taxonomic constraints such as *a hyponym synset inherits the hypernym and meronym synsets of its direct hypernym*. Thus, the addition of any new hypernym relation to a preexisting taxonomy will usually necessitate the addition of a set of other novel relations as implied by the taxonomic constraints. Each relation is associated with a probability of being part of the taxonomy based on the observed evidences in a given text corpus. Thus, the objective of the algorithm is to augment an existing hierarchy while maximizing the probability of the final product.

The most inventive work in this section is probably the one proposed by Ohshima & Tanaka (2009). Like others, they use pre-coded patterns together with search engine results. Their motivation is that the close semantic relations are symmetric and the constructions that involve words in such relations are symmetric as well. Such constructions can be appositives, conjunctions or coordination constructions. In order to discover related terms, they instantiate and send to a search engine a number of patterns filled only with one possible candidate, for example *or Los Angeles* and *Los Angeles or*. The

[1] `http://www.altavista.com/` [*13th July, 2011*]

search results are collected and the pattern is sought through the snippets. Where it was found, the corresponding counterpart is collected, thus constructing two sets of left and right contexts. The closely related terms are capable to appear in both sets. Those contexts that appear in both sets are taken to be the desired terms and they are not limited to single words but can be multiwords as well. An especially interesting property of this method is that it is capable to discover unexpected related terms with unspecified length. This property is particularly important in languages, like Japanese or Chinese, where the words are not separated by intervals. This method can discover asymmetric relations if both asymmetric patterns are available. For example, the detection of *painting* / *painter* pairs is possible by pairs of patterns, such as the followings:

- A was painted by B
- B painted A

where either the name of the painter B or the title of the painting A is known.

2.1.1 Discussion

Despite the variety of different approaches, a common feature of all the reviewed works is the necessity of manual effort needed to compose the patterns. Although there are early works that propose to exploit MRD and to thoroughly interpret the available data, later approaches make heavy use of unstructured texts or even search engines results coupled with other shallow information extraction techniques.

Most are oriented to purely semantic relations discovery (Berland & Charniak, 1999; Caraballo, 1999; Hearst, 1992; Lin *et al.*, 2003; Maynard *et al.*, 2009) and employ a variety of different filters in order to improve accuracy. However, they all root in IE and thus nonclassical relations can be spanned as shown in (Ohshima & Tanaka, 2009).

In contrast, the method proposed in Dias *et al.* (2010) and reviewed in Chapter 5 aims to find close semantic relations through a technique that bears certain similarities with pattern recognition, but is more prone to find pairs of words in paradigmatic semantic relations. In fact, by evidencing interchangeable words through paraphrase alignment, we propose a new way to discover local lexical patterns. Further, we validate the candidates in a similar manner to Lin *et al.* (2003) with the aid of a contextual similarity measure. We argue that our method avoids manual search of suitable patterns.

2.2 Document-Based Approaches

2.2.1 Distribution over documents

In this section, we describe an application of a formal method to the language acquisition problem. It was first employed in attempt to improve recall in IR. Only later, it became evident that the same representation could be viewed as a general cognitive model and was used for synonymy discovery. Here, we follow this same order of exposition.

2.2.1.1 Information retrieval approach to synonymy

In the context of IR, documents are represented by a set of features selected to describe the documents in a multidimensional space. A basis vector of unit length is assigned to each feature.

The role of features is usually played by the lexical units that make up the documents. As such can serve the words. Collocations are used in case of very frequent words, while thesaurus entries are helpful in frequency aggregation for rare terms, as stated in (Salton *et al.*, 1975). This paradigm of representation construction avoids introducing any additional ambiguity since it does not need interpretation of the input while at the same time partially solves the polysemy of very frequent words. This particular property of the direct manipulation of the raw textual material is widely exploited in all directions of NLP.

Thus, the representation of a document is the direct sum of basis vectors corresponding to the features extracted from the document. A document collection boils down to a matrix $M(m_{ij})$ with vector-rows corresponding to the features and vector-columns corresponding to the indexed documents viewed as bag of words. Each cell m_{ij} of the matrix M contains the frequency of word w_i in document d_j.

In IR, Deerwester *et al.* (1990) applied an algebraic method called Singular Value Decomposition (SVD) to a matrix of a collection of documents. SVD is the general method for linear decomposition of a matrix into independent principal components. Their specific settings require that the values in the matrix are association strengths between documents and features. The intention of this calculation is to reveal the latent relations between the documents of the collection and user search queries in the cases when they do not overlap. The method is called Latent Semantic Indexing (LSI). How important is the ability of LSI to reveal the latent commonalities between user search queries and relevant documents that do not share lexical units is illustrated by Furnas *et al.* (1983) who found that people choose the same key word for a single well-known object less than 20% of the time. Thus, the probability that a relevant information is formulated the way a user expects to is rather low for a satisfactory retrieval recall. The synonymy at lexical and discourse level is the problem addressed by LSI. The effect of LSI is realized through appropriate dimensionality reduction of the

representation space. The new space has orders of magnitude less dimensions in comparison with the case where the dimensions are the index terms. The new dimensionality is chosen experimentally in order to maximize the performance of the IR system with respect to precision, recall or F-measure on a specific task. In nature, this is a supervised learning process. A side effect of the reduced representation is significant computation reduction of search query processing. It is believed that this operation reveals the actual dimensionality of the data. The possibility to represent adequately the same data in a much more compact space is explained by the redundancy of the language, namely the variety of ways to express the same idea, i.e., synonymy at lexical and discourse level.

The main quality of the method, the ability to place close to each other documents that share only few or none lexical terms, is due to its ability to place close in the representation space lexical terms, that are similar in usage. The process of dimensionality reduction transforms initially orthogonal basis vectors into vectors that are proportionally more similar as the corresponding lexical features occur in similar contexts[1].

In the final reckoning, the improvement in IR performance achieved by LSI is due to better recall without lose of precision since in the new space, groups of closely related words are merged into joint dimensions. In this way, for a query containing the word *access* relevant documents containing the word *retrieval* would be returned too, as shown in (Deerwester *et al.*, 1990).

2.2.1.2 Cognitive model of synonymy

LSI is, however, symmetric with respect to the role of feature and described object. Thus, while the words are the units that convey the totality of the document meaning, the general meaning of the document in the collection can be viewed as specifying the consisting terms. This is the view taken in Landauer & Dumais (1997). They see each new word occurrence as a possible manifestation of a new meaning. Thus, the objective of the step of model training is to find a compact representation of the most salient word meaning aspects. The assumption that such an outcome is possible is based on the world knowledge that words have indeed relatively few distinct meanings. This application of SVD is called Latent Semantic Analysis (LSA).

The basic idea here is that if two words co-occur within a high number of documents, then we are bound to assume some important connection between them. Additional support to this hypothesis appears to be a special property of those words: they frequently occur in different documents that still share important vocabulary, i.e., they have similar co-occurrence patterns. Thus, SVD provides a coherent manner to explore both first and second order word associations. In a series of experiments on WSD, solving synonymy tests and representation of single digit arabic numerals described in Landauer & Dumais (1997), the model of merging distinct yet similar in usage terms proved its cognitive

[1] Here context is used in general sense of both co-occurring words and discourse.

plausibility. For those experiments, Landauer & Dumais (1997) extracted from the *Academic American Encyclopedia* a text corpus that consists of 30,473 excerpts of length of 151 words on average for a total of 4.6 million words and vocabulary of 60,768 unique words. They chose this corpus since it resembles in content and quantity the estimated reading done by an average 7^{th} grader. Next, they treated the text as bag of words without any syntactic preprocessing like lemmatization, POS tagging or syntactic phrase constituents extraction.

From the perspective of the current work, the most interesting is the evaluation of LSA against 80 items from the synonym portion of the TOEFL[1] test (Landauer & Dumais, 1996). Each item consists of a problem word and four alternative words from which the test taker is asked to choose which is the most similar in meaning to the stem. As they reported, the result of 64% correct answers is comparable to that of a large sample of applicants to U.S. colleges.

Although this result may seem modest compared to the subsequent developments, it is important to note that it was reached with the use of a rather small corpus at the expense of more complex model. These results show that the synonyms tend to occur in the same documents more often than by chance. Indeed, the strong result the method achieves is accounted for by its mathematical property to take into consideration multiple associations between words that share contexts, exploiting simultaneously first order and second order co-occurrences.

On the other hand, the first order associations have the side effect onto the model that it is prone to take as similar those words that co-occur frequently in texts on a determinate subject, but are not necessarily synonyms. This phenomena is rarely manifestable against TOEFL-like tests, however it would have unfavorable implications in pure synonymy discovery settings.

Landauer & Dumais (1997) introduced the synonymy detection in the form of TOEFL as an evaluation task. A clear distinction must be made between the notions of synonymy detection and synonymy discovery. Synonymy detection takes place when we are given a word and a list of possible alternatives that contains exactly one close synonym. In contrast, synonymy discovery is the task which can be defined as looking for the synonyms of a word within the entire vocabulary. Clearly, the later is a much more general and complex task. However, synonymy detection is not without merits. It allows straightforward comparison between different approaches, easy interpretation and provides human performance as a baseline.

Although synonymy detection is only a preliminary task to synonymy discovery, it is a necessary phase in order to justify or discontinue further development. Many authors, as well, took advantage of the opportunity to compare their results on a common ground (Ehlert, 2003; Freitag *et al.*, 2005; Rapp, 2004; Sahlgren, 2001; Terra & Clarke, 2003; Turney, 2001).

[1] Educational Testing Service, `http://www.ets.org/` [*13th July, 2011*]

2.2.2 Search engine hit counts

The techniques presented in this section bear certain similarity to the pattern-based ones (see Section 2.1) since they both use pre-coded patterns that are filled in with the words of interest, which is, indeed, an IE technique. On the other hand, the resulting statistics and the intuition of its interpretation is similar to the LSA, where it is supposed that similar words share common topics and tend to co-occur within the same documents[1].

The first work in this series, Turney (2001), introduces a simple unsupervised learning algorithm for recognizing synonyms. The task of recognizing synonyms is as defined in Landauer & Dumais (1997), i.e., given a problem word and a set of alternative words, choose the member from the set of alternative words that is most similar in meaning to the problem word.

The algorithm, called PMI-IR, uses Pointwise Mutual Information (PMI) (Church *et al.*, 1991; Church & Hanks, 1990) to analyze statistical data collected by issuing a number of queries to the AltaVista[2] search engine. Each candidate is assigned an association score with the problem word as in Equation 2.2.

$$score(choice) = \log_2 \frac{p(problem, choice)}{p(problem)p(choice)} \tag{2.2}$$

Here $p(problem, choice)$ is the probability that $problem$ and $choice$ co-occur in one document or within 10 words window when *NEAR* directive is applied. The algorithm picks the candidate that has the highest association score with the target as the true synonym. This is formally expressed in Equation 2.3

$$\arg\max_{choice} score(choice) = \arg\max_{choice} \log_2 \frac{p(problem, choice)}{p(problem)p(choice)} \tag{2.3}$$

In order to avoid high scores for antonyms, a number of sophisticated techniques, taking advantage of the special features of the query syntax of AltaVista, are proposed. The method has a success rate of 72.5% on a test set of 80 TOEFL synonymy problems. A number of applications of the method are given, among which the most interesting ones are query expansion for search engines and WSD using one strongly associated context.

Baroni & Bisi (2004) extended on this work and assessed its properties and evaluated the capability of the method to discover specific relations when strongly semantically associated rivals are present along with the correct synonym. In particular, they confirm the strong result achieved on the original

[1] Which is not always quite true. Consider stylistically colored words like *break* and *fracture*. Although *broken bone* and *bone break* are common colloquialisms for a bone fracture, *break* is not a formal orthopedic term.

[2] `http://www.altavista.com/` [*13th July, 2011*]

TOEFL test set. It is characteristics for the TOEFL that the tests consist of one correct response among 3 decoys that have very different meanings and usage.

However, when the method is faced with the choice between a synonym and a topical counterpart, for example antonym, it tends to fail. This situation is easy to understand when considered the findings of Justeson & Katz (1991) that the exact antonyms are frequently encountered in syntactic proximity. The problem is even harder when restricted technical domains are considered, since all the words are topically related and tend to co-occur within the same set of documents.

Further, Baroni & Bisi (2004) compared the performance of the PMI-IR with the Cosine measure of contextual similarity (see Section 3.4.1). The PMI-IR performed significantly better at all recall levels, however, it is important to note that they collected the contextual statistics from a small corpus and suboptimal windows size and did not apply feature weighting.

The results suggest that, since it relies exclusively on first order co-occurrence, PMI-IR is more appropriate to detect syntagmatic relations in general and not synonyms in particular.

Cilibrasi & Vitanyi (2007) define an information theoretic similarity measure and its proof of optimality. Their definition, similarly to Lin (1998b), estimates the ratio between the amount of common information to the amount of whole available information about the compared symbols. In this case, the significance of the value is a normalized similarity. However, there are important differences. First, the quantity of information is measured with respect to the number of Web pages indexed by Google[1], that contain the symbol opposed to syntactic contexts in Lin (1998b) and every page is counted with equal weight. Although Google calculates internally such information for page ranking purposes, it is impossible to be obtained for all the hits and it is unknown how does it relates to the semantics of the constituent tokens.

A great number of evaluation tasks are performed, among which clustering of color names and numbers as well as novelists and painters and their respective work titles, where the objective is to create the most specific categories. Certainly the most interesting experiments are those of classifying words according to a number of WordNet concepts where they achieved high accuracy levels despite the difficulty of the task.

Other works that make use of search engine hit counts, though not for semantic relation discovery, aim to obtain syntactic information. For example, Keller & Lapata (2003) assess the potential use of Web as a source of statistics about certain syntactic constructions as for example adjective-noun, noun-noun and verb-noun. Their conclusion is that the Web affords more reliable frequency estimates than an average sized off-line corpus and they correlate well with human plausibility judgements. Similarly, Nakov & Hearst (2005) apply a number of heuristics and Web search hit counts to the prob-

[1] `http://www.google.com/` [*13th July, 2011*]

lem of noun compound bracketing. They make use of very robust estimates of first order associations in order to determine the strongest association within the considered noun grouping.

All those methods can be criticized as they use Web pages counts instead of occurrence counts. However, the empirical data shows that the approximation is justified and some studies prove significant correlation between Web page counts and co-occurrence counts extracted from static off-line corpus (Zhu & Rosenfeld, 2001).

Back to the semantic relations problem, the similar rationale under both LSA and the methods described here brings about similar problems, i.e., semantically related words are recognized, however distinction between synonyms and antonyms or other relations is not possible. On the other hand, although similar to the pattern-based methods for semantic relation extraction, they use conditions and advanced queries, which are interpreted and processed in an unknown manner at the site of the search engine and thus loose control over specific details (Kilgarriff, 2007) that help to establish relation direction and range, i.e., *synonymy* versus *Is a* or less specific, symmetric or asymmetric relations.

It is characteristic of the systems described in this section that due to the manner they collect statistical information they first consider the tokens of interest as dictionary entries with all their possible meanings and usages in the estimation of its information content and consequently partially solve ambiguity by issuing pairwise queries and thus manipulating them as specific concepts. This procedure allows to catch relations between the most represented senses of the instances in question and in this are similar to the works of finding the predominant word sense in a corpus (McCarthy *et al.*, 2004, 2007; Mohammad & Hirst, 2006). However, they fail in discovering rare events or rare language synonyms.

2.3 Attributional Similarity

Here, we consider a number of works that demonstrate the feasibility of the purely statistical approach to word meaning similarity. The variable parameters in those works include objectives, forms of evaluation, manners of corpus preprocessing, concepts of context, weighting schemes and similarity measures. Due to the great attention that this direction of research received, the survey is necessarily incomplete and limited just to some of the most widely cited papers in the literature[1].

Next, this section is more voluminous as our main contribution is towards this same direction and it is thus most meaningfully comparable. On the other side, although here there are presented more distinct endeavors, we consider them in short, since we are going to give more details in Chapter 3. At the same time, it is important to note, that due to the great attention that this direction of work

[1] This is not to say that the other sections claim to be exhaustive.

received, it is very difficult to achieve significant performance improvement.

In a broad sense of attributional similarity, all the works reviewed belong to this section, as they estimate in some sense the quantity of the features shared by the objects of interest. What is different about the works in this section is the idea that semantically similar words are recognizable by the similarities in their usage (see Section 3.1) rather than by some concrete semantic aspect signaled by a specific construction.

2.3.1 Vector space model

Suppose we have statistics about (first order) co-occurrences of the words in the vocabulary. The co-occurrence counts that belong to one word, with zeros for that pairs that were not observed in the corpus, form a vector. This vector effectively specifies a location in a multi-dimensional space. The linguistic proposition that the meanings of the words can be guessed by the company they keep is a language model. This model coupled with specific geometric representation and a distance or similarity measure is called vector space model in the context of NLP. The distances between the points in this space give an estimation of the semantic distance of the corresponding words. Below we review various works that implement and evaluate different flavors of this model. The variable parameters are the definition of context, the weighting of the features, the similarity measures and the corpus used to obtain those statistics.

(Rubenstein & Goodenough, 1965) is probably the first work to estimate the correlation between corpus-based similarity and human perception of word meanings. They elicited 2,880 similarity judgments of 65 pairs of words with the collaboration of 51 human subjects plus a corpus of 4,800 sentences, 100 for each of the 48 nouns. The correlation between the quantity of overlapping contexts and human similarity judgments confirmed that most of the time reliable distinction between synonyms and non-synonyms could be made. However, the contextual similarity was only able to highlight those terms that are highly similar and failed to make clear distinction in the middle of the subjective scale. Apparently, this work required much manual effort. More statistically sound results would require larger corpora and adequate automated processing.

Hindle (1990) pointed out that the contextual similarity approach to language acquisition was impossible until computational power developed sufficiently to store and process adequate amounts of text and technologies like POS tagging and parsing become robust enough. Also, a number of persistent problems were identified, as for example polysemy, which causes evidences of multiple senses to be indiscriminately conflated together. As well, many word meanings and usages were not represented in the corpus of 6 million words of the Associated Press (AP) news stories. These observations indicate that much bigger and more diverse corpora would be required in order to produce general purpose semantic resources. At the same time, Hindle pointed out that the subject-verb-object relation was

a very strong indicator of noun semantics, which affords the analysis to be performed even with the aid of error-prone deep syntactic parsers.

Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996; Lund *et al.*, 1995) is a compact representation of the words by their contexts, which takes into account the distance and the direction of co-occurrence. In effect, each word is represented in a space with dimensionality double the size of the vocabulary. However, they found that retaining only about the 200 most variant dimensions results in an optimal trade-off between expressive power and computational cost. The work aimed to establish general properties and limitations of the distributional hypothesis proposed by Harris (1968), not at any specific application. They conducted a priming experiment that confirmed previous findings, that contextual similarity estimates for semantically related words, e.g., *table - bed*, are higher than for associatively related words such as *coffee - cup* or for unrelated ones.

Grefenstette (1994) implemented a number of knowledge-poor techniques in attempt to build a draft of a thesaurus of domain specific nouns. The produced entries consist of a number of contextually similar words, salient verb contexts and few frequent expressions with the entry word. The word similarities are calculated based on syntactic contexts, extracted after the corpus is morphologically analyzed, POS-tagged and finally parsed. It was found in Grefenstette (1996) that the syntactic contexts are more reliable and less noisy when frequent enough words have to be represented. On the other side, more reliable results for typical phrases are pursued by taking contexts from sentence long span of words. This work relies on a selection of heuristics that are believed to be optimal for the task.

It was noted that due to the lack of a tight definition for the concept of distributional similarity, many different measures were studied for a variety of applications (Weeds & Weir, 2005; Weeds *et al.*, 2004) and when a new algorithm is developed for a specific problem, one is obliged to evaluate a great number of component combinations in order to achieve optimal performance. In order to mitigate this problem and to afford more informed choices with respect to the requirements of the application being developed, an analysis of the statistical and linguistic properties of sets of distributionally similar words returned by different measures was proposed in Weeds & Weir (2005). They analyzed the behavior of a number of measures with respect to word's frequency and concluded that there exist two classes of similarity measures. First, the Cosine, the Jensen-Shannon divergence (Lin, 1991; Manning & Schütze, 1999), the α-skew divergence (Lee, 1999) and the confusion probability (Sugawara *et al.*, 1985) all tend to attribute similarity proportional to the frequency of the candidates. Second, Jaccard's coefficient (Salton & McGill, 1986) and Lin's measure (Lin, 1998b) show stronger affinity to candidates with low frequency. They show particularly interesting applications of this property to the problem of arrangement of words along a generality level scale, which is useful for the construction of an *Is a* hierarchy of the kind of WordNet for example.

Further, in Weeds & Weir (2005), a general framework of contextual similarity is introduced. It is founded on the notions of Precision and Recall from IR and, by analogy with document retrieval, they call it Co-occurrence Retrieval (CR). They show, in a formal fashion, that a number of common similarity measures can be cast into the framework while others, like the α-skew divergence can be closely approximated.

Another kind of characterization, proposed in Freitag *et al.* (2005), considers the difficulty of the problem in its TOEFL form. They created a large number of TOEFL-like test cases taking a pair of words from one synset and three more randomly selected words from other places in the WordNet hierarchy. What they found is that with the increase of the number of senses of the target word and the correct answer it is less likely that a contextual similarity measure makes a correct guess. Their intuitive conclusion is that the polysemy is indicator of the difficulty of the test. However, the findings in Weeds *et al.* (2004) and Weeds & Weir (2005) and the statistical evidence of strong correlation between word polysemy and word frequency (see Table 4.6) suggest that the effect observed in Freitag *et al.* (2005) might be a characteristic of the measure rather than of the tests only.

Likewise Weeds *et al.* (2004) and Weeds & Weir (2005), Heylen *et al.* (2008) undertake lexical similarity measure characterization with respect to the frequency of the words, as well as the level of generality or semantic specificity, measured in the Dutch EuroWordNet and 10 most frequent semantic classes of the *Is a* hierarchy, namely *object*, *event*, *property*, *situation*, *group*, *part*, *utterance*, *substance*, *location* and *thought*. The most interesting observation is that the cosine of the context vectors of features weighed by PMI (see Section 3.3.2) has the tendency to find more synonyms, hyponyms, hypernyms and co-hyponyms with the increase of the candidates' frequency. Out of the three models tested (dependency model, window-based model with first order words association and window-based model with second order words association), the best performing model is the dependency model. This is in accord with the findings of previous work (Grefenstette, 1996).

Terra & Clarke (2003) conducted a comparative study of a number of measures of word similarity using 53 billion words of Web data. This work compares measures introduced in various preceding works: first order measures such as PMI (Church & Hanks, 1990), χ^2, Log-likelihood ratio (Dunning, 1993), Mutual Information, and second order measures such as Cosine of PMI, L_1 norm, Jensen-Shannon Divergence and several variations. Without any parameter fitting, they achieved 81% on the standard TOEFL test set of 80 synonymy questions. This seems to be far better than the more complex method used in Landauer & Dumais (1997). However, Terra and Clarke used a terabyte-sized corpus of Web data, about 10,000 the size of *Grolier's Academic American Encyclopedia* which was the corpus used by Landauer & Dumais (1997). On the other side is Turney (2001) where the corpus of Altavista was used. What accounts for the different results is the more precise control on the processing of the queries issued to the off-line corpus, compared to the relatively restricted syntax of the commercial search engine interface. Beside the standard TOEFL test, Terra & Clarke (2003)

evaluated on two more test sets where the target word was disambiguated by an example phrase. The results on those sets led to the conclusion that the given context was of little or no help or was improperly used. However, a more probable explanation of this finding is that the context is capable to activate some subtle word meaning, which is inadequately represented by statistics that converges to the representation of the most frequent word meaning.

In a series of works (Sahlgren, 2001, 2006a; Sahlgren & Karlgren, 2002), Sahlgren developed and studied a contextual similarity framework that uses a formal method for Random Indexing. In the standard vector space model each word is represented by the contexts in which it appears and each context is a feature to which corresponds a dimension in a multidimensional space with as many dimensions as are the possible features. In contrast, in the model proposed by Sahlgren, the number of the dimensions is *a priori* fixed, 1800 in the concrete case. The vectors corresponding to the features are the random sum of 3 to 6 basis vectors of the space. Under these conditions there are fewer truly orthogonal directions compared to the number of features, yet there are many nearly orthogonal directions. At first, the reduction might seem unjustified. However, it only contradicts the assumption in the standard vector space model that all the features are orthogonal to each other. However, this assumption is not arguable in the face of synonymy.

The major advantage of this model is its simplicity of implementation and reduced computational complexity. Once the optimal dimensionality of the space is determined, the RI is more scalable, it is easier to add new features when new textual material is to be processed compared to LSA.

Finally, Sahlgren (2006b) studied the influence of the used context over the relations that the model captures. The most interesting finding with respect to the current work is that the looser the context the looser the extracted semantic relations are and the contrary, when more specific syntactic contexts are used, tighter semantic relations can be caught. This was somewhat expected as multiword units are usually extracted based on specific syntactic context (Daille, 1995; Dias, 2003; Justeson & Katz, 1995).

2.3.2 Distributional profiles of concepts

The distributional hypothesis states that "*you shall know a word by the company it keeps*" (Firth, 1957) and it is usually interpreted and applied as if the words are monosemous. Apparently, this assumption holds for the most part of the vocabulary but it breaks down when naturally occurring text is considered (Miller *et al.*, 1994). The works surveyed so far only build distributional representations of lexical units regardless of their polysemy.

A number of works aim to calculate semantic similarity using distributional profiles of concepts as opposed to distributional profiles of words (Agirre & de Lacalle, 2003; Bordag, 2003; Mohammad & Hirst, 2010; Pantel, 2005; Rapp, 2003). Such profiles can be built from sense tagged corpora.

However, such resources are scarce and sense annotation is an expensive and a tedious task. Thus, other methods have been proposed.

(Agirre & de Lacalle, 2003; Mohammad & Hirst, 2010; Pantel, 2005) are all similar in that they build distributional profiles that are representative of the distinct word senses out of raw text corpora and MRD. Agirre & de Lacalle (2003) take the monosemous words from one WordNet synset and collect their contexts from a text corpus. The resulting topic signatures are said to be representative of the corresponding concept and can be used to augment WordNet. Pantel (2005) takes all the children of a concept and the contexts that are common for many words and are kept as a concept specific representation. Mohammad & Hirst (2010) create a word to concept co-occurrence matrix. Each time a word w_i co-occurs with a word that belongs to the concept c_j the matrix cell m_{ij} is incremented. If a word belongs to more than one concept, all the corresponding cells are incremented. They argue that while this might introduce noise, the consequent bootstrapping step reduces sufficiently the errors due to word ambiguity. In effect, each dictionary concept is associated with the most salient co-occurring words and a context receives more importance when it co-occurs with fewer distinct concepts.

The methods described here make use of some word sense inventory and attempt to augment it with concept level usage data. In contrast, we are interested in building word usage profiles strictly from discourse evidence, independently from any language specific resource other than a corpus. We show in the following chapter how we deal with word polysemy and that in effect we obtain profiles that are in nature distributional profiles of concepts.

2.3.3 Multi-lingual resources

Beginning with the assumption that similar words have similar translations in other languages, Dyvik (2004) manually identifies the lists of possible translations of the words from a sentence aligned bilingual corpus. The list of translations of a word w is called *(first) t-image*. Then, the translations of the first t-image form a second t-image. Semi-lattices for the words in each language are then constructed by careful analysis of the overlap patterns of the t-images. The main proposition of this work consists of a number of heuristics for this process. The resulting structures allow diverse semantic information about synonyms, hyponyms and hypernyms to be derived. The process is similar to clustering, but applies heuristics in order to decide whether certain features signify a polysemous word and thus give rise to several distinct meanings or a word with single wide meaning, which is more felicitous to be treated as a high level concept. This approach avoids contrastive and antonymous words to be mixed together with the semantically similar ones as it is only capable to catch relations of correspondence. However, quantitative evaluation is missing.

A variant of this work, using bilingual dictionary is elaborated in Priss & Old (2005). Their motivation

was that bilingual dictionaries are available for many languages while aligned bilingual corpora are difficult to obtain. However, the benefit of extracting thesaurus information from existing dictionaries is questionable. They note, as well, that the coverage of the semantic information is limited compared to what could be extracted from a corpus. Furthermore, the semantic data extracted from a corpus is richer compared to dictionaries, since the translation process allows more liberal usage of words as well certain amount of interpretation. Thus, the corpus provides information not only about synonyms but about hypernyms and hyponyms as well, while dictionary entries only contain words at the same generality level.

An original use of bilingual data as a source of semantic representation of words is proposed in (van der Plas & Tiedemann, 2006). They assume that words with similar meanings are necessarily translated similarly in other languages. Instead of using the regular syntactic or window-based co-occurrence context, they take the features to be the translations of the words into 10 different languages. For this purpose, they use the multilingual parallel corpus Europarl (Koehn, 2002) including 11 languages aligned at sentence level. For comparison purposes, they implemented the standard co-occurrence model and calculated precision and recall figures as performance indicators. Interestingly, the co-occurrence model shows precision of about 8.8%, which is significantly lower compared to the 14% reported by Heylen *et al.* (2008) although they both use the Dutch part of EuroWordNet as a gold standard. The discrepancy is probably due to the more sophisticated context used in Heylen *et al.* (2008). The rich feature space of van der Plas & Tiedemann (2006) affords strong results in difficult problem as is the synonymy extraction. Given the settings, automatic sentence and word level alignment, and exhaustive search of noun synonyms, obtaining precision of 22% and recall of 6.4% is quite an impressive result.

Lin *et al.* (2003) present a method to filter non-synonymous words from a list of contextually similar ones with the aid of a bilingual dictionary. The rationale is grounded on the observation that translations of a word to another language are often synonyms of one another. However, when those translations are not synonymous, this is because the word in the source language is polysemous. In this case, the distributional profiles of the translations are usually quite different. This rationale is similar to the one in Dyvik (2004) and Priss & Old (2005). Thus, a pair of distributionally similar words can be identified as synonymous if they share a translation in any other language. Apparently, this method is precise, however it is limited in terms of recall by the coverage of the used dictionary.

2.4 Statistical Models of Language

It was shown in Brown *et al.* (1992) that 14.7% of the 3-grams in a new sample of text are expected to have a maximum likelihood estimate of zero when the model is trained over a corpus of 365 million words. The search of solutions to the sparse data problem led to various developments of class

models for smoothing of the probabilities of unobserved co-occurrences in the training data.

In order to smooth n-gram language model, Brown *et al.* (1992) proposed to group together words that take part in the same (n - 1)-grams. Thus, when an unseen co-occurrence is encountered its probability can be calculated based on an observed similar event. Brown *et al.* (1992) proved that the model that maximizes the likelihood of a sample from previously unseen text is the one that maximizes the mutual information of the adjacent classes, i.e., the model that maximizes the average mutual information of the classes to which belongs the pair of words which are adjacent in text. Put yet another way, the model that best predicts the next word of an utterance is the one that uses the immediate context as a clue. Although they do not show formal evaluation other than marginal reduction in the model's perplexity compared to plane n-gram model, the resulting clusters manifest apparent semantic nature. What is peculiar about this work is that it is based solely on mathematical reasoning and although it does not refer to any linguistic theory it reaches the familiar conclusion that the contexts of the words can be used to find words with similar meanings.

Another work with very similar objectives is Pereira *et al.* (1993). They abandon the geometric metaphor (Sahlgren, 2006b) in favor of a formal probabilistic method for soft clustering of nouns. Out of the initial probability distribution of the nouns over the transitive verbs that take them as direct objects, a cluster membership distribution is calculated through a computational process called simulated annealing (Rose *et al.*, 1990). The process begins with a single cluster comprising all the nouns and splits one cluster at a time until a certain, experimentally established, number of clusters is achieved. The result is an unlabeled hierarchy of nouns. This approach to clustering is less computationally intensive than the method of Brown *et al.* (1992). The resulting clusters consist of semantically similar words. A more formal evaluation comes in a fashion similar to the one adopted in Weeds & Weir (2005), i.e., they remove from the training corpus all the occurrences of certain verb-object co-occurrence types. Afterwards, they calculate the error rate of the artificial task of pseudo-disambiguation as a quantitative method to estimate the ability of the model to recover missing data. This method has possible applications in thesaurus construction and similarity-based smoothing for language modeling.

Statistical models of language are widely useful for smoothing of sparse data. In order to satisfy this objective, they are trained on large corpora. Unavoidable problem of this approach is mixing of statistical evidences related to distinct word meanings, when polysemous words are involved. However, our proposal aims to alleviate this very effect of polysemy, hence available methods of smoothing are not applicable, rather we rely on repeated comparisons within restricted context, which under most circumstances render words unambiguous.

2.5 Information-based Similarity

The first set of relatedness measures represents a group of inventive manners to calculate semantic relatedness out of resources that do not directly provide such information. This is a set of five measures studied in Budanitsky & Hirst (2001) from the perspective of spelling correction system performance. In this case, what is important to be caught is the semantic relatedness rather than the semantic similarity. The five measures represent several distinct approaches to the relatedness in WordNet and rely on the various kinds of connections between the concepts.

The first one is proposed in Hirst & St-Onge (1998) and searches for the shortest path between the concepts and counts how often the path changes the direction as shown in Equation 2.4.

$$sim_{HS}(c_1, c_2) = C - len(c_1, c_2) - k \times d \tag{2.4}$$

where $len(c_1, c_2)$ is the path length between concepts c_1 and c_2, d is the number of changes of direction in the path, and C and k are constants.

Leacock & Chodorow (1998) also rely on the path length between the concepts, however they measure it only over the *Is a* hierarchy and ignore the other connections since it is not always clear how to interpret them. The formal expression of their measure is given in Equation 2.5.

$$sim_{LC}(c_1, c_2) = \log \frac{len(c_1, c_2)}{2D} \tag{2.5}$$

where the scaling factor D is the overall depth of the taxonomy.

Those measures assume that any connection has the same length or *transfers* the same quantity of meaning from one concept to another. However, the connections between a hypernym and hyponym at the higher levels of the hierarchy are indeed *longer* compared to the connections close to the leaves. This feature of the language is accounted for by a number of information theory-based measures.

Resnik (1995b) considers only the *Is a* hierarchy of WordNet and calculates the information content of each concept with respect to corpus statistics. The relatedness of two concepts c_1 and c_2 is proportional to the quantity of the information expressed by the most specific common subsumer $mscs(c_1, c_2)$. This is formally expressed in Equation 2.6.

$$sim_R(c_1, c_2) = -\log p(mscs(c_1, c_2)) \tag{2.6}$$

Later, Jiang & Conrath (1997) and Lin (1998b) develop on the same idea and introduce the path length. Equation 2.7 and Equation 2.8 account for the observation that a concept that subsumes bigger subtrees is more vague and less informative. This led to even more natural estimation of the

information content of a concept by looking at the density of the subsumed concepts.

$$sim_{JC}(c_1, c_2) = 2\log p(mscs(c_1, c_2)) - (\log p(c_1) + \log p(c_2)) \tag{2.7}$$

$$sim_{Lin}(c_1, c_2) = \frac{2\log p(mscs(c_1, c_2))}{\log(p(c_1)) + \log(p(c_2))} \tag{2.8}$$

Those measures that make use of corpus statistics, apparently calculate domain dependent relatedness and for this are useful as a work-around for the *tennis problem*,[1] because they are capable of accentuating at the domain specific relations.

Agirre *et al*. (2004) present a method to approximate the WordNet-based similarities. The intuition is again similar to the one in Resnik (1995a) which states that groups of words tend to disambiguate each other. Thus, they query a search engine with the words from a particular WordNet synset, and gather the returned Web pages. The resulting text collection is expected to represent the specific word usages related to that concept. The intended application of the topic signatures, as they call it, is to augment the existing hierarchy with new words.

Those measures rely heavily on some semantic resource, and although sometimes corpus statistics is employed it has only secondary importance. Also, they are designed with some specific applications in mind while we are interested in the general task of language acquisition likewise Agirre *et al*. (2004).

2.6 Graph-based Methodologies

Bordag (2003) composes a graph of word co-occurrences. Each triangle in this graph is a group of words that renders them unambiguous most of the time (Resnik, 1995a). Next, the words that co-occur with each member of the triplet are taken as features and clustering is performed. In this manner, each word can be placed in one or more different clusters, one for each word meaning. Formal evaluation is not performed, however the method, depending on the type of context, seems to group together very similar words in sets similar to the synsets of WordNet. This method is designed with the language acquisition task in mind, however it suffers the common problem of misrepresentation of rare words' meanings.

Senellart & Blondel (2008) describe a method based on an algorithm that generalizes the HITS algorithm initially proposed in Kleinberg (1998) for searching the Web. Starting from a monolingual

[1] The *tennis problem* is a well know feature of WordNet. WordNet connects the concepts only by means of few semantic relations considered classical, and provides no ways to connect apparently related terms such as *nets*, *rackets* and *umpires* (Fellbaum, 1998).

dictionary, they first construct the associated dictionary graph G. Each word of the dictionary is a vertex of the graph and there is an edge from a word u to a word v if v appears in the definition of u. Then, for a given word w they construct a neighborhood graph G_w that is the subgraph of G which vertices are those pointed to by w or pointing to w. Finally, they look in G for subgraphs similar to the subgraph G_w and take the word corresponding to the node in the same position as w to be synonymous.

Esuli & Sebastiani (2007) used the graph structure of WordNet in order to classify words as positive or negative, given some seed information. The intuition is that given a word w with known positive or negative orientation, its orientation is transferred to the words defined by glosses containing w. This idea was further developed in Kozareva *et al*. (2008) in order to determine the membership to WordNet domains (Magnini & Cavaglia, 2000) with respect to a given text in attempt to calculate word similarities within a document. However, this later work is focussed on word sense disambiguation, rather than on the acquisition of general semantic information.

2.7 Heterogenous Models

Turney *et al*. (2003) combine various modules that draw information from diverse sources in order to solve the synonymy part of TOEFL and analogy tests: two lexical semantic models, i.e., LSA (Landauer & Dumais, 1997) and PMI-IR (Turney, 2001), that share the property to classify as similar words that are topically related. The third information source is the Wordsmyth[1] thesaurus. From it, they collected any word listed in the *Similar Words*, *Synonyms*, *Crossref*. *Syn.*, and *Related Words* sections for each of the case words. This module created lists of synonyms for the target and for each choice. Then they scored word similarity according to their feature sets overlap. At last, the connector module used 20 search result snippets from Google for pairs of words to assign a similarity score to the pair by taking a weighted sum of both the number of times the words appear separated by any one of the symbols {*[*, *"*, *:*, *,*, *=*, */*, **, *(*, *]*}[2] or tokens {*means*, *defined*, *equals*, *synonym*, whitespace, *and*}, and the number of times *dictionary* or *thesaurus* appear anywhere in the snippets. The result on the original set of 80 TOEFL cases is as high as 97.50% and they point out that a more challenging test set is required. However, this approach is not eligible for the language acquisition task since it requires pre-existing thesauri and a number of ad-hoc rules.

2.8 Discussion

In order to cover as thoroughly as possible the area of semantic relations discovery, we surveyed a number of fundamental works towards each of the existing directions. As those works attracted

[1] `http://www.wordsmyth.net/` [*13th July, 2011*]

[2] Right bracket *)* is missing in the original text.

much attention, they received much criticism as well. A number of problems are commonly pointed out.

In the recent years, it became very common to use Web search engines as a source of co-occurrence data. Initially, it seems that this source does not have disadvantages. It is easy to exploit, i.e., the commercial search engines have easy-to-use interface, frequency numbers are fast to obtain, and the experimental results are relatively reproducible. The search engines maintain huge, up-to-date indices of what is on Web. This comes as a promise of reliable frequency estimates even for rare linguistic events. However, it turns out that the frequency values they return are rough estimates. At the same time, those numbers should be carefully interpreted, since the user does not have control and knowledge about how they are calculated (Kilgarriff, 2007), while for strong conclusions one needs to exercise great control over the source of statistical data. In contrast, we gather from the Web an off-line corpus and compare it to a standard corpus for text categorization research.

At a glance, the practical value of a method for extraction of semantic information from existing dictionary is questionable and some systems were criticized on this ground. Although, for humans it is obvious what a dictionary entry conveys, for a computer application a detailed algorithm is necessary in order to utilize this information adequately. Indeed, those techniques are developed with some application in mind and they do not have the language acquisition as objective. Here go most of the measures mentioned in Section 2.5. Thus, direct comparison between both classes of systems, i.e., semantic resource utilization and semantic resource construction is not meaningful. It only could serve as an upper bound of what could be achieved in an automated manner.

One problem that is more difficult to address is the usage of any handcrafted resource. In order to accomplish language acquisition based on textual corpus and a set of rules, it is needed a corpus that covers the entire diversity of language and an exhaustive set of rules that will catch it. These two requirements imply a huge corpus that alongside the useful information inevitably contains significant amount of noise. As a consequence, rules have to filter out the irrelevant data while gleaning all instances of interest. The compilation of such a corpus and a set of rules and the manual revision of the output from the system in all likelihood equals the manual effort required to build the corresponding semantic resource.

From the point of view of the current work, the most significant criticism that applies to most of the systems is the inadequate treatments of the polysemous words. In the following chapters, we propose and evaluate a method designed to deal with this specific problem of the contemporary NLP, and attempt to avoid as much as possible the other problematic areas.

Chapter 3

Word Similarity Based on Syntagmatic Context

In this chapter, we give definitions of the basic concepts of lexical sematic analysis. We begin with the assumption that underlies the entire theoretical and the empirical work. Then, we motivate some criticism against the interpretation of this hypothesis and its applications. At last comes the interpretation that we propose, the first original contribution of this work.

3.1 Distributional Hypothesis

Any of the works surveyed in Sections 2.2 and 2.3 rely heavily on a common assumption. Those models can be constructed, possibly without human intervention and without any prior knowledge about word meaning and word meaning similarity. Rather, the source of empirical data is the language itself and its usage, i.e., the distributional properties of the words. The intuition behind the use of this information source is the observation that words with similar meanings appear in similar surrounding contexts.

One of the early formulations, attributed to Gottfried Leibniz, states that "*two words are synonymous if they are interchangeable in statements without a change in the truth value of the statements where the substitution occurs*". However, it is accepted that synonymy is indeed a continuous rather than dichotomous relation, i.e., for all synonyms there are statements whose meanings can be changed by the substitution of one word for another. Thus, a more realistic view requires to see synonymy as a continuous scale (Charles, 2000).

Ferdinand de Saussure takes a more general view. He claims that the meaning of a word emerges from its relations with the other words in the language. Further, he argues that all the concepts are completely *negatively defined* that is, defined solely in terms of other concepts. He maintains that "*language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others*" and that "*concepts are purely differential and defined not in terms of their positive content but negatively by their relations with other terms in the*

system" (de Saussure, 1959).

Through a simulation, Goldstone *et al.* (2005) demonstrate that it is feasible to deduce concept correspondence between two systems by knowing the relations between the concepts. The results of the general experiment that they conduct give strong support to the supposition that by knowing the association patterns between the units of a vocabulary it is possible to map them to an appropriate conceptual ontology. Thus, the assumption that the semantic similarity of two terms can be deduced by observing their association patterns seems to be weaker and less demanding. Its plausibility is first supported by psycho-linguistic research (Charles, 2000; Kaplan, 1950; Rubenstein & Goodenough, 1965) and recently by numerous empirical studies (Grefenstette, 1996; Justeson & Katz, 1991; Landauer & Dumais, 1997).

Hereafter, we adopt the concept that by analysis of word usage data we can only study lexical and semantic relations between words and in particular word meaning similarity. We do not attempt to understand how meaning is defined by the respective relations, nor to extract that meaning. We are only interested in learning semantic relations that hold between a set of vocabulary entries. To this end, we focus our attention on how to model word meaning similarity based on evidence from naturally occurring text.

3.2 Definition of Context

A context in general sense signifies surrounding circumstances. More specifically, a linguistic context can be a discourse or short distance neighbors from a paragraph or a sentence. In the area of lexical analysis, the context is usually taken to be the words that are within a predefined distance from the target, e.g., discourse, paragraph, sentence, window, or that are in some specific syntactic dependency relation with the target. In particular, the adopted definition of co-occurrence determines the properties of the contextual information and the ability to discriminate semantic relations (Kilgarriff & Yallop, 2001).

Context can be used in different manners for the sake of characterization. Syntagmatic use of context, in the terms of Sahlgren (2006b), arises when relations are sought among co-occurring symbols, as for example in Landauer & Dumais (1997) and Turney (2001). This kind of context use proved to result in robust representations, though with limited discriminative power.

The paradigmatic use of contexts emerges when the direct co-occurrence events of a word with its contexts are considered only as features. The representation constructed by those features allows for more precise search of paradigmatic relations. As we are concerned with synonymy, which is the prime example of tight paradigmatic relation, the definition of context adopted hereafter is that of paradigmatic use of context. Below are introduced the two most common approaches to extract

contextual information for the sake of paradigmatic word characterization.

3.2.1 Window-based

Window-based statistics are gathered from text excerpts where the word of interest is found. Its contexts are taken to be the preceding and succeeding words within a predefined distance. For example, Lund *et al.* (1995) and Lund & Burgess (1996) used a window of size 10 with the target word in the middle[1]. Rapp (2003) noted that approximately half of the running text words are closed class words. Thus, he takes a window of size 8 in order to obtain on average 4 open class context words.

As the window-based lexical analysis does not require any syntactic annotation, its first phase is less computationally intensive compared to more sophisticated contexts. At the same time, it can be applied to languages for which syntactic analyzers are not yet developed. Further, window-based lexical analysis is beneficial for rare words (Grefenstette, 1996). Indeed, the early attempts on lexical analysis were bound to extract contextual data indiscriminately of any intra-sentential relations since automatic syntactic analysis was not yet functional. Even more importantly, the volume of the available machine readable textual corpora and the limited possibility to manually annotate text with syntactic information imposed a trade-off in favor of quantity against quality of the contextual data.

3.2.2 Syntax-based

Consider a sentence in the active voice where a modified direct object follows a transitive verb, e.g., the current sentence. The lack of syntactic information in the window-based context extraction would result in that the modifier *transitive* of the direct object *verb* is taken, inappropriately, as a context of the verb *follow*. This indirect relation is in general rather vague. In order to include the correct context, the window is arbitrarily extended to include a number of tokens to both sides of the target word. Thus, the uninformed context extraction introduces significant amount of noise into the statistical data and leads to severe computational inefficiency.

The aforementioned study by Grefenstette (1996) proved that the syntax-based contexts are more selective and provide better overlap with manual thesaurus classes for common words compared to windows-based methods. Hindle (1990) advocates that the subject-verb-object relations are very strong indicators of the noun semantics, which afford the analysis to be performed even with the aid of an error-prone syntactic parser. These observations are confirmed and extended by Otero *et al.* (2004) and Otero (2008), where a syntactic analysis strategy is applied to syntactic attachment and co-hyponymy induction. Heylen *et al.* (2008) found that dependency-based models are more

[1] The analysis of the results of these studies led to the conclusion that most of the meaningful relations between words of a sentence are within a distance of 5. Although it is unrealistic to think that the language is constrained by such a constant, this simplification proved to be practical.

reliable in finding semantically similar words, i.e., synonyms, hyponyms, hypernyms and siblings, compared to windows-based models. Further, the application of stemming or lemmatization reduces the sparseness data problem by conflating together the inflected forms of the same stem word.

Since synonymy is known to hold between words of the same part of speech, the availability of syntactic information eliminates inappropriate candidates and thus helps to avoid unnecessary calculations and promptly improves precision.

In the following section, we introduce several context weighting schemas and specify the exact kind of contexts used in the following chapters.

3.3 Weighting Schemes

Now that we have defined the objectives and the theoretical basis of our work, we are going to give the concrete method of word sense comparison. In order to estimate the semantic similarity of words, the distributional hypothesis suggests to estimate the level of contextual similarity. In this context, we must evaluate the similarity between two nouns which are represented by their respective sets of observations X_{ip} on p variables (or attributes). For this purpose, the attributional representation of a noun consists of ordered pairs $\langle n, c \rangle$ where c is a context of the noun n. The eligible contexts are those that stand within a given distance or in a direct syntactic relation with the noun n. Precisely, c is an ordered pair $\langle w, r \rangle$ where w is a context word and r is the number of words between n and w or is a syntactic relation that holds between n and w. In the case of windows-based context, we take $r \leq 4$ and in the case of syntax-based context we restrict the set of the values for c to those that are most reliably identified in an automatic manner:

1. $\langle verb, subject \rangle$: n is subject of the verb
2. $\langle verb, direct\ object \rangle$: n is the direct object of the verb
3. $\langle adjective, modifier \rangle$: n is modified by the adjective
4. $\langle noun, modifier \rangle$: n is modified by another noun

For example, if the noun *controversy* appears with the verb *surrounding* in a subject relation, we will have the following triple $\langle$*controversy, surround, subject*$\rangle$[1] and the tuple $\langle$*surround, subject*$\rangle$ will be an attribute of the word context vector associated to the noun *controversy*.

In order to give appropriate importance to the more informative words compared to the closed class terms such as pronouns, auxiliary verbs and semantically empty words, e.g., the nouns *entity*, *object* and *thing*, the adjectives *little* and *real*, and the verb *exist*, a context weighting must be applied.

[1] For the statistical representation, we take the lemmatized forms of the words.

3.3.1 Inverse document frequency

The inverse document frequency (Spärck-Jones, 1972) was introduced in order to weight index terms in Information Retrieval. In the context of the syntactic attributional similarity paradigm, we define it as in Equation 3.1 where n is the target noun, c is a given attribute, N is the set of all the nouns, $|.|$ is the cardinal function and $tf(n,c)$ is the frequency of the noun n in context c.

$$tf.idf(n,c) = tf(n,c) \times \log_2 \frac{|N|}{|\{n_i \in N | \exists (n_i,c)\}|} \tag{3.1}$$

In IR, this value is used to measure how important is a given indexing term in distinguishing relevant documents. When a term is very common in the corpus its weight converges to 0, while more importance is given to infrequent terms. In the context of semantic relations extraction, we have a context in place of indexing term. Similarly, a very common context, one that co-occurs with many words of the vocabulary, e.g., the definite article *the*, is of little use to tell similar from dissimilar words, while specific contexts highlight important relations.

3.3.2 Pointwise mutual information

The value of each attribute c can also be seen as a measure of association between the noun being characterized and the context c. There is vast literature on Pointwise Mutual Information (PMI) applied to collocation mining (Manning & Schütze, 1999), and it is known that PMI, computed using large co-occurrence window, detects topically related words (Brown *et al.*, 1992) and important collocations (Church & Hanks, 1990). On this ground, Turney (2001) and Terra & Clarke (2003) proposed to use the PMI as defined in Equation 3.2 where n is the target noun and c is a given attribute.

$$PMI(n,c) = \log_2 \frac{P(n,c)}{P(n) \times P(c)} \tag{3.2}$$

It is a transformation of the independence expression, $P(n,c) = P(n)P(c)$, into a ratio. Positive values indicate that words occur together more than would be expected under an independence assumption. Negative values indicate that one word tends to appear only when the other does not. Values close to zero indicate independence. Thus, PMI is capable to indicate contexts that are strongly associated to the characterized noun, such that are particular of its usage.

3.3.3 Conditional probability

Another way to look at the relation between a noun n and a context c is to estimate their conditional probability of co-occurrence as in Equation 3.3. In our case, we are interested in knowing how strongly a given attribute c may evoke the noun n. Thus, when we compare two nouns n_1 and n_2 in respect to a given context c, we are interested to know how much inclined are they to appear in this context while the behavior of the context itself is ignored as it is common factor.

$$P(n|c) = \frac{P(n,c)}{P(c)} \tag{3.3}$$

Weeds *et al.* (2004) proposed the conditional probability of a word knowing a context as a weighting factor of the features in a vector space.

There exist many other association measures as shown in Pecina & Schlesinger (2006), but we will only focus on the TfIdf, PMI and the conditional probability in this work.

3.4 Measures of Similarity

There exist many similarity measures in the context of the attributional similarity paradigm (Weeds *et al.*, 2004). They can be divided into two main groups: (1) metrics in a multi-dimensional space also called vector space model and (2) measures which calculate the correlations between probability distributions.

Theoretically, an attributional similarity measure can be defined as follows. Suppose that $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip})$ is a row vector of observations on p variables associated with a label i, the similarity between two units i and j is defined as $S_{ij} = f(X_i, X_j)$ where f is some function of the observed values. In this context, we must evaluate the similarity between two nouns which are represented by their respective word context vectors.

As similarity measures are based on real-value attributes, we must define an appropriate function f that will accurately evaluate the similarity between two word context vectors. The following sections introduce one geometric and two probabilistic models studied in the literature.

3.4.1 Vector space model

The distributional hypothesis states that words with similar meanings tend to occur in similar contexts. So, if a correspondence is established between each observed context and a dimension in a multi-dimensional space, this hypothesis can be restated as words with similar usage are close to each other and dissimilar words are distant. There exist many possible measures of distance in

multi-dimensional spaces. In the literature, the most common one is the Cosine similarity measure.

To quantify the similarity between two words, the Cosine of the angle between their respective context vectors is commonly applied. The Cosine is the normalized scalar product of two vectors, i.e., its value is between 0 and 1, thus it is invariant with respect to the vector length. It estimates up to what extent two vectors point along the same direction. It is defined in Equation 3.4.

$$cos(n_1, n_2) = \frac{\sum_{k=1}^{p} c_{1k} \times c_{2k}}{\sqrt{\sum_{k=1}^{p} c_{1k}^2} \times \sqrt{\sum_{k=1}^{p} c_{2k}^2}} \tag{3.4}$$

where c_{1k} and c_{2k} are the context vectors that belong respectively to word n_1 and n_2.

The common practice to build a vector space in NLP is to assign a new perpendicular basis vector to each new feature. While this practice is justified by the geometric point of view, as the Cosine assumes that the basis vectors of the space are two by two perpendicular, it is not always compatible with the linguistic side of the problem. For example, in synonymy discovery, we assume that synonymous words have similar contexts and in consequence the Cosine of the angle between them is close to 1. However, the Cosine between every two basis vectors is always 0 by construction, although they may correspond to synonymous contexts[1]. This problem could be at least partially addressed by a number of approaches. For example, LSA (Landauer & Dumais, 1997) partially deals with this as it smoothes missing and erroneous observations through the least square approximation method. Similarly, Random Indexing (Sahlgren, 2001) assigns random, not necessary perpendicular, vectors to the features. This technique bears the potential to deal with the problem if assignment of vectors was performed in an informed manner using previously built knowledge bases or an iterative process.

3.4.2 Probabilistic models

3.4.2.1 Lin's measure

Lin (1998b) proposed an information theory-based similarity measure. The rationale behind it is that the similarity between two entities n_1 and n_2 can be measured by the ratio between the amount of information needed to state the commonality of n_1 and n_2 and the information needed to fully describe n_1 and n_2. In the context of word similarity it is defined as in Equation 3.5.

$$Lin(n_1, n_2) = \frac{2 \sum_{c \in A} \log_2 P(c)}{\sum_{c \in B} \log_2 P(c) + \sum_{c \in C} \log_2 P(c)} \tag{3.5}$$

[1] Similar argument is given in Senellart & Blondel (2008), though for documents.

where

$$A = B \cap C, \tag{3.6}$$

$$B = \{c | \exists (n_1, c)\}, \tag{3.7}$$

$$C = \{c | \exists (n_2, c)\}. \tag{3.8}$$

According to the characterization made by Weeds *et al.* (2004) this similarity measure tends to prefer words that have frequency similar to the frequency of the target word.

3.4.2.2 Ehlert model

The confusion probability, Equation 3.9, introduced in Sugawara *et al.* (1985) for the purpose of speech recognition and later explored as a lexical similarity measure in Ehlert (2003), estimates the probability that one word can be substituted by another one. As the words are more likely to be seen in identical contexts, they are more likely to be substitutable for one another and they are more likely to be confused when the missing word is to be predicted from a given context. The rationale behind this measure is further extended in Charles (2000) who takes the entire sentence as a single context.

$$Ehl(n_1|n_2) = \sum_{c \in A} \frac{P(n_1|c) \times P(n_2|c) \times P(c)}{P(n_2)} \tag{3.9}$$

According to the classification of Weeds *et al.* (2004), the measure is similar to the Cosine as both take as more similar candidates with higher frequency.

Once again, there exist many other similarity measures (Dias, 2010), but we will only focus on the Cosine, Lin's and Ehlert's measures in this work.

3.5 Global Similarity and Local Similarity

3.5.1 Global similarity

The approaches reviewed so far which build context attributional representations of words do so from a corpus as one huge text and do not respect the document limits. We call *Global similarities* the similarity estimations obtained in this manner.

However, this approach poses many problems for polysemous words as contexts which are pertinent to different meanings are gathered into a single global representation when they should be differ-

entiated. In this context, Freitag *et al*. (2005) found high correlation between polysemy and error level and concluded that polysemy level is characteristic of the difficulty of a test.

3.5.2 Local similarity

The idea presented in this section is the first contribution of our work. It is based on and mainly inspired by the works of Landauer & Dumais (1997), Turney (2001) and Agirre & de Lacalle (2003).

According to Gale *et al*. (1992) "*[…] if a polysemous word such as "sentence" appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense*". From this assumption follows that if a word representation is built out of single discourse evidences, it most probably describes just one sense of that word. Therefore, by obeying document borders, mixing different word senses can be avoided. Turney (2001) also demonstrates that synonyms tend to co-occur in texts more often than by chance. Similar supposition is made by Landauer & Dumais (1997) and numerous other works that treat the document as context or use search engine hit counts, where synonyms are sought among words that co-occur in the same set of documents.

According to Mohammad & Hirst (2010) "*[…] words that occur together in text tend to refer to senses that are closest in meaning to one another*". This comes to confirm again the rationale behind Landauer & Dumais (1997) and Turney (2001) that if between words holds a synonymy relation, they tend to appear as synonyms, when occurring within the same discourse. This is true in extent reciprocally related to the level of polysemy, as the results confirm.

To summarize, when a pair of words is encountered in the same document, (1) they behave as if they are nearly monosemous, and (2) when they have similar meanings they tend to signify this common idea. What makes these two observations useful is the tendency of the synonyms to co-occur within the same documents[1].

As a consequence, we apply the *one sense per discourse* paradigm and compare nouns only within a single document based on the attributional similarity paradigm.

Apparently, statistics gathered from a unique, possibly short, text may not be reliable as we will assess in Chapter 4. As a consequence, in order to obtain more stable results, we average attributional similarity values over the set of documents in which two nouns occur and introduce the $Local(.,.)$ function as in Equation 3.10, where $sim(.,.)$ is any function from Section 3.4 and d is any document of a set of documents D in which both n_1 and n_2 occur. Such an approach uses similarity measures in a local, i.e., per document, context and we call it *Local* similarity.

[1] This can be seen as the will of the writer to avoid repetition and produce more fluent text.

$$Local(n_1, n_2) = \frac{\sum_{d \in D} sim(n_1, n_2)}{|D|} \tag{3.10}$$

This approach to the polysemy problem was first proposed in Moraliyski & Dias (2007) and thoroughly evaluated in Dias & Moraliyski (2009). A claim is made that in most cases *Local* similarity compares statistical representations of word meanings as opposed to words and thus it is similar to the measures of concept-distance proposed in Agirre & de Lacalle (2003) and Mohammad (2008). However, *Local* similarity differs from those approaches as we do not make use of any preexisting knowledge source in order to build the distributional representations of the concepts, rather we rely uniquely on an automatically acquired corpus[1].

3.5.3 Combined similarity

Given a pair of words in a discourse, we can assume that they are semantically related since they most probably serve one and the same communication intent. Further, when contextual overlap is present, then we may confidently assume that the words share meaning in the given context. However, these conditions appear to be too demanding in the case of rare words. Statistical evidence show a strong positive correlation between frequency and polysemy (see Table 4.6). Thus, for rare words with respectively low polysemy or words with few related meanings *Local* similarity does not show any advantage over *Global* similarity. In such cases, the *Global* similarity measure should be preferred. Another obstacle in front of *Local* similarity are words with two acceptable spelling variants, e.g., *defense* and *defence*, and stylistically colored words, e.g., *begin* and *commence*, that are rarely used within the same discourse. This makes us to believe that *Global* and *Local* approaches have properties that complement each other. Moraliyski & Dias (2007) introduced the *Product* measure, a multiplicative combination of both *Local* and *Global* similarities as defined in Equation 3.11 and Dias & Moraliyski (2009) further studied it.

$$Product(n_1, n_2) = Global(n_1, n_2)^{\gamma} \times Local(n_1, n_2)^{(1-\gamma)}, \gamma \in [0, 1]. \tag{3.11}$$

In fact, Equation 3.11 is a generalization of both similarity measures. Indeed, when $\gamma = 0$, only the *Local* similarity is taken into account while when $\gamma = 1$ only the *Global* similarity is applied.

[1] This issue is discussed in details in Chapter 4.

3.6 Summary

In this chapter, we motivated and introduced our first main contribution, i.e., the *Local* similarity and the *Product* similarity measures. The measures are intended to alleviate the problems to lexical analysis caused by polysemy. The rationale behind it is the *one sense per discourse* hypothesis (Gale *et al.*, 1992). The *Local* similarity measure achieves its purpose by exploiting the corpus structure i.e, the document limits. With this respect, our method is comparable to Latent Semantic Analysis (LSA) and other works that treat the document as a context as both require first order co-occurrence. However, in contrast with LSA, here the co-occurrence is not an evidence of association, but rather a means of an implicit Word Sense Disambiguation (WSD).

The comparisons within a document of few hundreds of words is practically performed on a very small number of observations. One common possibility to mitigate the data sparseness problem would be to merge together all the documents that contain a pair of nouns and build one statistics for each word. However, the observations made in Mohammad (2008) and Erk & Padó (2008), that when two polysemous synonymous words occur in one discourse they tend to co-occur as synonyms, highlight only a tendency and not an absolute rule. Thus, the effect of merging all relevant texts together would be a step back to *Global* and its drawbacks in treating polysemy.

Since *Local* similarity imposes a set of requirements that limit its application, we proposed the *Product* similarity, a generalization of the common *Global* and the new *Local* similarities, that through suitable fitting parameter can cover the specter of polysemy better than each of the composing measures alone.

In the following chapter, we examine our reasoning. We give an evaluation of the proposed measures against a set of TOEFL-like tests derived in Freitag *et al.* (2005) and a comparison with the *Global* mode of contextual similarity.

Chapter 4

Empirical Studies in Closed Environment

A clear distinction must be made between two areas of application of contextual similarity measure as a measure of similarity of word meaning discussed in the literature. On one side is the task of solving a synonymy test, where it is known that synonymy relation holds for one and only one pair of a set of words. We refer to this problem as *synonymy detection*. This task is usually performed under the form of a test, where the only present synonym have to be detected, hence synonymy detection. As it is an artificial task, it has mostly an evaluative purpose and provides a common base for comparison between different methodologies.

On the other side is the task which occurs when for a given target word we have to find a set of close, in terms of meaning, words among all the words of the vocabulary. We refer to this task as *synonymy discovery*. Statistical evidences (see an experiment described in Section 1.5) support the intuition, that synonymy discovery is more difficult compared to synonymy detection. Indeed, both tasks are, the extremes of one and the same problem. Many authors take this direction, (Curran & Moens, 2002; Grefenstette, 1993; Lin *et al.*, 2003), however their results are difficult to compare since they evaluate on different bases.

In this chapter we, adopt the former technique of evaluation, i.e., synonymy detection. The goal of the first set of experiments is to provide a proof of concept and to demonstrate the viability of the *Local* and *Product* similarities. The evaluation framework used for this purpose is the commonly used TOEFL test set. Landauer & Dumais (1996) adopted this exercise for human language command assessment as means of evaluation in NLP and it was afterwards employed in many works in the field (Ehlert, 2003; Freitag *et al.*, 2005; Rapp, 2004; Sahlgren, 2001; Terra & Clarke, 2003; Turney, 2001). As we argue in the Introduction, we are concerned with the noun part of the vocabulary as it provides a reasonably difficult task, yet it restricts the task to gathering statistics for only one Part-Of-Speech (POS). Verbs, adjectives and adverbs each have different syntactic relations and would have required each one a specific implementation of the statistics collection phase.

4.1 Test Set

In order to illustrate the results of the proposed methodology, we used all 23 noun questions taken from the English as a Second Language (ESL) multiple choice test (Tatsuki, 1998), all 19 noun cases from the TOEFL[1], and the subset of all 103 noun questions out of the 301 test cases manually collected from Reader's Digest. This set was used in Turney *et al.* (2003). The biggest set consists of 355 randomly selected noun tests from the set used in Freitag *et al.* (2005). Thus, the test set rounds the 500 cases[2].

These 4 test sets are used in a number of works, taking various approaches to the problem of synonymy detection. Any comparison between their published results is difficult and biased as the sets vary widely in the adopted notion of synonymy, polysemy and perceived difficulty. For example, ESL is taken from an educational site that intends to teach two thousand basic English words. ESL set consists of 50 cases where the target word is given in a context of a sentence. Thus, the synonymous pairs are tight synonyms, that can replace each other in the given context and the decoys share significant semantic component with the correct answer, e.g., *barrel* | *cask* | *bottle* | *box* | *case*. In contrast, TOEFL tests contain one strong pair of synonymous words while the other candidates are usually rather distant in meaning, e.g., *task* | *job* | *customer* | *material* | *shop*. While this property might not make much difference for a human taker of the test, it significantly influences the performance of a system that takes decision based on word usage statistics. The set collected from Reader's Digest, for brevity we refer to it as RD, contains a fraction of tests in which the pair of words that make the correct answer are indeed associatively related, e.g., *cardiology* | *heart* | *gambling* | *disorder* | *sound*. Such a test case is probably better solved by LSA or PMI-IR, but is misleading for a system that aims at tight semantic relations. The set compiled by Freitag *et al.* (2005) covers the entire specter of polysemy, close synonyms, pairs that are synonyms through some rare sense of one or both of the candidates, e.g., *mass* | *raft* | *trouble* | *reputation* | *fitzgerald*, and cases where the synonymy relation is established through figurative interpretation, e.g., *police* | *law* | *ka* | *sandwich* | *megawatt*.

The success over synonymy tests does not guarantee success in real-world applications and the tests also show problematic issues as shown in Freitag *et al.* (2005). However, the scores have an intuitive appeal, they are easily interpretable and the expected performance of a random guesser (25%) and of a typical non-native speaker are both known (64.5%). Thus, TOEFL-like tests provide a good basis for evaluation.

[1] Educational Testing Service, `http://www.ets.org/` [*13th July, 2011*]

[2] The 80 TOEFL test cases were provided by Thomas Landauer, the 50 ESL cases and the 301 RD cases were provided by Peter Turney. The set of Freitag *et al.* (2005) is available at `http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz` [*13th July, 2011*]

4.2 Corpus

4.2.1 Reuters

Any work based on the attributional similarity paradigm depends on the corpus used to determine the attributes and to calculate their values. Terra & Clarke (2003) used a terabyte of Web data that contain 53 billion words in 77 million documents, Sahlgren & Karlgren (2002) used a 10 million words balanced corpus with a vocabulary of 94 thousand words and Freitag *et al.* (2005) and Ehlert (2003) both used the 256 million words North American News Corpus (NANC) (Graff, 1995).

For the experiments in this chapter we first considered the freely available Reuters Corpus Volume 1 (RCV1) (Lewis *et al.*, 2004). It is known that a domain corpus outperforms general purpose ones in terms of accuracy for domain specific terminology. However, high precision comes at cost of recall, which might be severely limited (Baroni & Bisi, 2004). The proposed *Local* similarity measure requires both candidate words to appear few times each within a single document. We observed that a substantial proportion of word pairs had zero occurrence in RCV1 although the RCV1 consists of more than 800 thousand stories produced by Reuters journalists between August 20, 1996 and August 19, 1997. This situation can be explained by the fact that TOEFL tests consist of unusual vocabulary for news stories. As a consequence, we proposed a methodology to automatically build a corpus that adequately covers the vocabulary under study.

4.2.2 Web as a corpus

Regarding the Web as a live corpus has become an active research topic (Bollegala *et al.*, 2007) and a number of works benefit from the vast coverage of Web. In particular, Keller & Lapata (2003) consider the application of co-occurrence counts of verb-object pairs obtained from the Web for semantic relations discovery, Nakov & Hearst (2005) apply search engine hit counts to noun bracketing resolution and Zhu & Rosenfeld (2001) use Web-based n-gram counts to improve language modeling.

As we did not want to reduce the test set to fit a standard off-line corpus, we decided to build a corpus suitable to the problem at hand. For this purpose, we used the Google Application Programming Interface (API).[1] For each test case, the search engine was queried with all the pairs that consist of the target word and one of the candidates. For example, for the test case *fall* | *autumn* | *franc* | *island* | *banquet* the queries *fall autumn*, *fall franc*, *fall island* and *fall banquet* were sent to the search engine. Subsequently, all of the seed results were collected and a set of selected links were followed to gather more textual information about the queried pairs.

We used a set of heuristics in order to choose which links to follow. We defined the *Text Quality* function as in Equation 4.1. If the $TQ(t)$ of text t is low, then it is useless to follow the links in t.

[1] As of November 2010 Goolge API is not available any more.

Otherwise, we should follow the links in t until enough textual data has been gathered.

$$TQ(t) = \frac{\sum_{p_j \in t} tf(p_j, t) \times idf(p_j)}{\max_{p_j \in t}(tf(p_j, t) \times idf(p_j)) \times card(\{p_j | tf(p_j, t) > 0\})} \tag{4.1}$$

where

$$idf(p) = \log_2 \frac{card(T)}{card(\{t_i \in T | tf(p, t_i) > 0\})}$$

where t_i is a Web page and p_j is a pair of words, $tf(p_j, t_i)$ is the occurrence frequency of the pair p_j in the text t_i, and T is the set of texts retrieved so far.

The basic idea of the *Text Quality* function is to give preference to texts where only the rarest pairs occur. Indeed, if t_i contains one rare pair p_j with high $tf(p_j, t_i) \times idf(p_j)$ and many others for which we already have many texts, i.e., with low idf, then the $TQ(.)$ value will be low. As a result, this will lead to choose only a few links from this page for further crawling as the new textual material would bring more of the same.

One of the problems with Web pages is that some of them only consist of link descriptions and do not contain meaningful text. In order to ensure that the retrieved Web pages will provide useful text material as well as useful links for further crawling, we propose a simple heuristic defined in Equation 4.2 which integrates the $TQ(.)$ function. We call it the *Page Quality* function and denote it $PQ(.)$ where c is a tuning constant. In our experiments, we arbitrary set $c = 300$. This requires that the Web page contains at least 300 characters of text for each link on the page.

$$PQ(t) = \frac{TQ(t) \times TextLengthInCharacters}{c \times LinksCount} \tag{4.2}$$

In order to take advantage of syntactic information, to reduce the noise of the statistics and the data sparseness problem, the overall collection of Web pages was then lemmatized and POS-tagged using the MontyLingua software package (Liu, 2004). At this stage, the corpus was ready to extract contextual data. Next, we extracted the verb and modifier predicate structures as described in Section 3.3, to which statistics we refer as syntactic-based contextual data. Thus, the corpus consists of 500 million words in 96 thousand documents in which each sentence is a predicate structure. This corpus we refer to as Web Corpus for Synonym Detection (WCSD)[1]. The benefit of such a corpus is to maximize the ratio of the observed instances to the volume of the processed text.

[1]This corpus is available at `http://hultig.di.ubi.pt/~rumen`

4.3 Comparative Results

Table 4.1 shows the differences in terms of accuracy obtained by comparing the RCV1 and WCSD, where Cos TfIdf stands for performance of Cosine measure over a vector space with features weighted by TfIdf, Cos PMI means Cosine combined with features weighted by Pointwise Mutual Information (PMI), and Cos Prob means Cosine with conditional probability as feature weight. Both Lin and Ehlert stand for the corresponding probabilistic models overviewed in Section 3.4.2.1 and Section 3.4.2.2. The results are obtained over the set of 500 tests, described in Section 4.1. The first two columns of Table 4.1 show the performance of the *Global* mode of the similarity measures and the following two columns contain the corresponding *Local* mode performance figures. All differences between RCV1 and WCSD are statistically significant at 95% confidence level.

As expected, the WCSD allows significantly higher accuracy. These results clearly show that the corpus used to compute the measures influences drastically the performance of any experiment and comparisons of different methodologies should always be made based on the same statistical evidences. As a consequence, the results given hereafter are only indicative as better or worse results may be obtained on different experimental frameworks.

Table 4.1: Comparison between RCV1 and WCSD.

	Global		Local	
	RCV1	**WCSD**	**RCV1**	**WCSD**
Cos TfIdf	38%	67%	42%	61%
Cos PMI	40%	68%	42%	60%
Cos Prob	36%	61%	40%	65%
Ehlert	41%	63%	44%	70%
Lin	38%	60%	41%	56%

4.3.1 Window versus syntactic dependencies statistics

After we have performed tests with syntactic-based contexts, we repeated the respective calculations with windows-based contextual data. It was initially only a test to verify the viability of the model. However, it turned out it was impossible to compute the required similarities for the entire test set with the same corpus, the reason being the excessive amount of statistics. For comparison, the summarized statistics extracted from the shallow parsed corpus consisted of 28 million observations, while the corresponding data extracted by window of size 4 consisted of 230 million observations. Due to the prohibitive volume of the computation for windows-based statistics, the results provided in the following sections are obtained only using shallow parsed corpora.

4.3.2 Statistical difference between measures

Dias & Moraliyski (2009) show that there does not exist any single methodology capable to accomplish synonymy detection alone. Depending on the type of the multiple choice question set, different measures and weighting schemas may be applied to improve overall performance. However, it is well known that some measures are highly correlated (Weeds *et al.*, 2004). In this case, the combination of these measures will not contribute to the overall performance. For that purpose, we propose to study the correlation between pairs of similarity measures with the Pearson Product-Moment Correlation test (Fisher, 1915). The figures occupy the first column of Table 4.2.

Table 4.2: Inter-measure correlation.

	Pearson	**Overlap**	**Optimal**
Cos TfIdf & Cos PMI	0.56	58%	77%(386)
Cos TfIdf & Cos Prob	0.36	50%	79%(396)
Cos TfIdf & Ehlert	0.11	45%	85%(427)
Cos TfIdf & Lin	0.77	58%	69%(347)
Cos PMI & Cos Prob	0.43	52%	78%(389)
Cos PMI & Ehlert	0.22	47%	83%(415)
Cos PMI & Lin	0.51	52%	76%(378)
Cos Prob & Ehlert	0.28	45%	79%(395)
Cos Prob & Lin	0.31	44%	77%(386)
Ehlert & Lin	0.06	39%	84%(418)

Additionally, we compute the overlap of correct answers in the second column of Table 4.2 and finally calculate the possible optimal performance that could be obtained by combining two measures both in percentage and number of possible correct answers.

The results are clear as they show that all similarity measures would benefit the most from their association with the Ehlert's measure. In particular, the optimal case could achieve 85% accuracy, i.e., 427 correct test cases by combining the Cos TfIdf and the Ehlert's models. Indeed, both measures share 45% of correct test cases, and the second smallest correlation, i.e., 0.11.

4.3.3 Global versus local similarity

Table 4.3 presents the overall comparative results for two modes of similarity measures, i.e., *Global* and *Local*. The Ehlert as a *Local* similarity evidences the overall best result with accuracy of 70%. The figures show that the *Local* similarity approach improves over the *Global* similarity for Cos Prob and for Ehlert's measures. In parallel, the worst results were obtained by the Lin model reaching 56%. The differences between *Global* and *Local* measures are statistically significant at 95% confidence

level for all the measures.

Table 4.3: Accuracy by measure without *Product*.

	Vector Space Model			Probabilistic	
	Cos TfIdf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Global	67%	68%	61%	63%	60%
Local	61%	60%	65%	70%	56%

4.3.4 The advantage of local similarity measures

Freitag *et al*. (2005) introduced a measure to evaluate the difficulty of a test based on its polysemy. Here, we elaborate on the same idea by considering the behavior of a similarity measure with respect to the level of polysemy of pairs of synonymous and pairs of non-synonymous words. For that purpose, we take polysemy of a pair of words as defined in Freitag *et al*. (2005), i.e., the sum of the number of different senses in WordNet of both words. When a word is not found in WordNet, we take it as if it is monosemous, the rationale being that WordNet lists the most frequent words and the very rare ones usually have only one meaning.

While more polysemous tests are definitely perceived as more puzzling, we prefer to see the performance of a measure with respect to the polysemy as a characteristic of the measure itself. We are interested in how the similarity score changes with respect to the polysemy in two cases - when the words are synonymous and when they are not. Figure 4.1 illustrates a major drawback of the *Global* similarity, namely that the similarity value is proportional to polysemy and respectively to the number of contexts, indiscriminately to whether the words are related or not. We performed a one-way Analysis of variance (ANOVA) (Chambers & Hastie, 1992) to confirm the effect of polysemy on *Global* similarity. The results are presented in Table 4.4. For example, with Cos TfIdf for synonyms, we obtain $F(1, 498) = 40.9$, and for non-synonyms, respectively $F(1, 1447) = 80.8$, both significant at 95% confidence level. This result means that similarity scores change significantly as a function of polysemy. In this particular case similarity scores display higher variance for the non-synonymous than for the synonymous pairs. Similar relation holds for Ehlert measure. For Cos PMI both synonyms and non-synonyms decrease in contextual similarity and the effect of polysemy is comparable in both cases with $F(1, 498) = 28.3$ and $F(1, 1447) = 23.8$, respectively, both statistically significant at 95% confidence level. More plots of other measures in *Global* mode are given in Appendix A.1.1.

In summary, these observations mean that *Global* similarity changes towards the same direction for both synonymous and non-synonymous pairs of words as a function of the number of senses of the words involved. Section 4.3.6 demonstrates how the significant variance of similarity scores of non-synonyms prevents *Global* to discern synonyms from non-synonyms at certain levels of polysemy.

These results are in accord with the findings made in Weeds *et al.* (2004).

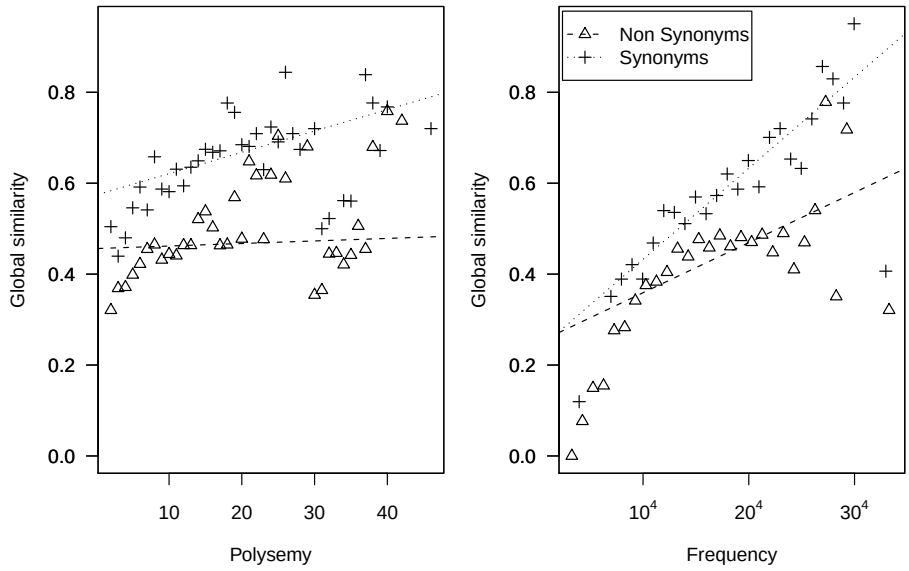

Figure 4.1: *Global* Cos TfIdf by Polysemy and Frequency.

Table 4.4: ANOVA on *Global* similarities.

	Synonyms		**Non-synonyms**	
	$F(1,498)$	$p-value$	$F(1,1447)$	$p-value$
Cos TfIdf	40.9	$< 10^{-9}$	80.8	$< 10^{-9}$
Cos PMI	28.3	$< 10^{-6}$	23.8	$< 10^{-5}$
Cos Prob	1.1	< 0.3	0.9	< 0.3
Ehlert	2.3	0.13	97.9	$< 10^{-9}$
Lin	34.5	$< 10^{-8}$	24.3	$< 10^{-6}$

On the other hand, *Local* similarity achieves a better differentiation between synonymous and non-synonymous pairs for the left half of the polysemy specter (see Figure 4.2). The reason why the *Local* similarity performs better than the *Global* similarity is that in most of the time it compares monosemous representations and thus it is less influenced by polysemy. As a result, *Local* similarity achieves partial separation between the two classes for the left half of the polysemy scale. However, with the increase of polysemy, the words shall begin to co-occur with increasing probability in non-synonymous relations. Consequently, the numerator of equation 3.10, p. 46, increases slower than the denominator and the similarities of synonymous and non-synonymous pairs converge.

Table 4.5 presents the results of a one-way ANOVA of the effect of polysemy on *Local* similarities. For example, Cos TfIdf scores for synonyms follow a significant trend to decrease with the increase of polysemy $F(1,498) = 19.4$ with $p-value < 0.05$, while for non-synonyms this tendency is rather weak, $F(1,1447) = 5.7$ with $p-value = 0.02$, i.e., *Local* similarities for non-synonyms are relatively

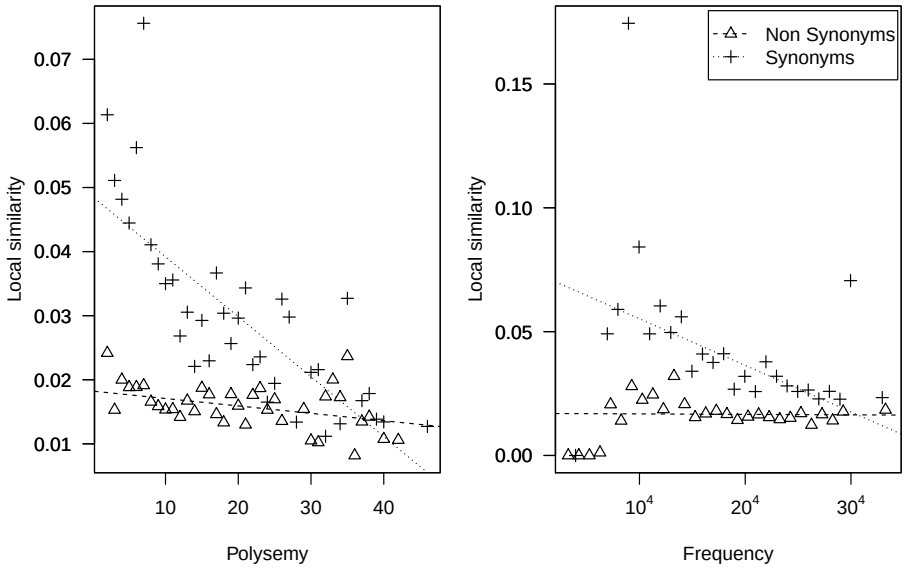

Figure 4.2: *Local* Cos TfIdf by Polysemy and Frequency.

constant across the polysemy scale and lower than *Local* similarities for synonyms as confirmed by Wilcoxon signed-rank test (Hollander & Wolfe, 1973) with $W = 538765$ and $p - value < 10^{-9}$. In contrast with *Global* similarity, for *Local* similarity, this pattern is consistent across all the measures, namely clear-cut separation between synonyms of polysemy less than 20 and all non-synonyms. More plots of other measures in *Local* mode are given in Appendix A.1.2.

Table 4.5: ANOVA on *Local* similarities.

	Synonyms		**Non-synonyms**	
	$F(1,498)$	$p - value$	$F(1,1447)$	$p - value$
Cos TfIdf	19.4	$< 10^{-4}$	5.7	0.02
Cos PMI	35.3	$< 10^{-8}$	3.7	0.06
Cos Prob	14.3	$< 10^{-3}$	0.4	0.6
Ehlert	15.6	$< 10^{-4}$	0.0	0.9
Lin	34.7	$< 10^{-8}$	0.2	0.7

4.3.5 Polysemy, frequency and number of contexts

The results presented here illustrate the relations between polysemy and the performance of contextual similarity as a measure of word meaning similarity. In reality, there exist a close bond between polysemy, frequency and the number of distinct contexts that a word accepts. For example, the correlation between word polysemy and the number of distinct contexts per word varies between 0.53 and 0.62 depending on the syntactic class of the contexts. The figures in Table 4.6 show the strong

Pearson product-moment correlation (Becker *et al.*, 1988) between the three quantities. These figures are interesting because they give a connection between the observations made here and previous results. The characteristics given by Freitag *et al.* (2005) show the positive correlation between the polysemy and the difficulty of a test. In particular, Weeds *et al.* (2004) conclude that in general contextual similarity estimates increase with the increase of frequency of the involved words. In contrast, *Local* similarities show consistent behavior with low estimates for non-synonyms and high scores for low to mid polysemy part of the vocabulary.

Table 4.6: Correlation between polysemy, corpus frequency and contexts counts.

	Polysemy	**Number of Contexts**	**Frequency**$^{\frac{1}{2}}$
Polysemy	1	0.60	0.57[1]
Number of Contexts		1	0.98
Frequency$^{\frac{1}{2}}$			1

4.3.6 Classification confidence

Now, we treat the same set of 500 TOEFL-like tests as a more general classification problem. In other words we take them as a set of 500 synonymous pairs and 1449 non-synonymous ones[2].

We divided both synonyms and non-synonyms in samples S_i and N_j respectively, where the index signifies the polysemy level of the pairs in the sample. We then performed Wilcoxon signed-rank test for each pair $\langle S_i, N_j \rangle$. Wilcoxon signed-rank test is a statistical hypothesis test for assessing whether one of two samples of related observations tends to have larger values than the other. Then, we made a plot where each column shows for a given polysemy level of synonyms S_i up to what polysemy level of non-synonyms N_j the measure is capable to distinguish between S_i and N_j with 95% confidence. The results show that both measures are capable to tell synonyms from non-synonyms provided that the candidates are of comparable level of polysemy. However, *Global* similarity fails when the correct synonym is of relatively lower polysemy compared to the other candidates. A graphical expression of these results is given in Figure 4.3. The figure shows, for example, that when we consider a sample of all synonymous pairs with polysemy 4, along the horizontal axis, *Global* can only reliably separate them from non-synonymous pairs of polysemy up to 5 senses. When we compare this sample of synonymous pairs with samples of non-synonyms of polysemy up to 5, we detect statistically significant location shift of samples mean values, $p - value < 0.05$. *Global* similarity fails to achieve statistically significant difference between synonymous pairs of polysemy 4 and non-synonymous candidates of polysemy higher than 5. *Global* Cos TfIdf and Lin's similarities are

[1] According to Zipf (1945), the number of senses of a word is related to the square root of its frequency, note however, that this is not the Zipf's law. Further, the Pearson product-moment correlation coefficient accounts for linear relations between two variables. Both conditions impose a suitable transformation of the data, in this case a square root of frequency.

[2] Few decoys appeared in more than one test, thus the unique non-synonymous pairs are less than 1500.

capable to reliably classify only synonyms of polysemy over 15 senses, while Cos PMI shows preference to low polysemy synonyms. More plots of classification confidence for *Global* mode measures are given in Appendix A.2.1.

On the other side, similar plot for *Local* similarity, Figure 4.4, shows that *Local* similarity is more successful along the low polysemy synonyms, while less confident with the increase of polysemy. This result is in accord with the results of ANOVA, which shows significant tendency of *Local* to decrease with polysemy for synonyms, while, at the same time, the influence of non-synonyms is limited within a relatively short interval of variance.

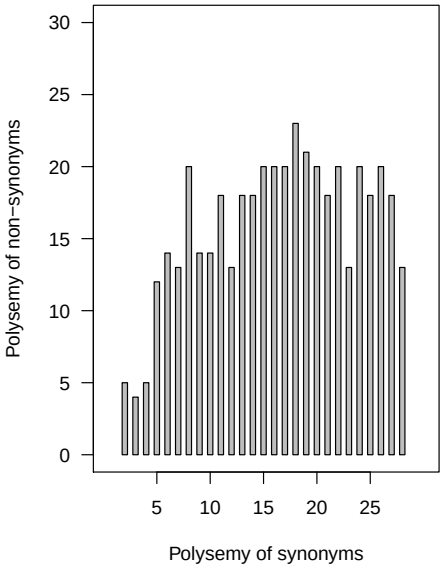

Figure 4.3: *Global* Cos TfIdf Classification Confidence.

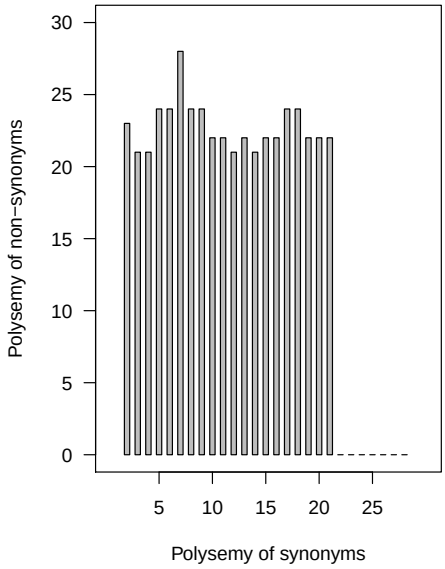

Figure 4.4: *Local* Cos TfIdf Classification Confidence.

4.3.7 Local similarity and free association norms

We were interested to know how does *Local* similarity behaves compared to the human perception of word meaning similarity. The University of South Florida Free Association Norms (USFFAN)[1] comprises over a million responses of more than 6,000 participants to 5,019 stimulus words. Participants were asked to write the first word that came to mind that was meaningfully related or strongly associated to the presented word on the blank shown next to each item. For example, if given *book*, they might write *read* on the blank next to it. This procedure is called a discrete association task because each participant is asked to produce only a single associate to each word (Nelson *et al.*, 2004).

Out of this set, we retained only those 39,660 stimulus - response pairs of which both words were found in WordNet and we labeled them as synonymous when both words appeared in one and the same synset. Next, we augmented the list with the corresponding polysemy levels. The ANOVA analysis on both USFFAN and *Local* shows a common pattern. Association strengths for synonyms for USFFAN

[1] Available at `http://web.usf.edu/FreeAssociation/` [13th July 2011]

weakly, $F(1, 1802) = 22.1$, but steadily, $p - value < 0.05$, decrease (see Figure 4.5), while for non-synonyms the null-hypothesis, that the mean similarity for all polysemy levels is the same, could not be rejected, i.e., mean values are relatively constant across the polysemy specter for non-synonyms. In contrast, in *Global* mode similarity for both synonyms and non-synonyms ANOVA shows significant variance.

These observations on USFFAN illustrate how polysemy relates to the human perception of synonymy test difficulty and allows us to hypothesize that *Local* similarity better mimics human similarity perception in comparison to *Global* similarity.

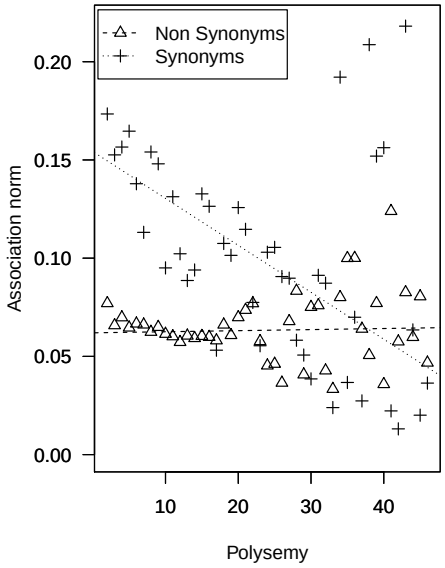

Figure 4.5: Free Association Norms by Polysemy.

4.4 Product

In Section 4.3.2, we saw that there is no single measure that successfully solves the synonymy problem. Similar motivation underlies another studies (Turney *et al.*, 2003). We saw, as well, that *Global* similarities only reliably distinguish synonyms at one end of the polysemy specter and although *Local* performs better in general it, likewise humans, grows confused when presented with more polysemous cases.

Here, we set to explore how performance of *Product* varies as a function of polysemy. To this end, we define the parameter γ of Equation 3.11, p. 46, to be a linear function of the polysemy of both involved words as shown in Equation 4.3, where p_1 and p_2 are the polysemy levels of both words, and $maxP$ stands for the maximal polysemy of the test set, here $maxP = 33$. The chosen function allows for importance of *Global* and *Local* on one side and polysemy on the other side to be varied independently. We optimized the values of the parameters a and k through a 10-fold cross validation

process. The data set was divided at random in 10 equal groups and each group was once held out as an evaluation set and the rest 90% of the tests served as a training set over which the parameters were optimized. Table 4.7 presents the results.

$$\begin{aligned} \gamma &= a + k \times \frac{p_1 + p_2}{2 \times maxP}, \\ a &\in [-1, 1], \\ -a &\leq k \leq 1 - a, \\ -a &\leq \frac{k}{maxP} \leq 1 - a. \end{aligned} \tag{4.3}$$

Table 4.7: Accuracy by measure with *Product*.

	Vector Space Model			Probabilistic	
	Cos TfIdf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Global	67%	68%	61%	63%	60%
Local	61%	60%	65%	70%	56%
Product	$72\% \pm 6$	$69\% \pm 6$	$70\% \pm 7$	$\underline{73\%}$ ± 4	$71\% \pm 8$
P - G	5%	1%	9%	10%	$\underline{11\%}$
a	0.14 ± 0.03	0.64 ± 0.10	0.37 ± 0.03	0.49 ± 0.32	0.16 ± 0.03
k	0.76 ± 0.16	-0.16 ± 0.38	0.22 ± 0.14	0.66 ± 0.28	0.35 ± 0.18

All *Product* - *Global* (*P* - *G*) differences are statistically significant at 95% confidence level with the exception of Cos PMI. On the other side, *Product* improves on *Local* for all measures except for Ehlert, where the difference is not statistically significant.

Product of Cos TfIdf uses *Local* for low polysemy cases, i.e., $a = 0.14$, and shows strong tendency to use *Global* for high polysemy cases, i.e., $k = 0.76$. This is to be expected as we saw that for low polysemy *Local* performs better, while with the increase of polysemy *Global* improves. *Product* of Cos PMI shows moderate initial preference to *Global* and is nearly insensitive to the changes in polysemy. This behavior is a result of a common property of *Global* and *Local* that both get equally confused at high polysemy levels as could be conferred by the classification confidence plots, respectively Figure A.17, Appendix A.2.1, and Figure A.22, Appendix A.2.2.

Ehlert's measure seems to be intractable by this optimization process as evidenced by the significant variance of both optimized parameters. This behavior is explained by the high variance of Ehlert's *Global* scores for non-synonyms. Further, considering plots of classification confidence, e.g., Figure A.19 and Figure A.24 we note significant overlap between *Global* and *Local*.

Both Lin and Cos TfIdf measures show relatively low similarities for low polysemy nouns for *Global*.

This is why both measures have low initial preferences to *Global*. *Global* Cos TfIdf however shows better separation of synonyms and non-synonyms at high polysemy in contrast to Lin. This is why Cos TfIdf is the measure that shows the strongest tendency towards *Global* as function of polysemy.

In summary, the gain is most significant for those pairs of measures that differ most. We saw that polysemy is a factor that strongly influences performance of all studied measures. At the same time, it was noted that in this respect the measures differ and this led to the hypothesis that optimization with polysemy as a parameter could bring positive results. Although $\gamma = f(p)$ for some non-linear function $f(p)$ of polysemy p could be more appropriate, e.g., in the case of Ehlert's measure, it is important to note that polysemy itself can not be taken as a clue to synonymy and this makes it only partially helpful to the optimization process.

4.5 Summary

In this chapter, we performed an exhaustive evaluation of three modes of measure of similarity, namely the common *Global* mode and the two newly proposed modes, i.e., *Local* and *Product*. We saw that *Local* improves on *Global* similarity for Cos Prob and Ehlert's measures and *Product* improves over *Global* for Cos TfIdf, Cos Prob, Ehlert's and Lin's models. The improved performance of *Product* is possible due to the wider area of confident separation of synonyms from non-synonyms that *Local* shows in comparison to *Global* similarity.

The analysis in this chapter complements the results of Freitag *et al*. (2005) and Weeds *et al*. (2004) by providing more detailed study of behavior of measures not only in respect with polysemy but also in respect with the semantic relation under study. Application of ANOVA to *Global* shows that for *Global* the variation of similarity scores of non-synonyms is commensurate to the variation of similarity scores of synonyms. As a result *Global* measures only classify synonyms under very specific conditions.

On the other hand, our results show that both *Local* and USFFAN produce high scores for synonyms for low to middle levels of polysemy, while non-synonyms are relatively constant and lower in relation to synonyms along the entire polysemy specter. Based on this evidence, we argue that *Local* mimics more closely human perception of word similarity compared to *Global*. Thus the new measure is more plausible from both linguistic and cognitive points of view.

Chapter 5

Discovery of Word Paradigmatic Relations

The syntagmatic relations within a sentence are usually between words of different syntactic classes. They hold together the various tokens to form sentences. Thus, the synonymy and antonymy are perceived as paradigmatic rather than as syntagmatic relations. However, closely related words of both kinds of relations (synonymy and antonymy) might form syntagmatic constructions as well. The antonyms co-occur in order to create contrast (Justeson & Katz, 1991), while synonyms and co-hyponyms co-occur in coordination and conjunction constructions.

If we only take coordination constructions, we are likely to extract good synonyms along with many other pairs of words that are perceived to have some properties in common but not similar in meaning. If, now, we impose a constraint that the pairs of words are interchangeable in context, then many words that can not be used as synonyms would be filtered out. But if only the paradigmatic information is taken into account, we can not distinguish between synonyms and antonyms, which only differ in their syntagmatic manifestation. Those observations suggest that in order to detect a given semantic relation, it is necessary to take advantage of both perpendicular directions, i.e., the syntagmatic and the paradigmatic, in order to highlight pairs of words that have specific properties in each semiotic dimension.

The literature on statistical word semantics shows that the synonyms can be detected, provided that a correct candidate is present. The weak point of such an approach is that disconnected features are used as a ground to judge about such a complex phenomena as word meaning. The proposed *Local* measure relates portions of this statistics, by taking them only from one document where in most of the cases the supposition that they belong to the same word meaning holds. Here, we propose a method that seeks for synonyms among words that satisfy the complex interplay between all the words within a sentence.

A WordNet synset has on average 2.3 members. This means that a monosemous word has about 1.3 synonyms. As the empirical evidences show, the exhaustive search has negligible chances to get good synonym ranked at the top. van der Plas & Tiedemann (2006) show that a syntactic context similarity measure manages to rank a good candidate at the top in about 8.8% of the time and does so for about 6.4% of the target words.

The problem is still more difficult when the relation between a pair of words is established not through their most common meaning but by some uncommon usage or they have related meanings only within a certain domain. For example, the words $\langle node, vertex \rangle$ are near-synonyms only in the context of graph theory while the pair $\langle spring, leap \rangle$ is synonymous only through the 4th most frequent meaning of *spring* according to WordNet. Discovering those relations is a very difficult task for purely statistical methods as the rare events are obscured by the most frequent ones (Bordag, 2003).

The difficulty of the statistical methods to find close relations are usually explained by the fact that the words are rather promiscuous with respect to the semantic frames in which they can fit. This specific behavior has primary origin in polysemy, the capacity of the words to have more than one meaning, and in the creative use of language. Locally, a single context is not enough to select a word sense. Rather, the semantic relations between the words within a sentence and a discourse select their meanings (Kaplan, 1950). Following the same idea, Charles (2000) collected a number of sentences, removed one content word from each sentence and asked two groups of human subjects to recover the missing words when presented alternatively with lists of sentences or lists of words taken from the same sentences. He observed that sentences impose stronger lexical preference than disconnected words and thus were more reliable evidence for measuring semantic similarity of pairs of words.

Therefore, in this chapter, we aim to find pairs of sentences in which one word is substituted by another one and then to make confident decisions whether both words share meaning or not. Detecting paraphrases provides an elegant solution to the first part of the problem. Paraphrases are sentences sharing an essential idea while written in different ways. As such, from paraphrases, we hope to learn TOEFL-like tests, i.e., clusters of words where there is a target word and a *short* list of semantically related candidates, predominantly in paradigmatic relations with the target. Then, the techniques proposed in Chapter 3 will be applied to solve the automatically created test cases.

In the next section, we introduce an unsupervised language-independent methodology to automatically extract, cluster and align paraphrases which will help creating the test cases in an automatic way.

5.1 Paraphrase Detection

Paraphrase is a restatement of a text or passage, using other words. This is often accomplished by replacing words with their synonyms, hyponyms or hypernyms and changing word order. For example, the sentences in Figure 5.1, taken from Web news stories excerpts, are paraphrases of a news about the release of a comic movie and show that *feature* can be substituted by *documentary*, *mockumentary* or *film* and as such may share common meanings. As a consequence, the extraction of

paraphrases can lead to the identification of semantically related words in a micro-world compared to a macro-world used by exhaustive search strategies which would seek for candidates in the entire vocabulary.

1. *Kazakhs are outraged by the wildly anticipated mock documentary feature Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*
2. *The news follows controversy surrounding the comedy film Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan which cut so close to the funny bone.*
3. *Meanwhile Borat is leaping to the big screen in the mockumentary Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*

Figure 5.1: A sample set of 3 paraphrases.

A few unsupervised metrics have been applied to automatic paraphrase identification and extraction (Barzilay & Lee, 2003; Dolan *et al.*, 2004). However, these unsupervised methodologies show a major drawback by extracting quasi-exact or even exact match pairs of sentences as they rely on classical string similarity measures such as the *Edit Distance* in the case of Dolan *et al.* (2004) and *Word N-gram Overlap* for Barzilay & Lee (2003). For these functions, the more similar two strings are the more likely they will be classified as paraphrases. At the extreme, the "best" pair will be precisely two exactly equal strings. This is clearly naïve and we may state that the more similar two strings are, the poorer will be the paraphrase quality they generate. It is desirable to identify paraphrases which have certain level of dissimilarity, because this is precisely what will open room for semantic relation discovery.

The *Edit Distance* is rather problematic for paraphrase identification as true paraphrase sentence pairs having a considerable amount of word reordering due to distinct syntactic structures are likely to be considered as non-paraphrases. For example, the sentences

1. *Due to high energy prices, our GDP may continue to fall, said Prime Minister, early morning.*
2. *Early morning, Prime Minister said that our GDP may continue to fall, due to growing energy values.*

are in fact paraphrases, however the *Edit Distance* by returning a high value would indicate great dissimilarity.

To overcome the difficulties faced by the existing functions, a new paraphrase identification functions have been investigated by Cordeiro *et al.* (2007b) such as Entropy functions (Equation 5.1), Parabolic functions (Equation 5.2), Trigonometric functions (Equation 5.3), Triangular functions (Equation 5.4) and Gaussian functions (Equation 5.5).

$$f_{Entropy}(x) = -xlog_2(x) - (1-x)log_2(1-x) \quad (5.1)$$

$$f_{Parabolic}(x) = 4x - 4x^2 \quad (5.2)$$

$$f_{Trigonometric}(x) = sin(\pi x) \quad (5.3)$$

$$f_{Triangular}(x) = 1 - 2 \times |x - 0.5| \quad (5.4)$$

$$f_{Gauss}(x) = ae^{-\frac{(x-b)^2}{2c^2}} \quad (5.5)$$

The x value represents some connection feature value, counted between the paraphrase candidate sentences, as for example word n-gram overlaps. In Cordeiro *et al.* (2007c), this value is calculated based on lexical exclusive links counts. For a given sentence pair, an exclusive link is a connection between two equal words from each sentence. When such a link holds then each word becomes bound and can not be linked to any other word. This is illustrated in Figure 5.2, where, for example, the definite article *the* in the first sentence has only one link to the first occurrence of *the* in the second sentence and the second occurrence of *the* remains unconnected.

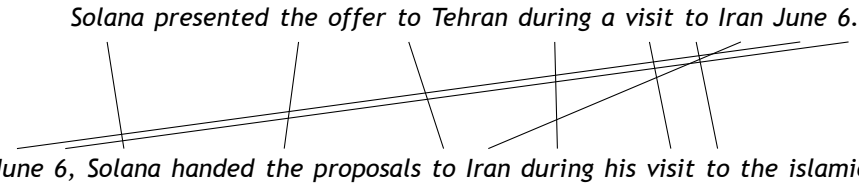

Figure 5.2: Exclusive links between a sentence pair.

In equations 5.1, through 5.5, x is defined as in Equation 5.6

$$x = \sqrt{\frac{\lambda}{m} * \frac{\lambda}{n}} \quad (5.6)$$

where the number of exclusive links binding two sentences is represented by λ, m is the number of words in the longer sentence and n the number from the shorter one. Unlike the classical functions, these ones share the common characteristic of having a hill shaped graph curve (see Figure 5.3), with zero or near zero values near the domain boundaries and a maximum value reached in between. The important property of this type of hill functions is not the exact form of how they are calculated but the general shape of their graphs. These graphs convey a common meaning, since the maximum value is reached strictly inside the $[0,1]$ interval, in some cases near the 0.5 value, which means, on one hand, that a certain degree of dissimilarity between the paraphrase sentences is "desirable" and, on the other hand, that either the excessive dissimilarity or similarity tend to be penalized as

we have the same property of zero approximation, on their boundaries, i.e.:

$$\lim_{x \to 0} f_{hill}(x) = 0 \tag{5.7}$$

and

$$\lim_{x \to 1} f_{hill}(x) = 0. \tag{5.8}$$

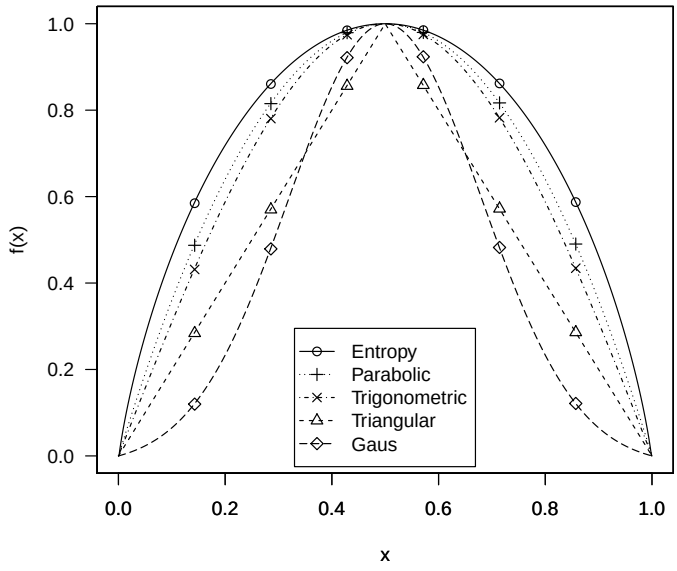

Figure 5.3: *Hill shape* functions for paraphrase identification.

The main difference with the classical paraphrase detection functions, e.g., *Edit Distance* and *Word N-gram Overlap*, is that these latter have

$$\lim_{x \to 1} f(x) = 1. \tag{5.9}$$

An example of a paraphrase that would have high value with classical functions and low value for f_{hill} functions, is shown in Figure 5.4. From the Distributional Hypothesis standpoint, this example contains obviously very low utility.

1. *The stock has gained 9.6 percent this year.*
2. *The stock has gained 9.6% this year.*

Figure 5.4: A too similar paraphrase example.

The results of Cordeiro *et al*. (2007c) suggest that the hill shaped functions (equations 5.1 to 5.5) perform better than the classical ones and better paraphrases were extracted. It is clear, however, that with the current techniques, paraphrase extraction is a difficult problem. In order to accomplish this task in a more dependable way Jing & McKeown (2000), Dolan *et al*. (2004) and others make use of parallel or aligned monolingual corpus. Following those works, Cordeiro *et al*. (2007c) evaluate a number of metrics on extraction of paraphrases from clusters of news stories. Indeed, clustering of complete stories is more robust than extracting just pairs of lexically similar sentences since it relies on more statistical evidence. Thus, extracting paraphrases from stories that are already known to deal with the same subject improves the probability that lexically similar sentences have the same focus.

Further, Cordeiro *et al*. (2007c) proposed another function, with similar characteristics as the functions in Equations 5.1 through 5.5, but performing even better than any other one in most of the standard corpora. This function is called the *Sumo-Metric*, and for a given sentence pair, where each sentence has m and n words respectively, and with λ exclusive links between the sentences, the *Sumo-Metric* is defined as in Equation 5.10 and Equation 5.11.

$$S(S_a, S_b) = \begin{cases} S(m,n,\lambda) & if \;\; S(m,n,\lambda) < 1.0 \\ 0 & if \;\; \lambda = 0 \\ e^{-k*S(m,n,\lambda)} & otherwise \end{cases} \tag{5.10}$$

where

$$S(m,n,\lambda) = \alpha \log_2(\frac{m}{\lambda}) + \beta \log_2(\frac{n}{\lambda}), \alpha, \beta \in [0,1], \alpha + \beta = 1. \tag{5.11}$$

In particular, Cordeiro *et al*. (2007c) show that the *Sumo-Metric* outperforms all state-of-the-art functions over the tested corpora and allows to identify similar sentences with high probability to be paraphrases by defining a non-continuous function of paraphrase similarity as shown in Figure 5.5 compared to the hill shape curve functions.

5.2 Paraphrase Clustering

In literature, it was shown that there are two main reasons to apply clustering for paraphrase extraction. On the one hand, as Barzilay & Lee (2003) evidence, clusters of paraphrases can lead to better learning of text-to-text rewriting rules compared to just pairs of paraphrases. On the other

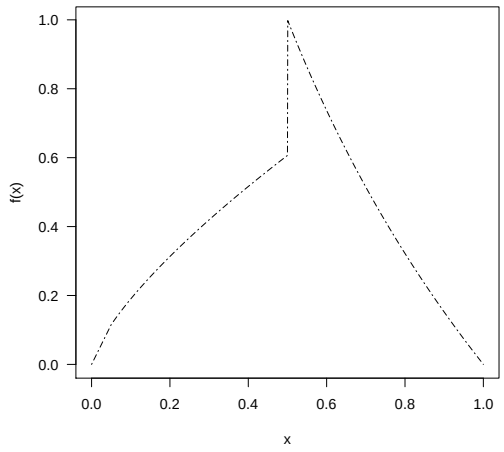

Figure 5.5: *Sumo-Metric* for paraphrase identification.

hand, clustering algorithms may lead to better performance than stand-alone similarity measures as they may take advantage of the different structures of sentences in the cluster to detect new similar sentences.

Thus, instead of extracting only sentence pairs from corpora, one may consider the extraction of paraphrase clusters. There are many well-known clustering algorithms, which may be applied to a corpus of sentences $S = \{s_1, \ldots, s_n\}$. Clustering implies the definition of a similarity or (distance) matrix $A_{n \times n}$, where each element a_{ij} is the similarity (distance) between sentences s_i and s_j. In our context, the similarity measure is the *Sumo-Metric* between two sentences extracted from automatically crawled news stories.

The conclusion from Cordeiro *et al.* (2007a), is that clustering tends to achieve worse results than simple paraphrase pair extraction, in terms of precision. In their work, a cluster of sentences is evaluated as a correct one, if and only if each pair of sentences in the cluster is a correct paraphrase pair. The baseline they adopted was to simply extract all sentence pairs through a similarity function $\sigma(.,.)$, e.g., *Edit Distance*, *Word N-gram Overlap*, *Sumo-Metric*, under the condition that the similarity between sentences exceeded a given threshold value ϵ, which meant that two sentences, s_i and s_j, were considered as a paraphrase pair, if and only if $\sigma(s_i, s_j) > \epsilon$. In this case, a paraphrase pair may be viewed as a *trivial* paraphrase cluster with only two sentences.

However, among the clustering algorithms evaluated by Cordeiro *et al.* (2007a) for sentence clustering, the Quality Threshold (QT) algorithm (Heyer *et al.*, 1999) achieved better precision (64%) than other clustering algorithms (57%) that do not need in advance the expected number of clusters. Therefore, in this work, the QT clustering algorithm is applied to the similarity matrix based on the

Sumo-Metric to obtain clusters of paraphrases. The QT algorithm was originally conceived to tackle the problem of gene clustering and despite the fact that it requires more computational power than other clustering algorithms, like hierarchical clustering, it enables more control directed towards the specific problem. For the particular task of paraphrase clustering, Cordeiro *et al*. (2007a) manually selected a number of true paraphrases and random sentence pairs and empirically established the optimal threshold at 0.2. Since *Sumo-Metric* takes values in the $[0,1]$ interval, this means that in each generated cluster, and for each sentence pair in that cluster, we will have $S(s_i,s_j) > 0.8$.

QT creates the largest possible non-overlapping clusters. It was designed to avoid a sometimes inadequate behavior of other clustering algorithms, K-means for example, that do not take into account whether a unit of a set to be clustered possibly belongs to any cluster, but force each unit into the nearest one. This property of QT coupled with the *Sumo-Metric* avoids grouping together very dissimilar sentences, but it avoids also grouping together sentences that are too similar and leaves out from the cluster the sentences that do not contribute new information. Thus, QT with *Sumo-Metric* naturally deal with possible redundant processing.

Finally, once paraphrase clusters have been extracted, we need to align the sentences in the clusters in order to extract lists of interchangeable words from which to create TOEFL-like test cases in an automatic way. For that purpose, we implement the methodology of Doucet & Ahonen-Myka (2006) who propose to extract Maximal Frequent Sequence (MFS) of a cluster of arbitrary sequences.

5.3 Alignment

In this section, our goal is to align the paraphrases inside each cluster, detecting their common parts in order to identify what differentiates them. Our approach considers sentences as word sequences and therefore reduces the resulting problem to that of multiple sequence alignment (Notredame, 2007). In the field of bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations, for instance. The alignment of sequences is performed to evidence their common and distinctive parts, possibly taking gaps into account.

Similarly, in the field of natural language processing, sequence alignment allows to observe variations in language use, and is particularly useful for similar text fragments, such as paraphrases (Barzilay & Lee, 2003). But while there are several efficient techniques for multiple sequence alignment in the field of bioinformatics, they actually aim at slightly different problems. Indeed, to-be-aligned biosequences are typically few, very long and with limited vocabulary, e.g., there are only 20 amino acids, and only 4 nitrogenous bases present in the nucleic acids DNA and RNA, designated by the

letters A, C, G and T. In comparison, paraphrases are more numerous, shorter, with a larger vocabulary and very few words are repeated within the same sentence. As a consequence, the techniques optimized for biosequences alignment turn out to be inappropriate for paraphrases (Barzilay & Lee, 2003).

In the following sections, we present a 2-phase approach designed to efficiently align a set of paraphrases. In the first phase, we identify long common fragments, that are later used as pivots for the alignment of the paraphrases.

5.3.1 Maximal frequent sequences

Maximal Frequent Sequences were defined by Ahonen-Myka (1999). A frequent sequence is defined as a *non contiguous sequence* of words that must occur in the same order more often than a given sentence-frequency threshold. MFSs are constructed by expanding a frequent sequence to the point where the frequency drops below the threshold. This expansion is done through a greedy algorithm extensively described in Ahonen-Myka (1999). It is worth noting that this technique does not require any preprocessing, i.e., neither stemming nor stop words removal are necessary. In this way, we can assign a set of MFSs to each set of paraphrases. In the following of this section, we formally define the notions of MFS.

Definition 1. *A sequence* $p = a_1 \cdots a_k$ *is a* subsequence *of a sequence* q *if all the items* $a_i, 1 \leq i \leq k$*, occur in* q *and they occur in the same order as in* p*. If* p *is a* subsequence *of* q*, we also say that* p occurs *in* q *and that* q *is a* supersequence *of* p.

For instance, the sequence *Glorious Nation of Kazakhstan* can be found in all of the three sentences in Figure 5.1 and as such is a *subsequence* of each sentence.

Definition 2. *A sequence* p *is* frequent *in a set of fragments* S *if* p *is a* subsequence *of at least* σ *fragments of* S*, where* σ *is a given frequency threshold.*

Definition 3. *A sequence* p *is a* maximal frequent (sub)sequence *in a set of fragments* S *if there does not exist any sequence* p' *in* S *such that* p *is a* subsequence *of* p' *and* p' *is frequent in* S.

As a consequence, in the example presented in Figure 5.1, the sequence *Glorious Nation of Kazakhstan* is not maximal, since it is a subsequence of the frequent sequence *Borat: Cultural learnings of America for Make Benefit Glorious Nation of Kazakhstan*. This latter sequence is maximal. With this simple example, we already get a glimpse of the compact descriptive power of MFSs. We do not restrict ourselves to the extraction of word pairs. Indeed, the 12-word sequence *Borat: Cultural learnings of America for Make Benefit Glorious Nation of Kazakhstan* would need to be replaced by $\binom{12}{2} = 66$ word pairs. With MFSs, no restriction is put on the maximal length of the phrases. Thus, we can obtain a very compact representation of the regularities of texts. So, by extracting the

MFSs of a cluster of paraphrases, we obtain a compact sequential description of the corresponding paraphrases, i.e., a "skeleton" of the cluster that may be used for alignment.

5.3.2 Multiple sequence alignment

Given the corresponding set of MFSs, we can extract the commons and specifics of a set of sentences, very efficiently, in one pass. For instance, let us assume we are to align the 3 paraphrases in Figure 5.1.

This set contains one MFS of frequency 3: *Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan*. Once this MFS has been extracted, we can rely on the sequential property of MFSs to know that, passing in parallel through each of the paraphrases, we are bound to encounter the word *Borat:*, and that any word encountered before is not common to all sentences. Once *Borat:* is encountered, we know that we are bound to encounter the second word of the MFS, *Cultural*. Since the MFS allows gaps, any word encountered between two successive words of the MFS is not common to all sentences. So, the process continues until the last word of the MFS is reached. This way the resulting alignment presented in Figure 5.6 is obtained in only one pass over the word sequences.

> [{1: *Kazakhs are outraged by the wildly anticipated mock documentary feature*} {2: *The news follows controversy surrounding the comedy film*} {3: *Meanwhile Borat is leaping to the big screen in the mockumentary*}] Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan [{2: *which cut so close to the funny bone*}]

Figure 5.6: The alignment corresponding to the sentences of Figure 5.1. Word sequences without brackets are common to both sentences. The sequences in curly brackets are specific to the sentences with the corresponding numbers.

5.4 Forming the Test Cases

The final step is to form TOEFL-like test cases from the aligned segments in the clusters. The notion of test implies one word in a specific position, or target word, for which a match is sought among a list of candidates. In this section, we show how to create tests that consist of words with high probability of being in a paradigmatic relation.

So far, we have clusters of sentences saying nearly the same thing, but slightly differing in expression and with the corresponding parts aligned. We now need to search for candidates among the words which appear out of the MFS, the different parts of the paraphrases, i.e., words in between the brackets in Figure 5.6.

In order to extract lists of interchangeable words, we first lemmatize and assign part-of-speech tags to the aligned paraphrases with MontyLingua Liu (2004). This step is necessary since we are interested in paradigmatic semantic relations and only open class words with the same part of speech are eligible candidates.

Those parts of the paraphrases that lie between two successive parts of a MFS have different orthographic appearance, nevertheless, we assume that they have similar meanings since they are both parts of paraphrase sentences and share left and right MFS contexts. Therefore, here is the place where we search for word substitutions. Precisely, the construction of the candidate tests goes as in Algorithm 1.

Algorithm 1 Construction of the candidate tests.

```
For each aligned sub-segment
    For each open class word
        Create a list of candidates from
        the rest of the segments that share
        left and right MFS contexts.
    End
End
```

For example, from the words from the first aligned paraphrase in Figure 5.6, we extract two test cases for the target words *kazakh* and *feature* as shown in Figure 5.7. Six more test cases would be extracted from this paraphrase cluster for the nouns *news*, *controversy*, *film*, *Borat*, *screen* and *mockumentary*.

1. *kazakh* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*
2. *feature* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*
3. *news* | *kazakhs* | *feature* | *borat* | *screen* | *mockumentary*
4. *controversy* | *kazakhs* | *feature* | *borat* | *screen* | *mockumentary*
5. *film* | *kazakhs* | *feature* | *borat* | *screen* | *mockumentary*
6. *borat* | *kazakhs* | *feature* | *news* | *controversy* | *film*
7. *screen* | *kazakhs* | *feature* | *news* | *controversy* | *film*
8. *mockumentary* | *kazakhs* | *feature* | *news* | *controversy* | *film*

Figure 5.7: TOEFL-like test cases.

5.5 Summary

In this chapter, we introduced our second main contribution. We propose to make use of a method for paraphrase extraction and alignment as an unsupervised manner to learn paradigmatic relations. We described how to create TOEFL-like tests with a short list of candidates that are predominantly in paradigmatic relations with the target word. Eventually, a candidate word can be interchangeable with the target word in context. In particular, this methodology is language independent and completely unsupervised which may allow the study of different languages. In the following chapter, we provide the results of an exhaustive study of *Global*, *Local* and *Product* similarity measures over three flavors of Vector Space Model and two Probabilistic Models to identify the semantic relations between the words inside the TOEFL-like tests, discovered based on paraphrase detection and alignment. Also, we give a thorough error analysis of the various steps of the algorithm.

Chapter 6

Results and Discussion in Open Environment

In this chapter, we combine both processes proposed so far in an attempt at a real world synonymy discovery task. First, we consider the output of the process that creates TOEFL-like test cases. In this, we perform manual annotation of over 1000 tests with a set of close semantic relations, wherever present. This effort provides a number of insights about the various sources of spurious tests and ways of improvement. Consequently, the method of contextual similarity, described in Chapter 3 is applied to the set of annotated tests and an analysis of the quantitative results is reported.

6.1 Creating TOEFL-like Test Cases

In this section, we evaluate the methodology, proposed in the previous chapter, over a set of Web news stories automatically extracted on a daily basis. This environment proves to be very fruitful for paraphrase extraction, since many sentences convey the same message but in a different form.

For this experiment, 3 days of news were collected from the Google News website[1] in November 2006. From these texts, 178 thousand sentences were extracted as paraphrase candidates which formed 27 thousand clusters of sentences. Finally, 183 thousand alignments were produced[2] which, then, yielded a set of 22 thousand TOEFL-like test cases with an average of 4.6 candidates. In order to keep the evaluation manageable, we retained at random 1000 clusters of sentences and from them extracted 1058 noun test cases[3].

6.1.1 Paraphrase extraction

The paraphrase extraction and clustering methodology that we adopted in this work was proposed and formally evaluated in Cordeiro *et al.* (2007a). Here, we examine its properties from the perspective

[1] `http://news.google.com/` [13th July 2011]

[2] `http://www.di.ubi.pt/~jpaulo/` [13th July 2011]

[3] Few clusters yielded more than one test.

of the semantic relations discovery problem.

Bad preprocessing, which means that HTML or XML tags were taken as tokens and incomplete sentences extraction accounted for the most part of wrong paraphrase classification. Indeed, the *SumoMetric* is very "optimistic" with respect to short sequences. For example, the following pair of sentences

1. *He is a superstar Texas senior cornerback Aaron Ross said.*
2. *What he is doing now is being a great leader Texas coach Mack Brown said.*

is classified as a paraphrase based on the pronoun *he*, the verb *is*, the toponym *Texas* and a very common citation frame in the news writing style, i.e., the verb *said*.[1]

Few wrong paraphrases could have been avoided with the assistance of named entity recognition or multiword unit extraction that would give a single count to a unique reference or concept. For example, in the following cluster the 12-words movie name and the collocations *kazakh authorities* and *legal action* should be treated as atomic concepts, as evidenced in Grigonytė *et al*. (2010).

1. *Many months later the funny bruised fruits of his labor Borat: cultural learnings of America for make benefit glorious nation of Kazakhstan are poised to hit the collective American conscience with a juicy splat.*
2. *Borat: cultural learnings of America for make benefit glorious nation of Kazakhstan opens in the United States on Friday but the run in with kazakh authorities who even threatened legal action generated huge pre release publicity.*

This results in the following spurious test: *splat* | *action* | *authority* | *kazakh* | *publicity* | *release*.

Even a relatively large number of overlaps do not always guarantee that the pair of sentences have the same communication intent as shown below.

1. *Luke broke onto the screen under Washington's direction in Antwone Fisher, then went on to Friday Night Lights and Glory Road.*
2. *Luke, best known for his work in Antwone Fisher and Friday Night Lights, is the versatile and commanding young leading man Hollywood needs.*

Although this mode of paraphrasing might seem very productive, a much more common source of

[1] Other examples are *confirmed* or *said in a statement*.

cases without any perceivable semantic relation are perfectly aligned paraphrases which make accent on different details such as in the following paraphrase.

1. *Federline released his debut CD on October 31.*
2. *Federline released his debut CD in which he raps about his rise from obscurity.*

Although the essential information is the same, deeper interpretation would be necessary in order to make clear that the adverbial and the subordinate clauses are not subject to semantic alignment. This kind of paraphrasing was the major source of tests void of semantic relations. Although the average sentence length in the news corpus is 24 words and most of them are correctly classified, the paraphrase discovery step alone is responsible for about 35% of the wrong test cases which represents 23% of all the extracted tests.

6.1.2 Paraphrase alignment

The alignment phase is based on finding an as long as possible sequence that is common for both sentences. However, the alignment failed in a number of cases when the paraphrasing effect is achieved through word order change. For example, although the following aligned sentences

[{1: *The median price of an existing single family home dropped* 2.5 *percent from September 2005, the biggest year on year drop since record keeping began in 1969*}] the national association of realtors said [{1: *in Washington*} {2: *existing home sales declined for the sixth consecutive month in September while the median price fell 2.5% year over year, the biggest decline on record*}]

are perfect paraphrases, the only possible alignment results in the test *washington* | *home* | *sale* | *month* | *september* | *price* | *year* | *decline* | *record* while from these sentences one could infer similarity between the nouns *drop* and *decline*. This is a common case when two long sentences are aligned around a single sequence that refers to the common agent. Even when the alignment is anchored in many points still there are possible conjunction rearrangements or even syntactic structure alterations such as in the following alignment.

[{1: *He found*} {2: *This revealed*}] that [{2: *their*}] sperm [{1: *count, viability, motility*} {2: *declined steadily in number, quality*}] and [{1: *shape declined*} {2: *ability to swim*}] as mobile phone usage increased.

Paraphrase classification and alignment can occur based on secondary details as well. For example, the following aligned sentences

[{1: *The gloomy prediction follows*} {2: *Marine species are disappearing at an accelerating rate posing a serious threat to human health and wellbeing*}] a four year [{1: *multinational*}] study of the state of the world's [{1: *seas and*}] oceans [{2: *has concluded*}].

are paraphrases and correctly aligned. However, the essential information is not explicitly present in the first one. In order to avoid this kind of alignment and the consequent bad test cases, discourse analysis is necessary as well as reconstruction of the intended message by means of anaphora resolution (Mitkov *et al.*, 2007) and noun-ellipsis resolution (Branco & Costa, 2006). This will be left for future work.

6.1.3 TOEFL-like tests

In order to keep the evaluation manageable, we retained at random 1000 clusters of sentences and from them extracted 1058 noun test cases. Then, we manually classified them. The reason to undertake the manual annotation instead of taking the semantic relation from WordNet or another resource is that news exhibit very creative use of language and rare synonymy relations like *leader | godfather*, that are not present in WordNet, foreign words, such as *madrassa*, when narrating about a religious school in Pakistan, and named entities from a wide variety of categories appear regularly in texts. During the process of classification, we found that four classes suffice to cover all the cases. The four classes are *Synonymy*, *Co-hyponymy*, *Is a* and *Instance of*. A test receives a label with respect to whether it contains a pair of words in one of the given relations. If there exist more than one semantic relation, i.e., the test belongs to more than one class, different test cases are created and labeled. All the cases with no perceivable paradigmatic semantic relation receive a class *Other*.

It is important to note that in order to classify a test, we first disambiguated all the words in the contexts of the source paraphrase cluster and the original news story, when necessary. If there was a candidate that referred to the same concept as the target, we subsequently classified the test with respect to the perceived relation. For example, the *Instance of* relation between *Aisawa | legislator* was extracted from the following paraphrase.

1. *Aisawa declined to elaborate.*
2. *The Japanese legislator declined to elaborate.*

In Table 6.1, we present the distribution of the tests per category. It is interesting to observe, that the *Synonymy* together with *Co-hyponymy* are more populous than the other two categories. It is no surprise, though, that words of the same level of generality are preferred substitutes for the sake of paraphrasing.

Table 6.1: Classification of the Test Cases.

Synonyms	Co-hyponyms	Is a	Instance of	Other	Overall
117	108	61	86	686	1058

One interesting finding is the fact that out of the 117 pairs that we found to be in synonymy role, 22 were not present in WordNet as such. An excerpt of the annotated tests is given in Table 6.2 and Table 6.3. They all contain a pair of words that could be regarded in a given semantic relation in the context they were observed.

Table 6.2: Manually annotated tests. The respective relations hold between the first and the second words of each test.

Synonyms:	*body* | *panel*
	michael | *mike*
	child | *infant*
	administration | *government*
	collaboration | *cooperation*
	article | *study* | *vivo*
	condition | *disease* | *treatment*
	veteran | *vet* | *congress* | *mark* | *war*
	seat | *place* | *american* | *congress* | *election*
Co-hyponyms:	*idea* | *plan*
	amazon | *ebay*
	black | *hispanic*
	department | *government*
	journalist | *videographer*
	culture | *habit* | *life*
	blaze | *wildfire* | *santa*
	reality | *point* | *campaign*
	draft | *resolution* | *director* | *sabliere*

The manual annotation and disambiguation process was instructive as for the strength of the semantic relations. Although the *Co-hyponyms* in the *Is a* hierarchy of WordNet are connected by longer paths, they tend to be perceived as more similar to each other than are the words in the *Is a* relation. This subjective judgement seems to be confirmed by the persistently higher contextual similarity between the former compared to the latter category (see Section 6.2).

We also tried to understand how much noise was introduced by our methodology to create TOEFL-

Table 6.3: Manually annotated tests. The respective relations hold between the first and the second words of each test.

Is a:	*conspiracy* | *obstruction*
	capability | *repair*
	sheik | *cleric* | *australia*
	status | *fame* | *fortune*
	baseball | *game* | *ball* | *innings*
	game | *play* | *room* | *sideline*
	allegation | *statement* | *admission* | *family* | *friday*
	deal | *agreement* | *afternoon* | *side* | *sunday*
	investigator | *agent* | *unit* | *revenue* | *service*
Instance of:	*july* | *month*
	graham | *coach*
	community | *un*
	aisawa | *legislator* | *japanese*
	patriot | *team* | *right*
	fedex | *company* | *order*
	george | *bush* | *plan* | *war*
	schwarzenegger | *star* | *film* | *terminator*
	nirvana | *group* | *catalogue* | *company*

like test cases. For that purpose, we show the distribution of the test cases over 9 categories with respect to the number of the candidate words in Table 6.4. Thus, the first column represents those cases, which come from paraphrase pairs in which only one noun is substituted by another one, the second column represents the cases where one noun is supposedly substituted by two nouns and so on. A substantial part of the tests have 4, 5 or 6 candidates. However, this does not indicate the most common mode of paraphrasing because these same sets contain the lowest ratio of paradigmatic semantic relations (the first line of Table 6.4). It is natural to expect, that the more candidates a given test has, the higher the probability that any of them will be in any of the specified semantic relations with the target word. The tests with more candidates come from paraphrases with greater absolute number of differences. For such a pair to be taken as a paraphrase, it also needs a greater number of common subsequences, thus implying that more information is shared between the sentences. This is why, after the level of the greatest linguistic variety in the middle of the specter, test extraction improves. In particular, we were able to only extract 10 and 6 tests with 10 and 11 candidates respectively.

Although many instances of semantic relations can be encountered, 65% of the tests do not belong to

Table 6.4: Proportion of good tests by test size.

	1	2	3	4	5	6	7	8	9
Good	41%	37%	36%	30%	31%	34%	41%	41%	35%
All	120	87	119	204	155	159	100	74	40

any of these semantic categories, as shown in Table 6.1. Some of them are due to wrong alignments as in *understanding* | *Lipunga* | *onion* | *tomato* | *village*. Another 25% of the tests are void of any perceivable semantic relation. Further, bad POS tagging caused a set of good candidates to be lost and replaced by words that were actually used in another POS role. For example, from the following aligned paraphrase that reports on an infant disease

[{1: *right now the*}{2: *currently the brain*}] defects can not be detected until after death

the test *right* | *brain* was extracted while *right* is indeed an adverb.

Finally, the rest of the tests in this group could be classified in some more loose semantic category such as in the following case: *study* | *caution* | *Washington* | *finding*. Tables A.1, A.2, A.3 and A.4 in Appendix A.3 show sets of extracted pairs with their corresponding categories, which should now be discovered by the resolution of the test cases with the similarity measures proposed in Chapter 3.

6.2 Solving TOEFL-like Test Cases

Now that we have the list of test cases, as shown in Figure 5.7, we need a method to select the best candidates. For this purpose, we employ the contextual similarity measures introduced and studied in Chapter 3 and Chapter 4. The contextual statistics required for this comparison were collected from a new Web corpus gathered for these specific experiments[1], following the methodology introduced in Section 4.2.

In order to quantify the feasibility of the methodology, we retained only the 372 test cases labeled with a specific semantic relation and performed a comparative study. For all the similarity measures and the respective weighting schemes, i.e., (1) the Cosine similarity measure associated to the TfIdf, the Pointwise Mutual Information and the Conditional probability, and (2) the Ehlert and Lin models, we solved each test using the *Global*, *Local* and *Product* similarities. The results are summarized in Table 6.5, Table 6.6 and Table 6.7.

The first observation that we can make from Table 6.5 is that the combination of Cosine with PMI

[1]This corpus and the test set are available at `http://hultig.di.ubi.pt/~rumen`

Table 6.5: Accuracy of *Global* on 372 tests.

	Vector Space Model			Probabilistic	
	Cos TfIdf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Synonymy	50%	75%	42%	58%	42%
Co-hyponymy	47%	65%	53%	29%	65%
Is a	54%	58%	46%	29%	42%
Instance of	26%	30%	23%	26%	14%
Overall	44%	59%	41%	37%	42%

Table 6.6: Accuracy of *Local* on 372 tests.

	Vector Space Model			Probabilistic	
	Cos TfIdf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Synonymy	58%	50%	71%	58%	54%
Co-hyponymy	41%	53%	47%	47%	59%
Is a	46%	42%	42%	42%	38%
Instance of	49%	40%	35%	42%	40%
Overall	49%	47%	51%	49%	49%

in *Global* mode is nearly sufficient to extract the closest semantic relations. However, none of the *Global* measures achieves results statistically different from random guessing for the category *Instance of*. This is no surprise since, in order to be solved, most of the cases in this category reduce to a problem of finding the most salient property associated to a proper name, which in most cases results in a highly polysemous test (see Table A.4). For example, the pair *president* | *Luiz* refers to *Luiz Inácio Lula da Silva*. However, *Luiz* is a common name and as such, used to refer to many different agents in a number of human activity areas. Here is where the *Local* similarity comes to play. Since it always compares monosemous representations, it is bound to associate *president* with *Luiz* in those documents where the president of Brazil is the subject. As a result, the performance of the *Local* similarities shows statistically significant, at 95% confidence level, improvement over the *Global* similarity measures for the *Instance of* test cases. In Section 4.3.6, we saw that *Local* similarity is not always that successful with highly polysemous words. The new data, however, does not contradict that observation, since the settings of our synonymy discovery experiment imply a relatively restricted domain of news stories, which results in reduced overall polysemy, especially for names of significant events, political and sports figures, celebrities and locations. Also, the association between a named entity and the more general concepts used to refer to it, conveyed by the discourse boundaries, further reduces polysemy down to levels where a contextual similarity measure is capable to reasonably highlight important relations.

Table 6.7: Accuracy of *Product* on 372 tests.

	Vector Space Model			Probabilistic	
	Cos TfIdf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Synonymy	63%	58%	46%	58%	54%
Co-hyponymy	65%	65%	59%	47%	71%
Is a	46%	58%	46%	46%	46%
Instance of	42%	42%	33%	42%	37%
Overall	56%	56%	47%	49%	54%

In summary, *Local* similarities perform better in comparison to *Global* similarities for those pairs that would be connected by the shortest paths in an *Is a* hierarchy, such as WordNet. Namely, *Local* Cos TfIdf, Cos Prob and Lin model improve over *Global* for the synonymous tests and all *Local* measures perform better than *Global* measures for the *Instance of* category. Accordingly, overall *Local* improves over *Global*, at 95% confidence level, for all measures except for PMI.

In order to optimize the *Product* measures performance we defined the parameter γ of Equation 3.11, p. 46, to be a linear function, as explained in Section 4.4. Here, we resort to the number of contexts as a substitute for polysemy, as both quantities are strongly correlated and since not all of the words in our test set are presented in WordNet. We optimized the values of the parameters a and k using the 500 tests described in Section 4.1 as a training set. The results of this process are given in Table 6.7. In summary, *Product* similarity measures improve, at 95% confidence level, on *Global* for all measures except for PMI. Further, *Product* similarity measures improve on *Local* similarity for Cos TfIdf, Cos PMI and Lin model.

The results evidence that a single measure can not solve the entire problem. The *Synonym* relation is best treated by *Global* values, the *Instance of* relation is best treated by *Local* values, while the Lin model is the one that deals best with the *Siblings* for the *Product* values. We summarize these results by kind of semantic relation and measure in Table 6.8.

6.3 Summary

The method proposed in Chapter 5 and evaluated here gleans the thesaurus entry candidates from their use in language. Since the purpose of paraphrasing is to change the form while preserving the information, a corpus of paraphrases, aligned at word level, provides information about possible sentence modifications and word substitutions that have occurred during the paraphrasing process.

In the current chapter, we saw that the proposed method is capable to spot word substitutions about

Table 6.8: Best methodology by category.

	Vector Space Model			Probabilistic	
	Cos Tfldf	**Cos PMI**	**Cos Prob**	**Ehlert**	**Lin**
Synonymy	-	Global	-	-	-
Co-hyponymy	-	-	-	-	Product
Is a	-	Global or Product	-	-	-
Instance of	Local	-	-	-	-
Overall	-	Global	-	-	-

35% of the time. This amounts to 372 out of 1058 test cases that manifest a tight semantic relation. Further, we identified a number of possible directions to improve the process, among which are more robust POS tagging, integration of syntactic and multiword unit information in the processes of paraphrase detection and sentence alignment.

Applying measures of contextual similarity succeeded in identifying 75% of the synonyms and 59% of all 372 tightly related word pairs. The data highlights the significant advantage of the *Local* mode of similarity measures in the case of *Instance of* relation compared to *Global* similarities. This result is important since *Instance of* relation emerges between a named entity and a more general concept, where the named entities are the most dynamic part of the current vocabulary and demand constant effort in maintaining up to date inventory.

Previously, we demonstrated that exhaustive search is counterproductive both from computational and performance perspectives. Here, we proposed a solution to the exhaustive search showing that it can be avoided and the results of the application of contextual similarity measures on the set of 372 tests show that automatic discovery of close semantic relations is feasible. Our method limits the search space such that the studied contextual similarity measures are capable to discover infrequently manifesting relations. The proposed method discovers *Synonymy* relations at a rate of 8%, *Co-hyponymy* relation at a rate of 7%, *Is a* relation at a rate of 3%, *Instance of* relations at a rate of 4% of the entire set of 1058 test cases and 15% overall on tight semantic relations.

While the definition of synonymy of Leibniz, that states

> Two words are synonymous if they are interchangeable in statements without a change in the truth value of the statements where the substitution occurs.

is mostly correct, we saw that in a third of the occasions the paraphrasing effect is achieved by substitution of a word for a word with a lower level of generality. From this follows that synonymy is a loose notion of words that refer to the same thing in the context of a given discourse.

Chapter 7

Conclusions and Future Works

7.1 The Objectives

In the introduction, we briefly discussed the concept of synonymy. In particular, we adopted a definition that roots in the actual view about word meaning and word meaning similarity and states that synonymy is a continuous scale of word meaning similarity and the judged sense similarity is proportional to the fraction of the occasions when both words can be used in order to express the same idea. This definition served as a departure point for the proposals made in Chapter 3 and Chapter 5 as well as for the collection and the analysis of empirical data.

In this dissertation, we sought to answer why contextual similarity fails to solve the problem of semantic relations discovery. It is accepted that the main difficulty faced by lexical analysis is the capacity of the words to signify different, frequently disparate concepts. Many studies are devoted to the subject in order to explain polysemy and still in search of adequate approaches to deal with its effects. Here, we showed how exactly it does influence the performance of the contextual similarity measures as indicators of similarity of meaning.

A previous study (Weeds *et al.*, 2004) shows that in general the similarity measures are highly dependant on the frequencies of the compared words and tend to prefer more frequent words over the less frequent ones. This is again indirectly confirmed by the observation that only the most frequent word senses might be successfully characterized.

The correlation between word frequency and word polysemy is indisputably established and in the light of these results and the empirical data from Chapter 4, we can affirm that when *Global* similarity is used to compare contextual representations of words, strong preference is given to the candidates of one end of polysemy specter irrespectively to whether they do really relate to the same concept. In fact, the contextual similarity estimates produced by *Global* are up to a great extent due to a net effect of spurious overlaps of contexts pertaining to disconnected meanings. Thus, the first objective we set in this work was to find a formal method to produce word similarity estimates that are as little influenced by polysemy as possible.

Our second objective was to avoid the exhaustive search for the specific task of noun semantic relations discovery. Indeed, the inefficiency of the exhaustive search strategy is twofold. Algorithmic complexity in order of $\mathrm{O}\left(n^3\right)$ makes it difficult to scale to the entire vocabulary. Further, according to WordNet a word has on average 1.3 near synonyms and this single synonym must be singled out of the entire vocabulary. This is a nearly impossible task for the exhaustive approach. Rather, the best n candidates are usually kept. In order to classify the relations or to filter out unwanted candidates, manually selected patterns are applied at cost of recall. These concerns motivated the new methodology proposed in Chapter 5. To the best of our knowledge this question has never been addressed so far, neither considered as a problematic area.

7.2 Contributions

In Chapter 3, relying on the results of previous works, we motivated a similarity measure that performs comparison of monosemous word usage profiles. In contrast to the common vector space model, which creates single representations for each word and mingles in contexts belonging to all possible senses of the given word, here, we proposed to make use of corpus structure and discourse boundaries to build per-meaning statistical descriptions. *One sense per discourse* hypothesis (Gale *et al.*, 1992) suggests that building statistical representation out of a single document significantly increases the probability that only concrete word meaning is characterized. On the other side, according to Turney (2001) and Mohammad & Hirst (2010) synonyms co-occur in texts more often than expected by chance and tend to refer to the same concept even in case of polysemous words. These properties of language make possible to calculate *Local* as an average of multiple similarities of pairs of words across a set of documents where both words occur. This new similarity is our first contribution and aims to reduce the influence of polysemy on word similarity estimates.

In Chapter 4, applying the *Local* measures on a set of TOEFL-like test cases, we showed their capacity to find out the correct synonym, at least as reliably as does the *Global* strategy. The advantage of *Local* over *Global*, which justifies its further study is demonstrated in Section 4.3.4. The fact that *Local* compares monosemous representations in most cases allows it to reliably tell synonyms from non synonyms for polysemy levels of up to 13 senses. Although this is only about a third of the polysemy specter, it covers about two thirds of the examined cases.

The results presented in Chapter 4 clearly demonstrate why the *Global* similarity strategy is hindered by polysemy. The results make it clear that the estimates produced by *Global* depend to greater extent on the level of polysemy compared to the level of semantic similarity between words. On the contrary, the *Local* similarities are nearly immune to variations of polysemy.

In Section 4.3.2, we demonstrated the significant difference between different measures of contextual similarity. The complementing qualities of *Global* and *Local* similarities were highlighted

in various occasions throughout the work, the most salient example being the consistently better performance of *Local* compared to *Global* similarities on the set of *Instance of* relations, reported in Section 6.2. These observations prompted us to study a multiplicative combination of *Global* and *Local* similarity, hence we call it *Product* similarity. We optimized its performance against various properties of words that could be learned from WordNet or from corpora, polysemy and number of word contexts among others. Performance improvement was possible when the behavior of the base measures differed most with respect to polysemy. *Product* measure was also successful to improve on both *Global* and *Local* measures even when we trained the parameters against word context counts on the 500 test set described in Section 4.1 and evaluated on the fairly different data set of 372 tests automatically extracted from a news corpus. However, polysemy and word contexts count do not correlate in any way with word meaning similarity, thus, a more meaningful predictor of similarity is desirable.

In Chapter 5, we explored the possibility to learn semantic information from paraphrases. Indeed, paraphrases are rich sources of information for any language acquisition task. Paraphrases that convey variable levels of details were used in automatic summarization and sentence compression tools (Cordeiro *et al.*, 2009). In contrast, we are interested in those cases where the entire information is present but is expressed in different forms. This is where one can learn different manners to state one and the same idea, that is, synonymy at sentence and word level. Having these small chunks of information, it is not needed anymore to perform comparisons of every possible pair of words in the vocabulary in order to find synonyms. What needs to be done is only to confirm, by means of lexical analysis, the correct correspondence within a list of on average 4.6 candidates. In fact, this is a hybrid methodology between pattern recognition and lexical analysis technics for automatic language acquisition. The benefits of this pathway to synonymy is to avoid the expensive exhaustive search, but even more important is the ability to single out the unique synonym in a list of thousands of possible candidates. The fact that most of the improbable candidates are discarded on an early stage brings to light rarely manifested relations, that are otherwise obscured by the most frequent ones or even by spurious candidates.

It was repeatedly noted that it is extremely difficult to characterize infrequent word senses as their specific semantic relations and usage patterns are obscured by more frequent meanings. The method proposed in Chapter 5 deals with this problem by reducing the set of possible candidates to a very short list, usually including alternatives that are in some syntagmatic relation with the target. This virtually eliminates the concurrency of more common relations and the correct candidate stands out.

Furthermore, the method seems to discover a small set or relations that are found to be organizing the noun part of the vocabulary, namely *Synonymy*, *Is a*, *Co-hyponymy* and *Instance of*. Pairs of words between which these relations hold, might under certain conditions, be used as synonyms. This is in contrast with the methods that perform exhaustive search as they frequently find relations that are

difficult to classify (Heylen *et al.*, 2008).

In particular, we avoid manual work to find constructions that indicate specific semantic relations by exploiting the fact that occurrences of different words in the same sentence frame probably have the same communication intent. Viewed from this perspective, every sentence may be used to learn from when it is paired with a paraphrase in contrast to the limited coverage of a set of predefined constructions.

The evaluation of the automatically created test cases provided new important evidences of the ability of *Local* similarity to deal with polysemous words. Indeed, the results of the *Global* strategy on the *Instance of* test cases is not statistically different from random guessing, while the *Local* one scored 49% of the time. Some categories are too dynamic to be included in printed dictionary, e.g., proper names are rarely included in common language dictionary. However, names of geographical categories and historical personalities are good candidates. Probably more named entity categories will be included in dictionaries with the advance of methods capable to reliably detect and classify such information.

The system was able to find 280 word pairs that are already included in WordNet, further it proposed 22 pairs in *Synonymy* relation, 27 pairs in *Co-hyponymy* relations, 9 pairs in *Is a* relation, 34 pairs in *Instance of* relation and 41 words, most of which proper names, not included in WordNet.

This work is also in the line of a number of endeavors that exploit Web as a corpus. Unlike the systems that rely only on search engine results to access aggregated statistics, we implemented an ad-hoc algorithm to gather textual corpora by crawling the Web. Then, we extracted detailed syntactic information from this collection. Web as a corpus is a viable and probably an inevitable alternative to off-line corpora. The technical difficulties to collect and process Web textual data are paid out by the advantages of a representative and up-to-date sample of the current language.

7.3 Future Works

In Section 3.4.1, we noted that the assumption of the vector space model that every two basis vectors are orthogonal is inadequate for word similarity modeling. The Cosine between every two basis vectors is always 0 by construction, although they may correspond to otherwise similar in meaning or even synonymous contexts. A more appropriate treatment of such contexts presumes that their similarity estimate reflects their perceived similarity. The InfoSimba measure (Dias & Alves, 2005) provides an alternative approach as it takes under account the association between distinct features by augmenting the Cosine formula with an Equivalence Index (EI).

This measure alone provides various directions for further explorations. For example, it could be

integrated in an iterative process, where at the odd numbered steps verbs are contexts used to characterize nouns and at even steps nouns become contexts characterizing verbs. At each consecutive step the values of EI are taken to be the similarities calculated at the previous iteration. This variation is called Recursive InfoSimba in Dias (2010)

$$InfoSimba(n_1, n_2) = \frac{\sum_{j=1}^{p}\sum_{k=1}^{p} c_{1j} \times c_{2k} \times EI(c_{1j}, c_{2k})}{\left(\begin{array}{l}\sum_{j=1}^{p}\sum_{k=1}^{p} c_{1j} \times c_{1k} \times EI(c_{1j}, c_{1k}) + \\ \sum_{j=1}^{p}\sum_{k=1}^{p} c_{2j} \times c_{2k} \times EI(c_{2j}, c_{2k}) - \\ \sum_{j=1}^{p}\sum_{k=1}^{p} c_{1j} \times c_{2k} \times EI(c_{1j}, c_{2k})\end{array}\right)} \tag{7.1}$$

Another interesting direction is the possibility to make use of general-to-specific ranking as produced in Dias *et al*. (2008a) for EI values. It is supposed that when the predominant level of generality of the contexts of a given word is known, we will be able to correctly predict the general-to-specific ranking of the specified word. Thus, apart of being able to construct horizontal sets of similar words by the conventional similarity, we will be able to construct the vertical structure of an *Is a* hierarchy.

In the same thread of thoughts, Caraballo & Charniak (1999) observed that adjectives are good indicator of whether two nouns belong to the same subtree of general-to-specific hierarchy of nouns. This finding is in accord with a note in Miller (1990) reflecting on the structure of a dictionary, which gives as a definition of a word a superordinate concept together with distinguishing features. The adjectival contextual information might be combined with noun coordination and conjunctive constructions data in order to organize nouns in a hierarchy. Another possibility is suggested by Dias *et al*. (2008b) who proposed a knowledge-poor technique to rank set of nouns with respect to generality level. Such a method might be applied to the pairs returned by the paraphrase alignment method examined in Chapter 5 and Chapter 6. Generality level information combined with contextual similarity through a multiple-view clustering algorithm, e.g., Cleuziou *et al*. (2009) might allow automatic word pairs classification in one of the categories *Synonymy*, *Co-hyponymy*, *Is a*, *Instance of* when they are close in respect with contextual overlap or generality level or both. If both directions show significant dissimilarity then no semantic relation is present.

The method proposed in Chapter 3 requires statistical data of pairs of words that appear within the same text. Under this condition, we are interested in their common contexts. This data could be conveniently stored in a graph structure and traversed in order to calculate similarity. A number of compact representations and efficient traversal algorithms are proposed in the area of computer graphics, e.g. (Silva & Gomes, 2003), that could allow for improved computational efficiency.

In Chapter 5, we proposed a method that makes use of aggregations of news stories. The news is naturally aligned by the time of the described events. When two texts are lexically similar, as well,

the clustering of the texts becomes very robust, due to the high probability that they concern the same subject. Even though, in Chapter 6, we saw that only about 35% of the paraphrases provide useful information. Thus, this dependency turns out to be a clear drawback since not every possible domain affords this convenient source of paraphrasing data. As a consequence, more reliable algorithms for paraphrase extraction are required. One necessary direction for improvement is to perform multiword unit discovery following Dias *et al*. (1999) and Dias (2003) prior to paraphrase extraction. This preprocessing will help to avoid paraphrase detection and alignment based on long collocations, that are more properly treated as atomic units.

We feel especially curious to investigate the possibility to integrate syntactic and dependency information in the sentence alignment phase. In Section 6.1.1, we noted that a number of wrong tests could be avoided if the algorithm was aware of the syntactic dependencies of the various parts of the sentence, the concrete example being alignment of an adverbial and a subordinate clause. In general, it seems that syntactic information can be helpful at any stage of the algorithm.

Common alignment techniques do not deal with sentence reordering which is one of the important causes of wrong alignments. As a future work, we aim to test a new alignment technique proposed by Cordeiro *et al*. (2007c) who use a combination of local and global biology-based alignment algorithms which deal with sentence reordering in an elegant way.

Although the performed calculations and the results provide enough evidence to prove the viability of the proposed model, they are far from comprehensive coverage of the language. Nouns provided fruitful ground for evaluation. However verbs, adjectives and adverbs partake in synonymy relations too. Verbs seem to be another world. It is known that verbs exhibit higher level of polysemy than nouns. Thus, the application of *Local* similarities to verbs is not only curiosity but an important test for the new measure. However, Levin (1993) proposed a radically different characterization of verbs, based mainly on syntactic features. Whether lexical analysis is really applicable to verbs is an open question.

References

Agirre, E. & de Lacalle, O.L. (2003). Clustering WordNet word senses. In *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003.*, 11–18.

Agirre, E., Alfonseca, E. & Lacalle, O.L.D. (2004). Approximating hierarchy-based similarity for WordNet nominal synsets using topic signatures. In *Proceedings of the Second Global WordNet Conference*, 15–22.

Ahlswede, T. & Evens, M. (1988). Parsing vs. text processing in the analysis of dictionary definitions. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, 217–224, Association for Computational Linguistics, Morristown, NJ, USA.

Ahonen-Myka, H. (1999). Finding all frequent maximal sequences in text. In *Proceedings of ICML-99 Workshop on Machine Learning in Text Data Analysis*, 11–17.

Baroni, M. & Bisi, S. (2004). Using cooccurrence statistics and the Web to discover synonyms in a technical language. In *Proceedings of LREC 2004*, 1725–1728.

Barzilay, R. & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 16–23, Association for Computational Linguistics, Morristown, NJ, USA.

Becker, R.A., Chambers, J.M. & Wilks, A.R., eds. (1988). *The new S language*. Wadsworth & Brooks/Cole, Monterey, CA, USA.

Berland, M. & Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of ACL*, 57–64, Association for Computational Linguistics, Morristown, NJ, USA.

Bollegala, D., Matsuo, Y. & Ishizuka, M. (2007). Measuring semantic similarity between words using Web search engines. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, 757–766, Association for Computing Machinery, New York, NY, USA.

Bordag, S. (2003). Sentence co-occurrences as small-world graphs: A solution to automatic lexical disambiguation. In *CICLing'03: Proceedings of the 4th international conference on Computational linguistics and intelligent text processing*, 329–332, Springer-Verlag, Berlin, Heidelberg.

Branco, A. & Costa, F. (2006). Noun ellipsis without empty categories. In *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, 81–101, Stanford, CSLI Publications.

Brent, M.R. (1991). Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 209–214, Association for Computational Linguistics, Morristown, NJ, USA.

Brown, P.F., Pietra, V.J.D., deSouza, P.V., Lai, J.C. & Mercer, R.L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**, 467–479.

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.

Caraballo, S.A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 120–126, Association for Computational Linguistics, Morristown, NJ, USA.

Caraballo, S.A. & Charniak, E. (1999). Determining the specificity of nouns from text. In *Proceedings SIGDAT-99*, 63–70.

Cederberg, S. & Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 111–118, Association for Computational Linguistics.

Chambers, J.M. & Hastie, T.J. (1992). *Statistical Models in S*. Wadsworth & Brooks/Cole, Monterey, CA, USA.

Charles, W.G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, **21**, 505–524.

Church, K., Gale, W., Hanks, P. & Hindle, D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115–164, Erlbaum.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**, 22–29.

Cilibrasi, R.L. & Vitanyi, P.M.B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, **19**, 370–383.

Cleuziou, G., Exbrayat, M., Martin, L. & Sublemontier, J.H. (2009). CoFKM: A centralized method for multiple-view clustering. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, 752–757, IEEE Computer Society, Washington, DC, USA.

Cordeiro, J., Dias, G. & Brazdil, P. (2007a). Learning paraphrases from WNS corpora. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, 193–198, AAAI Press.

Cordeiro, J., Dias, G. & Brazdil, P. (2007b). New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software*, **2**, 12–23.

Cordeiro, J., Dias, G. & Cleuziou, G. (2007c). Biology based alignments of paraphrases for sentence compression. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing (ACL-PASCAL / ACL2007)*, 177–184, Association for Computational Linguistics.

Cordeiro, J., Dias, G. & Brazdil, P. (2009). Unsupervised induction of sentence compression rules. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, UCNLG+Sum '09, 15–22, Association for Computational Linguistics, Stroudsburg, PA, USA.

Curran, J.R. & Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, 59–66.

Dagan, I. & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, **20**, 563–596.

Daille, B. (1995). Combined approach for terminology extraction: lexical statistics and linguistic filtering. Tech. Rep. 5, UCREL, Lancaster University.

de Saussure, F. (1959). *Course in General Linguistics*. Philosophical Library.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. & Harshman, R.A. (1990). Indexing by latent semantic analysis. *JASIS*, **41**, 391–407.

Dias, G. (2003). Multiword unit hybrid extraction. In *Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003)*, 41–49, Association for Computational Linguistics, Stroudsburg, PA, USA.

Dias, G. (2010). Information digestion. HDR Thesis, University of Oleans (France). 10 December. `http://www.di.ubi.pt/~ddg/publications/Thesis-HDR.pdf`.

Dias, G. & Alves, E. (2005). Discovering topic boundaries for text summarization based on word co-occurrence. In *International Conference On Recent Advances in Natural Language Processing, RANLP 2005*, 187–191.

Dias, G. & Moraliyski, R. (2009). Relieving polysemy problem for synonymy detection. In *Progress in Artificial Intelligence, 14th Portuguese Conference in Artificial Intelligence, EPIA 2009*, 610–621, Aveiro, Portugal.

Dias, G., Guilloré, S. & Lopes, J.G.P. (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of 6th Annual Conference on Natural Language Processing*, 333–339.

Dias, G., Mukelov, R. & Cleuziou, G. (2008a). Unsupervised graph-based discovery of general-specific noun relationships from Web corpora frequency counts. In *12th International Conference on Natural Language Learning (CoNLL 2008)*, Manchester, UK.

Dias, G., Mukelov, R. & Cleuziou, G. (2008b). Unsupervised learning of general-specific noun relations from the Web. In *21th International FLAIRS Conference (FLAIRS'2008)*, 147–152.

Dias, G., Moraliyski, R., Cordeiro, J., Doucet, A. & Ahonen-Myka, H. (2010). Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Journal of Natural Language Engineering, Special Issue on Distributional Lexical Semantics*, **16**, 439–467.

Dolan, B., Quirk, C. & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING '04: Proceedings of 20th International Conference on Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, USA.

Doucet, A. & Ahonen-Myka, H. (2006). Probability and expected document frequency of discontinued word sequences, an efficient method for their exact computation. *Traitement Automatique des Langues (TAL)*, **46**, 13–37.

Dunning, T.E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**, 61–74.

Dyvik, H. (2004). Translations as semantic mirrors: from parallel corpus to WordNet. *Language and Computers*, **49**, 311–326(16).

Ehlert, B. (2003). *Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics*. Master's thesis, University of California, San Diego.

Erk, K. & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 897–906, Association for Computational Linguistics, Morristown, NJ, USA.

Esuli, A. & Sebastiani, F. (2007). Pageranking WordNet synsets: An application to opinion mining. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, 424–431.

Fellbaum, C. (1990). English verbs as a semantic net. *International Journal of Lexicography*, **3**, 278–301.

Fellbaum, C., ed. (1998). *WordNet: an electronic lexical database*. MIT Press.

Firth, J.R. (1957). *A synopsis of linguistic theory 1930-1955*, 1–32. Philogical Society, Oxford.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. In *Biometrika*, vol. 10, 507–521.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R. & Wang, Z. (2005). New experiments in distributional representations of synonymy. In *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, 25–32, Association for Computational Linguistics, Morristown, NJ, USA.

Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1983). Statistical semantics: analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, **62**, 1753–1806.

Gale, W., Church, K.W. & Yarowsky, D. (1992). One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, 233–237, Morristown, NJ, USA.

Giovannetti, E., Marchi, S. & Montemagni, S. (2008). Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text. In *Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives*, Rome, Italy.

Goldstone, R.L., Feng, Y. & Rogosky, B.J. (2005). *Connecting concepts to the world and each other*, 292–314. Cambridge University Press, Cambridge.

Graff, D. (1995). North american news text corpus. Linguistic Data Consortium.

Grefenstette, G. (1993). Automatic thesaurus generation from raw text using knowledge-poor techniques. In *Making Sense of Words. Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and text Research*.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Grefenstette, G. (1996). *Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches*, 205–216. MIT Press, Cambridge, MA, USA.

Grigonytė, G., Cordeiro, J., Dias, G., Moraliyski, R. & Brazdil, P. (2010). Paraphrase alignment for synonym evidence discovery. In *COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics*, 403–411, Tsinghua University Press, Beijing, China.

Harris, Z.S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 92: Proceedings of the 14th International Conference on Computational Linguistics*, 539–545, Association for Computational Linguistics, Morristown, NJ, USA.

Heyer, L.J., Kruglyak, S. & Yooseph, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, **9**, 1106–1115.

Heylen, K., Peirsman, Y., Geeraerts, D. & Speelman, D. (2008). Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 3243–3249.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting of ACL*, 268–275, Association for Computational Linguistics, Morristown, NJ, USA.

Hirst, G. & St-Onge, D. (1998). *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*, chap. 13, 305–332. MIT Press.

Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. John Wiley and Sons, New York.

Huffman, S.B. (1996). Learning information extraction patterns from examples. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, 246–260, Springer-Verlag, London, UK.

Inkpen, D. & Hirst, G. (2006). Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, **32**, 223–262.

Jarmasz, M. & Szpakowicz, S. (2004). Roget's thesaurus and semantic similarity. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP)*, 212–219.

Jiang, J.J. & Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*.

Jing, H. & McKeown, K.R. (2000). Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 178–185, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Justeson, J.S. & Katz, S.M. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, **17**, 1–19.

Justeson, J.S. & Katz, S.M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 9–27.

Kaplan, A. (1950). An experimental study of ambiguity and context. *Mechanical Translation*, **2**, 39–46.

Keller, F. & Lapata, M. (2003). Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, **29**, 459–484.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, **33**, 147–151.

Kilgarriff, A. & Yallop, C. (2001). What's in a thesaurus? In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC)*, 1371–1379.

Kleinberg, J.M. (1998). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604–632.

Koehn, P. (2002). Europarl: A multilingual corpus for evaluation of machine translation, unpublished draft, available from `http://www.statmt.org/europarl/`.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics*, **14**, 241–257.

Kozareva, Z., Moraliyski, R. & Dias, G. (2008). Web people search with domain ranking. In P. Sojka, A. Horák, I. Kopecek & K. Pala, eds., *Text, Speech and Dialogue*, vol. 5246 of *Lecture Notes in Computer Science*, 133–140, Springer-Verlag.

Landauer, T.K. & Dumais, S.T. (1996). How come you know so much? From practical problem to theory. *Basic and applied memory: Memory in context*, 105–126.

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

Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, ed., *WordNet: an electronic lexical database*, 265–283, MIT Press.

Lee, L. (1999). Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, 25–32.

Levin, B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press.

Lewis, D.D., Yang, Y., Rose, T.G. & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, **5**, 361–397.

Lin, D. (1994). Principar - an efficient, broad-coverage, principle-based parser. In *COLING '94: Proceedings of the 15th International Conference on Computational Linguistics*, 482–488.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *COLING 98: Proceedings of the 17th International Conference on Computational Linguistics*, 768–774.

Lin, D. (1998b). An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304, Morgan Kaufmann, San Francisco, CA, San Francisco, CA, USA.

Lin, D., Zhao, S., Qin, L. & Zhou, M. (2003). Identifying synonyms among distributionally similar words. In G. Gottlob, T. Walsh, G. Gottlob & T. Walsh, eds., *IJCAI'03: Proceedings of the 18th*

international joint conference on Artificial intelligence, 1492–1493, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, **37**, 145–151.

Liu, H. (2004). Montylingua: An end-to-end natural language processor with common sense. Available at: `http://web.media.mit.edu/~hugo/montylingua/`.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, **28**, 203–208.

Lund, K., Burgess, C. & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Cognitive Science Proceedings, LEA*, 660–665.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press, London.

Magnini, B. & Cavaglia, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, 1413–1418.

Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marcus, M.P., Marcinkiewicz, M.A. & Santorini, B. (1993). Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, **19**, 313–330.

Markowitz, J., Ahlswede, T. & Evens, M. (1986). Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, 112–119, Association for Computational Linguistics, Morristown, NJ, USA.

Maynard, D., Funk, A. & Peters, W. (2009). Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proceedings of the Workshop on Ontology Patterns (WOP2009) collocated with ISWC2009*, 39–52.

McCarthy, D., Koeling, R., Weeds, J. & Carroll, J. (2004). Finding predominant word senses in untagged text. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 279, Association for Computational Linguistics, Morristown, NJ, USA.

McCarthy, D., Koeling, R., Weeds, J. & Carroll, J. (2007). Unsupervised acquisition of predominant word senses. *Computational Linguistics*, **33**, 553–590.

Miller, G.A. (1990). Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, **3**, 245–264.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K.J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, **3**, 235–244.

Miller, G.A., Chodorow, M., Landes, S., Leacock, C. & Thomas, R.G. (1994). Using a semantic concordance for sense identification. In *HLT '94: Proceedings of the workshop on Human Language Technology*, 240–243, Association for Computational Linguistics, Morristown, NJ, USA.

Mitkov, R., Evans, R., Orăsan, C., Ha, L.A. & Pekar, V. (2007). Anaphora resolution: To what extent does it help NLP applications? In A. Branco, ed., *Anaphora: Analysis, Algorithms and Applications, Lecture Notes in Artificial Intelligence (LNAI 4410)*, 179–190, Springer Berlin Heidelberg.

Mohammad, S. (2008). *Measuring Semantic Distance using Distributional Profiles of Concepts*. Ph.D. thesis, University of Toronto, Toronto, Canada.

Mohammad, S. & Hirst, G. (2006). Determining word sense dominance using a thesaurus. In *Proceedings of the 11th conference of the European chapter of the Association for Computational Linguistics (EACL-2006)*, 121–128.

Mohammad, S. & Hirst, G. (2010). Measuring semantic distance using distributional profiles of concepts. *Submitted*.

Mohammad, S., Dorr, B. & Hirst, G. (2008). Computing word-pair antonymy. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 982–991, Association for Computational Linguistics, Morristown, NJ, USA.

Moraliyski, R. & Dias, G. (2007). One sense per discourse for synonymy extraction. In *International Conference On Recent Advances in Natural Language Processing, RANLP 2007*, 383–387.

Morris, J. & Hirst, G. (2004). Non-classical lexical semantic relations. In *In Proceedings of HTL-NAACL Workshop on Computational Lexical Semantics*, 46–51.

Muslea, I. (1999). Extraction patterns for information extraction tasks: A survey. In *In AAAI-99 Workshop on Machine Learning for Information Extraction*, 1–6.

Nakamura, J. & Nagao, M. (1988). Extraction of semantic information from an ordinary english dictionary and its evaluation. In *COLING '88: Proceedings of the 12th Conference on Computational Linguistics*, 459–464, Association for Computational Linguistics, Morristown, NJ, USA.

Nakov, P. & Hearst, M. (2005). Search engine statistics beyond the n-gram: application to noun compound bracketing. In *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, 17–24, Association for Computational Linguistics, Morristown, NJ, USA.

Navigli, R., Litkowski, K.C. & Hargraves, O. (2007). Semeval-2007 task 07: coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, SemEval '07, 30–35, Association for Computational Linguistics, Stroudsburg, PA, USA.

Nelson, D.L., McEvoy, C.L. & Schreiber, T.A. (2004). The University of South Florida word association, rhyme, and word fragment norms. http://web.usf.edu/FreeAssociation/. *Behavior research methods instruments computers a journal of the Psychonomic Society Inc*, **36**, 402–407.

Noord, G.V. (2006). At last parsing is now operational. In *In TALN 2006*, 20–42.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, **3**, 1405–1408.

Ohshima, H. & Tanaka, K. (2009). Real time extraction of related terms by bi-directional lexico-syntactic patterns from the Web. In *ICUIMC '09: Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*, 441–449, ACM, New York, NY, USA.

Otero, P.G. (2008). Comparing window and syntax based strategies for semantic extraction. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR '08, 41–50, Springer-Verlag, Berlin, Heidelberg.

Otero, P.G., Lopes, J.G.P. & Agustini, A. (2004). The role of optional co-composition to solve lexical and syntactic ambiguity. *Procesamiento del lenguaje natural*, **33**, 73–80.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 125–132, The Association for Computer Linguistics.

Pecina, P. & Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 651–658, Association for Computer Linguistics, Morristown, NJ, USA.

Pereira, F., Tishby, N. & Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 183–190.

Pradhan, S.S., Loper, E., Dligach, D. & Palmer, M. (2007). Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, SemEval '07, 87–92, Association for Computational Linguistics, Stroudsburg, PA, USA.

Priss, U. & Old, L.J. (2005). Conceptual exploration of semantic mirrors. In *Formal Concept Analysis: Third International Conference, ICFCA 2005*, Springer-Verlag.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, 315–322.

Rapp, R. (2004). Utilizing the one-sense-per-discourse constraint for fully unsupervised word sense induction and disambiguation. In *Proceedings of Forth Language Resources and Evaluation Conference (LREC 2004)*..

Resnik, P. (1995a). Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the third workshop on very large corpora*, 54–68.

Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, 448–453, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Riloff, E. & Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 117–124.

Roark, B. & Charniak, E. (1998). Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *COLING 98: Proceedings of the 17th international conference on Computational linguistics*, 1110–1116, Association for Computational Linguistics, Morristown, NJ, USA.

Rose, K., Gurewitz, E. & Fox, G.C. (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, **65**, 945–948.

Rubenstein, H. & Goodenough, J.B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**, 627–633.

Ruge, G. (1997). Automatic detection of thesaurus relations for information retrieval applications. In *In Foundations of Computer Science: Potential - Theory - Cognition, Lecture Notes in Computer Science, volume LNCS 1337*, 499–506, Springer-Verlag.

Sahlgren, M. (2001). Vector-based semantic analysis: Representing word meanings based on random labels. In *ESSLI Workshop on Semantic Knowledge Acquistion and Categorization*, Kluwer Academic Publishers, Helsinki, Finland.

Sahlgren, M. (2006a). Towards pertinent evaluation methodologies for word-space models. In *Proceedings of LREC 2006: Language Resources and Evaluation*.

Sahlgren, M. (2006b). *The Word-Space Model*. Ph.D. thesis, Stockholm University, Stockholm, Sweden, online `www.sics.se/~mange/TheWordSpaceModel.pdf`.

Sahlgren, M. & Karlgren, J. (2002). Vector-based semantic analysis using random indexing for cross-lingual query expansion. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, 169–176, London, UK.

Salton, G. & McGill, M.J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Salton, G., Yang, C.S. & Yu, C. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, **26**, 33–44.

Senellart, P. & Blondel, V.D. (2008). Automatic discovery of similar words. In M.W. Berry & M. Castellanos, eds., *Survey of Text Mining II: Clustering, Classification and Retrieval*, 25–44, Springer-Verlag.

Silva, F.G.M. & Gomes, A.J.P. (2003). Adjacency and incidence framework: a data structure for efficient and fast management of multiresolution meshes. In *Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, GRAPHITE '03, 159–166, ACM, New York, NY, USA.

Snow, R., Jurafsky, D. & Ng, A.Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 801–808, Association for Computational Linguistics, Morristown, NJ, USA.

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**, 11–21.

Sugawara, K.M., Nishimura, M., Toshioka, K., Okochi, M. & Kaneko, T. (1985). Isolated word recognition using hidden markov models. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85*, 1–4.

Tang, J., Hong, M., Zhang, D., Liang, B. & Li, J. (2006). Information extraction: Methodologies and applications.

Tatsuki, D. (1998). Basic 2000 words - synonym match 1. In Interactive JavaScript Quizzes for ESL Students, `http://a4esl.org/q/j/dt/mc-2000-01syn.html`.

Terra, E. & Clarke, C. (2003). Frequency estimates for statistical word similarity measures. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 165–172, Association for Computational Linguistics, Morristown, NJ, USA.

Turney, P.D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, 491–502, Springer-Verlag, London, UK.

Turney, P.D., Littman, M.L., Bigham, J. & Shnayder, V. (2003). Combining independent modules in lexical multiple-choice problems. In *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003.*, 101–110.

van der Plas, L. & Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 866–873, Association for Computational Linguistics, Morristown, NJ, USA.

Weeds, J. & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistic*, **31**, 439–475.

Weeds, J., Weir, D. & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, 1015–1021, Association for Computational Linguistics, Morristown, NJ, USA.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 189–196, Association for Computational Linguistics, Morristown, NJ, USA.

Zhu, X. & Rosenfeld, R. (2001). Improving trigram language modeling with the World Wide Web. In *Acoustics, Speech, and Signal Processing, ICASSP'01*, 533–536.

Zipf, G.K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, **33**, 251–266.

Appendices

A.1 Similarity Graphics by Polysemy

A.1.1 Global similarity graphics

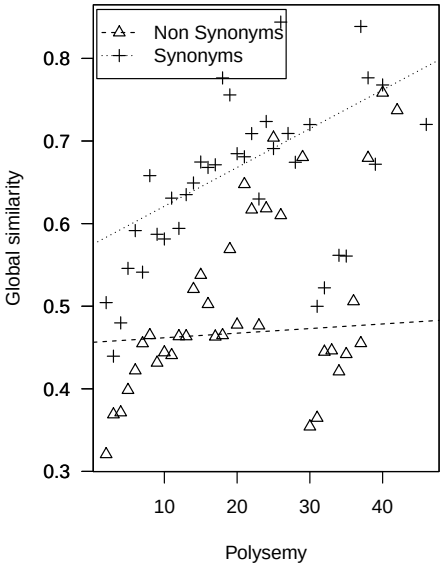

Figure A.1: *Global* Cos TfIdf by polysemy

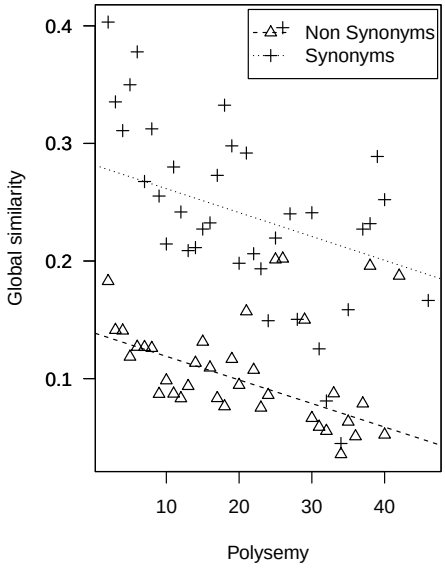

Figure A.2: *Global* Cos PMI by polysemy

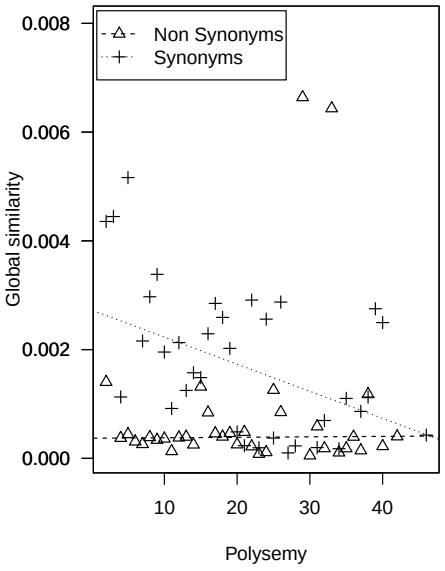

Figure A.3: *Global* Cos Prob by polysemy

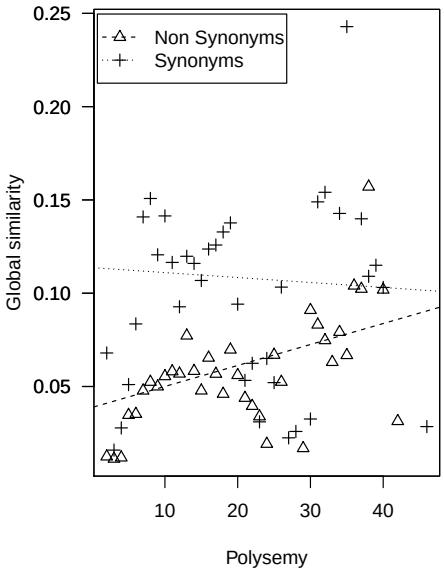

Figure A.4: *Global* Ehlert by polysemy

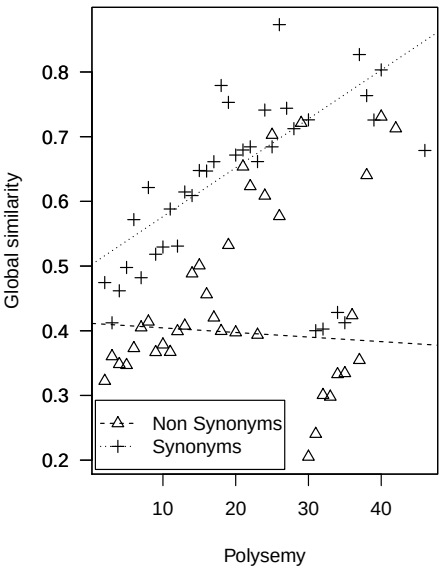

Figure A.5: *Global* Lin by polysemy

A.1.2 Local similarity graphics

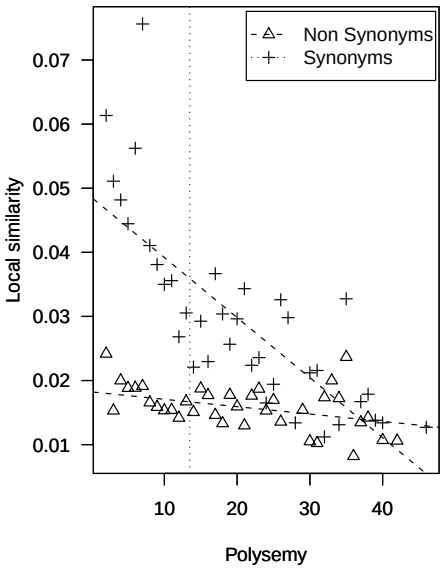

Figure A.6: *Local* Cos TfIdf by polysemy

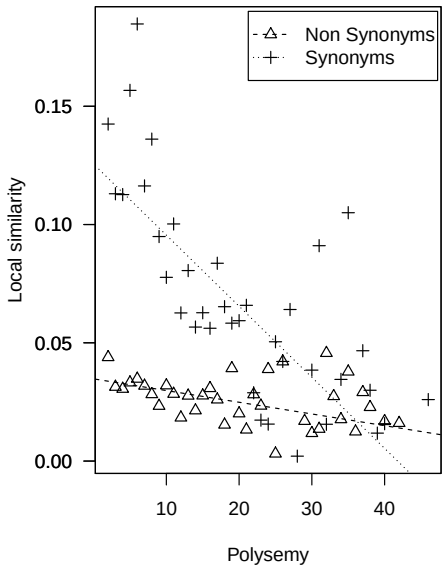

Figure A.7: *Local* Cos PMI by polysemy

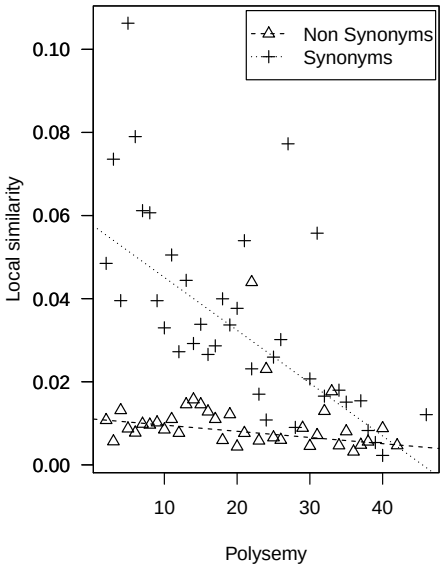

Figure A.8: *Local* Cos Prob by polysemy

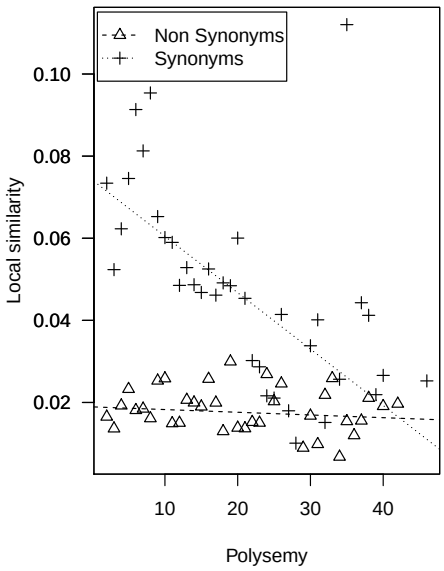

Figure A.9: *Local* Ehlert by polysemy

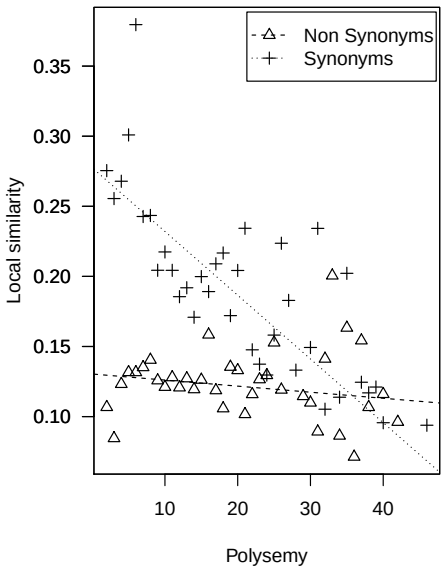

Figure A.10: *Local* Lin by polysemy

A.1.3 Product similarity graphics

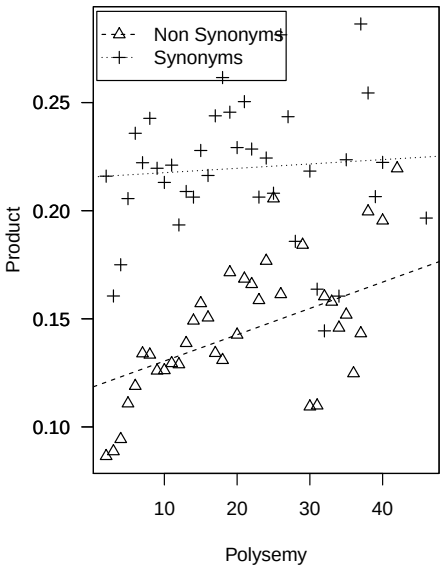

Figure A.11: *Product* Cos Tfldf by polysemy

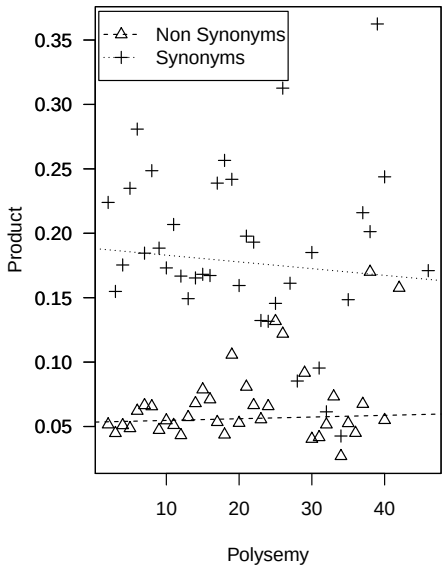

Figure A.12: *Product* Cos PMI by polysemy

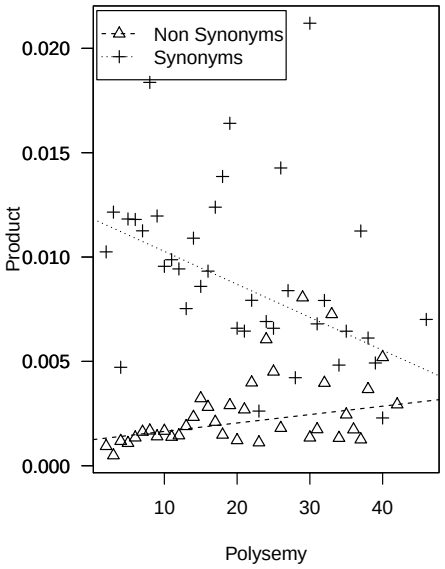

Figure A.13: *Product* Cos Prob by polysemy

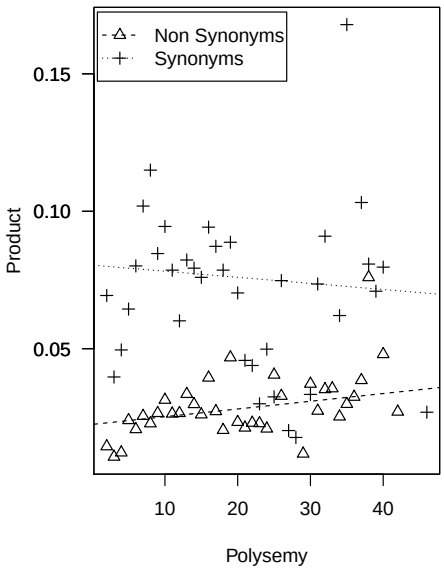

Figure A.14: *Product* Ehlert by polysemy

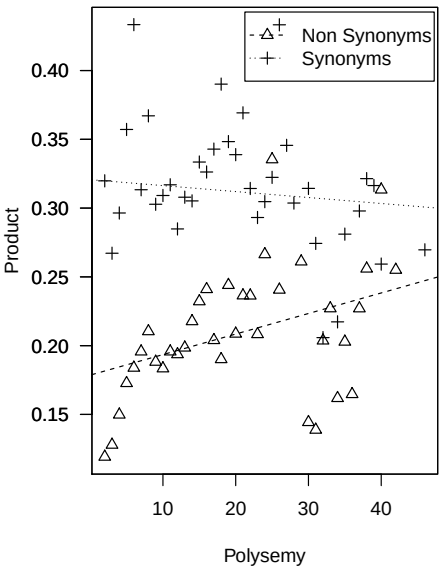

Figure A.15: *Product* Lin by polysemy

A.2 Classification Confidence Graphics

A.2.1 Global similarity classification confidence graphics

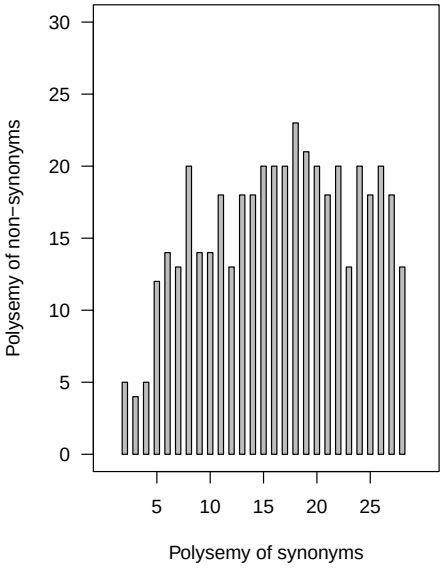

Figure A.16: *Global* Cos TfIdf Confidence

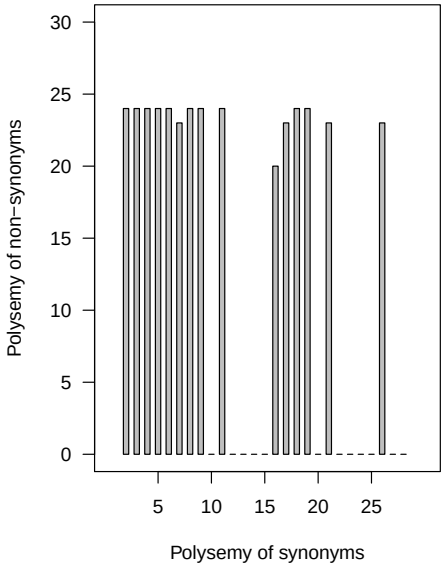

Figure A.17: *Global* Cos PMI Confidence

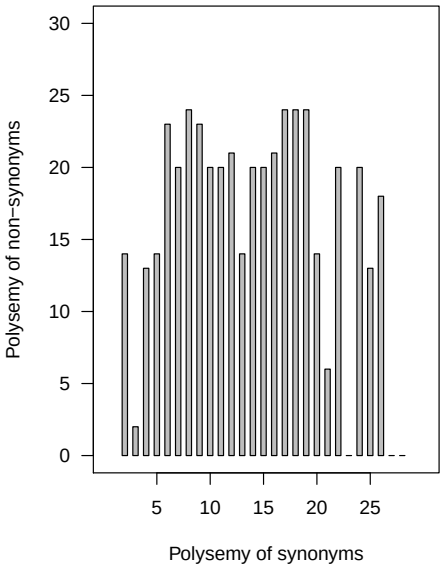

Figure A.18: *Global* Cos Prob Confidence

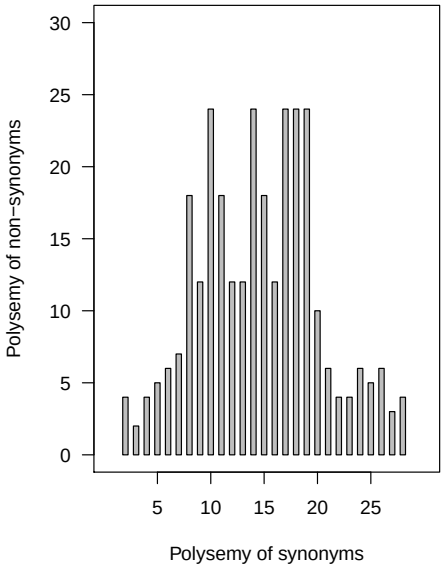

Figure A.19: *Global* Ehlert Confidence

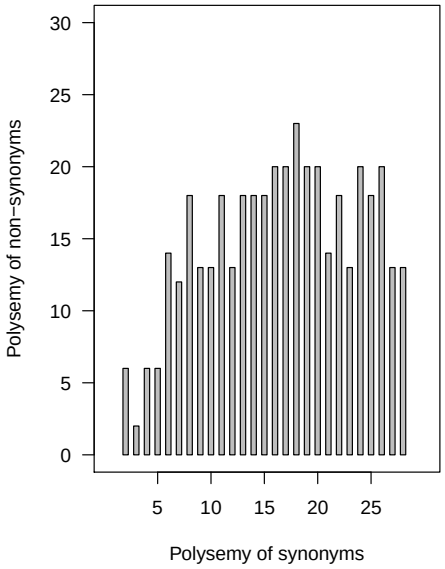

Figure A.20: *Global* Lin Confidence

A.2.2 Local similarity classification confidence graphics

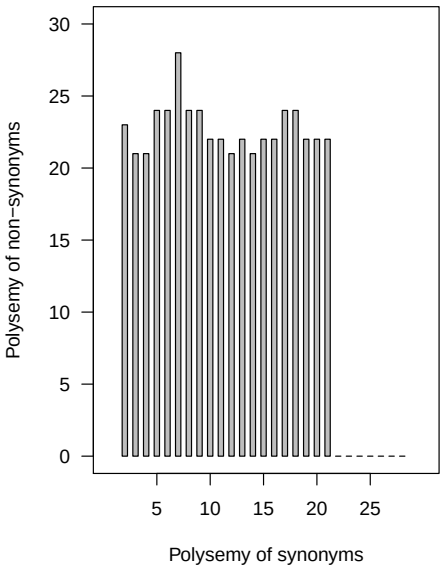

Figure A.21: *Local* Cos TfIdf Confidence

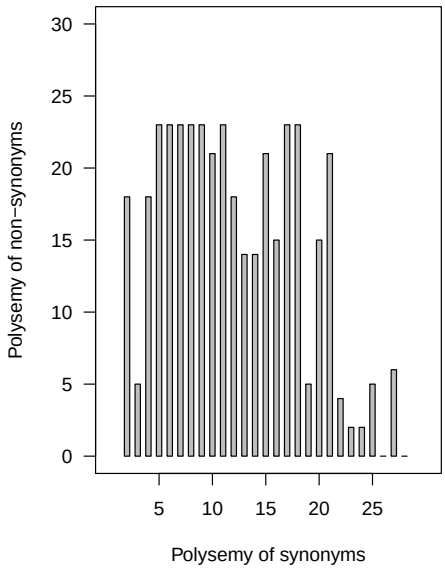

Figure A.22: *Local* Cos PMI Confidence

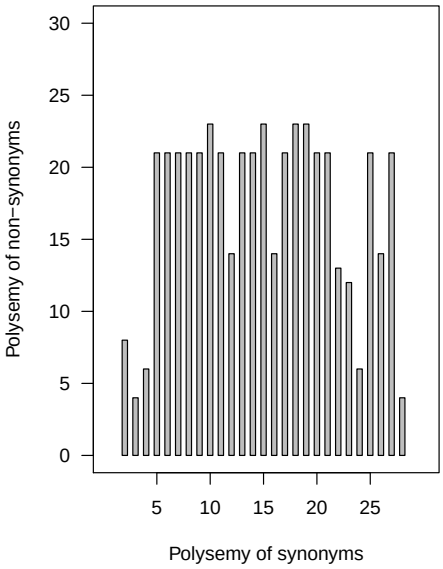

Figure A.23: *Local* Cos Prob Confidence

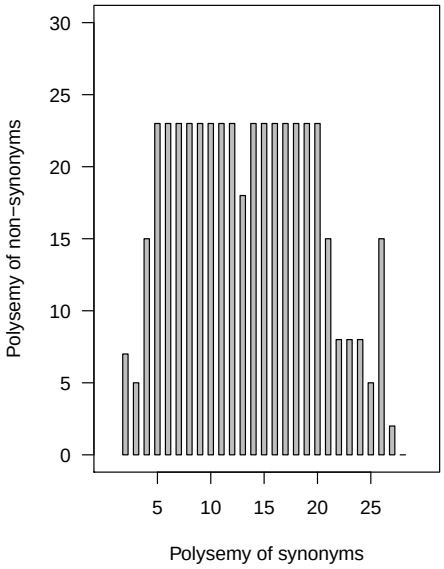

Figure A.24: *local* Ehlert Confidence

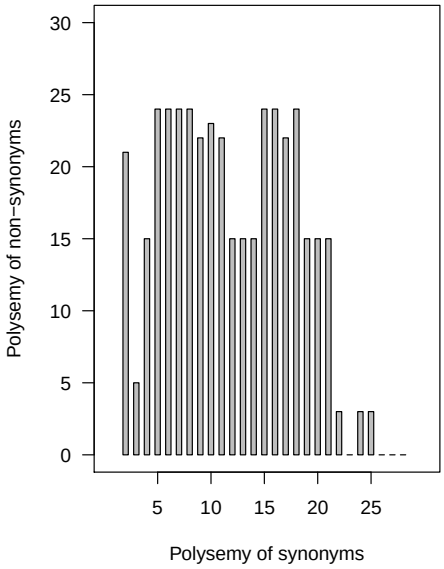

Figure A.25: *Local* Lin Confidence

A.3 Candidate Thesaurus Relations

Table A.1: Candidate thesaurus relations (*Synonymy*).

Synonymy			
Missouri	Mo	participant	party
subject	topic	planning	policy
choice	option	professor	scholar
Sept	September	Rudolph	Rudy
honor	honour	sen	senator
defense	defence	Shia	Shiite
nation	country	Addou	Addow
Nov	November	collaboration	cooperation
Oct	October	LTTE	tiger
veteran	vet	Mahdi	Mehdi
program	programme	percent	cent
seat	place	Wolverine	Michigan
world	globe	Haniya	Haniyeh
accord	treaty	corporation	firm
spot	position	Bande	Banda
manufacturer	maker	Howard	Goward
Maryland	Md	Butler	Butner
proceedings	proceeding	body	panel
measure	bill	child	infant
organization	organisation	cabinet	government
tv	television	company	firm
percent	percentage	city	stronghold
infant	baby	worker	laborer
child	baby	feature	film
home	family	breakup	break
purpose	intent	savings	synergy
case	lawsuit	article	study
state	nation	Paterson	Patterson
role	function	Youni	Yune
agency	way	accomplice	group

Table A.1: (continued)

Synonymy			
issue	matter	area	district
commitment	dedication	body	widebody
show	appearance	cofounder	founder
price	cost	newspaper	paper
voting	vote	holding	company
consumption	expenditure	election	vote
union	marriage	researcher	scientist
percentage	share	fee	tuition
part	role	college	school
butchery	carnage	board	panel
death	destruction	negotiator	envoy
administration	government	petroleum	oil
advertising	campaigning	revolt	rebellion
agency	association	battle	struggle
authority	government	reporter	newsman
case	trial	animation	excitement
CAT	CT	moon	month
commander	gen	voter	resident
concertgoer	fan	guide	cookbook
condition	disease	contender	challenger
corps	force	discussion	dialog
godfather	leader	discussion	forum
gunman	sniper	chamber	committee
imagery	model	earnings	profit
madrasa	madrassa	award	honor
madrassa	school	attorney	lawyer
marine	navy	plan	program
meeting	talk	couple	pair
Michael	Mike		

Table A.2: Candidate thesaurus relations (*Co-hyponymy*).

Co-hyponymy			
Amazon	eBay	Sunday	Monday
DaimlerChrysler	Mercedes	tomorrow	Monday
Thursday	yesterday	estimate	expectation
today	Monday	Monday	Friday
battle	race	chase	race
black	hispanic	Afghanistan	Pakistan
Chrysler	NYSE	database	data
department	government	leadership	government
flight	trip	Mo	Carpenter
idea	plan	official	spokeswoman
barrel	gallon	organization	effort
journalist	videographer	phrase	description
lady	wife	researcher	doctor
protester	woman	scenario	intent
publicist	spokesperson	week	year
report	review	finding	observation
blaze	wildfire	fraud	scandal
Boeing	Tractor	plot	proposal
condition	argument	study	test
country	family	time	globe
culture	habit	war	invasion
feedback	note	Oct	Friday
flaw	issue	UK	Argentina
information	matter	Aug	September
pastor	head	reporter	journalist
wife	mother	today	Sunday
year	century	study	research

Table A.2: (continued)

Co-hyponymy			
Brady	Bruschi	Washington	Tehran
day	Wednesday	Missouri	Arkansas
Dollar	Lenox	Bernardini	Albertrani
draft	resolution	Spain	Austria
embryo	life	newspaper	journal
hour	day	editorial	periodical
parent	family	capital	centre
Patriot	Dolphin	day	year
Pittsburgh	Steeler	day	age
Prado	Albertrani	Monday	year
profit	revenue	army	navy
reality	point	Syria	Lebanon
representative	publicist	SID	COT
softie	gentleman	singer	actress
stake	cup	typhoon	landslide
stewardship	charge	Jose	Seattle
Sunday	Friday	week	time
unrest	horror	security	justice
Wednesday	today	stock	fund
week	weekend	oil	gas
yesterday	Monday	fighting	guerrilla
album	cd	conscience	culture
challenger	candidate	contender	player
command	domination	remark	reference
course	strategy	target	commitment
Golen	Mccauley	challenge	task
performance	role	pressure	temperature

Table A.3: Candidate thesaurus relations (*Is a*).

Is a			
repair	capability	confrontation	crisis
company	user	administration	leader
conspiracy	obstruction	warming	action
sheik	cleric	city	area
slum	district	fat	food
fame	status	program	system
baseball	game	custom	agency
foe	opponent	brain	tissue
game	play	brainstem	tissue
minibus	bus	breathing	ability
yeast	organism	arrest	operation
allegation	statement	news	interview
blaze	fire	food	source
deal	agreement	marine	force
firefighter	crew	navy	corps
aircraft	plane	holding	company
coach	person	computer	property
documentary	film	university	institution
investigator	agent	grant	aid
overfishing	fishing	autopsy	study
dictator	strongman	feedback	note
sultan	figure	penalty	sanction
mail	message	poll	survey
cancer	disease	CEO	executive
diabetes	disease	oil	fuel
protest	resistance	brent	oil
negotiation	discussion	song	music
indictment	charge	animation	production
website	source	deadlock	clash
race	contest	commissioner	ambassador
state	region		

Table A.4: Candidate thesaurus relations (*Instance of*).

Instance of			
Australian	national	Limbaugh	commentator
Graham	coach	Courtney	singer
UN	community	Cobain	singer
Ruiz	governor	Cimaron	typhoon
July	month	Oracle	system
Mbeki	president	Office	software
Aisawa	legislator	Downer	minister
Corzine	democrat	Howard	minister
Banda	farmer	Qarase	minister
Lville	school	Witherspoon	winner
Silva	president	Reese	winner
Michael	Somare	Argentina	state
Patriot	team	Ryder	publicist
Germany	country	China	state
FedEx	company	Caremark	manager
IAEA	agency	Sean	administrator
Worm	researcher	Dalhousie	university
Schwarzenegger	star	Mohammed	sheik
Ethiopia	state	Paul	legend
China	country	Triple	holding
David	child	Triple	company
Baur	investigator	Alwaleed	prince
MSNBC	radio	Isadore	CEO
Nicaraguan	population	Washington	headquarters
Russia	state	Citigroup	bank
Kean	contender	Delhi	capital
Monday	day	Microsoft	company
Guardian	newspaper	Jose	city
Ortega	sandinista	Stanford	university
Shaukat	pakistani	Kinney	researcher
voter	floridian	LLC	trust

Table A.4: (continued)

Instance of			
Holmes	actress	IAEA	inspector
Dec	month	Amazon	stock
Haggard	rev	SID	syndrome
Sony	company	Google	engine
UK	country	Gibson	reporter
Bush	president	Shrek	animation
Nirvana	group	Moon	minister
Paterson	researcher	Talent	politician
Barrow	rep	Sardenberg	ambassador
Joshua	researcher	Cohen	comedian
Winfrey	host	Cordovez	envoy
Canada	state	Wii	console