Concept-Based Approach to Word-Sense Disambiguation

Ariel Raviv
Concept-Based Approach to Word-Sense Disambiguation

Research Thesis

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Ariel Raviv

Submitted to the Senate of the Technion — Israel Institute of Technology

Sh’vat, 5772 Haifa February 2012
The research thesis was done under the supervision of Prof. Shaul Markovitch in the Computer Science Department.

I would like to express my gratitude to my advisor, Prof. Shaul Markovitch, for his guidance during my studies. I am thankful to Shaul for assisting me in my first steps in the world of research, for the fruitful and inspiring discussions that often led to new directions and better results, and for always finding the time to meet me, even under challenging constraints. I would also like to thank Sonya Liberman, Kira Radinsky, Omer Levy, Anat Hashavit, Erez Karpas, Haggai Toledano, Assaf Glazer and Gal Naimark, my colleagues from the Learning and Reasoning lab, for many valuable discussions. Finally, I would like to thank my spouse Elizabeth, for her love and support during my studies.

The generous financial support of the Technion is gratefully acknowledged.
## Contents

Abstract

1 Introduction

2 Concept-Based Word-Sense Disambiguation
   1 The CBD Algorithm
   2 The Sense Retriever
   3 The Semantic Interpreter
      3.1 Explicit Semantic Analysis
      3.2 Compact Hierarchical Semantic Representation
      3.3 Anchored Explicit Semantic Analysis
   4 The Similarity Estimator

3 Empirical Evaluation
   1 Experimental Setup
   2 The Performance of the CBD Algorithm
   3 Comparison with State-of-the-Art Methods
   4 Domain Specific Word Sense Disambiguation
   5 Learning to Improve CBD
   6 Qualitative Analysis of CBD
      6.1 Benefits of World Knowledge
      6.2 Pitfalls of Knowledge-Based Algorithms

4 Related Work

5 Conclusions

Bibliography

Abstract in Hebrew
List of Figures

2.1 General illustration of the system .................................................. 8
2.2 Procedure for finding the most appropriate sense ............................. 9
2.3 Generation of an ESA model from Wikipedia articles. The articles and words in them are processed to build a weighted inverted index, representing each word as a vector in the space of all Wikipedia concepts (articles). ........... 11
2.4 The compact hierarchical semantic representation generated by CHESA for the word money ............................................................... 13
2.5 A partial view of the CHESA representation for the word apple ........... 16
2.6 The top categories of the Anchored-CHESA representation for the word apple anchored by fruit ......................................................... 16
2.7 The top categories of the Anchored-CHESA representation for the word apple anchored by computer ................................................. 17
3.1 A partial view of the ESA (top) and the Anchored-ESA (bottom) representations for the context of the word program as appears in the test corpus. 27
3.2 A partial view of the ESA (top) and the Anchored-ESA (bottom) representations for the third sense of the word program, which refers to a computer program ................................................................. 28
3.3 A partial view of the CHESA (top) and the Anchored-CHESA (bottom) representations for the context of the word balloon as appears in the test corpus. ................................................................. 29
3.4 A partial view of the CHESA (top) and the Anchored-CHESA (bottom) representations for the second sense of the word balloon, which refers to a hot-air balloon. ................................................................. 30
Abstract

The task of automatically determining the correct sense of a polysemous word has remained a challenge to this day. It is crucial in many natural language processing (NLP) applications such as speech recognition, information retrieval, machine translation and computational advertising. In our research, we introduce Concept-Based Disambiguation (CBD), a novel framework that utilizes recent semantic analysis techniques to represent both the context of the word and its senses in a high-dimensional space of natural concepts. The concepts are retrieved from a vast encyclopedic resource, thus enriching the disambiguation process with large amounts of domain-specific knowledge. In such concept-based spaces, more comprehensive measures can be applied in order to pick the right sense. Additionally, we introduce a novel representation scheme, denoted anchored representation, that builds a more specific text representation associated with an anchoring word. We evaluated our framework using two recent text representation schemes, Explicit Semantic Analysis (ESA) and Compact Hierarchical Explicit Semantic Analysis (CHESA) and their two anchored counterparts, and showed that the anchored representation is more suitable to the task of word sense disambiguation (WSD). Finally, we show that our system is superior to state-of-the-art methods when evaluated on domain-specific corpora, and is competitive to recent methods when evaluated on a general corpus.
Chapter 1

Introduction

Human language is innately ambiguous. In all natural languages, many words can be interpreted in a variety of ways, in accordance with their context. For instance, the English noun *bar* can mean *saloon*, *horizontal rod*, or *block of solid substance*. For us humans, choosing the correct sense of an ambiguous word in a given context is usually a trivial task. For example, it is obvious to any reader that the word *bar* in the sentence, “*He ate a whole bar of chocolate*”, refers to a *block of solid substance* rather than a *horizontal rod* or a *saloon*. However, because computers do not have the benefit of a human’s vast experience of the world and language, the task of automatically determining the correct sense of a polysemous word becomes a difficult problem. In fact, word sense disambiguation (WSD) has been described as an AI-complete problem by Mallery (1988): a problem whose difficulty is equivalent to solving central problems in artificial intelligence (AI), such as the Turing Test.

Disambiguating a word in context is crucial in many natural language processing (NLP) applications such as:

- **Speech Recognition.** WSD aids the processing of homophones, words which are spelled differently but pronounced the same. For example: “sun” and “son”.

- **Information Retrieval.** WSD helps improve term indexing and query expansion in information retrieval. IR systems can eliminate irrelevant hits by using different index entries for each sense of a keyword. WSD can also aid in expansion of short queries by adding related keywords, allowing more accurate results.

- **Machine Translation.** WSD assists in capturing the true essence of communication, thus allowing ideas to be transferred from one language to another.

- **Computational Advertising.** Computational advertising is a new scientific sub-discipline, at the intersection of information retrieval, machine learning, optimization, and microeconomics. Its central challenge is to find the best ad to present to a user engaged in a given context, such as querying a search engine (“sponsored search”) or reading
a Web page. WSD identifies the appropriate meanings of the main terms in the given context, thus ultimately assisting in finding the best ad to match the given query or page.

The importance of WSD was recognized in the late 1940s by Weaver (1955), who suggested its main components: the context in which a target word occurs, statistical information about words and senses, and knowledge resources. However, implementations of WSD systems during the 60’s and 70’s yielded inadequate results, due to a dearth of machine-readable knowledge. In this respect, work on WSD reached a turning point in the 1980s, with the release of large-scale lexical resources, which enabled automatic methods for knowledge extraction (Wilks, Fass, Ming-Guo, McDonald, Plate, & Slator, 1990). Since then, WSD methods can be classified into two types: knowledge-based and machine learning.

The machine learning approach utilizes information gathered from training on a corpus (Véronis, 2004; Mihalcea, 2007; Chan, Ng, & Zhong, 2007; Cai, Lee, & Teh, 2007). It usually includes building a classifier and using it to assign senses to unseen examples. It involves mainly collocation and co-occurrence features. In order to achieve high performance, supervised approaches need large training sets hand-annotated with the most appropriate word senses. This kind of knowledge is extremely expensive to create, space-wise and time-wise (Edmonds, 2000). Furthermore, as large as a training corpus may be, it will always lack full coverage of all senses for all the words in the lexicon, leading to inaccurate results.

The knowledge-based approach, on the other hand, does not rely on sense-annotated corpora, but takes advantage of the information contained in large lexical resources, such as WordNet (Fellbaum, 1998; Agirre, De Lacalle, & Soroa, 2009). The Lesk (1986) algorithm, is a classic example of the knowledge-based approach. A simplified version of the algorithm counts the number of words that are in both the neighborhood of the ambiguous word and in the definition of each sense in a dictionary. It then chooses the sense with the larger number of words. While simple and intuitive, Lesk’s approach is very sensitive to the exact wording of the definitions, so the absence of a certain word can radically change the results. This is a significant limitation as dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions. These drawbacks are common to most dictionary-based methods, as they have not realized the potential of combining the relatively limited information in such definitions with the abundant co-occurrence information extractable from text corpora (Cuadros & Rigau, 2006).

In the last decade, many approaches for enrichment of existing resources have been developed in order to deal with the problem of insufficient information (Hearst, 1992; Cimiano, Handschu, & Staab, 2004; Hearst, 1992; Cimiano et al., 2004; Harabagiu, Miller, & Moldovan, 1999; Girju, Badulescu, & Moldovan, 2006; Pennacchiotti & Pantel, 2006; Snow, Jurafsky, & Ng, 2006; Agirre & Martinez, 2001; McCarthy & Carroll, 2003; Navigli & Velardi, 2005; Navigli, 2009). Whereas some of these methods represent state-of-the-art approaches for enriching lexical resources, none of them make use of world knowledge, which is necessary for the WSD task.
In recent years, Wikipedia has become an external source of knowledge for WSD tasks. Wikipedia supplies vast amounts of common-sense and domain-specific world knowledge, which increase daily. There are two basic approaches for employing Wikipedia in disambiguation tasks: machine learning (Mihalcea, 2007; Bunescu & Pasca, 2006; Cucerzan, 2007) and similarity models (Strube & Ponzetto, 2006; Milne, 2007; Turdakov & Velikhov, 2008).

These approaches, while successfully incorporating the restricted information in dictionary definitions with the abundant information extractable from Wikipedia, still suffer from several limitations. Some require the word to appear as ambiguous in Wikipedia; any word that does not have a disambiguation page in Wikipedia cannot be disambiguated. Moreover, some methods, such as Bunescu and Pasca (2006) and Cucerzan (2007), are only suitable for named-entity disambiguation, as each entity has to have a separate article in the Wikipedia in order to be disambiguated. Some methods employ mapping algorithms that try to map each sense to a corresponding article in Wikipedia, either manually (Mihalcea, 2007) or automatically (Ponzetto & Navigli, 2010). The mapping relies mainly on textual overlap of titles and content, causing the process to become brittle.

Moreover, methods based on mapping of dictionary senses are less suitable in cases where named entities and domain-specific terms are involved, which are common in domain-specific corpora. As these entities do not appear in the dictionary to begin with, the mapping procedures do not take them into account. As a result, common knowledge regarding these entities is ignored during the WSD process, a fact that can impair performance. Domain-specific corpora also pose much difficulty for learning-based methods, in which the choosing of the most appropriate sense is deeply influenced by the frequency of the senses in the training corpora. As opposed to a general corpus, in a domain-specific corpus the distributions of the senses of words are often highly skewed. This property can severely impair the performance of supervised methods when evaluated across different domains (Agirre et al., 2009).

In this paper, we will introduce Concept Based Disambiguation (CBD), a novel framework which utilizes recent semantic analysis techniques to represent both the context of the word and its senses in a space of natural concepts grounded in human cognition. It then picks the sense that is most similar to the word’s context in that space. The concepts are retrieved from a vast encyclopedic resource, thus enriching the disambiguation process with a large amount of domain-specific knowledge. This approach has several advantages: (1) it can disambiguate any word as long as it exists somewhere in the knowledge source, not only in titles; (2) it is suitable for named entities and domain-specific terms (3) the process is completely automatic (4) any knowledge source that connects topics with texts is suitable: it need not have a specific structure or include “special” pages for ambiguous words. Additionally, as this approach relies on natural concepts, it should be able to capture the main gists of each sense as perceived by humans and therefore will better agree with human annotators.

We evaluate our system in coarse-grained general settings and show that it outperforms state-of-the-art learning-based methods which exploit large annotated corpora and is also comparable to recent state-of-the-art knowledge-based methods that make use of encyclopedic knowledge. Additionally, we prove that our system is superior to recent state-of-the-art methods when evaluated on domain-specific corpora.

The rest of this paper is organized as follows: In Section 2 we first describe the general CBD framework. We then present several methods for semantic representation and assessing relatedness that we have evaluated in this paper. In Section 3 we report our experimental results, comparing the different algorithms and evaluating our system against recent state-of-the-art learning-based and knowledge-based methods in both general settings and domain settings. We also present a qualitative analysis of CBD to demonstrate our system’s strengths and weaknesses. We discuss related work in Section 4, where we elaborate on the differences between our methods and recent knowledge-based methods. We conclude in Section 5.
Chapter 2

Concept-Based Word-Sense Disambiguation

Given a word in a text, word sense disambiguation (WSD) addresses the task of associating that word with an appropriate definition or meaning that is distinguishable from other meanings attributable to that word. In practice, the task involves three steps: retrieving a list of senses from a sense inventory (a dictionary), identifying the context of the word in the given text, for instance, the sentence in which it appears, and assigning the appropriate sense to the given word in the given context. We will now describe in detail how our approach handles these three steps. First, we will present a general framework for WSD that is entirely concept based and identify the semantic components and data sources it requires. Next, we will describe the sense inventory that our algorithm relies on in this paper and the methods we have implemented to exploit it. Finally, we will elaborate on the semantic components we chose to evaluate our framework, as these components lie at the core of our system.

1. The CBD Algorithm

Our algorithm follows the traditional Lesk (1986) algorithm in the way that it chooses the sense most similar to the context, but whereas Lesk relies on simple word overlap, our algorithm computes the semantic relatedness of texts in a high-dimensional space of concepts. It relies on the notation of a concept space, denoted $C$, of size $n$, where each concept $c_i \in C$ is associated with a natural concept or topic in the given knowledge source. In addition, it relies on a given dictionary, denoted $D = \{\langle w_i, S_i \rangle \}$, where for each word $w_i$ there exists a list of senses $S_i = \{\langle s_{ij} \rangle | j = 1, \ldots, n_i \}$.

The algorithm’s input consists of the word to disambiguate, denoted $w$, the context in which it appears, denoted $Ctx$, and a list of senses, $S$, which consist of senses of $w$ as found in the dictionary. We call the component responsible for associating each sense with a textual fragment the sense retriever, denoted $SR$, which naively matches each sense to its gloss. The algorithm can also benefit from a hierarchically-structured dictionary, as we
will describe in the following subsections. In addition, the algorithm relies on the existence of two semantic components. One is a \textit{semantic interpreter}, denoted \textit{SEM}, which is able to represent text fragments in a high-dimensional space of natural concepts. The other is a \textit{similarity estimator}, denoted \textit{SIM}, that given a pair of representations outputs their semantic distance.

The algorithm uses the sense retriever to associate each sense $s_i \in S$ with a textual fragment $t_i$. Then it uses the semantic interpreter to convert both the context, $Ctx$, and sense’s associated texts, $t_i$, to their concept-based representation by the semantic interpreter. The result is a weighted vector of concepts, denoted $\langle w_{1^c}, w_{2^c}, \ldots, w_{n^c} \rangle$ and $\langle w_{1^i}, w_{2^i}, \ldots, w_{n^i} \rangle$ respectively. In the next stage, the algorithm employs the similarity estimator to measure the semantic distance between the context’s representation and the representation of each of the meanings. Finally, it chooses the sense whose representation maximizes the similarity measure.

Figure 2.1 illustrates the system in general while Figure 2.2 describes the main procedure. First, textual data must be gathered in order to serve as input to the semantic interpreter before the representation is applied. In the following subsections we will describe this data in detail, as well as other knowledge sources and methods from which the algorithm benefits.

2. The Sense Retriever

Our algorithm relies on a machine-readable dictionary, which for each word holds its meanings described by fragments of text. The \textit{sense retriever} component is responsible for extracting a list of senses from the dictionary, and associating each sense with text that will later serve as input to the semantic interpreter. Naturally, a basic implementation of a sense retriever will associate each sense with its gloss in the dictionary. But the limited fragments of natural text in these glosses are inadequate for high-performance WSD. If, however, the dictionary has an intra-linked hierarchical structure, a more extensive sense retriever can be used, to the benefit of our algorithm.

The WordNet\textsuperscript{1} dictionary is an example of such a dictionary, and we will use it to evaluate our algorithm throughout this paper. WordNet is a widely used computational lexicon for the English language developed at the Cognitive Science Laboratory of Princeton University. WordNet distinguishes between words as they appear literally in a text and their actual sense. A set of words that share one sense is called a \textit{synset}. Thus, each synset identifies one semantic concept. Words with multiple meanings (ambiguous words) belong to multiple synsets. As of version 3.0, WordNet contains 82,115 synsets for 117,798 unique nouns, as well as other types of words like verbs and adjectives. WordNet also provides relations between synsets such as hypernymy/hyponymy (the relation between a sub-concept and a super-concept) and holonymy/meronymy (the relation between a part and the whole). In this paper, we will focus on hypernyms/hyponyms. The hypernymy

\textsuperscript{1} \url{http://wordnet.princeton.edu/}
Figure 2.1: General illustration of the system
**Procedure CBD**(\(w, Ctx, S\))
\[
\langle w_1^c, w_2^c, \ldots, w_n^c \rangle \leftarrow SEM(Ctx)
\]
\[
maxSim \leftarrow 0
\]
\[
pickedSense \leftarrow \text{nil}
\]
**Foreach** \(s_i \in S\):
\[
t_i \leftarrow SR(s_i)
\]
\[
\langle w_1^i, w_2^i, \ldots, w_n^i \rangle \leftarrow SEM(t_i)
\]
*If* \(SIM(\langle w_1^c, w_2^c, \ldots, w_n^c \rangle, \langle w_1^i, w_2^i, \ldots, w_n^i \rangle) > maxSim:*
\[
maxSim \leftarrow SIM(\langle w_1^c, w_2^c, \ldots, w_n^c \rangle, \langle w_1^i, w_2^i, \ldots, w_n^i \rangle)
\]
\[
pickedSense \leftarrow s_i
\]
**Return** \(pickedSense\)

---

**Gloss**

As the gloss contains a textual definition of \(s\), we include all of the words of the gloss after stemming, and after stop-words are removed. For instance, given the first sense of the noun *bank* from the Wordnet dictionary with its gloss, *a financial institution that accepts deposits and channels the money into lending activities*, the lemmas *financial institution*, *accept*, *deposit*, *channel*, *money*, *lending*, *activity* will be stemmed and added to the textual data.

**Synonymy**

We include the lemmas of all synonyms of \(s\) (including \(s\) itself). For the aforementioned example, the synonyms *banking concern* and *banking company* will be included.

**Hypernymy**

We include the lemmas and the glosses of all synonyms in the sense \(h\) such that \(h\) is a hypernym of \(s\). As \(h\) is a generalization of \(s\), lemmas from its gloss supply a vague description of \(s\) which can be valuable for disambiguation. In the aforementioned instance, the hypernyms *financial institution* and *financial organization* will be included, as well as lemmas from their gloss *an institution (public or private) that collects funds (from the public or other institutions) and invests them in financial assets.*

---

Figure 2.2: Procedure for finding the most appropriate sense

relation in WordNet can be conceived as spanning a directed acyclic graph (DAG) with a single root node called an *entity*.

WordNet glosses are limited in content. As explained earlier, for high performance WSD, the text of each sense needs to be extended using its semantic relations with other synsets (Banerjee & Pedersen, 2003). We will now describe an extensive sense retriever that utilizes the dictionary’s inter-linked structure to output textual data that will serve as input to our semantic interpreter. This data originates from several sources:

- **Gloss**
  - As the gloss contains a textual definition of \(s\), we include all of the words of the gloss after stemming, and after stop-words are removed. For instance, given the first sense of the noun *bank* from the Wordnet dictionary with its gloss, *a financial institution that accepts deposits and channels the money into lending activities*, the lemmas *financial institution*, *accept*, *deposit*, *channel*, *money*, *lending*, *activity* will be stemmed and added to the textual data.

- **Synonymy**
  - We include the lemmas of all synonyms of \(s\) (including \(s\) itself). For the aforementioned example, the synonyms *banking concern* and *banking company* will be included.

- **Hypernymy**
  - We include the lemmas and the glosses of all synonyms in the sense \(h\) such that \(h\) is a hypernym of \(s\). As \(h\) is a generalization of \(s\), lemmas from its gloss supply a vague description of \(s\) which can be valuable for disambiguation. In the aforementioned instance, the hypernyms *financial institution* and *financial organization* will be included, as well as lemmas from their gloss *an institution (public or private) that collects funds (from the public or other institutions) and invests them in financial assets.*
Hyponymy
We include the lemmas of all synonyms in the senses of such that is a hyponym of . Because is a specialization of , it can contribute to its representation. The aforementioned example has many hyponyms, such as: credit\_union, Federal\_Reserve\_Bank, acquirer, commercial\_bank.

3. The Semantic Interpreter
The main part of our algorithm consists of utilizing a semantic interpreter. In this paper, we experimented with two methods for semantic representation of text, Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2006) and Compact Hierarchical Semantic Representation (CHESA) (Liberman & Markovitch, 2009). Both methods output a semantic representation of texts in a high-dimensional space of concepts, but while ESA supplies a flat representation, CHESA produces a hierarchical one. In addition, we introduce two novel representation schemes that modify the aforementioned methods for the WSD task. In the remainder of this subsection, we describe these methods in further detail.

3.1 Explicit Semantic Analysis
Explicit Semantic Analysis (ESA) was introduced by Gabrilovich and Markovitch (2005, 2006, 2007, 2009) as a method for semantic representation of natural language texts. Inspired by the desire to augment text representation with massive amounts of world knowledge, ESA represents meaning in a high-dimensional space of concepts, automatically derived from large-scale human-built repositories such as Wikipedia. Since it was first proposed, ESA has been successfully applied to text categorization (Gabrilovich & Markovitch, 2006; Gupta & Ratinov, 2008; Chang, Ratinov, Roth, & Srikumar, 2008), semantic relatedness calculation (Gabrilovich & Markovitch, 2007; Gurevych, Müller, & Zesch, 2007; Radinsky, Agichtein, Gabrilovich, & Markovitch, 2011), cross-language information retrieval (Potthast, Stein, & Anderka, 2008; Cimiano, Schultz, Sizov, Sorg, & Staab, 2009), and concept-based information retrieval (Egozi, Gabrilovich, & Markovitch, 2008).

Wikipedia is the largest encyclopedia in the world. It is more than 25 times larger than the next largest English-language encyclopedia, Encyclopedia Britannica. Wikipedia provides a comprehensive source of world knowledge, organized within a taxonomy-like structure determined by its articles and categories. Every article in Wikipedia contains a textual description of a single topic. We thus view each Wikipedia article as defining a concept corresponding to that topic. For example, the article that is titled Dog and describes the domestic animal dog corresponds to the concept Dog.

In Wikipedia-based ESA, the semantics of a given word are described by a vector storing the word’s association strengths to Wikipedia-derived concepts. A concept is generated from a single Wikipedia article and is represented as a vector of words that occur in this article, weighted by their TFIDF score (Salton & McGill, 1986). Once these concept vectors are generated, an inverted index is created to map back from each word to the concepts.
it is associated with. Thus, each word appearing in the Wikipedia corpus can be seen as triggering each of the concepts it points to in the inverted index, with the attached weight representing the degree of association between that word and the concept. The inverted index is also used to discard insignificant associations between words and concepts by removing those concepts whose weights for a given word are too low. The process is illustrated in Figure 2.3.

For example, here are the top ten concepts triggered by the word *investor*: 1. *Investment* 2. *Angel investor* 3. *Stock trader* 4. *Mutual fund* 5. *Margin (Finance)* 6. *Modern portfolio theory* 7. *Equity investment* 8. *Exchange-traded fund* 9. *Hedge fund* 10. *Ponzi scheme*. Even without reading the Wikipedia articles associated with these concepts, it will be intuitively clear to most readers that these concepts are relevant to the input word. The concepts’ labels also exhibit a degree of semantic similarity and relatedness to the input term that extends simple synonymy. As a result, computing relatedness between words based on their Wikipedia-ESA representation was shown to be more effective than any other method at the time of publication (Gabrilovich & Markovitch, 2007).

ESA can also be used to generate a semantic interpretation of given a text fragment in a centroid based manner. Given a text fragment, we first represent it as a vector using the TFIDF scheme. The semantic interpreter iterates over the text words, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts that represents the given text. Let $T = \{w_1, \ldots, w_l\}$ be input text, and let $\langle v_1, \ldots, v_m \rangle$ be its TFIDF vector, where $v_i$ is the weight of word $w_i$ (where $m$ is the total number of words in the vocabulary). Let $\langle k_1, \ldots, k_n \rangle$ be an inverted index entry for word $w_i$, where $k_j$ quantifies the strength of association of word $w_i$ with Wikipedia concept $c_j$, $\{c_j \in c_1, \ldots, c_n\}$ (where $n$ is the total number of Wikipedia concepts). Then, the semantic interpretation...
vector \( V \) for text \( T \) is a vector of length \( n \), in which the weight of each concept \( c_j \) is defined as \( \sum_{w_i \in T} v_i \cdot k_j \). Entries of this vector reflect the relevance of the corresponding concepts to text \( T \).

### 3.2 Compact Hierarchical Semantic Representation

Experimental evaluation of ESA on semantic relatedness as well as text categorization tasks showed considerable improvements over previous approaches. Nevertheless, ESA has three notable drawbacks.

1. **ESA representation is excessive.** It represents word semantics as a weighted combination of all Wikipedia articles, which can amount to millions of concepts. This can greatly impair the efficiency of the semantic interpreter. While it is possible to make ESA compact by selecting the top highly associated concepts (as done when using ESA for text categorization (Gabrilovich & Markovitch, 2006)), it can omit important semantic aspects of the word for which the concepts are being generated. For example, consider the top twenty concepts generated by ESA for the word *money*:

   1. Money
   2. Money, Money, Money
   3. Money creation
   4. Money for Nothing/Beverly Hillbillies
   5. Make Money Fast
   6. Money Supply
   7. Moneyness
   8. Cash Money Records
   9. Eddie Money
   10. Money Laundering
   11. Money Mark
   12. The Color of Money
   13. Making Money
   14. Take the Money and Run
   15. Money Market
   16. I Get Money
   17. The Money Programme
   18. Money (That’s What I Want)
   19. Electronic Money
   20. Money Wars.

Observe that eleven out of the top twenty concepts represent relatively minor meanings of the word, such as novels (*The Color of Money*), songs (*I Get Money*) and television programs (*The Money Programme*), while neglecting important aspects of the main meaning, such as *Monetary Economics* and *Bank*.

2. **ESA is noisy.** Even in its compact form, ESA contains many redundant and over-specific concepts. ESA is not able to differentiate between concepts representing the primary gists of the word (such as *Money* for the word *money* in the example above), and over-specific, noisy concepts (such as *Money, Money, Money* that refers to a ABBA song), as the association scores for both types of concepts are high. As ESA generates each concept independently of the other concepts in the representation, redundancies often occur. For example, the top-ten concepts generated by ESA for the word *car* contain seven different car types (*Concept Car, Sports Car, Armored Car, Executive Car, City Car, Compact Car and Full-Size Car*) and two types of car number plates (*Polish Car Number Plates* and *Greek Car Number Plates*).

3. **ESA is flat.** ESA views Wikipedia as a flat conceptual ontology and constructs flat semantic representations, thus disregarding the inner structure and inter-dependencies
of concepts. This is a notable weakness when addressing semantic relatedness of words and texts. When humans perform this task, they use their innate ability to generalize. For example, a human would easily determine that the words *money* and *wealth* are related as they both trigger high-level concepts related to economics and sociology.

In order to overcome these problems, we chose to employ Compact Hierarchical Explicit Semantic Representation (CHESA) (Liberman & Markovitch, 2009). CHESA is a novel approach that leverages structured encyclopedic knowledge encoded within Wikipedia articles and categories and uses the conceptual hierarchy inferred from this knowledge resource to represent text semantics. Figure 2.4 shows the CHESA representation for the word *money*. The representation expresses the primary gist of the word, showing both specific (e.g., MACROECONOMICS and MONETARY ECONOMICS) and high-level (e.g., SOCIETY) concepts triggered by it.

CHESA represents the semantics of an input text as a weighed sub-hierarchy within the global ontology. Namely, it draws a virtual separating curve on top of the hierarchy: this curve determines which concepts are included into the semantic representation (those above the separating curve) and which are excluded (those below the separating curve).

Formally, we can say that the CHESA algorithm, in order to build a semantic representation for input text $t$, performs a DFS top-down traversal over the nodes of the hierarchy. The algorithm starts with a representation containing the root $r$ only. When a node $n$ is reached, for every child $d$ of $n$, the algorithm applies a conditional over-representation criterion to determine whether $d$ is significantly more associated with text $t$ than its parent $n$. If so, $d$ is included in the representation and DFS traversal proceeds within the subtree of $d$. The conditional over-representation criterion is based on a hypergeometric test, and its value quantifies the marginal contribution of the child concept $d$ given that its parent concept $n$ is included into the semantic representation.
Intuitively, we can say that CHESA performs specification by demand: it starts by representing the word or text on a very high conceptual level and iteratively reveals more specific meanings of the word, until a more specific description would no longer contribute to its semantic interpretation. Moreover, every concept within the representation is assigned a weight which represents the strength of association between its textual content and the input text.

3.3 Anchored Explicit Semantic Analysis

Like other similarity based methods, our approach chooses the appropriate sense of an ambiguous word by measuring semantic relatedness between that text and the textual description of each sense. However, these textual fragments often include words that are ambiguous in themselves, with different meanings that span beyond their role in the given text. Those meanings, while unrelated to the word in question, are still included by the aforementioned representation schemes.

For example, consider the sentence The Family Tree of King Alfred the Great shows 37 generations and over 3000 individuals including kings and queens of England, Scotland, Denmark, and Spain, with the word tree to disambiguate. The word tree has two main meanings according to the Wordnet dictionary:

1. a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown

2. a figure that branches from a single root

The context of the word for the disambiguation process can vary in size, but will surely include the word king, which appears twice in the sentence and also most adjacent to tree. Notice the word crown, which appears in the description of the first sense. In the context of that gloss, the word crown refers to the upper leaves of a tree. However, the word crown usually stands for a symbol of monarchy, and is thus highly related to the word king, which is found in the context. In terms of representation, both interpretation vectors of the words crown and king share monarchy-related concepts (e.g., British Monarchy, Crown dependency), which should also be included in the interpretation vectors of the texts in which the words appear. As a result, the interpretation vector of the first sense becomes more similar to the interpretation vector of the context than the first one. This phenomenon could result in choosing the first sense over the second, while the second sense is obviously the right one.

In order to deal with this problem, we must change the representation of each sense to include only meanings associated with the word of the task at hand. A naive approach to this problem would be to include the ambiguous word itself, such as tree in the aforementioned example, in each sense’s semantic representation. As the representation of a textual fragment consists of a centroid of the semantic representation of its words, this modification will surely intensify the concepts of the appropriate meanings of rest of the words in
the text. However, this approach has several drawbacks. First, this method affects all of
the senses in a similar manner, and thus does not contribute to the goal of distinguishing
between them. Second, while unrelated concepts would get weaker in this process, they
would not be eliminated, and would still be taken into account upon measuring similarity.

As the naive approach does not seem to solve the problem, we introduce a novel approach
to represent the semantics of a text fragment _t_ in the context of a word. This word is denoted
as _anchoring word_ or simply _anchor_. The Anchored Representation of a text is achieved by
retaining only concepts that are related to the anchoring word. We designed two algorithms
to construct these anchored representations: one that is a modification of ESA and one that
is a modification of CHESA.

For Explicit Semantic Analysis, we construct an anchored representation for each of
_t_’s words and combine them, as usual, as a centroid. The anchored representation of a
word _w_ with respect to its _anchor_ is obtained by eliminating from _w_’s interpretation vector
concepts that are not included\(^2\) in the _anchor_’s interpretation vector. This process can also
be perceived as a projection of _w_’s interpretation vector onto a subspace spanned by concepts
related to the _anchor_. As a result, _w_’s semantic representation will include only meanings
associated with the anchoring word. This way, the unrelated concepts will not be taken into
account upon measuring similarity, while the rest of the concepts still hold their original
weights. This result is very different from the naive approach presented earlier, where the
original weights are modified and no longer signify _w_’s association with each concept. We
call this process Anchored Explicit Semantic Analysis, denoted Anchored-ESA.

To see the benefits of anchoring, let us examine how ESA and Anchored-ESA handle
the above example. ESA would find the words _crown_ and _king_ to be very similar, with
many concepts in common, such as British monarchy, Spanish monarchy and Crown
dependency. Consequently, when using cosine-based similarity to measure the similarity
of the two vectors, the resulting score is very high. However, this high score stems mainly
from the monarchy-related meaning of the word _crown_, which is not relevant to its role
in association with the word _tree_. But if we anchor the representation of the word _crown_
with the word _tree_, we get only specific, nature-related concepts, such as Eucalyptus
and Sequoia. As a result, the similarity score for _crown_ and _king_ is much lower, as it should
be in that context.

We also designed an anchored version of CHESA, called Anchored-CHESA. The algo-
rithm identifies concepts and categories that are not included in the representation graph
of the anchoring word, and deletes them from the full Wikipedia hierarchy. Then it builds
the CHESA representation, top down, in the regular manner. The benefits of anchoring on
CHESA are easily demonstrated. The different meanings of a word will often be represented
by different branches of the CHESA representation tree. So, for example, the category tree
of the word _apple_ will include both nature related categories, derived from one meaning,
and technology related categories, derived from the other, as can be seen in Figure 2.5.

\[^2\text{More precisely, we eliminate concepts that have a weight of 0 (or below some predefined threshold) in the anchor representation vector.}\]
Figure 2.5: A partial view of the CHESA representation for the word *apple*

Figure 2.6: The top categories of the Anchored-CHESA representation for the word *apple* anchored by *fruit*

Anchoring the representation of the word *apple* to the word *fruit* would retain only the nature related branches, while anchoring it to the word *computer* would retain only the technological ones. The results are illustrated in Figures 2.6 and 2.7.

4. The Similarity Estimator

Our algorithm employs a *similarity estimator* to compute the semantic relatedness of a pair of text fragments represented in a high-dimensional space of concepts. In this paper,
wherever ESA-based representation was applied, we followed the work of Gabrilovich and Markovitch (2007) and measured relatedness by comparing a pair of concept-based vectors using the cosine metric (Zobel & Moffat, 1998). For CHESA-based representations, we define the semantic relatedness between two input texts, $t_1$ and $t_2$, as the cosine similarity between their linearized CHESA representations. Formally,

$$\text{SIM}(t_1, t_2) = \frac{\sum_{c \in C} s(t_1, c)s(t_2, c))}{\sqrt{\sum_{c \in C} s^2(t_1, c)}\sqrt{\sum_{c \in C} s^2(t_2, c)}}$$

where $C$ is the set of all the concepts in the global hierarchy.

It seems on first glance that our definition of relatedness does not exploit the hierarchical structure of CHESA representations, but closer examination shows otherwise. Indeed, there exist other measures that use hierarchical information explicitly. Such measures, e.g., generalized-cosine similarity (Ganesan, Garcia-Molina, & Widom, 2003) and Earth-mover’s distance (Rubner, Tomasi, & Guibas, 2000; Wan & Peng, 2005), compute similarity between collections, whose elements correspond to leaf nodes in some pre-defined global hierarchy. They then consider the relative proximity of the elements in that hierarchy to compute similarity. For example, the generalized-cosine similarity, extends the cosine-similarity measure, by dropping the assumption that different elements correspond to orthogonal dimensions. It defines a positive inner product between different dimensions, which is a function of the proximity of the corresponding elements in the hierarchy. Although it seems natural to use the aforementioned measures to compute similarity between CHESA representations, we find them unsuitable for this task. These measures assume a model where the collections they compare correspond to leaf nodes of some underlying hierarchical structure.
CHESA representations do not coincide with such a model. Internal nodes of the Wikipedia conceptual hierarchy are explicitly present within the representations. In fact, CHESA representations (in particular the compact ones) are mostly dominated by internal concepts and not by the leaf nodes of the conceptual hierarchy. Moreover, CHESA algorithms ensure that a concept cannot be part of the representation unless its ancestors are part of it. By considering both internal and leaf nodes, cosine similarity is capable of implicitly capturing the hierarchical structure.
Chapter 3

Empirical Evaluation

We evaluated our CBD algorithm on a common benchmark for disambiguation. We first describe the experimental setup. Next, we present the performance of our algorithms in their various versions. We then compare the performance of our method to several state-of-the-art systems. Finally, we provide a qualitative analysis to understand the strengths and weaknesses of our method.

1. Experimental Setup

We implemented our method using a Wikipedia snapshot as of October 18, 2007, which includes over 2 million Wikipedia articles. We follow the footsteps of Gabrilovich and Markovitch (2006) and discard overly-small articles having fewer than 100 words that are not stop-words, and articles having less than 5 incoming or outgoing links. We also exclude other types of articles and categories that are unlikely to be useful for describing semantics: for example, those referring to a specific date or an event occurring in a specific year (e.g., 1961 plays, 1820s in fashion) or categories which clearly do not express hypernymy relations (e.g. Writers by nationality and British musicians by instrument). We also disregard lists and stubs. At the end of this process our knowledge base contained 497,153 articles.

Our system was evaluated on the well-known SemEval-2007 coarse-grained all-words WSD task (Navigli, Litkowski, & Hargraves, 2007). We chose coarse-grained word sense disambiguation over fine-grained, since the latter often suffers from low inter-annotator agreement. For example, in the fine-grained Senseval-3 WSD task, inner-annotator agreement was only 72.5% among expert lexicographers (Snyder & Palmer, 2004). This is mainly because WordNet senses are full of distinctions which are difficult even for humans to judge. Coarse word senses allow for higher inter-annotator agreement of 86.4% (Snyder & Palmer, 2004; Chklovski & Mihalcea, 2002), and better reflect the average person’s perceptions of the different word senses. The senses were created semi-automatically using a clustering algorithm developed by the task administrators (Navigli, 2006), and then manually verified.
For an example of how the dictionary distinguishes different word senses, let us review the senses of the word “bank,” according to the WordNet dictionary.

1. a financial institution that accepts deposits and channels the money into lending activities;
2. sloping land (especially the slope beside a body of water);
3. a supply or stock held in reserve for future use (especially in emergencies);
4. a building in which the business of banking transacted;
5. an arrangement of similar objects in a row or in tiers;
6. a container (usually with a slot in the top) for keeping money at home;
7. a long ridge or pile;
8. the funds held by a gambling house or the dealer in some gambling games;
9. a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force;
10. a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning);

The dictionary distinguishes between ten different senses of the word, whereas most humans perceive the different meanings in a much coarser manner. Indeed, the clustering algorithm yields only three groups of meanings: financial (1,3,4,6,8), geographical (2,7,9,10), and geometrical (5). Obviously, these groups, while lacking some fine distinctions, are more compatible with the common intuition.

The test data set consists of 5,377 words of running text from five different articles: the first three were obtained from the WSJ corpus, the fourth was the Wikipedia entry for computer programming, the fifth was an excerpt from Amy Steedman’s Knights of the Art, biographies of Italian painters. In Table 3.1 we report the domain, number of running words, and number of annotated words for the five articles. Overall, 2,269 content words constituted the test data set, where 29.88% (678/2269) of the instances were monosemous according to the coarse sense inventory. The average polysemy of the test set with the coarse-grained sense inventory was 3.06 while the average polysemy with the WordNet inventory was 6.18. The inner-annotator agreement was 93.80%, a much higher number than of previous fine-grained tasks.

2. The Performance of the CBD Algorithm

First, we evaluated our CBD algorithm with four different semantic interpreters: ESA-based, CHESA-based and their anchored versions. We use here the nouns-only subset of

the test corpus, containing 1108 instances, since the Wikipedia articles are mainly focused on nouns. The context we used was the sentence in which the ambiguous word appears.

Traditionally, the performance of disambiguation systems is evaluated by the F1 measure. The evaluated algorithm either returns a sense, or returns “I don’t know.” The precision, recall and F1 are computed from these answers. In our system, the decision to reply “I don’t know” is determined by a threshold on the similarity scores. The optimal threshold for each algorithm was empirically estimated by maximizing the F1-measure on a development set of 1,000 randomly chosen noun instances from the SemCor corpus.

The results are presented in Table 3.2. Two baselines were calculated: a random baseline, in which senses are chosen at random, and the most frequent sense baseline (MFS), according to the frequencies in the SemCor corpus (Miller, Leacock, Tengi, & Bunker, 1993). The results strongly imply that the anchored versions of ESA and CHESA yield a consistent improvement against the unanchored algorithms, with +2.87% and +2.22% F1 respectively. This outcome verifies that the anchored representation is more suited for WSD tasks, as the texts of each of the senses are compared in the context of the ambiguous word at hand, rather than a general one.

Additionally, we can see that the CHESA-based algorithms perform better than their ESA counterparts, with +2.25% and +1.6% F1 for the unanchored and anchored version respectively. The superiority of CHESA over ESA in this task fits our prior assumptions that weighted hierarchical representation, which allows varying abstraction levels, is more suited to the human perception of different meanings. Moreover, the results clearly indicate that the anchored-CHESA algorithm outperforms the MFS baseline, which is notable for being a difficult competitor for unsupervised and knowledge-based systems.

### 3. Comparison with State-of-the-Art Methods

We compared our anchored-CHESA algorithm with several supervised and unsupervised state-of-the-art competitors. The supervised group consisted of NUS-PT (Chan et al., 2007), NUS-ML (Cai et al., 2007), LCC-WSD (Novischi, Srikanta, & Bennett, 2007), GPLSI (Izquierdo, Suárez, & Rigau, 2007), UPV-WSD (Buscaldi & Rosso, 2007). The unsupervised group consisted of TreeMatch (Chen, Ding, Bowes, & Brown, 2009), SUSSX-FR (Koeling & McCarthy, 2007), UOR-SSI (Navigli & Velardi, 2005), Degree (Ponzetto & Navigli, 2010)
Table 3.2: Performance on Semeval-2007 coarse grained all-words WSD (nouns only subset).

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS BL</td>
<td>77.40</td>
<td>77.40</td>
<td>77.40</td>
</tr>
<tr>
<td>Random BL</td>
<td>63.50</td>
<td>63.50</td>
<td>63.50</td>
</tr>
<tr>
<td>ESA</td>
<td>90.32</td>
<td>63.99</td>
<td>74.91</td>
</tr>
<tr>
<td>Anchored-ESA</td>
<td>89.06</td>
<td>69.04</td>
<td>77.78</td>
</tr>
<tr>
<td>CHESA</td>
<td>87.16</td>
<td>69.22</td>
<td>77.16</td>
</tr>
<tr>
<td>Anchored-CHESA</td>
<td>91.92</td>
<td>69.86</td>
<td>79.38</td>
</tr>
</tbody>
</table>

Table 3.3: System scores for nouns (1,108 instances) and all words (2,269 instances) with MFS adopted as a back-off strategy when no sense assignment is attempted.

<table>
<thead>
<tr>
<th>System</th>
<th>Nouns Only P/R/F1</th>
<th>All Words P/R/F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS BL</td>
<td>77.44</td>
<td>78.89</td>
</tr>
<tr>
<td>NUS-PT</td>
<td>82.31</td>
<td>82.50</td>
</tr>
<tr>
<td>NUS-ML</td>
<td>81.41</td>
<td>81.58</td>
</tr>
<tr>
<td>LCC-WSD</td>
<td>80.69</td>
<td>81.45</td>
</tr>
<tr>
<td>GPLSI</td>
<td>80.05</td>
<td>79.55</td>
</tr>
<tr>
<td>UPV-WSD</td>
<td>79.33</td>
<td>78.63</td>
</tr>
<tr>
<td>SUSSX-FR</td>
<td>81.10</td>
<td>77.00</td>
</tr>
<tr>
<td>TreeMatch</td>
<td>N/A</td>
<td>73.60</td>
</tr>
<tr>
<td>UOR-SSI</td>
<td>84.12</td>
<td>83.21</td>
</tr>
<tr>
<td>ExtLesk</td>
<td>81.00</td>
<td>79.10</td>
</tr>
<tr>
<td>Degree</td>
<td>85.50</td>
<td>81.70</td>
</tr>
<tr>
<td>Anchored-CHESA</td>
<td>85.02</td>
<td>82.68</td>
</tr>
</tbody>
</table>

and ExtLesk (Ponzetto & Navigli, 2010). Three of the above competitors, NUS-PT, NUS-ML and LCC-WSD, were the top systems in the Semeval-2007 coarse-grained all-words Task. One competitor–SUSSX-FR–was the best unsupervised system that participated in that task and two, more recent unsupervised systems, Degree and ExtLesk, achieved the best performance in the literature. For a further discussion of these methods see Section 4.

Since most WSD methods, especially knowledge-based ones, resort to the MFS strategy when the confidence regarding the correct sense is low, we added such a back-off strategy to our anchored-CHESA algorithm. We evaluated our system against the aforementioned methods both on the complete test set and also on a nouns-only subset. The results are detailed in Table 3.3.

The results indicate that on the nouns-only subset, our system’s performance is comparable with state-of-the-art unsupervised systems, namely Degree and UOR-SSI, and is
much better than the best supervised and unsupervised systems, NUS-PT and SUSSX-FR respectively, which participated in SemEval-2007 (+2.71% and +3.92% F1 respectively). On the entire dataset, it is proven to be competitive with state-of-the-art supervised and unsupervised systems 3.

4. Domain Specific Word Sense Disambiguation

As described earlier, our system employs vast amounts of knowledge. In particular, it utilizes domain-specific information, such as named entities and domain-specific terms, which makes it naturally suitable for domain-specific WSD. We used the evaluation dataset published by Koeling, McCarthy, and Carroll (2005). The dataset consists of examples retrieved from the Sports and Finance sections of the Reuters corpus. 41 words related to the Sports and Finance domains were selected, with an average polysemy of 6.7 senses, ranging from 2 to 13 senses. Around 100 examples for each word were annotated by three reviewers with fine-grained senses from WordNet, yielding an inter-tagger agreement of 65%. The “correct” sense is the one chosen by the majority of taggers. Overall, each domain consists of nearly 3500 examples.

For evaluation, we used the best configuration of our system found in the general settings, namely Anchored-CHESA. Since the distributions of the senses of words are highly skewed in each domain, the thresholds that were previously used to decide when to resort to the MFS back-off strategy are no longer applicable. As no training data is available, we tuned these hyper-parameters using leave-one-out cross validation scheme. The results in comparison with the latest results by Agirre et al. (2009)4 and Ponzetto andNavigli (2010) are detailed in Table 3.4.

The table includes $k$-NN, a fully supervised system, which employs SemCor to train a k-nearest-neighbors classifier for each word in the dictionary. Additionally, it includes two recent unsupervised systems introduced by Agirre et al. (2009), namely Static PageRank and Personalized PageRank. These systems use WordNet as the Lexical Knowledge Base (LKB), and employ graph-based methods, such as PageRank (Brin & Page, 1998) and Personalized PageRank to perform WSD. Finally, we compare or system’s performance to recent state-of-the-art unsupervised systems, namely ExtLesk and Degree using WordNet++, introduced by Ponzetto andNavigli (2010).

The results indicate that in domain-specific settings, our system’s performance is superior to state-of-the-art systems, both supervised and unsupervised. They also strongly imply that our system is competitive to recent methods in fine-grained settings. It outperform by a large margin the best supervised system, K-Nearest Neighbors (k-NN) trained on SemCor, which have been used extensively in public evaluation exercises, and have succeeded in gaining high ranks in both lexical-sample and all-words tasks (Snyder & Palmer,

---

3. The differences between the results in bold in each column of the table are not statistically significant at $p < 0.05$

4. We compare only with token-based WSD systems, i.e. systems that disambiguate each instance of a target word separately.
<table>
<thead>
<tr>
<th>System</th>
<th>Sports P/R/F1</th>
<th>Finance P/R/F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS BL</td>
<td>19.6</td>
<td>37.1</td>
</tr>
<tr>
<td>Random BL</td>
<td>19.5</td>
<td>19.6</td>
</tr>
<tr>
<td>k-NN</td>
<td>30.3</td>
<td>43.4</td>
</tr>
<tr>
<td>Static PageRank</td>
<td>20.1</td>
<td>39.6</td>
</tr>
<tr>
<td>Personalized PageRank</td>
<td>35.6</td>
<td>46.9</td>
</tr>
<tr>
<td>ExtLesk</td>
<td>40.1</td>
<td>45.6</td>
</tr>
<tr>
<td>Degree</td>
<td>42.0</td>
<td>47.8</td>
</tr>
<tr>
<td>Anchored-CHESA</td>
<td><strong>46.8</strong></td>
<td><strong>50.0</strong></td>
</tr>
</tbody>
</table>

Table 3.4: System scores on the Sports and Finance sections of the dataset from Koeling et al. (2005)

2004; S. Pradhan & M.Palmer, 2007). This outcome supports the arguments presented earlier regarding the lack of robustness of supervised systems across different domains.

Additionally, Our systems achieves much better results than the best unsupervised system introduced by Agirre et al. (2009), namely Personalized PageRank, on both Sports and Finance corpora (+11.2% and +3.1% F1 respectively). We can see that the margin is much higher in the Sports domain than in the Finance one. These findings can be explained by the frequent usage of named entities, such as teams names and players’ names, in the Sports domain. These entities, while having vast coverage in Wikipedia, lack any reference in common dictionaries, thus imper the performance of graph-based methods that use WorNet as a knowledge base, such as Personalized PageRank.

Finally, we can note that our system achieves better results than recent state-of-the-art unsupervised systems, ExtLesk and Degree using WordNet++, introduced by Ponzetto and Navigli (2010), on both Sports and Finance corpora (+4.8% and +2.2% F1 respectively). While these systems performance is competitive to ours in general settings, our system outperforms them in domain-specific settings. A justification of this outcome can be associated with the different language model of domain-specific corpora, where usage of named entities and domain-specific terms is much more frequent. This model poses difficulty for methods which rely on a mapping between WordNet senses and Wikipedia pages, such as ExtLesk and Degree. As this terms lack full coverage in WordNet, the mapping procedure will not always succeed in associating them with information from Wikipedia. Whereas our algorithm, which can be applied to any word in the knowledge source, does not suffer from that limitation.

Let us review some examples that illustrate our claims. Consider the sentence Wide receiver Alvin Harper signed with the Washington Redskins on Wednesday from the Sports domain, with the word receiver to disambiguate. According to WordNet, the word receiver has six meanings that span across various domains such as technology, law, and sports. Here the correct meaning is the one associated with football. Since the sentence is mainly
composed by named entities, gathering sufficient information regarding them is crucial to this task. As the terms Alvin, Harper and Redskins, which relate to a football player and a football team respectively, are not located nor referred to in WordNet to begin with, the mapping procedure is unable to enrich the context with the relevant information from Wikipedia. On the contrary, our algorithm is able to choose the correct sense through representation of named entities, as well as plain words, in a concepts space composed of all of Wikipedia’s pages. It that space, these entities are highly correlated with sports-related concepts and categories, enabling our system to pick the correct sense.

A similar example can be found in the Finance corpus. Consider the sentence The department is currently reviewing U.S. imports of Hyundai and LG semicon chips in the period from May 1995 through April 1996, with the word chip to disambiguate. The word chip has 9 meanings according to WordNet, where here the correct sense is the one referring to electronic equipment. Again, there is no other indication for picking that sense but the presence of the named entities Hyundai and LG. As these entities are not referred to in WordNet, the relevant information would not be retrieved from Wikipedia by enrichment procedures, a fact that can imper performance. While in the concept space we use, these entities are highly associated with industry and technology concepts, making the correct sense an easy pick. We provide further detailed observation of such examples later in this section.

5. Learning to Improve CBD

As explained earlier, we employ MFS as a back-off strategy in a similar manner to recent knowledge-based methods. The back-off strategy is used mostly in cases where the difference between the similarity scores of the top senses is below an empirically defined threshold, meaning none of them is significantly more related to the context of the ambiguous word. During our research, we experimented with more complex, learning-based criteria to decide whether to use our similarity-based method or the MFS method.

We used two types of features. One type is based on the dictionary (Wordnet) while the other is built on the semantic representation. Some of the dictionary-based features utilize direct attributes such as the length of the gloss and the number of senses while others build on Wordnet’s hierarchical structure, which characterize each sense’s direct and undirect hypernyms. These include the ratio between abstract hyponyms and concrete ones, as well as some specific entities.

Examples for features that are based on the semantic representation are the number of concepts, in the representation, as well as the mean and standard deviation of their association scores. The purpose of adding such features was to identify noisy representations, where the weights are distributed uniformly among the different concepts rather than assigned to specific ones.

These two groups of features, as well as a mixture between them, were used to train a classifier to identify the cases where the MFS method should be preferred over the CBD.
method. Several learning algorithms were applied, including SVM, decision trees and KNN. Unfortunately, empirical results showed no significant improvement over the current results.

6. Qualitative Analysis of CBD

To better understand the specific properties of CBD and illustrate its strengths and weaknesses in assessing semantic relatedness between the context and each of the senses, we present a qualitative analysis of our method using real examples.

6.1 Benefits of World Knowledge

CBD benefits from world knowledge especially in cases where no exact words are shared between the context and the right sense of the word to disambiguate. Consider the sentence: However most PC desktop applications such as word processors or image manipulation programs are written in more runtime and memory efficient languages like C, C++ and Delphi, with the word *programs* to disambiguate. According to the WordNet dictionary, there are several meanings to the lemma *Program*:

1. a system of projects or services intended to meet a public need;
2. a series of steps to be carried out or goals to be accomplished;
3. (computer science) a sequence of instructions that a computer can interpret and execute;
4. an integrated course of academic studies;
5. a radio or television show;
6. a performance (or series of performances) at a public presentation;
7. a document stating the aims and principles of a political party;
8. an announcement of the events that will occur as part of a theatrical or sporting event.

The right sense for this example is obviously the third one, which relates to a computer program, yet there is no word overlap between the context and the right sense. In order to pick the right sense, one must possess prior knowledge about computers, realizing *runtime* and *memory* are computer related terms, and that *C, C++* and *Delphi* are indeed computer programming languages. This is a classic example where CBD succeeds while simpler algorithms, with no world knowledge, fail. Here are the top ten concepts that are shared between the context representation and the right sense representation, according to the algorithm: 1. COMPUTER LANGUAGE 2. APPLICATION SOFTWARE 3. COMPUTER 4. COMPUTER PROGRAM 5. COMPUTER SCIENCE 6. COMPUTER SOFTWARE 7. MEMORY MANAGEMENT 8. VIRTUAL MEMORY 9. APPLICATION PROGRAMMING INTERFACE 10. MEMORY ADDRESS. As can be easily seen, CBD is able to capture the main ideas of the text and identify connections between words that stem from their semantics rather than syntax.
However, most PC desktop applications such as word processors or image manipulation are written in more runtime and memory efficient languages like C, C++, and Delphi.

This example also illustrates the benefits of anchoring. Figure 3.1 provides a partial view of the ESA representation of the context. The vector includes a mixture of meanings of the words in the text. We can see that the word language triggers both highly relevant concepts, such as Computer language, and non-relevant ones, such as Languages of India. Similarly, the word Delphi triggers both the concept Delphi, which refers to a town in Greece, and the concept Object Pascal, which refers to a programming language that is also known as Delphi. However, the anchored-ESA representation of the context is much more definite, as illustrated in Figure 3.1. Anchoring the original representation of the word program omits the non-relevant terms, Languages of India and Delphi, while retaining the relevant ones, Computer language, Application software, and Object Pascal.

A similar effect is achieved with each of the senses. Let us examine the ESA representation of the third sense, which is the correct one. Figure 3.2 provides a partial view of its ESA vector, which includes a mixture of meanings as well. We can see that the words computer and instructions trigger highly relevant concepts, such as Computer language, Application software and Instruction set. However, the word execute triggers non-relevant concepts, such as Executive producer and Executive car. Again, the anchored representation is much more precise, where the non-relevant concepts were omitted, as illustrated in Figure 3.2. As a result, the third sense’s representation is very similar to the context’s representation, making it an easy pick.

Another example for the necessity of world knowledge in the WSD process can be found in the d003 corpus, with regard to the sentence: When the balloon is cruising along at a steady altitude there is little sense of motion. The word to disambiguate is balloon. According to WordNet dictionary, there are two meanings for the lemma balloon:

1. small thin inflatable rubber bag with narrow neck; an artifact designed to be played with;
2. large tough nonrigid bag filled with gas or heated air; aircraft supported by its own buoyancy.

The second sense, which refers to a hot-air balloon, is the correct one. However, it shares no common words with the sentence in which the word appears. In order to pick the right sense, the algorithm must identify the words *cruising*, *altitude* and *motion* as highly related to the words *aircraft* and *heated air*, as they are all connected to the field of aviation. Indeed, according to the CBD algorithm, both the context representation and the representation of the second sense share aviation related categories such as: AVIATION, AERONAUTICS and TRANSPORTATION, making the second sense a predominant pick.

This example also highlights the benefits of anchoring in hierarchical representations. Figure 3.3 provides a partial view of the CHESA representation of the context. Its hierarchy includes several branches, as the meanings of the words span several fields. For instance, the word *sense* triggers the non-relevant category PHILOSOPHY. The word *cruising* triggers both relevant categories, such as AVIATION, and non-relevant ones, such as WATER TRANSPORT. However, we get a much more definite representation by anchoring the representation to the word *balloon*, as illustrated in Figure 3.3. The categories PHILOSOPHY, ASTRONOMY, WATER TRANSPORT and their descendants are removed while relevant ones, such as AVIATION and PHYSICS, are retained.

We get a similar effect in each of the senses. Let us review the CHESA representation of the second sense, which is the correct one. Figure 3.4 provides a partial view of its CHESA representation, which includes a mixture of meanings as well. We can see that the representation includes categories, such as AVIATION and LEVITATION, that describes the dynamics of a hot-air balloon. We can also find categories, such as TRANSPORTATION and Military technology, that imply its usages. However, the representation also includes non-relevant categories, such as CHEMICAL ENGINEERING and ENERGY, which are triggered by the words heat and gas. Again, we can see that the anchored representation, as illustrated in Figure 3.4, is more compact and precise, free of unrelated components that can otherwise
imper the process of WSD. Its includes all of the relevant categories, while omitting the non-relevant ones. As a result, the similarity between this representation and the one of the context is very high, making it an apparent choice.

6.2 Pitfalls of Knowledge-Based Algorithms

Knowledge is not helpful in all cases. Consider the sentence: *Most balloonists seldom go higher than 2,000 feet and most average a leisurely 5-10 miles an hour*, with the word *mile* to disambiguate, as taken from the d003 corpus presented earlier. According to WordNet dictionary, there are several meanings to the lemma *mile*:

1. a unit of length equal to 1760 yards;
2. a unit of length used in navigation; equivalent to the distance spanned by one minute of arc in latitude; 1,852 meters;
3. a large distance;
large tough non-rigid bag filled with gas or heated air; aircraft supported by its own buoyancy

Figure 3.4: A partial view of the CHESA (top) and the Anchored-CHESA (bottom) representations for the second sense of the word balloon, which refers to a hot-air balloon.
4. a former British unit of length once used in navigation; equivalent to 1828.8 meters (6000 feet);
5. a British unit of length equivalent to 1,853.18 meters (6,082 feet);
6. an ancient Roman unit of length equivalent to 1620 yards;
7. a Swedish unit of length equivalent to 10 km;
8. a footrace extending one mile.

The right sense in this example is the first, most common one. CBD has much difficulty disambiguating this example. As most of the senses contain the words unit, length, feet and yards, it is extremely hard to distinguish between them. Since all of the senses deal with distance related terms, CBD assigns them high yet similar scores, but has difficulty picking the right one. This phenomenon is not confined to CBD. It is typical to most knowledge-based algorithms, which rely solely on similarity between the content of sentence and the content of the right sense.

A more extreme example can be found earlier in the same corpus, in the sentence: A whole morning of ballooning and I had been off the ground barely minutes, with the word morning to disambiguate. Referring to the WordNet dictionary again yields several meanings of that word:

1. the time period between dawn and noon;
2. a conventional expression of greeting or farewell;
3. the first light of day;
4. the earliest period.

In contrast to the former example, where most senses shared similar content with the context, here there is no conceptual overlapping at all between the sentence and each of the word’s meanings. As a result, all of the senses get very low scores by the CBD algorithm, and the task of choosing among them becomes noisy. Obviously, the right sense is the first one, while the third one is also acceptable according to the coarse-grained clustering algorithm. Notice the first sense is also the one that is the most commonly used (82%). Again, this drawback is shared between all knowledge-based and similarity-based algorithms, including CBD, and highlights the necessity of using MFS as a back-off strategy.
Chapter 4

Related Work

Many approaches have been introduced in the past for WSD. In this section we discuss the main ones and present some known algorithms that follow each approach. We focus on systems that participated in the Semeval-2007 coarse-grained all-words task, as well as recent state-of-the-art methods.

WSD approaches can be supervised, semi-supervised, or unsupervised. A supervised method requires a sense-annotated corpus from which syntactic and semantic features are extracted to create a classifier using machine learning techniques (Ng & Lee, 1996; Pedersen, 2001; Chklovskii & Mihalcea, 2002). The features involved are mainly collocation and co-occurrence features. Then, the classifier is used to assign senses to unseen examples. Consequently, such methods are deeply influenced by the frequency of each word in the training corpora. In order to perform well, supervised methods must acquire sufficient contextual information for every sense of a large number of words. Since training corpora are manually annotated, this information is very expensive and requires huge resources. To overcome this limitation, semi-supervised systems employ a small annotated corpus as seed data in a bootstrapping process (Hearst, 1991; Yarowsky, 1995; Ng, Wang, & Chan, 2003). However, the recent development of various large annotated corpora has placed the supervised systems at the top of the Semeval-2007 coarse-grained all-words task.

We will examine the top supervised systems from Semeval-2007 in greater detail. The top system, NUS-PT (Chan et al., 2007), employs a parallel-text approach which involved training on two large annotated corpora, SemCor and the Defense Science Organization (DSO), using a support vector learning algorithm. The NUS-ML system (Cai et al., 2007) employs a hierarchical Bayesian LDA model to cluster bag-of-words features. These features are learned from a locally-created collection of collocation features, in addition to part-of-speech tags and syntactic relations. The LCC-WSD system (Novischi et al., 2007) employs a maximum entropy classifier and support vector machines. To create its features, it uses a variety of corpora: SemCor, Senseval 2 and 3, and Open Mind Word Expert. Additionally, it makes use of WordNet glosses, extended WordNet, syntactic information, information on
compound concepts, part-of-speech tagging, and named entity recognition. These systems also utilize the MFS back-off strategy in cases where model confidence is low.

In contrary to the supervised approach, the unsupervised approach does not rely on a sense-annotated corpora. It hypothesizes that the correct sense is the one that is most related to the context of the ambiguous word. Most unsupervised methods are dictionary-based, yet some are dictionary-free. For instance, the HyperLex algorithm (Véronis, 2004) determines word uses in a text without recourse to a dictionary, by detecting “hubs” and high-density components in co-occurrence graphs. Such methods pose difficulty for comparison due to the different definitions of the senses.

Dictionary-based methods rely on lexical knowledge bases such as dictionaries and thesauri. For example, the Lesk (1986) algorithm assigns a word with the sense whose dictionary definition has the highest word overlap with the word’s context. Some methods also utilize the dictionary’s intra-linked structure to extend the glosses of the senses under consideration to include the glosses of other senses to which they are related according to the given hierarchy (Banerjee & Pedersen, 2003; Navigli & Velardi, 2005; Agirre et al., 2009; Agirre & Soroa, 2009; Ponzetto & Navigli, 2010).

In the last decade, many approaches for enrichment of existing resources have been developed. These resources, such as WordNet, usually hold limited information regarding each of the senses and are thus insufficient for high performance disambiguation. In order to deal with this problem, Hearst (1992), Cimiano et al. (2004) and (Girju et al., 2006) developed methods using lexico-syntactic patterns. Harabagiu et al. (1999) applied heuristic methods based on lexical and semantic regularities, while Pennacchiotti and Pantel (2006) and (Snow et al., 2006) implemented taxonomy-based ones. Other ways to enhance dictionaries include the extraction of semantic preferences from sense-annotated (Agirre & Martinez, 2001) and raw corpora (McCarthy & Carroll, 2003), in addition to the disambiguation of dictionary glosses using cyclic graph patterns (Navigli, 2009). Some approaches are based on collocations, either obtained from specialized learner’s dictionaries (Navigli & Velardi, 2005) or extracted via statistical methods (Cuadros & Rigau, 2008). Whereas some of these methods represent state-of-the-art approaches for enriching lexical resources, none can work on a very large scale and incorporate huge amount of encyclopedic knowledge into the sense disambiguation process.

In recent years, unsupervised methods have succeeded in employing vast knowledge sources with domain-specific information (Navigli & Velardi, 2005; Koeling & McCarthy, 2007; Chen et al., 2009). Knowledge sources provide data essential for connecting senses with words. They can vary from collections of raw texts (Navigli & Velardi, 2005; Koeling & McCarthy, 2007) to libraries of structured data such as Wikipedia (Milne, 2007; Turdakov & Velikhov, 2008; Ponzetto & Navigli, 2010).

Raw texts are mainly used in a graph-based approach, as demonstrated by SUSSX-FR (Koeling & McCarthy, 2007), the best unsupervised system in the Semeval-2007 coarse-grained all-words task. It uses a method described in McCarthy, Koeling, Weeds, and Carroll (2004) for finding predominant senses from raw text. The method employs a the-
saurosaurus obtained from the text using the distributional similarity metric described by Lin (1998), then ranks the senses of a word by means of a WordNet similarity score (Pedersen, Patwardhan, & Michelizzi, 2004).

More recent graph-based methods were introduced by Agirre et al. (2009). These methods treat WordNet as a Lexical Knowledge Base (LKB) and exploit its information and intra-liked structure to construct a graph of entities. Then, some version of PageRank of the graph is computed by concentrating the initial probability mass uniformly over the context nodes, then injecting mass into the concepts they are associated with, which thus become relevant nodes, and spread their mass over the LKB graph. Finally, the sense with the highest rank is returned. Agirre et al. (2009) findings also support the argument that knowledge-based systems exhibit a more robust performance than their supervised alternatives when evaluated across different domains, that was later encouraged by Ponzetto andNavigli (2010).

Another example for a knowledge-based system is Structural Semantic Interconnections (SSI) (Navigli & Velardi, 2005), introduced by one of the task organizers. SSI creates structural specifications of the possible senses for each word in a context and selects the best hypothesis according to a grammar describing relations between sense specifications. Sense specifications are created from several available lexical resources that are integrated in part manually, in part with the help of automatic procedures.

Over the past several years the importance of Wikipedia as an external source of knowledge for WSD tasks has been increasingly recognized, and many techniques for incorporating Wikipedia in the task of WSD have been developed. Some methods manually map Wikipedia pages to WordNet senses (Mihalcea, 2007), while others do so automatically (Ponzetto &Navigli, 2010). Degree and ExtLesk (Ponzetto &Navigli, 2010) employ WordNet++ and achieve the best performance in the literature. WordNet++ is an extension of WordNet, automatically enriched with encyclopedic relational knowledge from Wikipedia. The enrichment procedure relies on a mapping between Wikipedia pages and WordNet senses. Additionally, semantic relations from Wikipedia are transferred to WordNet using Wikipedia's intra-linked structure.

The use of Wikipedia for disambiguation tasks is not limited to unsupervised systems. A machine learning method, as described in Mihalcea (2007), uses Wikipedia link structure for generation of a sense-tagged corpus, which is employed for training a classifier. As many concepts mentioned in Wikipedia are explicitly linked to their corresponding article, these links can be regarded as sense annotations for concepts—a valuable property for ambiguous entities. There are several limitations to this method. First, words must appear as ambiguous in Wikipedia, meaning any word that does not have its own article in Wikipedia cannot be disambiguated. Second, Wikipedia annotations must be manually mapped to senses in WordNet for evaluation, resulting in a small amount of test data, limited to nouns.

Wikipedia is also employed by measures to compute semantic distance between words. These include path-based measures (Strube & Ponzetto, 2006), textual overlap (Mihalcea & Csomai, 2007; Ponzetto &Navigli, 2010), and data-driven algorithms that are based on
probabilities collected from large amounts of annotated data (Koeling & McCarthy, 2007; Mihalcea & Csomai, 2007; Chen et al., 2009). WikiRelate!, introduced by Strube and Ponzetto (2006), searches for a pair of Wikipedia articles that respectively contain the pair of given words in their titles. Semantic relatedness is then computed using various distance measures between the two articles. These measures rely either on the texts of the pages, or on path distances within the category hierarchy of Wikipedia. As a result, WikiRelate! can only process words that actually occur in titles of Wikipedia articles. Moreover, the WikiRelate! representation of a word is limited to the text of the article associated with it, or to the nodes in the category hierarchy.

A similar method is employed by Wikify!, introduced by Mihalcea and Csomai (2007). Wikify! identifies the important concepts in the text and automatically links them to the corresponding Wikipedia pages. Link disambiguation is resolved by two different algorithms. The first one is based on a measure of contextual overlap approximated with the corresponding Wikipedia pages. The second approach is a data-driven method based on Wikipedia links; it integrates both local and topical features into a machine learning classifier.

Milne (2007), Turdakov and Velikhov (2008) implemented a richer semantic representation for a word or text, utilizing Wikipedia’s interlink-structure. Given a word, a corresponding concept vector is built, starting with those articles whose titles match the given word, and adding the articles in the neighborhood of those articles. Turdakov and Velikhov (2008) also assigned different weights to different links corresponding to their role in Wikipedia, such as See also links, Category links, inverted links, and so on. Upon disambiguation, semantic distance between the context and a candidate term meaning is calculated as the number of common articles in the neighborhoods of the context and the term in Wikipedia. These methods, while efficient and easy to calculate, are restricted to terms that match a Wikipedia title. They also pose difficulties for evaluation using data from other sources.

Other ways to calculate semantic distance between the context of a word and each of its senses include textual overlap and graph-based algorithms. For instance, Ponzetto andNavigli (2010) apply one of two algorithms in order to pick the correct sense. The first algorithm is ExtLesk, a simplified version of the extended Lesk algorithm (Lesk, 1986) which, given a target word w, assigns to w the sense whose gloss has the highest overlap (i.e., most words in common) with the context of w (a sentence). The gloss of each sense s is extended to include all WordNet synsets which are directly connected to s, either by means of the semantic pointers found in WordNet or through the unlabeled links found in WordNet++. The second algorithm used is Degree, a graph-based approach that relies on the notion of vertex degree (Navigli & Lapata, 2010). Starting from each sense of the target word, it performs a depth-first search (DFS) of the WordNet++ graph and collects all the paths connecting s to senses of other words in context. A sentence graph is then produced and a maximum search depth established to limit its size. The sense of the target word with the highest vertex degree is selected.
While these knowledge-based methods, and WordNet++ based ExtLesk in particular, resemble our system in their motivation to augment WSD with vast encyclopedic knowledge, several differences exist. First, since our system analyzes the full contents of the articles rather than just their titles, links, and category labels, it can succeed in cases where these components alone lack sufficient semantic information. Another distinction lies in the methodology for assessing relatedness. For instance, WordNet++ based ExtLesk relies on bag-of-words (BOW) similarity when initially mapping senses into articles, and also measures relatedness in an unstructured manner by means of overlapping words. These words are often excessive, over-specific and noisy. Furthermore, humans do not judge text relatedness merely at the word level. Words trigger reasoning at the much deeper level of concept manipulation: concepts are the basic units of meaning that serve humans to organize and share their knowledge. Our CHESA-based approach allows us to measure relatedness through shared concepts, utilizing categories to represent semantics at varying abstraction levels and to avoid using unnecessary data.

Additionally, our method is suitable where named entities and domain-specific terms are involved, as common in domain-specific corpora. It treats named entities the same as plain words, thus utilizing the vast information that Wikipedia offers for named entities in the WSD process. As mentioned earlier, supervised systems that were trained on a general corpora exhibit a decline in performance when evaluated on a domain-specific corpus. The domain of a document has a strong influence on the sense distribution of words, but it is not feasible to produce large manually annotated corpora for every domain of interest. Additionally, domain-specific corpora can pose difficulty also for unsupervised methods based on mapping between dictionary senses and Wikipedia articles. Named entities, which are frequent in these corpora, do not appear initially in the dictionary, so the mapping algorithms do not take them into account. As a result, relevant knowledge regarding these entities cannot be reached during the WSD process, a fact that can impair performance.

Our representation of textual semantics using Wikipedia concepts follows the line of research used in other Wikipedia-based applications, including, among others, text categorization (Gabrilovich & Markovitch, 2006), computing semantic similarity of texts (Gabrilovich & Markovitch, 2007; Ponzetto & Strube, 2007a; Milne & Witten, 2008; Radinsky et al., 2011), co-reference resolution (Ponzetto & Strube, 2007b), multi-document summarization (Nastase & Strube, 2008), text generation (Sauper & Barzilay, 2009), and concept-based information retrieval (Egozi, Markovitch, & Gabrilovich, 2011).
Chapter 5

Conclusions

We presented a concept-based disambiguation framework (CBD) that employs large-scale algorithms for automatic representation of word senses using vast encyclopedic knowledge. This knowledge is successfully utilized without deep language understanding, specially crafted inference rules, or additional common-sense knowledge bases. Unlike other methods, CBD converts both the senses and the word’s context into a high dimensional space composed of natural concepts and categories which are grounded in human cognition. The use of concepts and categories is in keeping with the innate human ability to generalize, while capturing the main gists of each text. Thus, they can more suitably represent human perception of the different meanings. This is in contrast to the brittle process employed by many knowledge-based approaches, which use word overlap techniques to compute semantic relatedness. In the concept-based space, more comprehensive and sophisticated algorithms for assessing relatedness can be applied, yielding better performance.

We also introduced a novel Anchored Representation scheme that, given a text and an anchoring word, builds a semantic representation of the text. The semantics is generated so that it will be associated with the anchoring word. Meanings unrelated to the anchoring word are omitted, resulting in much more definite representation of a word or text. This scheme has proven to be more suitable to the task of WSD, where the ambiguous word plays a key role in anchoring the texts of its senses. As these texts are often ambiguous, we use the word in question to construct more specific text representations. As a result, the incidental similarities that stem from unrelated word meanings are ignored, similarities that would otherwise impair accuracy.

Our approach was proven to be highly beneficial both in a coarse-grained general setting and in a fine-grained domain-specific setting. It was shown to be superior to state-of-the-art supervised systems, and competitive with recent state-of-the-art unsupervised knowledge-based methods, when evaluated on a general dataset. Additionally, it was proven to be superior to both state-of-the-art supervised and unsupervised methods when evaluated on domain-specific corpora. We note that more complex algorithms for generating representations and assessing relatedness could yield even higher performance, and we intend to
research such algorithms in future work. Moreover, since our CBD approach does not rely on word overlap, it has a tremendous advantage in cases where the context of the ambiguous word and its senses do not share a common vocabulary. Its impact in a multilingual setting should thus be examined as well. Since Wikipedia is a multilingual semantic network, concept-based analysis can be applied to texts of different languages, allowing disambiguation of words using a dictionary of an entirely different language. Indeed, ESA has been used successfully for cross-lingual tasks (Hassan & Mihalcea, 2009; Potthast et al., 2008; Cimiano et al., 2009).

Many agree that complex disambiguation problems should be eventually solved using deep semantic analysis. We believe that the framework presented in this paper takes a step in this direction.
Bibliography


A new method for representing the meaning of a text is presented in this study, which is based on a single meaning representation of the text in accordance with the word of the text. Moreover, the text is represented in a specific context, in order to reflect the meaning of the text in a specific context. The method is tested on texts of various domains, showing that it is more accurate than existing methods.

The results show that the new method is more accurate than existing methods, and can be used as a basis for further research in the field of text analysis.
בעד לסק את הסמך על הפרטים של מילויים למחרת יידע, ואלאלגריתם של מחרת קורב מילוי המחרת בקצרה, ובכניסה למדיה הקורב בין סקסיים.

במרח הבר-רמי מי מעוניין, ובתהליך מתכון ומקף הוא.

האלגוריתם של נומחבר עלחפיפה של מילים לימדה והתאמה בין הטקסטים במרח מרחב-הממד של מקורות, בההליך מורכב ומגיף יותר. האלגוריתם של נומניח קיים של מרחב מקורות טבעיים, כאשר כל מושג מקושר לנושא כלשהו המוגדר, והמילון המכיל רשימה של פירושים עבור כל מילה. ה_votes לאלגוריתם מכיל את המילה שיש להגלה, והקשר שלה, ורשימת הפשלים.除此之外, האלגוריתם מסתמך על שלושה רכיבים סמיוטיים: estimator (משרך קירבה), semantic interpreter (מפרש סמיוטי), sense retriever (מאחזר פירושים). האלגוריתם תא למאחזר הפירושים כדיל קשר בין כל פירוש למשטח מסוים. 

לאחר מכן הוא מפעיל את המפרשת סמיוטית כדיל ייצוג את ההקשר של המילה והן את כל מקטעי הטקסט של הפירושים במרח מרחב-הממד של מקורות הטקסט של המילים. בנו האלגוריתםavid את אלגוריתם הסמיוטי בין כל פיזור פירושים בכל היצוג של הקירבה, חבר.

את הפירושים של כל אלה הקורב חבר.

אנו מתכינו לצרında את המונחים ואת השילוש הם לעת לפיתוח את האלאלגריתם של מחרת, לאכול אחד ממקורות קירבה. קיימים, ופרטים שבהם הממודים נמצאים בувеличен של מילויים. 

אנו מתכינו לצרída את המונחים ואת השילוש הם לעת לפיתוח את האלאלגריתם של מחרת, לאכול אחד ממקורות קירבה. קיימים, ופרטים שבהם הממודים נמצאים בувеличен של מילויים. 

במחקר זה, אנו מצפים בידיעות עדכניות ליצוג סמיוטית של הטקסט, כדי לקבל עדכניות בdda המפרשת סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).

שלט זה ממעודכן באמצעות חיבורים של כל מקורות, והם מצפים בידיעות עדכניות ליצוג סמיוטית של הטקסט, כדי לקבל עדכניות בdda המפרשת סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).

שלט זה ממעודכן באמצעות חיבורים של כל מקורות, והם מצפים בידיעות עדכניות ליצוג סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).

שלט זה ממעודכן באמצעות חיבורים של כל מקורות, והם מצפים בידיעות עדכניות ליצוג סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).

שלט זה ממעודכן באמצעות חיבורים של כל מקורות, והם מצפים בידיעות עדכניות ליצוג סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).

שלט זה ממעודכן באמצעות חיבורים של כל מקורות, והם מצפים בידיעות עדכניות ליצוג סמיוטית Компактный эксплицитный универсальный семантический анализ (Compact Hierarchical Explicit Semantic Analysis, CHESA).
A-Israeli PhD in Computer Science (MSC-2012-07 - 2012)

The study of language involves the study of words and word meanings, which are complex and multifaceted. In this context, dictionary-based methods (Besad Disam) provide a framework for understanding the meaning of words and their relationships within a semantic space. These methods aim to capture the subtle differences in meaning between words and their synonyms.

One of the challenges in word-related tasks is the need for a comprehensive dictionary that can handle the vast number of words and their meanings. However, even with the vast amount of information available online, there are limitations in the ability of machines to understand the meaning of words beyond their dictionary definitions. This is especially true in the context of machine translation, where the meaning of words can be deeply connected to the context in which they are used.

In this study, a new approach is proposed to improve the understanding of language for machine translation. The approach is based on the idea of concept-based disambiguation, which involves the use of a semantic space to capture the meaning of words and their relationships within that space.

The study proposes a new algorithm for concept-based disambiguation that is based on the idea of selecting the closest meaning to the context in which the word appears. This approach is then tested on a number of datasets, including those related to different domains.

The results show that the proposed approach is effective in improving the accuracy of machine translation, especially in cases where a simple dictionary-based approach would not be sufficient. The approach also allows for the automatic generation of semantic relationships between words, which can be used to improve the understanding of language in a variety of contexts.

In conclusion, the study demonstrates the potential of concept-based disambiguation for improving the understanding of language in machine translation. The approach provides a new way to capture the meaning of words and their relationships within a semantic space, which can be used to improve the accuracy of machine translation and other natural language processing tasks.
תקציר

השפה האנושית היא בר המשמשות המ眸ה. בכל שפה בעלת מילולי בולטים ניתן להפריש מספר פירושים שונים. על פי הרשון, בר את המילים ניתן להפריש בפן המחושל "כ"זרצ"א" דיון. הצהרה, בברית הפוסטוניק של מילה בר-משמעות התכשיט

כ"כָּלָו" כ"כָּלָו" "כ"כָּל" כ"כָּל" לעברון בו-ידאות, בתוכנית הפוסטוניק של מילה בר-משמעות התכשיט

כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּל" כ"כָּล
המחקר עשה בבניהית פרוフィיל מרקובני בפקולים למיפוי המ昉ים.

אנימודה הוכנה את התוכנית לת gratuita, פרוфиיל מרקובני, עבורי הדריכה בתפקודה למיפוי. אני מודה לaszal על סיפתי לبعثו של הדריך מרקובני, על שמכזה את קינון וברזים עם דינו בברזים עם הרובים ולסיפתי תובלה. אני מודה שלשודו של פילוסופיה בוביל למקהל וברזים מרתון תלמידים, קינון דריזמל, עמר ול, על השבטים, ארז כרפו, גיא טלדני, אסף נגר, יקיי מירק, עמית במעבידת תלמידה והששיכר, עмор דינו בברז שנספה פרי. לספר, האהבה לתורה לפליים
אליביט, על האהבה וה씨ית לכל אוצר תפוקת למיפוי.

אני מודה לตนינו על התמיכת המוספת והנימה בבריתلعبית.
גישה מבוססתמושגים להתרת רב-משמעות
של מילים

היברו על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
منهجי למדעי במדעי המחשב

אריאל רביב

הוגש לתכנית התענוגות – מרכז טכנולוגי לישראל
שבט תשע”ב

מרץ 2012
גישה מבוססת מושגים להתרב-משמעות של מילים

אריאל רבינב