Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Concept-Based Approach to Word-Sense Disambiguation

Ariel Raviv and **Shaul Markovitch**
Computer Science Department
Technion—Israel Institute of Technology
Haifa 32000, Israel
{*arielr, shaulm*}*@cs.technion.ac.il*

Abstract

The task of automatically determining the correct sense of a polysemous word has remained a challenge to this day. 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 evaluate our framework and show that the anchored representation is more suitable to the task of word-sense disambiguation (WSD). Additionally, we show that our system is superior to state-of-the-art methods when evaluated on domain-specific corpora, and competitive with recent methods when evaluated on a general corpus.

1 Introduction

Since 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. It is crucial in many natural language processing (NLP) applications such as speech recognition, information retrieval, machine translation and computational advertising. Word-sense disambiguation (WSD) methods can be classified into two types: knowledge-based and machine learning. The machine learning approach usually includes building a classifier with collocation and co-occurrence features and using it to assign senses to unseen examples (Chklovski and Mihalcea 2002; Ng, Wang, and Chan 2003). To perform well, it needs large training annotated sets that are extremely expensive to create (Edmonds 2000). Furthermore, the training set 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 (Banerjee and Pedersen 2003; Navigli and Velardi 2005). This approach usually picks the sense whose definition is most similar to the context of the ambiguous word, by means of textual overlap or using graph-based measures (Agirre, De Lacalle, and Soroa 2009). Consequently, most dictionary-based methods are sensitive to the exact wording of the definitions, as they have not realized the potential of combining the limited information in such definitions with the abundant information extractable from text corpora (Cuadros and Rigau 2006).

In the last decade, many methods for enrichment of existing resources have been developed in order to deal with this problem (Girju, Badulescu, and Moldovan 2006; Pennacchiotti and Pantel 2006) yet none of them made use of world knowledge, which is necessary for WSD. Recently, Wikipedia has become an external source of knowledge for WSD tasks, as it supplies vast amounts of common-sense world knowledge. Wikipedia is used both by machine learning algorithms (Bunescu and Pasca 2006; Cucerzan 2007) and similarity models (Strube and Ponzetto 2006; Milne 2007; Turdakov and Velikhov 2008). Additionally, recent enrichment methods utilize Wikipedia and map senses to corresponding articles (Mihalcea 2007; Ponzetto and Navigli 2010).

While these methods incorporate the limited dictionary definitions with vast common-sense knowledge, most are restricted to words that appear in titles of articles, or rely on textual overlap, causing the process to become brittle. Moreover, methods that rely on mapping procedures are less suitable where named entities and domain-specific terms are involved, as common in domain-specific corpora. As these entities do not appear in the dictionary to begin with, crucial knowledge regarding them can be ignored during the enrichment process. Domain-specific corpora also pose much difficulty for learning-based methods. As opposed to a general corpus, in a domain-specific corpus the distributions of the senses of words are often highly skewed, causing their performance to decline (Agirre, De Lacalle, and Soroa 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 retrieved from a vast encyclopedic resource. It then picks the sense that is most similar to the word's context in that space. This approach has several advantages: (1) it can disambiguate any word as long as it exists somewhere in the knowledge source; (2) it is suitable for named entities and

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

domain-specific terms;(3) the process is completely automatic and (4) any knowledge source that connects topics with texts is suitable. 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.

2 Concept-Based WSD

Given a word in a text, WSD addresses the task of associating that word with an appropriate definition or meaning that is distinguishable from other meanings attributable to that word. We will now describe in detail how our approach handles this task. We will present a general framework for WSD that is entirely concept based and elaborate on the semantic components and data sources it requires.

2.1 The CBD Algorithm

Our algorithm follows the traditional Lesk algorithm (Lesk 1986) 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 $S_i = \{\langle s_i \rangle \, | j = 1, \ldots, n_i\}$.

The algorithm's input consists of the word to disambiguate, denoted w, its context, denoted Ctx, and a list of dictionary senses, S. Additionally, it requires 3 components: (1) a *sense retriever*, denoted SR, that is responsible for associating each sense with a textual fragment; (2) a *semantic interpreter*, denoted SEM, which is able to represent text fragments in a high-dimensional space of natural concepts and (3) a *similarity estimator*, denoted SIM, that given a pair of representations outputs their semantic distance.

The algorithm uses SR to associate each sense $s_i \in S$ with a textual fragment t_i. Then it uses SEM to convert both the context, Ctx, and sense's associated texts, t_i, to their concept-based representation. The result is a weighted vector of concepts, denoted $\langle {w_1}^c, \ldots, {w_n}^c \rangle$ and $\left\langle {w_1}^i, \ldots, {w_n}^i \right\rangle$ respectively. Finally, the algorithm employs SIM to chooses the sense whose representation maximizes the similarity measure. Figure 1 illustrates the system in general while Figure 2 describes the main procedure.

2.2 The Sense Retriever

The *sense retriever* is responsible for 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 (Banerjee and Pedersen 2003). If, however, the dictionary has an intra-linked hierarchical structure, such as WordNet, a more extensive sense retriever can be used, to the benefit of our algorithm. In our implementation, we used textual data that originates from several sources in WordNet: the sense's gloss, its synonyms, its hypernyms and its hyponyms.

2.3 The Semantic Interpreter

The main part of our algorithm consists of utilizing a *semantic interpreter*. In this paper, we experimented with two

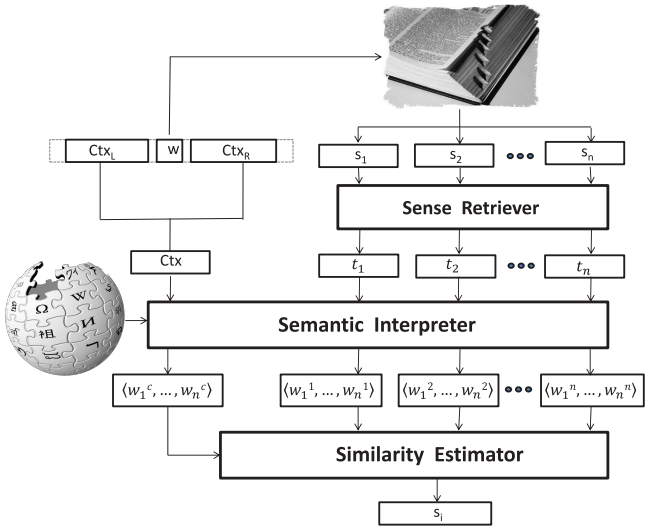

Figure 1: General illustration of the framework

Procedure CBD(w, Ctx, S)
  $\langle {w_1}^c, \ldots {w_n}^c \rangle \leftarrow SEM(Ctx)$
  $maxSim \leftarrow 0$
  $pickedSense \leftarrow nil$
  Foreach $s_i \in S$:
    $t_i \leftarrow SR(s_i)$
    $\left\langle {w_1}^i, \ldots, {w_n}^i \right\rangle \leftarrow SEM(t_i)$
    If $SIM(\langle {w_1}^c, \ldots, {w_n}^c \rangle, \left\langle {w_1}^i, \ldots, {w_n}^i \right\rangle) > maxSim$:
      $maxSim \leftarrow SIM(\langle {w_1}^c, \ldots, {w_n}^c \rangle, \left\langle {w_1}^i, \ldots, {w_n}^i \right\rangle)$
      $pickedSense \leftarrow s_i$
  Return $pickedSense$

Figure 2: Procedure for finding the most appropriate sense

methods for semantic representation of text, *Explicit Semantic Analysis* (ESA) (Gabrilovich and Markovitch 2009) and *Compact Hierarchical Explicit Semantic Analysis* (CHESA) (Liberman and Markovitch 2009). ESA was introduced as a method for semantic representation of natural language texts. In Wikipedia-based ESA, the semantics of a given word is 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. While ESA supplies a flat representation, CHESA produces a hierarchical one. It leverages the conceptual hierarchy inferred from Wikipedia's category system to represent text semantics. Namely, it draws a virtual separating curve on top of the global hierarchy, omitting redundant and over-specific components. We believe CHESA is more suitable for WSD as it is in keeping with the innate human ability to generalize when performing such tasks.

Anchored Explicit Semantic Analysis Sense definitions 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* with the word *tree* to disambiguate. The word *tree* has two main meanings according to WordNet: (1) *a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown* and (2) *a figure that branches from a single root*. Notice the word *crown* that appears in the first gloss and refers to the upper leaves of a tree. As it usually stands for a symbol of monarchy, it is highly related to the word *king* that appears in the context. Accordingly, both ESA vectors of the words *crown* and *king* share many monarchy-related concepts (e.g., BRITISH MONARCHY, CROWN DEPENDENCY). As a result, the first sense is considered by ESA to be more related to the context than the second one, and is wrongly picked.

In order to deal with this problem, we must change the representation of each sense to remove such ambiguity. Therefore, we introduce a novel approach to represent the semantics of a text t in the context of a word, denoted $anchor$. The *Anchored Representation* of a text is achieved by retaining only concepts that relate to the $anchor$.

For ESA, 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[1] in the $anchor$'s interpretation. Formally,

$$w_c^{anchored}(w) = \begin{cases} w_c(w) & \text{if } w_c(anchor) > 0 \\ 0 & \text{if } w_c(anchor) = 0 \end{cases} \quad (1)$$

where $w_c(w)$ is the association weight of concept c and word w obtained by ESA. 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 vector includes only concepts associated with the *anchor*, which maintain their original weights, thus still signify w's association with each concept. We call this process *Anchored Explicit Semantic Analysis* (*Anchored-ESA*). Referring to the aforementioned example, after anchoring to *tree*, *crown*'s representation includes 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.

Similarly, *Anchored-CHESA* initially identifies concepts and categories that are not included in CHESA's representation of the anchoring word, and deletes them from the full Wikipedia hierarchy. Then it builds the anchored representation, top down, in the regular manner. Its benefits 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* includes both nature related categories, derived from one meaning, and technology related categories, derived from the other. We can see that anchoring the representation to the word *fruit* retains only the nature related branches, while anchoring it to the word *computer* retains only the technological ones, as illustrated in Figure 3. A similar method was explored by Reddy et al. (2011) to represent the semantics of noun-noun compounds. Yet, while they used a BOW approach, we prune unrelated concepts, rather than words, and we do so in a structured manner.

[1]More precisely, we eliminate concepts that have a weight of 0 (or below some predefined threshold) in the $anchor$'s vector.

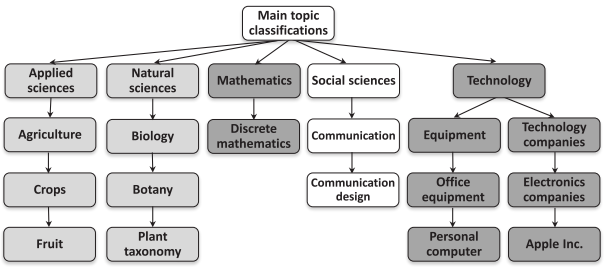

Figure 3: CHESA representation for the word *apple*. The results of anchoring by *fruit* and by *computer* are highlighted in light and dark gray, respectively.

2.4 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, we used the cosine metric to calculate similarity for both ESA and CHESA representations, where for CHESA-based representations, we used the linearized representations as Liberman and Markovitch (2009). Formally,

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

where C is the set of all the concepts in the global hierarchy, and $w_c(t_i)$ is the association weight of concept c and text t_i.

3 Empirical Evaluation

We evaluated our CBD algorithm on a common benchmark for WSD. We now present the performance of our various algorithms, then compare to several state-of-the-art systems.

3.1 Experimental Setup

We implemented our algorithms using a Wikipedia snapshot as of October 18, 2007, which includes over 2 million Wikipedia articles. We follow the footsteps of Gabrilovich and Markovitch (2009) and discard overly-small or isolated articles. We also disregard over-specific categories, lists, and stubs. At the end of this process our knowledge base contained 497,153 articles. Similarly, we implemented our algorithms using the parameters proposed in the original papers (Gabrilovich and Markovitch 2009; Liberman and Markovitch 2009).

Our system was evaluated on the well-known SemEval-2007 coarse-grained all-words WSD task (Navigli, Litkowski, and Hargraves 2007). We chose coarse-grained word- sense disambiguation over fine-grained, since the latter often suffers from low inter-annotator agreement. 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 (Snyder and Palmer 2004), and better reflect the average person's perceptions of the different word senses. Overall, 2,269 content words constituted the test data set, where the average polysemy with the coarse-grained sense inventory

System	P	R	F1
MFS BL	77.40	77.40	77.40
Random BL	63.50	63.50	63.50
ESA	90.32	63.99	74.91
Anchored-ESA	89.06	69.04	77.78
CHESA	87.16	69.22	77.16
Anchored-CHESA	91.92	69.86	79.38

Table 1: Performance on Semeval-2007 coarse grained all-words WSD (nouns only subset).

was 3.06. The inner-annotator agreement was 93.80%, a much higher number than of previous fine-grained tasks.

3.2 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 "don't know." The precision, recall and F1 are computed from these answers. In our system, the decision to reply "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 1. 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 et al. 1993). The results strongly imply that the anchored versions of ESA and CHESA yield a consistent improvement against the unanchored versions, 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.

To further exhibit the strengths of our algorithm , let us review an example from the dataset. 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 WordNet, there are 8 meanings to the lemma *Program*. The right sense obviously relates to a computer program: *a sequence of instructions that a computer can interpret and execute*. However, 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 some of the top concepts that are shared between the context and the right sense, according to the algorithm: 1. COMPUTER LANGUAGE 2. APPLICATION SOFTWARE 3. COMPUTER PROGRAM 4. VIRTUAL MEMORY. As can be easily seen, CBD is able to identify connections between words that stem from their semantics rather than syntax.

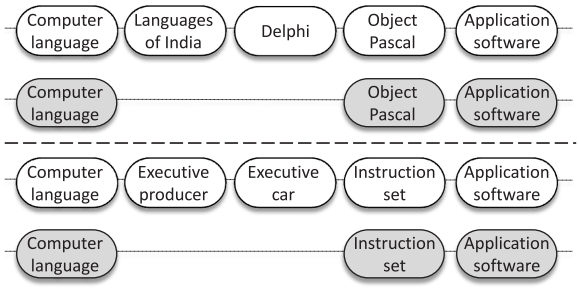

Figure 4: ESA and Anchored-Esa representations for the context of the word *program* (top) and its correct sense (bottom). The anchored versions are highlighted in gray.

This example also illustrates the benefits of anchoring. Figure 4 provides a partial view of the ESA representation of the context and of the correct sense, before and after anchoring. The context's original vector includes concepts triggered by a mixture of meanings of the words in the text. For instance, the word *Delphi* triggers both the non-relevant concept DELPHI, which refers to a town in Greece, and the relevant concept OBJECT PASCAL, which refers to a programming language that is also known as Delphi. However, the anchored-ESA vector is much more definite, retaining only the relevant concepts. A similar effect is achieved with each of the senses. The ESA vector of the correct sense also includes a mixture of meanings. We can see that the word *execute* triggers non-relevant concepts, such as EXECUTIVE PRODUCER and EXECUTIVE CAR. Again, these are omitted in the Anchored-ESA vector. Consequently, the correct sense's representation is very similar to the context's representation, making it an easy pick.

3.3 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, Ng, and Zhong 2007), NUS-ML (Cai, Lee, and Teh 2007), LCC-WSD (Novischi, Srikanth, and Bennett 2007), GPLSI (Izquierdo, Suárez, and Rigau 2007), UPV-WSD (Buscaldi and Rosso 2007). The unsupervised group consisted of SUSSX-FR (Koeling and McCarthy 2007), UOR-SSI (Navigli and Velardi 2005), Degree and ExtLesk (Ponzetto and Navigli 2010). Three of the above competitors, NUS-PT, NUS-ML and LCC-WSD, were the top sys-

System	Nouns Only	All Words
MFS BL	77.44	78.89
NUS-PT	82.31	**82.50**
NUS-ML	81.41	81.58
LCC-WSD	80.69	81.45
GPLSI	80.05	79.55
UPV-WSD	79.33	78.63
SUSSX-FR	81.10	77.00
UOR-SSI	**84.12**	**83.21**
ExtLesk	81.00	79.10
Degree	**85.50**	**81.70**
Anchored-CHESA	**85.02**	**82.68**

Table 2: System scores (P/R/F1) with MFS adopted as a back-off strategy

tems 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 2 as reported in Ponzetto and Navigli (2010).

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 [2].

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 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. Overall, each domain consists of nearly 3500 examples manually annotated with fine-grained senses from WordNet. The average polysemy of the senses is 6.7, and the inter-tagger agreement is 65%.

We used here the best configuration of our system that was 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. Therefore, we avoided using MFS back-off strategy in domain specific settings. Table 3 details the results compared to recent systems [3].

System	Sports	Finance
MFS BL	19.6	37.1
Random BL	19.5	19.6
k-NN	30.3	43.4
Static PageRank	20.1	39.6
Personalized PageRank	35.6	46.9
ExtLesk	40.1	45.6
Degree	42.0	47.8
Anchored-CHESA	**46.5**	**49.3**

Table 3: System scores (P/R/F1) on domain-specific corpora

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, De Lacalle, and Soroa (2009), namely Static PageRank and Personalized PageRank. These systems view WordNet as the Lexical Knowledge Base (LKB), and employ graph-based methods to perform WSD. Finally, we compare our system's performance to recent state-of-the-art unsupervised systems, namely ExtLesk and Degree (Ponzetto and Navigli 2010).

The results indicate that in domain 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 outperforms 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 and Palmer 2004; S. Pradhan and M.Palmer 2007).

Additionally, our system achieves much better results than Personalized PageRank, on both Sports and Finance corpora (+10.9% and +2.4% F1 respectively). The large margin in Sports compared to the one in Finance can be explained by the frequent usage of named entities, such as teams and players. These entities, while having vast coverage in Wikipedia, lack any reference in common dictionaries, thus imper the performance of dictionary-based methods.

Finally, we can note that our system achieves better results than recent state-of-the-art unsupervised systems, ExtLesk and Degree using WordNet++, on both Sports and Finance corpora (+4.5% and +1.5% 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 also be associated with the frequent usage of named entities and domain-specific terms in these corpora, which poses difficulty for such methods, which rely on a mapping between WordNet senses and Wikipedia pages. As these terms are not referred to in WordNet, the mapping procedure cannot associate them with information from Wikipedia. However, our algorithm, which can be applied to any word in the knowledge source, does not suffer from that limitation.

[2] The difference between the top result in each column and non-bold results is statistically significant at $p < 0.05$, while for other results in bold it is not (using the one-sided hypothesis).

[3] We compare with token-based WSD systems, i.e. systems that disambiguate each instance of a target word separately.

Let us review an example 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, in our concept space, these entities are highly correlated with sports-related concepts and categories, enabling our system to pick the correct sense.

4 Related Work

In recent years, unsupervised methods have succeeded in employing vast knowledge sources with domain-specific information (Chen et al. 2009). Knowledge sources provide data essential for connecting senses with words. They can vary from collections of raw texts to libraries of structured data such as Wikipedia. Raw texts are mainly used in a graph-based approach (Koeling and McCarthy 2007). For instance, SSI (Navigli and Velardi 2005) 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 that are integrated from several resources manually and automatically. Additionally, Agirre, De Lacalle, and Soroa (2009) used WordNet to construct a graph of entities and compute the PageRank of the graph by concentrating the initial probability mass uniformly over the context nodes, and finally returning the sense with the highest rank. Their findings also support the argument that knowledge-based systems exhibit a more robust performance than their supervised alternatives when evaluated across different domains.

Over the past several years the importance of Wikipedia as an external source of knowledge for WSD tasks has been increasingly recognized. Wikipedia is mainly used to compute semantic distance between words. These include path-based measures (Strube and Ponzetto 2006), textual overlap (Mihalcea and Csomai 2007), and data-driven algorithms that are based on annotated data (Mihalcea and Csomai 2007; Chen et al. 2009). Still, most of these measures restrict a word's representation to one corresponding article. Milne (2007), Turdakov and Velikhov (2008) implemented a richer semantic representation for a word or text, utilizing Wikipedia's interlink-structure. However, as most of the aforementioned methods, they can only be applied to words that actually occur in titles of Wikipedia articles.

Recently, Ponzetto and Navigli (2010) employed Wikipedia to create an enriched version of WordNet, namely WordNet++, by using mapping procedures. Then, they 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, which, given a target word w, assigns to w the sense whose gloss has the highest overlap with the context of w (a sentence). The second algorithm used is Degree, a graph-based approach that relies on the notion of vertex degree (Navigli and Lapata 2010).

While these methods resemble our algorithm 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 where these components alone lack sufficient semantic information. Second, most of these methods rely on bag-of-words (BOW) similarity, either when mapping senses to articles or upon measuring relatedness. Words are often excessive, over-specific and noisy. Furthermore, for humans, words trigger reasoning at a much deeper level. We measure relatedness though shared concepts, utilizing categories to represent semantics at varying abstraction levels and to avoid using unnecessary data. 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 and Markovitch 2009), co-reference resolution (Ponzetto and Strube 2007), multi-document summarization (Nastase and Strube 2008), text generation (Sauper and Barzilay 2009), and information retrieval (Cimiano et al. 2009).

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. This is in contrast to the brittle process employed by many knowledge-based approaches, which use word overlap techniques to compute semantic relatedness.

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 shown to be more suitable to the task of WSD, where the ambiguous word plays a key role in anchoring the texts of its senses, thus preventing incidental similarities.

Our approach was shown to be competitive with recent methods when evaluated on a general dataset, and superior to state-of-the-art 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, its impact in a multilingual setting should be examined as well. Indeed, ESA has been used successfully in the past for cross-lingual tasks (Hassan and Mihalcea 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.

References

Agirre, E.; De Lacalle, O. L.; and Soroa, A. 2009. Knowledge-based wsd on specific domains: Performing better than generic supervised wsd. In *Proc. of AAAI-09*.

Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proc. of IJCAI-03*.

Bunescu, R. C., and Pasca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL-06*.

Buscaldi, D., and Rosso, P. 2007. Upv-wsd: Combining different WSD methods by means of fuzzy borda voting. In *Proc. of SemEval-07*.

Cai, J. F.; Lee, W. S.; and Teh, Y. W. 2007. Nus-ml: Improving word sense disambiguation using topic features. In *Proc. of SemEval-07*.

Chan, Y. S.; Ng, H. T.; and Zhong, Z. 2007. Nus-pt: Exploiting parallel texts for word sense disambiguation in the English all-words tasks. In *Proc. of SemEval-07*.

Chen, P.; Ding, W.; Bowes, C.; and Brown, D. 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proc. of NAACL HLT-09*.

Chklovski, T., and Mihalcea, R. 2002. Building a sense tagged corpus with open mind word expert. In *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*.

Cimiano, P.; Schultz, A.; Sizov, S.; Sorg, P.; and Staab, S. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proc. of AAAI-09*.

Cuadros, M., and Rigau, G. 2006. Quality assessment of large scale knowledge resources. In *Proc. of EMNLP-06*.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proc. of EMNLP-07*.

Edmonds, P. 2000. Designing a task for senseval-2. Technical report, University of Brighton, U.K.

Gabrilovich, E., and Markovitch, S. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*.

Girju, R.; Badulescu, A.; and Moldovan, D. 2006. Automatic discovery of part-whole relations. *Comput. Linguist.* 32:83–135.

Hassan, S., and Mihalcea, R. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proc. of EMNLP-09*.

Izquierdo, R.; Suárez, A.; and Rigau, G. 2007. Gplsi: Word coarse-grained disambiguation aided by basic level concepts. In *Proc. of SemEval-07*.

Koeling, R., and McCarthy, D. 2007. Sussx: Wsd using automatically acquired predominant senses. In *Proc. of SemEval-07*.

Koeling, R.; McCarthy, D.; and Carroll, J. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proc. of EMNLP-05*.

Lesk, M. E. 1986. Automatic word sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proc. of SIGDOC-86*.

Liberman, S., and Markovitch, S. 2009. Compact hierarchical explicit semantic representation. In *Proc. of the IJCAI-09 Workshop on User-Contributed Knowledge and Artificial Intelligence*.

Mihalcea, R., and Csomai, A. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of ACM CIKM-07*.

Mihalcea, R. 2007. Using Wikipedia for automatic word sense disambiguation. In *Proc. of NAACL HLT-07*.

Miller, G. A.; Leacock, C.; Tengi, R.; and Bunker, R. T. 1993. A semantic concordance. In *Proc. of ACL HLT-93*.

Milne, D. 2007. Computing semantic relatedness using Wikipedia link structure. In *Proc. of NZCSRSC-07*.

Nastase, V., and Strube, M. 2008. Decoding Wikipedia categories for knowledge acquisition. In *Proc. of AAAI-08*.

Navigli, R., and Lapata, M. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.* 32:678–692.

Navigli, R., and Velardi, P. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.* 27:1075–1086.

Navigli, R.; Litkowski, K. C.; and Hargraves, O. 2007. Semeval-2007 task 07: Coarse-grained English all-words task. In *Proc. of SemEval-07*.

Ng, H.; Wang, B.; and Chan, Y. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proc. of ACL-03*.

Novischi, A.; Srikanth, M.; and Bennett, A. 2007. Lcc-wsd: System description for English coarse grained all words task at semeval-2007. In *Proc. of the 4th SemEval-07*.

Pennacchiotti, M., and Pantel, P. 2006. Ontologizing semantic relations. In *Proc. of COLING-06 and ACL-06*.

Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proc. of ACL-10*.

Ponzetto, S. P., and Strube, M. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *J. Artif. Int. Res.* 30:181–212.

Reddy, S.; Klapaftis, I.; McCarthy, D.; and Manandhar, S. 2011. Dynamic and static prototype vectors for semantic composition. In *Proc. of IJCNLP-11*.

S. Pradhan, E. Loper, D. D., and M.Palmer. 2007. Semeval-2007 task-17: English lexical sample srl and all words. In *Proc. of SemEval-07*.

Sauper, C., and Barzilay, R. 2009. Automatically generating Wikipedia articles: a structure-aware approach. In *Proc. of the ACL-IJCNLP-09*.

Snyder, B., and Palmer, M. 2004. The English all-words task. In *In Proc. ACL-04 SENSEVAL-3 Workshop*.

Strube, M., and Ponzetto, S. P. 2006. Wikirelate! computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*.

Turdakov, D., and Velikhov, P. 2008. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proc. of SYRCODIS-08*.