Multilingual Word Sense Discrimination: A Comparative Cross-Linguistic Study

Alla Rozovskaya
Department of Linguistics
Univ. of Illinois at Urbana-Champaign
Urbana, IL 61801
rozovska@uiuc.edu

Richard Sproat
Department of Linguistics
Univ. of Illinois at Urbana-Champaign
Urbana, IL 61801
rws@uiuc.edu

Abstract

We describe a study that evaluates an approach to Word Sense Discrimination on three languages with different linguistic structures, English, Hebrew, and Russian. The goal of the study is to determine whether there are significant performance differences for the languages and to identify language-specific problems. The algorithm is tested on semantically ambiguous words using data from Wikipedia, an online encyclopedia. We evaluate the induced clusters against sense clusters created manually. The results suggest a correlation between the algorithm's performance and morphological complexity of the language. In particular, we obtain FScores of 0.68 , 0.66 and 0.61 for English, Hebrew, and Russian, respectively. Moreover, we perform an experiment on Russian, in which the context terms are lemmatized. The lemma-based approach significantly improves the results over the word-based approach, by increasing the FScore by 16%. This result demonstrates the importance of morphological analysis for the task for morphologically rich languages like Russian.

1 Introduction

Ambiguity is pervasive in natural languages and creates an additional challenge for Natural Language applications. Determining the sense of an ambiguous word in a given context may benefit many NLP tasks, such as Machine Translation, Question Answering, or Text-to-Speech synthesis.

The *Word Sense Discrimination* (WSD) or *Word Sense Induction* task consists of grouping together the occurrences of a semantically ambiguous term according to its senses. Word Sense Discrimination is similar to Word Sense Disambiguation, but allows for a more unsupervised approach to the problem, since it does not require a pre-defined set of senses. This is important, given the number of potentially ambiguous words in a language. Moreover, labeling an occurrence with its sense is not always necessary. For example, in Information Retrieval WSD would be useful for the identification of documents relevant to a query containing an ambiguous term.

Different approaches to WSD have been proposed, but the evaluation is often conducted using a single language, so it is difficult to predict performance on another language. To the best of our knowledge, there has not been a systematic comparative analysis of WSD systems on different languages. Yet, it is interesting to see whether there are significant differences in performance when a method is applied to several languages that have different linguistic structures. Identifying the reasons for performance differences might suggest what features are useful for the task.

The present project adopts an approach to WSD that is based on similarity measure between context terms of an ambiguous word. We compare the performance of an algorithm for WSD on English, Hebrew, and Russian, using lexically ambiguous words and corpora of similar sizes.

We believe that testing on the above languages

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 82–87,
Prague, June 2007.

might give an idea about how accuracy of an algorithm for WSD is affected by language choice. Russian is a member of the Slavic language group and is morphologically rich. Verbs, nouns, and adjectives are characterized by a developed inflectional system, which results in a large number of wordforms. Hebrew is a Semitic language, and is complex in a different way. In addition to the root-pattern morphology that affects the word stem, it also has a complex verb declination system. Moreover, function words, such as prepositions and determiners, criticize, thereby increasing the number of wordforms. Lastly, cliticization, coupled with the absence of short vowels in text, introduces an additional level of ambiguity for Hebrew.

There are two main findings to this study. First, we show that the morphological complexity of the language affects the performance of the algorithm for WSD. Second, the lemma-based approach to Russian WSD significantly improves the results over the word-based approach.

The rest of the paper is structured as follows: first, we describe previous work that is related to the project. Section 3 provides details about the algorithm for WSD that we use. We then describe the experiments and the evaluation methodology in Sections 4 and 5, respectively. We conclude with a discussion of the results and directions for future work.

2 Related Work

First, we describe several approaches to WSD that are most relevant to the present project: Since we are dealing with languages that do not have many linguistic resources available, we chose a most unsupervised, knowledge-poor approach to the task that relies on words occurring in the context of an ambiguous word. Next, we consider two papers on WSD that provide evaluation for two languages. Finally, we describe work that is concerned with the role of morphology for the task.

2.1 Approaches to Word Sense Discrimination

Pantel and Lin (2002) learn word sense induction from an untagged corpus by finding the set of the most similar words to the target and by clustering the words. Each word cluster corresponds to a sense. Thus, senses are viewed as clusters of words.

Another approach is based on clustering the occurrences of an ambiguous word in a corpus into clusters that correspond to distinct senses of the word. Based on this approach, a sense is defined as a cluster of contexts of an ambiguous word. Each occurrence of an ambiguous word is represented as a vector of features, where features are based on terms occurring in the context of the target word. For example, Pedersen and Bruce (1997) cluster the occurrences of an ambiguous word by constructing a vector of terms occurring in the context of the target. Schütze (1992) presents a method that explores the similarity between the context terms occurring around the target. This is accomplished by considering feature vectors of context terms of the ambiguous word. The algorithm is evaluated on natural and artificially-constructed ambiguous English words.

Sproat and van Santen (1998) introduce a technique for automatic detection of ambiguous words in a corpus and measuring their degree of polysemy. This technique employs a similarity measure between the context terms similar in spirit to the one in (Schütze, 1992) and singular value decomposition in order to detect context terms that are important for disambiguating the target. They show that the method is capable of identifying polysemous English words.

2.2 Cross-Linguistic Study of WSD

Levinson (1999) presents an approach to WSD that is evaluated on English and Hebrew. He finds 50 most similar words to the target and clusters them into groups, the number of groups being the number of senses. He reports comparable results for the two languages, but he uses both morphologically and lexically ambiguous words. Moreover, the evaluation methodology focuses on the success of disambiguation for an ambiguous word, and reports the number of ambiguous words that were disambiguated successfully.

Davidov and Rappoport (2006) describe an algorithm for unsupervised discovery of word categories and evaluate it on Russian and English corpora. However, the focus of their work is on the discovery of semantic categories and from the results they report for the two languages it is difficult to infer how the languages compare against each other.

We conduct a more thorough evaluation. We also control cross-linguistically for number of training examples and level of ambiguity of selected words, as described in Section 4.

2.3 Morphology and WSD

McRoy (1992) describes a study of different sources useful for word sense disambiguation, including morphological information. She reports that morphology is useful, but the focus is on derivational morphology of the English language. In the present context, we are interested in the effect of inflectional morphology on WSD, especially for languages, such as Russian and Hebrew.

Gaustad (2004) proposes a lemma-based approach to a Maximum Entropy Word Sense Disambiguation System for Dutch. She shows that collapsing wordforms of an ambiguous word yields a more robust classifier due to the availability of more training data. The results indicate an improvement of this approach over classification based on wordforms.

3 Approach

Our algorithm relies on the method for selection of relevant contextual terms and on distance measure between them introduced in (Sproat and van Santen, 1998) and on the approach described in (Schütze, 1998), though the details of clustering differ slightly. The intuition behind the algorithm can be summarized as follows: (1) words that occur in the context of the ambiguous word are useful for determining its sense; and (2) contextual terms of an ambiguous word belong to topics corresponding to the senses of the ambiguous word. Before describing the algorithm in detail, we give an overview of the system.

The algorithm starts by collecting all the occurrences of an ambiguous word in the corpus together with the surrounding context. Next, we build a symmetric distance matrix D, where rows and columns correspond to context terms, and D[i][j] is the distance value of term *i* and term *j*. The distance measure is supposed to reflect how the two terms are close semantically (whether they are related to the same topic). For example, we would expect the distance between the words *financial* and *money* to be smaller than the distance between the words *financial* and *river*: The first pair is more likely to occur in the same context, than the second one. Using the distance measure, the context terms are partitioned into sense clusters. Finally, we group the sentences containing the ambiguous word into sentence clusters using the context term clusters.

We now describe each step in detail:

1. We collect contextual terms of an ambiguous word *w* in a context window of 50 words around the target. Each context term *t* is assigned a *weight* (Sproat and J. van Santen, 1998):

$$w_t = \frac{CO(t|w)}{FREQ(t)} \tag{1}$$

   $CO(t|w)$ is the frequency of the term in the context of w, and $FREQ(t)$ is the frequency of the term in the corpus. Term weights are used to select context terms that will be helpful in determining the sense of the ambiguous word in a particular context. Furthermore, term weights are employed in (4) in sentence clustering.

2. For each pair t_i and t_j of context terms, we compute the distance between them (Sproat and J. van Santen,1998):

$$D_w[i][j] = 1 - \frac{[\frac{CO_w(t_i|t_j)}{FREG(t_i)} + \frac{CO_w(t_j|t_i)}{FREQ(t_j)}]}{2} \tag{2}$$

   $CO_w(t_i|t_j)$ is the frequency of t_i in the context of t_j, and $FREQ(t_i)$ is the frequency of t_i in the training corpus. We assume that the distance between t_i and t_j is inversely proportional to the semantic similarity between t_i and t_j.

3. Using the distance matrix from (2), the context terms are clustered using an agglomerative clustering technique:

   - Start by assigning each context term to a separate cluster
   - While stopping criterion is false: merge two clusters whose distance [1] is the smallest.[2]

[1]There are several ways to define the distance between clusters. Having experimented with three - Single Link, Complete Link and Group Average, it was found that Complete Link definition works best for the present task. (*Complete Link* distance between clusters *i* and *j* is defined as the maximum distance between a term from cluster *i* and a term from cluster *j*).

[2]In the present study, the clusters are merged as long as the

The output of step (3) is a set of context term clusters for the target word. Below are shown select members for term clusters for the English word *bass*:

Cluster 1: songwriter singer joined keyboardist

Cluster 2: waters fishing trout feet largemouth

4. Finally, the sentences containing the ambiguous word are grouped using the context term clusters from (3). Specifically, given a sentence with the ambiguous word, we compute the score of the sentence with respect to each context word cluster in (3) and assign the sentence to the cluster with the highest score. The score of the sentence with respect to cluster *c* is the sum of weights of sentence context terms that are in *c*.

4 Experiments

The algorithm is evaluated on 9 ambiguous words with two-sense distinctions. We select words that (i) have the same two-sense distinction in all three languages or (ii) are ambiguous in one of the languages, but each of their senses corresponds to an unambiguous translation in the other two languages. In the latter case, the translations are merged together to create an artificially ambiguous word. We believe that this selection approach allows for a collection of a comparable set of ambiguous words for the three languages. An example of an ambiguous word is the English word *table*, that corresponds to two gross sense distinctions (*tabular array*, and *a piece of furniture*). This word has two translations into Russian and Hebrew, that correspond to the two senses. The selected words are presented in Table 1.

The words display different types of ambiguity. In particular, disambiguating the Hebrew word *gishah* (access; approach) or the Russian word *mir* (peace; world) would be useful in Machine Translation, while determining the sense of a word like *language* would benefit an Information Retrieval system. It should also be noted that several words possess additional senses, which were ignored because they rarely occurred in the corpus. For example, the Russian word *yazyk* (language) also has the meaning of *tongue* (body part).

number of clusters exceeds the number of senses of the ambiguous word in the test data.

The corpus for each language consists of 15M word tokens, and for the same ambiguous word the same number of training examples is selected from each language. For each ambiguous word, a set of 100-150 examples together with 50 words of context is selected from the section of the corpus not used for training. These examples are manually annotated for senses and used as the test set for each language.

5 Evaluation Methodology

The evaluation is conducted by comparing the induced sentence clusters to clusters created manually. We use three evaluation measures : *cluster purity*, *entropy*, and *FScore*. [3]

For a cluster C_r of size q_r, where the size is the number of examples in that cluster, the dominating sense S_i in that cluster is selected and *cluster purity* is computed as follows:

$$P(C_r) = \frac{n_r^i}{q_r}, \tag{3}$$

where n_r^i is the number of examples in cluster C_r with sense S_i.

For an ambiguous word *w*, cluster purity *P(w)* is the weighted average of purities of the clusters for that word. [4]. Higher cluster purity score corresponds to a better clustering outcome.

Entropy and *FScore* measures are described in detail in Zhao and Karypis (2005). *Entropy* indicates how distinct senses are distributed between the two clusters. The perfect distribution is the assignment of all examples with sense 1 to one cluster and all examples with sense 2 to the other cluster. In such case, the entropy is 0. In general, a lower value indicates a better cluster quality. Entropy is computed for each cluster. Entropy for word *w* is the weighted average of the entropies of the clusters for that word.

Finally, *FScore* considers both the coverage of the algorithm and its ability to discriminate between the two senses. *FScore* is computed as the harmonic

[3] Examples whose scores with respect to all clusters are zero (examples that do not contain any terms found in the distance matrix) are not assigned to any cluster, and thus do not affect cluster purity and cluster entropy. This is captured by the FScore measure described below.

[4] In the present study, the number of clusters and the number of senses for a word is always 2

Senses	English	Hebrew	Russian
access;approach	access;approach	gishah	dostup;podxod
actor;player	actor;player	saxqan	akter;igrok
evidence; quarrel	argument	vikuax;nimuq	argument
body part; chief	head	rosh	golova;glava
world;peace	world; peace	shalom;olam	mir
furniture; tabular array	table	shulxan;tavlah	stol;tablitza
allow;resolve	allow;resolve	hershah;patar	razreshat'
ambiance; air	atmosphere	avira;atmosfera	atmosfera
human lang.;program. lang.	language	safah	yazyk

Table 1: Ambiguous words for testing: The first column indicates the senses; unambiguous translations that were merged to create an ambiguous word are indicated by a semicolon

mean of *Precision* and *Recall*, where recall and precision for sense S_i with respect to cluster C_r are computed by treating cluster C_r as the output of a retrieval system for sense S_i .

6 Results and Discussion

We show results for two experiments. Experiment 1 compares the algorithm's performance cross-linguistically without morphological analysis applied to any of the languages. Experiment 2 compares the performance for Russian in two settings: with and without morphological processing performed on the context terms.

Table 2 presents experimental results. Baseline is computed by assigning the most common sense to all occurrences of the ambiguous word. We observe that English achieves the highest performance both in terms of cluster purity and FScore, while Russian performs most poorly among the three languages. This behavior may be correlated with the average frequency of the context terms that are used to construct the distance matrix in the corpus (cf. 7 for English and 4.2 for Russian). In particular, the difference in the frequencies can be attributed to the morphological complexity of Russian, as compared to English and Hebrew. Hebrew is more complex than English morphologically, which would account for a drop in performance for the Hebrew words vs. the English words. Furthermore, one would expect a higher degree of ambiguity for Hebrew due to the absence of short vowels in text.

It is worth noting that while both Hebrew and Russian possess features that might negatively affect the performance, Hebrew performs better than Russian. We hypothesize that cliticization and the lack of vowels in text are not as significant factors for the performance as the high inflectional nature of a language, such as Russian. We observe that the majotity of the context terms selected by the algorithm for disambiguation belong to the noun category. This seems intuitive, since nouns generally provide more information content than other parts of speech and thus should be useful for resolving lexical ambiguity. While an English or a Hebrew noun only has several wordforms, a Russian noun may have up to 12 different forms due to various inflections.

The morphological complexity of Russian affects the performance in two ways. First, cluster purity is affected, since the counts of context terms are not sufficiently reliable to accurately estimate term distances. Incorrect term distances subsequently affect the quality of the term clusters. Second, the percentage of default occurrences (examples that have no context terms occurring in the distance matrix) is the least for English (0.22) and the highest for Russian (0.27). The default occurrences affect the recall.

The results of experiment 2 support the fact that morphological complexity of a language negatively affects the performance. In that experiment, the inflections are removed from all the context terms. We apply a morphological analyzer [5] to the corpus and replace each word with its lemma. In 10% of the word tokens, the analyzer gives more than one possible analysis, in which case the first analysis is selected. As can be seen in Table 2 (last row), removing inflections produces a significant improvement both in recall and precision, while preserving the cluster purity and slightly reducing cluster entropy. Moreover, the performance in terms of recall, precision, and coverage is better than for English and

[5] Available at *http://www.aot.ru/*

Language	Baseline	Coverage	Precision	Recall	FScore	Purity	Entropy
English	0.73	0.78	0.77	0.61	0.68	0.79	0.61
Hebrew	0.72	0.79	0.76	0.58	0.66	0.82	0.59
Russian	0.72	0.73	0.70	0.54	0.61	0.81	0.62
Russian(lemma)	0.72	0.80	0.77	0.66	0.71	0.82	0.61

Table 2: Results: Baseline is the most frequent sense; coverage is the number of occurrences on which the decision was made by the algorithm

Hebrew.

7 Conclusions and Future Work

We have described a cross-linguistic study of a Word Sense Discrimination technique. An algorithm based on context term clustering was applied to ambiguous words from English, Hebrew, and Russian, and a comparative analysis of the results was presented. Several observations can be made. First, the results suggest that the performance can be affected by morphological complexity in the case of a language, such as Russian, specifically, both in terms of precision and recall. Second, removing inflectional morphology not only boosts the recall, but significantly improves the precision. These results support the view that morphological processing is beneficial for WSD.

For future work, we plan to investigate more thoroughly the role of morphological analysis for WSD in Russian and Hebrew. In particular, we will focus on the inflectional morphology of Russian in order to determine whether removing inflections consistently improves results for Russian ambiguous words across different parts of speech. Further, considering the complex structure of the Hebrew language, we would like to determine what kind of linguistic processing is useful for Hebrew in the WSD context.

Acknowledgments

We are grateful to Roxana Girju and the anonymous reviewers for very useful suggestions and comments. This work is funded in part by grants from the National Security Agency and the National Science Foundation.

References

Dmitry Davidov and Ari Rappoport 2006. Efficient Unsupervised Discovery of Word Categories Using Symmetric Patterns and High Frequency Words. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 297–304. Sydney, Australia.

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Tanja Gaustad. 2004. A Lemma-Based Approach to a Maximum Entropy Word Sense Disambiguation System for Dutch. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)*, 778-784. Geneva.

Dmitry Levinson. 1999. Corpus-Based Method for Unsupervised Word Sense Disambiguation. *www.stanford.edu/ dmitryle/acai99w1.ps*.

Susan Weber McRoy. 1992. Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 18(1): 1–30.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text In *In Proceedings of ACM SIGKDD*, pages 613-619. Edmonton.

Ted Pedersen and Rebecca Bruce. 1997. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 197-207. Providence, RI, August.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Richard Sproat and Jan van Santen. 1998. Automatic ambiguity detection. In *Proceedings of International Conference on Spoken Language Processing* . Sydney, Australia, 1998.

Ying Zhao and George Karypis. 2005. Hierarchical Clustering Algorithms for Document Datasets *Data Mining and Knowledge Discovery*, 10(2):141–168.