An Unsupervised Method for Multilingual Word Sense Tagging Using Parallel Corpora: A Preliminary Investigation

Mona Diab
Linguistics Department & UMIACS,
University of Maryland, College Park, MD 20742
mdiab@umiacs.umd.edu

Abstract

With an increasing number of languages making their way to our desktops everyday via the Internet, researchers have come to realize the lack of linguistic knowledge resources for scarcely represented/studied languages. In an attempt to bootstrap some of the required linguistic resources for some of those languages, this paper presents an unsupervised method for automatic multilingual word sense tagging using parallel corpora. The method is evaluated on the English Brown corpus and its translation into three different languages: French, German and Spanish. A preliminary evaluation of the proposed method yielded results of up to 79% accuracy rate for the English data on 81.8% of the SemCor manually tagged data.

Keywords
Unsupervised; multilingual; alignments; parallel corpora; word sense tagging

1. Introduction

With the term "globalization" becoming the theme of current political and economic discourse, communications technology – exemplified by the World Wide Web (WWW) - has become a source of an abundance of languages. Language researchers are faced with an ever so present challenge and excitement of being able to study and process these languages and create the appropriate NLP applications for them. Yet, a major bottleneck for many NLP applications such as machine translation, cross language information retrieval, natural language understanding, etc, is word sense ambiguity. The problem escalates as we deal with languages that are scarce in processing resources and knowledge bases. The availability of large scale, accurately, sense tagged data should help alleviate the problem.

It has been acknowledged that best way to acquire sense tags for words in a corpus is manually, which has proven to be a very expensive and labor intensive endeavor. In an attempt to approximate the human effort, both supervised [Bruce & Weibe, 1994; Lin, 1999;etc.] and unsupervised methods [Resnik 1997; Yarowsky, 1992&1995; etc.] have been proposed to solve the problem automatically. On average supervised methods report higher accuracy rates, but they are faced with the problem of requiring large amounts of sense tagged data as training material. Most of the methods, to date, aim at solving the problem for one language, namely the language with the most available linguistic resources. Moreover, most of the proposed approaches report results on a handful of the data, rendering them solutions for a small scale of the data.

Many researchers in the field have looked at language translations as a source for sense distinctions [Dagan & Itai, 1994; Dyvik, 1998; Ide, *in press*; Resnik & Yarowsky, 1999; etc.]. The idea is that polysemous words in one language can be translated as distinct words in a different language. The problem has always been the availability of large corpora in translation, i.e. parallel corpora. Resnik [1999] proposed a method for facilitating the acquisition of parallel corpora from the WWW. Potentially, we can have parallel corpora in a myriad of languages, yet the downside is the scarcity of linguistic knowledge resources and processing tools for less widely represented/studied languages. Consequently, we decided to bootstrap the process of word sense tagging for both languages in a parallel corpus using the translations as a source of word sense distinction. Thereby, attaining sense tagged data for languages with scarce resources as well as creating a supply of large-scale, automatically sense tagged data for a the language with more knowledge resources –albeit noisy - to be utilized by supervised algorithms.

In this paper, we propose an unsupervised method for word sense tagging of both corpora automatically. The algorithm assumes the availability of a word sense inventory in one of the languages. The preliminary evaluation of the method on the nouns in an English corpus, yielded accuracy rates in the range of 69-77% against the polysemous nouns in a hand tagged test set, which contrasts with a random baseline of 25.6%, and a baseline of the most frequent sense of 67.6%.

In the following section we describe the proposed method, followed by a preliminary evaluation of the method. Section 4 discusses related work and we conclude with some thoughts on future directions in section 5.

2. Proposed method

We propose a method that utilizes translations as filters for sense distinctions. The method is unsupervised since it does not rely on the availability of sense tagged data. As an illustration, if we look up the canonical ambiguous word ***bank*** in the Oxford Hachette English-French dictionary, we find that it translates to several words indicating its possible senses. ***Bank***, as a noun, translates to the French words ***banque***, ***rive***, ***bord***, etc. If we reverse the French translations into English, we get the original word ***bank*** as well as other English equivalents. Accordingly, ***rive*** translates back into English as ***bank*** and ***shore***; ***bord*** translates into ***bank***, ***edge***, and ***rim***. Therefore, given a parallel corpus with a source and target language, if there exists a method of finding word alignments from the source language corpus to words in the target language corpus, one can create a set of all the words in the target corpus that are aligned with a word in the source corpus. For example, given a French/English parallel corpus, we would expect the word ***rive****,* on the French side*,* to align with the words ***bank*** and ***shore****,* on the English side, in the correct contexts with a high probability. This approach essentially hinges upon the diversity of contexts in which words are translated.

We will refer to the English side of the parallel corpus as the **target** language corpus since we assume the knowledge resources exist for English. The foreign language side is referred to as the **source** corpus.

The required linguistic knowledge resource is a lexical ontology that has the words in the target language and a listing of their associated senses. There are several databases of that sort available for language researchers, among which is WordNet [Fellbaum, 1998; Miller et al., 1990]. WordNet is a lexical ontology - a variant on semantic networks with more of a hierarchical structure, even though some of the nodes can have multiple parents - that was manually constructed for the English language. It comprises four taxonomies for four parts of speech: nouns, verbs, adverbs and adjectives.

Accordingly, given a taxonomy like WordNet for the target language, and an appropriate distance measure between words with their associated senses, the distance between all the senses for both ***shore*** and ***bank*** is calculated. In WordNet 1.6, ***bank*** has 10 senses, the 3 topmost frequent senses are:

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 especially for future use (especially in emergencies)*

shore has two senses listed:

1. *the land along the edge of a body of water (a lake or ocean or river)*
2. *a beam that is propped against a structure to provide support*

One would expect that the distance between sense #2 of ***bank*** and sense #1 of ***shore*** to be smaller than the latter's distance from the other two senses of ***bank***. Accordingly, with an appropriate optimization function over the distance measures between all the senses of the two words, sense #2 for ***bank*** and sense #1 for ***shore*** are assigned as the correct tags for the words, respectively. In effect, we have assigned sense tags to ***rive*** in its respective alignments, in the appropriate contexts. Therefore the instances where ***rive*** is aligned with ***bank*** gets assigned sense #2 for the noun ***bank***; instances where ***rive*** is aligned with ***shore*** is assigned sense #1 for ***shore***. Furthermore, we created links automatically in WordNet for the French word ***rive***. Our approach is described as follows:

- Preprocessing of corpora
  - Tokenize both corpora
  - Align the sentences of the corpora such that each sentence in the source corpus is aligned with one corresponding sentence in the target corpus.
- For each source and corresponding target sentence, find the best token level alignments. Methods for automating this process have been proposed in the literature. [Al Onaizan et al., 1999; Melamed, 2000; etc.]
- For each source language token, create a list of its alignments to target language tokens, **target set**
- Using the taxonomy, calculate the distance between the senses of the tokens in the target set; assign the appropriate sense(s) to each of the tokens in the target set based on an optimization function over the entire set of target token senses
- Propagate the assigned senses back to both target and source corpora tokens, effectively, creating two **tag sets**, one for each the target and source corpus
- Evaluate the resulting tag sets against a hand tagged **test set**.

3. Preliminary Evaluation

3.1. Materials

We chose the Brown Corpus of American English [Francis & Kučera, 1982] – of one million words - as our target language corpus. It is a balanced corpus and it has more than 200K words that are manually sense tagged as a product of the semantic concordance (SemCor) effort using WordNet [Miller et al. 1994]. The SemCor data is tagged in running text – words of varying parts of speech are tagged in context – using WordNet 1.6. Hence, we used WordNet 1.6 taxonomy as the linguistic knowledge resource. [Fellbaum, 1998] For purposes of this preliminary investigation, we only explored nouns in the corpus, yet there are no inherent restrictions in the method for applying it to other parts of speech. Accordingly, we used part of speech tags that were available in the Penn Tree Bank for the Brown Corpus.

The test set was created from the polysemous nouns in SemCor. The nouns were extracted from the Brown corpus with their relative corpus and sentence position information. The test set comprised 58372 noun instances of 6824 polysemous nouns. The nouns were not lemmatized.

Two baselines were constructed. A random baseline (RBL), where each noun instance in the test set was assigned a random sense from the list of senses pertaining to that noun in the taxonomy. And a default baseline (DBL), where each noun instance in the test set is assigned its most frequent sense according to WordNet 1.6.

The Brown Corpus only exists in English; therefore, we decided to automatically translate it into three different languages using two commercially available machine translation (MT) packages, Systran Professional 2.0 (SYS) and Globalink Power Translator Pro v.6.4 (GL). We used two different translation packages to maximize the variability of the word translation selection, in an attempt to approximate a human translation. The idea is that different MT packages use different bilingual lexicons in the translation process. Moreover, we decided to use more than one language since polysemous words can be translated in different ways in different languages, i.e. an ambiguous word that has two senses could be translated into two distinct words into one language but into one word in another language. We translated the Brown Corpus into French, German and Spanish, since these are considered the most reliable languages for the translation quality of the MT packages. Furthermore, the fact that EuroWordNet exists for these languages facilitates the process of evaluating the source language tag set.

3.2. Experiments

Once we had the translations available, the seven corpora – namely, English Brown corpus, French GL, German GL, Spanish GL, French SYS, German SYS, and Spanish SYS - were tokenized and the sentences were aligned[1]. For

[1] This was a relatively easy task since the corpora are artificially created, therefore there was a one to one

token level alignments, we used the GIZA program [Al Onaizan et al. 1999]. GIZA is an intermediate program in a statistical machine translation system, EGYPT. It is an implementation of Models 1-4 of Brown et al. [1993], where each of these models produces a Viterbi alignment. The models are trained in succession where the final parameter values from one model are used as the starting parameters for the next model. We trained each model for 10 iterations. Given a source and target pair of aligned sentences, GIZA produces the most probable token-level alignments. Multiple token alignments are allowed on the target language side, i.e. a token in English could align with multiple tokens in the foreign language. Tokens on either side could align with nothing, designated as a null token. GIZA requires a large corpus in order to produce reliable alignments, hence, the use of the entire Brown corpus: both the SemCor tagged data without the tags and the untagged data. Therefore, we produced the alignments for the 6 parallel corpora – a parallel corpus comprises the English corpus and its translation into one of the three languages using one of the MT packages - with English as the target language.

The Brown Corpus has 52282 sentences. Due to processing limitations, GIZA ignores sentences that exceed 50 words in length, therefore it ignored ~3000 sentences on average per parallel corpus alignment. GIZA output was converted to an internal format: sentence number followed by all the tokens[2] in the sentence represented as token positions in the target language aligned with corresponding source language token positions in the aligned foreign sentence.

All the token positions were replaced by the actual tokens from the corresponding corpora. Tokens that were aligned with null tokens on either side of the parallel corpus were ignored. All the tokens were tagged with the sentence number and sentence position. In order to reduce the search space, we reduced the list to the nouns in the corpus. We created a list of the source language words that were aligned to nouns in the target language, thereby creating a source-target noun list for each source word. We removed punctuation marks and their corresponding alignments; also, we filtered out stop words from the source language. Finally, we compressed the source-target list to have the following format:

$\boldsymbol{Src\_wd_i} \qquad \boldsymbol{trgt\_nn_1, trgt\_nn_2, \ldots, trgt\_nn_n}$

> *where* $\boldsymbol{Src\_wd_i}$ *is a word*[3] *in the source corpus and* $\boldsymbol{trgt\_nn_j}$ *is the noun*[4] *it aligned to in the target corpus.*

Source words that were aligned with one target word only throughout the corpus were excluded from the final list of words to be tagged in our tag set. Each resulting set – a set had to include at least 2 nouns - of English target nouns, corresponding to a source word, was passed on to the distance measure routine.

We used an optimization function over the senses of the nouns in a set. The function aims at maximizing a similarity of meaning over all the members of a set based on a pair wise similarity calculation over all the listed senses in WordNet 1.6. The algorithm, **disambiguate_class,** which is implemented by Resnik and described in detail in [Resnik, 1999], calculates the similarity between all the words' senses of words in a set. It assigns a confidence score based on shared information content of the sense combinations, which is measured via the most informative subsumer in the taxonomy. The senses with the highest confidence scores are the senses that contribute the most to the maximization function for the set. The algorithm expects the words to be input as a set for calculating the confidence scores. In many instances, we observed considerable noise in the target noun set. For example, the French source word *accord* was aligned with the English nouns *accord, agreement, signing, consonance, and encyclopaedia* in the target corpus. All but the last word in the target set seem to be related to the word *accord* in French except *encyclopaedia.* The source of noise can be attributed to the specific translation system, or to the alignment program, or in other cases to the

correspondence between the source and target sentences.

[2] Tokens include punctuation

[3] Parts of speech are not necessarily symmetric in alignments, i.e. nouns could very well map to verbs or other parts of speech.

[4] Note that the nouns at this point are types not tokens, i.e. not instances in the corpus rather a conflation of instances

fact that the source language word itself is ambiguous.

Consequently, we conducted three types of experiments in an attempt to reduce the noise in the target sets: **Class_sim, Pair_sim_1** and **Pair_sim_all.** They essentially varied in input format to disambiguate_class.

For **Class_sim**, the target noun data was produced directly from the source-target list and input to the distance measure routine with no special formatting. Each of the target nouns was assigned the sense(s) that had the maximum confidence level from among the senses listed for it in the taxonomy. Thereby creating the tag set for the target language, English. If a noun does not have an entry in the taxonomy, it is assigned a null sense.

On the other hand, for both **Pair_sim_1** and **Pair_sim_all** the nouns in the target list for each source word were formatted into all pair combinations in the set and then sent to disambiguate_class. The idea was to localize the noise to the pair level comparison, since disambiguate_class optimizes over the entire set of nouns. The senses that were selected were the ones with the maximum confidence score from the noun pair sense comparison. All the senses with a maximum confidence score for a noun were aggregated into a final list of senses for that noun and duplicates were removed.

In **Pair_sim_1**, only the senses that had a confidence score of 100% were considered, i.e. if disambiguate_class is agnostic as to whether the senses of the target noun pair are similar, each noun in this pair comparison is assigned a null sense, for the noun pair in the local comparison, respectively. That does not necessarily mean that either noun will have a final null sense in the aggregate list, it rather depends on the sum total of comparisons for each of them with all the nouns in the set.

In **Pair_sim_all**, the same conditions apply as in Pair_sim_1, yet there is no threshold of a 100%. A pair of nouns in a local comparison is assigned a null sense if one of the nouns in the pair is not in WordNet or all the senses get a confidence score of 0%.

Once we had the tag set for each of our parallel corpora, we evaluated it against the manually tagged test set. So far, we only evaluated the tag set for the target language, English. Evaluation of the source tag set is in progress; a serious hurdle is that EuroWordNet is interfaced with WordNet 1.5 only. The preliminary evaluation metric is:

$$acc = \frac{num\_correct\_taggedsenses\_found}{total\_num\_testsenses} * 100 \quad [1]$$

We only considered the first sense assigned in the test set for any noun instance in the process of our evaluation. The system was not penalized if it assigned more than one sense to the noun in the tag set if the correct sense was among the senses assigned.

We conducted the three types of experiments on the 6 parallel corpora. In the following section, we present the results for GL translations for the three languages and the SYS translation for Spanish, since we found no significant difference in the results across the two translation systems for the three experiment types. Furthermore, we wanted to test the effect of merging the token alignments of the two MT systems on the accuracy rates. For all the experiment conditions, the noun instances that were excluded from the tag set and were in the test set were sense tagged using the default baseline of 67.6%, in order to report the results at 100% coverage for the test set, the results of which are presented in table 2 below.

3.3. Results and Discussion

The investigation yielded the following results

	Class_sim		Pair_sim_1		Pair_sim_all	
	Cov	***Acc.***	***Cov***	***Acc.***	***Cov***	***Acc.***
FG	62.4	45.0	55.4	60.1	60.1	73.9
GG	49.0	48.2	41.6	57.1	48.0	70.7
SG	57.2	47.2	50.7	57.1	56.1	72.8
SS	56.8	46.0	50.6	55.7	55.5	72.9
MSp	83.6	45.5	75.8	63.0	81.8	79.0

Table 1: *Results for the different experiment types at various coverage levels of the test set*

Table 1. presents the results at different coverage percentages of the test set data for the English target corpus. The first column has the 5 experiment conditions used as source language filters of the English target corpus, and the first row has the three experiment types. *FG* is the French translation of the Brown corpus rendered by the MT system GL; *GG* is the German translation by GL; *SG* is the Spanish translation by GL; *SS* is the Spanish translation by the MT system SYS; and *MSp* is the merged Spanish translations from both MT systems. All the results are presented as percentages, where the ***Cov***. indicates the percentage covered by the tag set of the test set. ***Acc***. is the percent correct at the coverage level based on the evaluation measure in [1].

Across the board, the results from Pair_sim_all for all the experiment conditions are higher than the results from Pair_sim_1, which in turn are higher than Class_sim results. The results do not seem to suggest any significant difference in the results from the two Spanish translations SG and SS across the three experiment types. On the other hand, results from MSp outperform the individual Spanish translation systems for the Pair_sim_1 and Pair_sim_all experiments by a margin of ~25% more in coverage and ~6% in accuracy rates. In the Class_sim experiment, the individual Spanish translations outperform the MSp condition. We also note that coverage is higher for all the experiment conditions.

	Class_sim	Pair_sim_1	Pair_sim_all
FG	53.5	63.4	71.4
GG	58.1	63.2	69.1
SG	55.9	62.3	70.5
SS	55.3	61.6	70.6
MSp	49.1	64.1	76.9
RBL	25.6 %		
DBL	67.6 %		

Table 2: *Results at 100% coverage of the test set*

Table 2 reports the results at 100% coverage of the test set data for the target tag set. *FG, GG, SG, SS, MSp*, are the same as in table 1. *RBL* is the random baseline, while *DBL* is the default baseline. All the experimental conditions significantly outperformed the random baseline. None of the conditions outperformed the default baseline, DBL, in both Class_sim and Pair_sim_1 experiments. Pair_sim_1 had a higher accuracy rate than Class_sim for all the experiment conditions. Similar to the observations in table 1, Pair_sim_all outperformed the other two experiment types for all the experiment conditions. Pair_sim_all also outperformed the default baseline with an improvement of 1.4 (marginal in this case) to 9%. It is worth noting that there was no significant difference between the experimental conditions SG and SS across the experiment types. As in Table 1, the results from MSp are significantly higher than those obtained from the individual Spanish translation conditions for both Pair_sim_1 and Pair_sim_all, while the results for Class_sim were much lower than the individual Spanish conditions. This can be attributed to the fact that while combining evidence from both translations, we aggregated the noise in the target set from both translations. The noise causes disambiguate class to get trapped into assigning higher confidences to irrelevant senses.

In terms of the overall performance of the different conditions, the results suggest that merging the two translation systems yields the best results, with an improvement of 6% over the individual translations independently for Spanish in Pair_sim_all. Examining the results across the three languages, it seems there were slight variations in the accuracy rates in the Pair_sim_1 and Pair_sim_all experiments at full coverage, exemplified in table 2. Yet we note the low relative coverage of the test data in the German, GG condition, as shown in table 1. This can be explained as a result of the nature of the German language, which is highly agglutinative, thereby affecting the quality of the alignments. Also it could be a reflection of the quality of the GL MT system for the German language.

The most interesting result is the result of the MSp condition in table 1, which indicates that 81.8% of the target data can be sense tagged with an accuracy of 79%, significantly higher than chance (25.6%) as well as it is higher than the default tagging of 67.6%. We have yet to investigate the source tag set in order to see how many of these source words can transparently acquire the target noun senses. The fine graininess of WordNet leads us to suspect that the appropriate level of evaluation will be at the most informative subsumer level in the taxonomy (a coarser grain) as opposed to the actual sense tagged for the corresponding aligned target noun.

The low accuracy rates over the full test set (table 2) may be attributed to the cascading of different sources of noise in the evaluation method, starting off with a less than perfect translation[5] and an automated alignment program with a reported accuracy rate of ~92% for word alignments, English to German. [Och & Ney, 2000] The latter result has to be considered with caution in the present experimental design context since the evaluation of the alignments was done with a human translation on a closed domain corpus, for only one of the languages under consideration in the current investigation. A large-scale multilingual evaluation of the alignment program is much needed. By qualitatively looking at some of the automatic alignments, some of the cases had very tight alignments in the target language. For instance, the French word ***abri*** was aligned with ***cover*** and ***shed***; ***agitation,*** in French, was aligned with the nouns ***agitation, bustle, commotion, flurry, fuss, restlessness,*** and ***turmoil***.

Word ambiguity in the source language could have contributed to the low accuracy rates attained. In many cases, we noticed that the source language seemed to preserve the ambiguity found in the target language. For example, (a) the French word ***canon*** was aligned with the target nouns: ***cannon, cannonball, canon, theologian***; (b) the French word ***bandes*** was aligned with the target nouns: ***band, gang, mob, streaks, strips, tapes, tracks***. In both examples we see at least two clusters in the target noun sets, in (a), ***cannon*** and ***cannonball*** are one cluster and ***canon*** and ***theologian*** form the other cluster; in (b), the word ***band*** is ambiguous, we can see that ***band, gang*** and ***mob*** can form a cluster, while ***band, streaks, strips, tapes*** and ***tracks*** could form another. We are currently investigating the effect of incorporating co-occurrence information as a means of clustering the words in the target set, aiming at delineating the senses for the source language word. Another source of noise is the metaphoric as well as slang usage of some of the words in the target language, for instance, ***bébés,*** in French, was aligned with ***babes*** and ***babies*** in the target language.

We expect the results to improve the more distant the language pair. Moreover, combining different language sources simultaneously could yield improved results due to the fact that languages will differ in the manner in which they conflate senses.

We would like to explore different evaluation metrics for the target language, which are fine-tuned to the fine granularity of WordNet. As well as, devise methods for obtaining a quantitative measure of evaluation for the source tag set.

4. Related Work

There are many proposed unsupervised methods in the literature addressing the problem of sense ambiguity in language. All the reported unsupervised methods use monolingual materials, therefore comparable to the results obtained on the target tag set of our preliminary investigation. Moreover, due to differences in the knowledge resources and evaluation material it is hard to establish a direct comparison. For instance, Yarowsky [1992&1995] reports the highest accuracy rates, to date, for an unsupervised method of a mean of 92%, yet his evaluation was measured using a knowledge resource, Roget's thesaurus, which has a coarser granularity in its sense representation than WordNet.

The most comparable results to our preliminary results are those reported by Resnik [1997] since he used the same corpus and evaluated against the same test set. He did not restrict his evaluation to nouns only. Resnik proposed an unsupervised method for sense disambiguation using selectional preference information, thereby using grammatical relations between words in a corpus in order to arrive at the correct sense for a word. He reports accuracy rates in the range of 40.1% on average for five grammatical relations. Yet, Resnik explores a different dimension of meaning that uses a linguistically motivated context window which we expect will be very useful if combined with our approach for examining the verb data, for example.

The most related work reported in the literature is that of Ide [*in press*]. Ide explores the question of whether using cross-linguistic information for sense distinction is worth pursuing. She reported a preliminary analysis of translation equivalents in four different languages of George Orwell's *Nineteen-Eighty-four*. The translations were human translations, i.e. natural parallel corpora. In her study, only 4 words were considered. Native speakers of the four respective languages

[5] We do not know of any formal evaluation on the quality of the two translation packages used

aligned the chosen English words to their foreign translations manually. The goal of her research was to explore the degree to which words are lexicalized differently in translated text. Ide classifies translation types based on how much they vary in what they align with in translation, for example, if a word aligns with a single word or a phrase or nothing, etc. She reports that in *Nineteen-Eighty-Four,* only 86.6% of the English words have a single lexical item used in the translation. This suggests that with using alignment methods that target single word to single word alignments the upper bound that the approach can yield is 86.6% for this specific corpus. It will be interesting to conduct a similar study here of the Brown corpus.

5. Conclusion and Future Directions

We presented an unsupervised method for word sense tagging for both the source and the target languages in a parallel corpus. The method relies on translations as a source of sense distinction. The goal of the proposed algorithm is to bootstrap the process of word sense tagging on a large scale for a language with linguistic knowledge resources as well as for languages with scarce resources. As a proof of concept, we evaluated the approach on 6 artificially created translation corpora. The preliminary evaluation yielded accuracy rates of up to 79% for 81.8% of the test set in the target language. The source language tag set is yet to be evaluated. Future directions include devising methods for reducing the noise in the target sets. Moreover, testing the approach on other parts of speech. Furthermore, it would be interesting to test the method on naturally created parallel corpora.

Acknowledgements

This work has been has been supported in part by DARPA contract N6600197C8540 and DARPA co-operative agreement N660010028910. We would like to thank Philip Resnik for his support and feedback throughout the different phases of this investigation. We would like also to acknowledge an anonymous reviewer for his/her useful comments.

References

Al-Onaizan, J. Y.Curin, M. Jahr, K. Knight, J. Laferty, D. Melamed, F. Och, D. Purdy, N. Smith, & D. Yarowsky (1999). *Statistical Machine Translation, Final Report*, JHU workshop. http://www.clsp.jhu.edu/ws99/projects/mt/final.report/mt-final-report.ps

Brown, P. F., S. S. Della Pietra, V. J. Della Pietra, and R. L. Mercer (1993). *The mathematics of statistical machine translation: Parameter estimation*. Computational Linguistics, 19(2): 263-311.

Bruce, Rebecca & Janyce Wiebe (1994). *Word-sense Disambiguation Using Decomposable Models*. Proc. of 32nd Association of Computational Linguistics, Las Cruces, NM.

Dagan, Ido & Alon Itai (1994). *Word Sense Disambiguation Using a Second Language Monolingual Corpus.* Computational Linguistics 20, pp. 563-596

Dyvik, Helge (1998). *A Translational Basis for Semantics.* In Stig Johansson and Signe Oksefjell (eds.): Corpora and Cross-linguistic Research: Theory, Method and Case Studies, 51-86.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Francis, W. & H. Kučera (1982). *Frequency Analysis of English Usage.* Houghton Mifflin Co: New York.

Ide, Nancy (in press). *Cross-lingual sense determination: Can it work?* Computers and the Humanities, 34.

Lin, Dekang (1999). *A Case-base Algorithm for Word Sense Disambiguation*. Pacific Association for Computational Linguistics, Waterloo, Canada.

Melamed, I. Dan (2000). *Models of Translational Equivalence among Words*, Computational Linguistics 26(2), pp. 221-249, June.

Miller, G., M. Chodorow, S. Landes, C. Leacock, and R. Thomas (1994). *Using a Semantic Concordance for Sense Identification*. ARPA Human Language Technology Workshop, San Francisco, CA.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller (1990). *WordNet: An on-line lexical database.* International Journal of Lexicography, 3(4), 235-244.

Och, Franz J. & Hermann Ney (2000). *A Comparison of Alignment Models for*

Statistical Machine Translation. 8th Int. Conference on Computational Linguistics, Saarbrücken, Germany, July.

Resnik, Philip (1999). *Mining the Web for Bilingual Text*, 37th meeting of Association for Computational Linguistics, College Park, Maryland, USA, June.

Resnik, Philip (1997). *Selectional Preference and Sense Disambiguation*, SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, D.C., USA, April.

Resnik, Philip (1999). *Semantic Similarity in a Taxonomy: An information-based Measure and its Application to Problems of Ambiguity in Natural Language*. Journal of Artificial Intelligence Research, 11, 95-130.

Resnik, Philip & David Yarowsky (1998). *Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation.* Natural Language Engineering, 1, 1-25.

Yarowsky, David (1992). *Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* Proc. of 14th International Conference on Computational Linguistics, Nantes, France, July.

Yarowsky, David (1995). *Unsupervised Word Sense Disambiguation Rivalling Supervised Methods*. 33rd meeting of Association for Computational Linguistics, Cambridge, MA.