Knowledge sources for disambiguating highly ambiguous verbs in machine translation

LUCIA SPECIA
University of Sao Paulo / University of Sheffield
L.Specia@dcs.shef.ac.uk

ABSTRACT. Word sense disambiguation (WSD) is one of the most challenging outstanding problems in the current machine translation systems. An effective proposal in this context will rely on the use relevant knowledge sources. Moreover, it must perform better than the current traditional approaches. We present some experiments with machine learning algorithms traditionally applied to WSD, aiming to discover both the best knowledge sources and the performance of these approaches. The results confirmed those already reported in monolingual WSD, indicating collocations and semantic word associations as the best word sense distinctive characteristics. In future work, we will use the best knowledge sources discovered, along with good rules produced by a symbolic algorithm, in a new WSD approach.

1 Introduction

Word Sense Disambiguation (WSD) in Machine Translation (MT) is required to carry out the lexical choice in the case of semantic ambiguity during the translation, i.e., the choice for the most appropriate translation for a source language word when the target language offers more than one option, with different meanings, but the same part-of-speech. For example, assuming English-Portuguese translation, the noun *bank* can be translated as *banco* (*financial institution*) or *margem* (*land along the side of a river*), and the verb *to run* can be translated as *correr* (*to move quickly*) and *ir* (*to travel* or *to go*). In this context, thus, “sense” means, in fact, “translation”.

Several WSD approaches have been proposed for MT. Most of them are knowledge-based, strongly dependent on the manual encoding of accurate linguistic knowledge and disambiguation rules, e.g., (Pedersen, 1997) and (Dorr and Katsova, 1998). To avoid excessive manual effort, we focus on corpus-based and hybrid approaches. The first make use of knowledge automatically acquired from text using machine learning techniques, while the latter merge characteristics from knowledge and corpus-based approaches, minimizing the knowledge acquisition bottleneck, but also concerning about the accuracy of the acquired knowledge.

In fact, recent approaches to WSD have converged to the use of corpus-based and hybrid techniques, e.g., (Zinovjeva, 2000) and (Lee, 2002). They have shown good results, in terms of accuracy and coverage, especially those following the supervised learning. As stressed by Wilks and Stevenson (1997), when considering MT, it is difficult to determine the effectiveness of unsupervised approaches, since the possible senses (translations) must be previously defined.

The approaches mentioned above deal with translation from English to different languages: Danish, Spanish, French, Swedish and Korean. Considering English-Portuguese MT, the language pair addressed in this paper, there is no worthy note work. As evidenced in (Specia, 2005), the lack of effective WSD mechanisms is one of the main reasons for the unsatisfactory results of the existent MT systems.

The work we present here is part of a major research project, which aims at the creation of a new hybrid symbolic approach to WSD, to be applied to English-Portuguese MT, as described in (Specia, 2005). The main innovative feature of this approach is the formalism to be used to represent instances and background knowledge. Unlike other WSD works, which use restricted propositional formalisms, the proposed approach will employ a relational formalism, based on first-order logic.

In this paper we describe a number of experiments developed in order to gain insight to the proposed approach, namely: (1) find out the accuracies of supervised propositional machine learning algorithms, considering several knowledge sources, to compare them to the ones to be obtained by the proposed approach; (2) figure out the best knowledge sources and filters to be used in the proposed approach; and (3) find in the produced models good symbolic rules that may be used as knowledge source in the proposed approach. Although we discuss all these goals, we focus on the second one. As we will illustrate, there are several works exploring the contribution of knowledge sources to WSD, however, with only one exception, all of them consider monolingual applications. Moreover, the multilingual work evaluates a few knowledge sources in a very limited experimental setting.

We experimented with seven highly ambiguous verbs and four algorithms of different learning paradigms: symbolic, statistical, functional and memory-based. In all the cases, we tried out features representing syntactic, semantic and topical knowledge, either individually or in combinations of two or three features. Although we analyze all the resulting data, we look more carefully into those resulting from the symbolic algorithm, since the proposed approach is also symbolic. Moreover, regarding the third goal, the symbolic algorithm is the only one that generates a comprehensible model, which can be in effect analyzed.

The rest of this paper is organized as follows. We first describe, Section 2, our experimental setting, including the knowledge sources used as features, our sample corpus, and the algorithms employed. In Section 3 we present and discuss the results of the experiments according to the three mentioned goals. In Section 4 we illustrate some related work, also aimed at evaluating knowledge sources to WSD. Finally, in Section 5 we conclude with some final remarks and future work.

2 Experimental setting

2.1 Knowledge sources, features and lexical resources

To explain the knowledge sources explored in our work, as well as the ones used by the related work presented in Section 4, we use the taxonomy defined by Agirre and Stevenson (2005). They distinguish between **knowledge sources**, **features**, and **lexical resources** used for WSD. Knowledge sources are the high-level abstract linguistic and semantic phenomena relevant to WSD. They can be of syntactic (e.g. part-of-speech and collocations), semantic (e.g. syntagmatic relations and selectional preferences) or pragmatic/topical (e.g. topical word associations and domain information) natures. Features are the ways of encoding the knowledge sources used by the systems. For instance, the domain of a word sense can be represented by the words co-occurring often with that word sense (bag-of-words features). Lexical resources are the resources used to extract the features in actual systems. For instance, bag-of-words features can be extracted from sense-tagged corpora. According to this taxonomy, we explore knowledge from the three sources, in the form of the following features:

1. Syntactic: part-of-speech (POS) and different kinds of collocations, encoded as local patterns:
   - Lemmas of content words in a ± 5 word window (F1).
   - POS tags of content words in a ± 5 word window (F2).
   - Set of 10 collocations defined by Stevenson and Wilks (2001) (first and second words to left and right, first noun, first adjective and first verb to left and right of the target word), plus the first preposition to the right, amounting to 11 patterns, represented by the lemmas of the words (F3).
2. Semantic: syntagmatic semantic word associations, encoded as the lemmas of the verb subject and object syntactic relations (F4).
3. Pragmatic/topical: topical word associations, encoded as bag-of-words of ± 0-100 lemmas of words surrounding the target word, considering all words (F5), and also the POS tags of all words in a ± 0-100 word window (F6).

The features are grouped in six classes, F1-F6. The features F1, F2, F3, F4 and F6 are multi-valued features: their possible values are the lemmas/POS in the sentence position that they represent. The features F5, on the other hand, are binary features, i.e., every lemma is encoded as a feature and analyzed according to its presence or absence in the sample sentences. The number of multi-valued features amounts to 233 (F1=10 + F2=10 + F3=11 + F4=2 + F6=200). The number of binary features will depended on the different words in the corresponding ± 0-100 positions in the sample sentences.

Regarding the lexical resources, all the features were extracted from a corpus (Section 2.2) annotated with the senses and all the other information (POS, etc.).

2.2 Sample data

Our sample corpus consists of English sentences containing the seven verbs under consideration: *to come*, *to get*, *to give*, *to go*, *to look*, *to make* and *to take*. The sentences were collected from the corpus Compara (Frankenberg-Garcia and Santos, 2003), which comprises fiction books texts. Each sentence has a sense tag, which corresponds to the translation of the verb in that sentence. The sense tagging process was first carried out automatically, as described in Specia et al. (2005), and then manually reviewed. Besides the sense tags, the corpus presents other information:

- POS of all words, produced by the tagger Mxpost (Ratnaparkhi, 1996).
- Lemmas of all words, produced by the parser Minipar (Lin, 1993).
- Subject-object syntactic relations, also produced by Minipar.

The original set of corrected sentences amounts to 1,420: about 200 sentences for each verb. However, some preliminary experiments with those sentences have shown that they were not adequate to be used as instances to machine learning algorithms, because of their sparseness with respect to the classes (senses): many senses had only one sentence as instance. So, we filtered the data in order to select only the instances whose sense occurred at least three times in the sample data. The initial number of senses and the number of remaining instances and senses after the filter are shown in Table 1, along with the resulting percentage of instances with the most frequent sense.

Verb	**Initial number of senses (200 instances)**	**Remaining instances**	**Remaining senses**	**% most frequent sense**
to come	26	183	11	50.3
to get	51	157	17	21
to give	27	180	5	88.8
to go	25	197	11	68.5
to look	16	191	7	50.3
to make	39	170	11	70
to take	63	142	13	28.5

Table 1: Sample data after the instance filter

In order to gather the features from the sample corpus, we created a feature extractor, which allows the choice of any of the six possible features (along with their parameters), either individually or in combination. The output of the system is the set of chosen features, with their headers, in the attribute-relation file format of the machine learning environment Weka[1], used to run our experiments.

Besides the instance filter applied before the feature extraction, we used two kinds of attribute filters: (1) Weka filters to convert the string features into multi-valued features (F1, F3 and F4) or into binary features (F5); and (2) a filter designed to cope with the feature sparseness. The latter is a very simple filter, also used by Lee and Ng (2002) (Section 4), which removes from all instances the features values (or the features themselves, in the case of binary features) that do not occur at least a given N number of times with a certain sense. We experimented with N=1 (i.e., no filter), N=2 and N=3. Higher values of N would cause too many features to be removed.

2.3 Algorithms

Based on the results reported for monolingual WSD (e.g. (Mooney, 1996), (Lee and Ng, 2002), and (Yarowsky and Florian, 2003)), we chose four algorithms of different learning paradigms to perform our experiments: **Naïve Bayes**, **Decision Rules**, **Support Vector Machines,** and **Memory-based**. We used implementations provided by the Weka environment (NaïveBayes, PART, SMO and LWL, respectively).

2.4 Combinations of features tested

At first we experimented with three verbs and 30 settings of features, including each feature individually and in combination with one or two other features, varying the number of words in the bag-of-word features, and trying to avoid combinations of

[1] http://www.cs.waikato.ac.nz/~ml/weka/

redundant features. Based on the results of these preliminary tests, we selected the best settings, shown in Table 2 with their respective numbers for further reference.

We experiment with these 11 settings considering the three possible parameters of the attribute filter (N=1, N=2 or N=3).

No.	Setting
S1	Lemmas of content words in a ± 5 word window (F1)
S2	Lemmas and POS tags of content words in a ± 5 word window (F1+F2)
S3	Lemmas of the first and second words to left and right, first noun, first adjective, and first verb to left and right, and first preposition to the right of the target word (F3)
S4	Lemmas of content words in a ± 5 word window and subject and object syntactic relations (F1+F4)
S5	POS tags of content words in a ± 5 word window and subject and object relations (F2+F4)
S6	Lemmas and POS tags of content words in a ± 5 word window and subject and object relations (F1+F2+F4)
S7	Lemmas of the first and second words to left and right, first noun, first adjective, and first verb to left and right of the target word, and first preposition to the right of the target word, and subject and object relations (F3+F4)
S8	Bag-of-words of ± 5 lemmas of words surrounding the target word (F5)
S9	POS tags of all words in a ± 100 word window (F6)
S10	Bag-of-words and POS tags of ± 5 lemmas of words surrounding the target word (F5+F6)
S11	Bag-of-words and POS tags of ± 5 lemmas of words surrounding the target word, and subject and object relations (F4+F5+F6)

Table 2: Features tested in the experiments

3 Results and discussion

The results were obtained in terms of the precision of each algorithm, with each subset of features, for each verb. Since we have few instances, the training and test sets are defined by resampling the set of instances, using a 10-fold cross validation strategy. In Table 3 we present the average precision for all verbs[2].

Algorithm Setting	No filter				Filter, N=2				Filter, N=3			
	PA	NB	LWL	SOM	PA	NB	LWL	SOM	PA	NB	LWL	SOM
S1	55.0	53.9	58.9	58.3	54.0	61.5	67.0	64.2	54.1	58.3	63.4	59.8
S2	56.3	54.6	57.9	57.0	56.0	65.2	66.4	65.2	55.9	64.2	67.3	66.0
S3	62.2	60.4	63.2	66.0	63.1	**71.8**	**74.0**	**74.7**	63.8	**71.2**	**73.6**	**73.5**
S4	55.4	52.5	59.2	61.2	54.3	62.6	68.7	66.6	54.1	61.1	65.1	61.5
S5	54.9	54.9	56.5	53.4	55.1	62.7	63.2	63.8	55.6	62.0	65.6	63.9
S6	56.2	55.2	60.1	57.6	55.3	67.0	67.5	66.4	55.9	66.1	68.4	66.2
S7	61.8	59.6	64.1	66.4	63.2	**72.5**	**75.0**	**75.3**	63.8	**72.5**	**74.5**	**73.3**
S8	59.0	55.0	59.8	64.1	65.3	66.2	64.3	**72.0**	**71.0**	62.5	66.3	**72.0**
S9	55.5	56.2	54.8	59.0	57.9	**71.6**	64.6	**70.0**	59.9	**70.6**	**74.7**	**70.5**
S10	59.7	56.0	60.6	63.3	65.4	65.9	60.7	**77.6**	**71.5**	63.4	64.0	**80.5**
S11	61.2	57.0	61.0	64.4	66.3	65.0	61.0	**78.8**	**70.9**	63.9	64.0	**80.8**

Table 3: Average results for the seven verbs

[2] PA = PART, NB = Naïve Bayes, LWL = Memory-based, SOM = Support Vector Machines.

3.1 Best knowledge sources and filters

We analyzed the best features and filters for each verb, since they have different numbers of instances and possible senses, and also different bias, given mainly by the distribution of senses among the instances. For example, if we consider the most frequent sense of each verb, *to get* has 21% of the instances with that sense, while *to give* has 88.8%. As a consequence, the results are different for different verbs. We discuss the average results, but also point out idiosyncratic cases, when the results for individual verbs present significant deviations with relation to the average.

As has already been revealed by previous work (Section 4), the best knowledge sources and filters depend not only on the target word, but also on the algorithm used. As shown in Table 4, in average, the best precisions (above 70%, in bold face), considering all the algorithms, are achieved by features S3, S7, S8, S9, S10, and S11, and the attribute filters with N=2 and N=3. As for the average precision for each algorithm and filter, the best precision features with no filter are the same for all the algorithms: S3 and S7. This shows that local patterns representing collocations (and collocations plus syntactic relations) are strong indicators of the correct sense even when the sample data are sparse.

Using the filters, on the other hand, the best precision feature varies among the algorithms: for Naïve Bayes, S3 and S7 with filter N=3, and S9 with N=2; for LWL, also S3 and S7 with N=2, and S9 with N=3; for SOM, S10 and S11 with both filters; for PART, S8, S10 and S11 with both filters. It seems to indicate that the filters must be removing some important, though not so frequent, collocations, making other features to become the stronger indicators of the senses. Nonetheless, if we consider the overall average results, the two features enclosing collocations (S3 and S7), mainly in combination with syntactic relations, are the strongest indicators of the target word senses, followed by S11 and S10.

The strengths of the features S3 and S7 have already been pointed out by works reporting experiments in monolingual contexts, such as Agirre and Stevenson (2005) and Mohammad and Pedersen (2004). The reason for that seems to be that the semantic word associations, represented here both as syntactic dependencies and implicitly as local patterns, play an important role in identifying the sense of a verb, even when the sense is the translation to other language. Hence, in this case the conclusions for monolingual contexts remain valid in our multilingual experiments. Zinovjeva (2000), as described in Section 4, experimenting with one verb, also shows similar results, but only with relation to the syntactic dependencies. Additionally, as pointed in the monolingual studies, the combination of knowledge sources yields better results than those obtained by each knowledge source individually.

Looking into each verb separately, we verify the same phenomenon. For most of the verbs, with exception of *to give* and *to make*, S3 and S7 are the best features. These two verbs are those with the highest irregular sense distributions: *to give* has 88.8% of the instances with only one sense, and the rest with other four senses, while *to make* has 70% of the instances with only one sense, and the rest with other ten senses. This strong bias to the most frequent sense seems to make the local patterns in S3 less strong sense indicators. As for the syntactic dependencies, they are kept in the best feature combinations in both cases: for *to give*, S11 and S10, while for *to make*, S11 and S6.

Regarding the filters, the average results show that using a filter is significantly better for all the algorithms and features, with exception of the decision rule algorithm, which in some cases present better accuracy without any of the two filters. This can be explained by the fact that this algorithm already uses a method to select the best features, based on the information gain concept. Between the two different filters, N=2 allows an average precision better than N=3, and it is also valid for the individual verbs, in most of the cases. However, again, the precision varies depending on the algorithm and feature. For Naïve Bayes, the filter with N=2 is better than (or equal to) N=3 for all the features; for LWL and SOM, in general, N=2 is the best filter if used with multi-valued features, while N=3 is the best with binary features or features comprising larger context windows. This seems to be reasonable, since the sparseness in data tends to be larger with binary and big context windows features, and hence a more restrictive filter may be more adequate.

The general conclusion about the filters is that N=2 is slightly better than the N=3, considering the reduced sample data employed. If we include more instances as sample data, it is likely that N=2 will prove to be the best filter, since the sparseness will tend to be minimized and then a less restrictive filter might be more appropriate.

3.2 Baseline results

To evaluate the results with respect to our first goal, we looked into the precision of the algorithms, which we intend to compare to those of our proposed approach under similar conditions (same instances, same or equivalent knowledge sources). We can elect the best algorithm and use its results as a baseline to the comparison.

The first conclusion about the accuracy of the algorithms is that it is quite satisfactory, outperforming the baseline of the most frequent sense, for all the verbs, in most of the cases. The exceptions are the PART algorithm, which in some cases presents the same precision as the baseline, and the Naïve Bayes, which presents lower precision than the baseline for some features when no filter is applied, but only for two verbs (*to come* and *to give*).

To get some clues about the best algorithm, we can consider the average precision for all verbs. According to this average, the best results are achieved by LWL, followed by SOM, when using filters. When no filter is used, SOM presents the best precision, followed by LWL. However, without filters, the differences among the algorithms are not as great as with the filters. Although LWL could be used as baseline to be compared to our approach, this will require further analysis, taking into account the features being used, as well as the precision for individual verbs.

3.3 Symbolic rules

Considering the models generated by the algorithm PART, a more qualitative evaluation can be carried out. Some objective measures, including the precision of the rule, as well as other not so objective measures, such as the specificity and novelty of the rules, can be automatically calculated. Additionally, some more subjective criteria can also be analyzed, such as the level of interest of the rules. This requires a manual examination, but can allow extracting important evidences to disambiguation. Furthermore, the examination allows changing the rules, as done by Zinovjeva (2000), in order to improve them. It can be extremely time consuming: in some cases, only two or three rules were produced; but in other cases, the models comprise more than 40 rules. In Figure 1, we show some examples accurate rules with wide coverage.

1_col = at: olhar (51.0/2.0)	into <= 0 AND	into <= 0 AND
1_col = like: parecer (24.0)	through <= 0 AND	through <= 0 AND
1_col = for: procurar (14.0/2.0) **(a)**	on > 0: continuar (7.0)	in <= 0: ir (161.0/20.0) **(b)**

Figure 1: Examples of rules generated for *to look* (a) and *to go* (b)

The three rules in Figure 1(a) were generated as part of the model for *to look* considering collocations (S3) as features, without any filter. The most significant local pattern is *1_col*, which corresponds to the first word after the verb. All the rules represent knowledge about the verb followed by prepositions or particles. In the first case, *at* is a preposition and the translation is *olhar* (*perceive with attention*). In the other cases, *like* and *for* are particles, forming phrasal verbs translated as *parecer* (*bear resemblance to*) and *procurar* (*search*). In all the cases, the coverage and accuracy of the rules is significant, as shown (covered instances / incorrectly covered instances).

The rules in Figure 1 (b) are part of the model for *to go* considering bag-of-words in a ± 5 context window (S8), with the filter N=3. The rules test the presence (> 0) or absence (<= 0) of words. For example, the first rule (first column) assigns the translation *continuar* (*proceed*) when *on* is amongst the 10 words surrounding the verb, but *into* and *through* are not. This rule covers correctly seven instances.

It is important to note that the symbolic rules analysis can also contribute to identify the knowledge sources effectively used in the disambiguation models, since the features tested are explicitly shown in the rules. For example, in Figure 2, even though the rules refer to bag-of-words, it is clear that they are working like collocations, since the kind of words used (prepositions and particles) does not represent the topic of the sentences. Instead, they are likely to refer always to the first word after the verb.

4 Related work

In what follows we describe some recent work related to our second goal: finding out the best knowledge sources. Different systems explore different knowledge sources. Many of them try to combine several sources in order to get better results (e.g., Hirst, 1987; McRoy, 1992); but only some recent works present systematic analyses and quantitative results regarding to the contribution of different knowledge sources to WSD. With exception of Zinovjeva (2000), the others are aimed at monolingual WSD.

Stevenson and Wilks (2001) carried out experiments which compared several knowledge sources extracted from LDOCE (Procter, 1978): POS, morphology, collocations, syntagmatic relations, topical word associations, selectional preferences, and domain information. The authors evaluated these sources in all Semcor content words. The system uses the output of processes applying such sources (except the collocations), in the form of filters or partial taggers, as input, along with the collocations, to a memory-based learning algorithm. The average accuracy of each process varies between 44.8 and 79.4%. The best isolated knowledge source is domain information, while the least successful is selectional preferences (except for the verbs). Collocations were tried out only together with other knowledge sources. The combination of all the knowledge sources yielded an average accuracy of 90.4%.

Agirre and Martínez (2001) experimented with six algorithms and several knowledge sources: frequency of senses, syntagmatic and paradigmatic relations, local context (POS, morphology, collocations, subcategorization and syntactic relations), global context (semantic and topical word associations and pragmatics), and selectional preferences. The algorithms were tested to disambiguate nouns in two subsets of Semcor (Miller et al., 1994). In both cases, the supervised algorithm using local context, syntagmatic relations or global context presented the best results. According to the authors, approaches combining knowledge sources from different resources can provide better results. They also emphasize that collocations, semantic associations, syntagmatic relations, and frequency of senses, when learned from sense tagged corpora , are strong indicators of the word senses.

Yarowsky and Florian (2002) experimented with a range of syntactic, semantic and topical knowledge sources, using local context features (n-grams, using the words, lemmas, and POS), syntactic dependencies features (depending on the POS of the target word), and bag-of-words features. They tested their system on the Senseval-2 evaluation data, employing five machine learning algorithms. They compare the accuracy for all the features combined to the accuracy when each of the features (or two of them) is omitted. The best accuracy is achieved when all the features are used, for all the algorithms and grammatical categories. In average, the features causing the most decrease in the accuracy if omitted are bag-of-words, and those causing the less decrease are the syntactic features, which the authors credit to the sparse instantiation rate of these features and also to the noise in the detection process. Nonetheless, analyzing the results for each grammatical category, the verb disambiguation accuracy appears to derive a significant benefit from syntactic features, mainly the object.

Lee and Ng (2002) evaluated four learning algorithms and several knowledge sources, through local patterns features (POS of three surrounding words, 11 n-grams of different lengths), some syntactic relations features, and bag-of-words of all words in the sentence. They also evaluated the use of a feature selection method that selects a feature only if it occurs at least in three instances of the ambiguous word with a given class in the training data. The algorithms were tested on the Senseval-1 and Senseval-2 data sets, considering each of the features individually and the combination of all features. Individual features presented a similar accuracy among algorithms and also among features (50.9-60.5%). With exception of one of the algorithms (decision tree), the rest benefit from the combination of all features (62.7-65.8%). In almost all the cases, the use of the feature selection method improved the results.

Martínez et al. (2002) evaluated the contribution of a set of syntactic features for supervised WSD. In addition to various commonly used local and topical features, such as collocations and bag-of-words, they experimented with syntactic features generated by a parser. These features include "instantiated grammatical relations", which are coded as triples containing the word-sense, the relation, and the value of the relation, and the grammatical relations themselves, which are coded as n-grams containing the word-sense and one or more relations. The syntactic features were analyzed in isolation and also in combination with the traditional features mentioned. The algorithms were tested using Senseval-2 data. In both algorithms, the syntactic features outperformed the traditional ones in terms of accuracy. In one of the algorithms, the combination of all features achieved a better accuracy, while in the other, the accuracy of the syntactic features is higher then the accuracy of the combination, except for verbs.

Mohammad and Pedersen (2004) accomplished some experiments with local patterns in the form of unigrams and bigrams, the target word form, the POS of the target word and of its two surrounding words, and four kinds of syntactic dependency relations. They used a decision tree algorithm, considering each feature individually and in combinations of two to five features. The experimental data were the Senseval-1 and Senseval-2 test sets, as well as four corpora with examples for one word each (line, hard, serve and interest). Each corpus presented best results using different features: Senseval-1 (68%), serve (75.7%) and interest (80.6%) – POS of the target word and of 1-2 surrounding words; Senseval-2 (55%) – unigrams or bigrams; line (74.5%) – unigrams; hard (89.5%) – bigrams.

Aiming to learn symbolic WSD rules for MT from English to Swedish, Zinovjeva (2000) experimented with a transformation-based algorithm and some knowledge sources: collocations (1-4 words), POS of the surrounding 1-4 words and subject, object and prepositions verbal syntagmatic relations. Only two nouns and one verb were considered. The first two knowledge sources were tested individually, yielding to the following accuracies for the three words: 92.1%, 95.2% and 73.1%, for collocations; and 93.6%, 95.4% and 80.8%, for POS tags. The third and fourth experiments were accomplished for the verb. The third considered the verbal syntactic relations and the fourth, the combination of the three sources of knowledge. The accuracies obtained were 83.3% and 84.6%, respectively. These experiments, although limited to only one word, shows that the combinations of several knowledge sources also seems to be appropriate to disambiguate verbs in multilingual contexts.

5 Conclusions

We presented the some experiments in which machine learning algorithms have been applied to WSD for MT, considering several knowledge sources and filters. The results confirm those reported in previous work focusing monolingual contexts, regarding the relevance of knowledge sources: collocations and semantic word associations are the strongest indicators of the senses, mainly for verbs. Also, they confirm that the combination of knowledge sources, when they are not redundant, can convey better results. Since we have a small and sparse sample data, the use of instance and attribute filters showed to be appropriate.

As for the accuracy of the algorithms, it can be considered satisfactory, given the high level of ambiguity of the verbs, the small number of instances and the sparseness in the sample data. With relation to the individual performance of the algorithms, the memory-based and support vector machines approaches achieved the best accuracies, in general.

Looking into the models generated by the symbolic algorithm, we found out rules with good coverage and precision, capturing interesting knowledge, which could be used as knowledge sources in our relational approach, but a further study, considering objective and subjective measures, will be necessary to select the best rules.

Acknowledgments

I'd link to thank Mark Stevenson, for reviewing the paper, and Syllas Freitas de Oliveira Neto, for helping me with the feature extraction module.

References

Agirre, E. and Martínez, D. (2001). Knowledge Sources for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference on Text Speech and Dialogue,* Plzen, Czech Republic.

Agirre, E. and Stevenson, M. (2005) (in press). Knowledge Sources for Word Sense Disambiguation. In *Word Sense Disambiguation: Algorithms, Applications and Trends*, Agirre, E. and Edmonds, P. (Eds.), Kluwer.

Bruce, R. and Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association. for Computational Linguistics (ACL)*, Las Cruces, pp.139-145.

Dorr, B. J. and Katsova, M. (1998). Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of AMTA'1998*, Langhorne, pp. 438-447.

Frankenberg-Garcia, A. and Santos, D. (2003). Introducing COMPARA: the Portuguese-English Parallel Corpus. *Corpora in translator education*, pp. 71-87.

Hirst, G. (1987). Semantic Interpretation and the Resolution of Ambiguity. *Studies in Natural Language Processing*. Cambridge University Press, Cambridge.

Lee, H. (2002) Classification Approach to Word Selection in Machine Translation. In *Proceedings of AMTA'2002*, Berlin, pp. 114-123.

Lee, Y.K.; Ng, H.T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, pp. 41-48.

Lin, D. (1993). Principle based parsing without overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, pp. 112-120.

Martínez D., Agirre E. and Màrquez L. (2002). Syntactic Features for High Precision Word Sense Disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei.

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

Miller, G.A.; Chorodow, M.; Landes, S.; Leacock, C. and Thomas, R.G. (1994). Using a Semantic Concordancer for Sense Identification. In *Proceedings of the ARPA Human Language Technology Workshop (ACL)*, Washington, pp. 240-243.

Mohammad, S. and Pedersen, T. (2004). Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, Boston.

Mooney, R.J. (1996). Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In *Proceedings of the Conference on Empirical Methods in NLP*, New Jersey, pp. 82-91.

Pedersen, B.S. (1997). *Lexical Ambiguity in Machine Translation: Expressing Regularities in the Polysemy of Danish Motion Verbs*. PhD Thesis, Center for Sprogteknologi, Copenhagen.

Procter, P. (ed.) (1978). *Longman Dictionary of Contemporary English*. Longman Group, Harlow, UK.

Ratnaparkhi, A. (1996). A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in NLP Conference*. University of Pennsylvania

Specia, L. (2005). A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation. In *Proceedings of the 8th Research Colloquium of the UK Special-interest Group in Computational Linguistics*, Manchester, pp. 71-78.

Specia, L.; Oliveira-Netto, S.; Nunes, M.G.V. and Stevenson, M. (2005). An Automatic Approach to Create a Sense Tagged Corpus for Word Sense Disambiguation in Machine Translation. In *Proceedings of the 2nd Meaning Workshop (Meaning-2005),* Trento, pp. 31-36.

Stevenson, M. and Wilks, Y. (2001). The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3):321-349.

Wilks, Y. and Stevenson, M. (1997). Sense Tagging: Semantic Tagging with a Lexicon. In *Proceedings of the SIGLEX Workshop 'Tagging Text with Lexical Semantics: What, why and how"*. Washington.

Yarowsky, D. and Florian, R. (2003). Evaluating Sense Disambiguation Across Diverse Parameter Spaces. *Journal of Natural Language Engineering*, 8(2):293-310.

Zinovjeva, N. (2000). *Learning Sense Disambiguation Rules for Machine Translation*. Master's Thesis in Language Engineering. Department of Linguistics, Uppsala University.