A Relational Approach for Word Sense Disambiguation

Lucia Specia[1], Maria das Graças Volpe Nunes[2], Mark Stevenson[3]

[1]Research Group in Computational Linguistics, University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
l.specia@wlv.ac.uk

[2]NILC - Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Caixa Postal 668, 13560-970, São Carlos, SP, Brazil
gracan@icmc.usp.br

[3]Department of Computer Science – University of Sheffield
Regent Court, 211 Portobello St., Sheffield, S1 4DP, UK
marks@dcs.shef.ac.uk

Abstract. We propose a novel approach for word sense disambiguation which makes use of corpus-based evidence combined with background knowledge. Using an inductive logic programming technique, it generates expressive models which exploit several knowledge sources and also the relations among them. The approach is evaluated in monolingual and multilingual tasks: identification of the correct English-Portuguese translation of verbs and disambiguation of verbs and nouns from official WSD competitions. The accuracy obtained in the multilingual task outperforms the alternative learning techniques investigated. The models also yielded significant improvement to the translation quality when integrated into a machine translation system. In the monolingual tasks, even though only some of our knowledge sources can be used and nouns are included, the approach performs as well as or very close to the state-of-the-art systems.

Keywords: Word Sense Disambiguation, Inductive Logic Programming.

1 Introduction

Word Sense Disambiguation (WSD) is concerned with the identification of the meaning of ambiguous words in context. For example, among the possible senses of the verb "run" are "to move fast by using one's feet" and "to direct or control". Sense ambiguity has been recognized as one of the most important obstacles to successful language understanding in a number of applications, such as Machine Translation and Question Answering.

A number of *knowledge-based approaches* have been proposed to this problem, making use of linguistic knowledge manually codified or extracted from lexical resources [1]. Recent approaches focus on the use of various lexical resources and corpus-based techniques, along with statistical or machine learning algorithms to induce disambiguation models and thus avoid the substantial effort required to codify linguistic knowledge [14], [22]. These approaches have shown good results, particularly those using supervised learning [2].

However, they rely on limited knowledge representation and modeling techniques: traditional machine learning algorithms and attribute-value vectors to represent disambiguation instances. Attribute-value vectors have the same expressiveness as propositional formalisms, that is, they only allow the representation of atomic propositions and constants. This has made it difficult to exploit deep knowledge sources in the generation of the disambiguation models, that is, knowledge that goes beyond simple features extracted directly from the corpus, like bag-of-words and collocations. For example, it is not possible to utilize relational information, such as semantic relations among the words in the sentence. As a consequence, the models produced reflect only the shallow knowledge that is provided.

In this paper we present a novel approach for WSD that follows a hybrid strategy, i.e. combines knowledge and corpus-based evidence, and employs a first-order formalism to allow the representation of deep knowledge about disambiguation examples together with a powerful modeling technique. This is achieved using Inductive Logic Programming (ILP) [7], which has not yet been applied to WSD. Our hypothesis is that by using a highly expressive representation formalism, a range of (shallow and deep) knowledge sources and ILP as learning technique, it is possible to generate models that, when compared to models produced by machine learning algorithms conventionally applied to WSD, are both more accurate for fine-grained distinctions, and more likely to convey potentially new knowledge, in a format that can be easily interpreted by humans.

WSD systems have generally been more successful in the disambiguation of nouns than other grammatical categories [6]. Disambiguation of verbs generally benefits from very specific knowledge sources, such as the verb's relation to other items in the sentence. We believe this is a task to which ILP is particularly well-suited. Therefore, we focus on the disambiguation of verbs, as opposed to most of the previous work.

WSD is usually approached as an independent task, however, it has been argued that different applications may have specific requirements. For example, in machine translation, WSD, or *translation disambiguation*, is responsible for identifying the correct *translation* of an ambiguous source word. This paper focuses on the application of our approach to the translation of verbs from English to Portuguese, although experiments with a monolingual task are also described.

In the remainder of this paper, we briefly introduce ILP and show how we apply this technique to WSD (Section 3) to then we describe our experiments and results (Section 4).

3 A hybrid relational approach to WSD

3.1 Inductive Logic Programming

Inductive Logic Programming [7] employs techniques from Machine Learning and Logic Programming to build first-order theories from examples and background knowledge, which are also represented by first-order clauses. It allows the efficient representation of substantial knowledge about the problem, which is used during the learning process, and produces disambiguation models reflecting this knowledge. The general approach underlying ILP can be outlined as follows. **Given:**

- a set of positive and negative examples $\boldsymbol{E} = \boldsymbol{E}^{+} \cup \boldsymbol{E}^{-}$
- a predicate $\boldsymbol{p}$ specifying the target relation to be learned

- knowledge $\boldsymbol{K}$ of the domain, described according to a language $\boldsymbol{L_k}$, which specifies which predicates $\boldsymbol{q_i}$ can be part of the definition of $\boldsymbol{p}$.

The goal is: to induce a hypothesis (theory or model) $\boldsymbol{h}$ for $\boldsymbol{p}$, with relation to $\boldsymbol{E}$ and $\boldsymbol{K}$, which covers most of the $\boldsymbol{E}^+$, without covering the $\boldsymbol{E}^-$, i.e., $\boldsymbol{K} \wedge \boldsymbol{h} \models \boldsymbol{E}^+$ & $\boldsymbol{K} \wedge \boldsymbol{h} \not\models \boldsymbol{E}^-$.

We use the Aleph ILP system [21], which provides a complete inference engine and can be customized in various ways.

3.2 Knowledge sources

An important step when designing ILP-based approaches is the appropriate identification, extraction and representation of relevant background knowledge for the problem. This is not a trivial process, but without carefully designed feature engineering the ILP characteristics that make it different from traditional learning algorithms cannot be truly exploited. The following sources of knowledge were automatically extracted from corpus and lexical resources and used by in our experiments. We used already existing NLP tools whenever possible, and implemented our own tools when necessary. We limit the context window to the size of the sentence containing the ambiguous word:

- **KS_1**. Bag-of-words consisting of 5 words to the right and left of the verb.
- **KS_2**. Frequent bigrams consisting of pairs of adjacent words in a sentence which occur more than 10 times in the corpus.
- **KS_3**. Narrow context containing 5 content words to the right and left of the verb, identified by the Mxpost Part-of-Speech (POS) tagger [13].
- **KS_4**. POS tags of 5 words to the right and left of the verb, given by Mxpost.
- **KS_5**. 11 collocations of the verb: 1st preposition to the right, 1st and 2nd words to the left and right, 1st noun, 1st adjective, and 1st verb to the left and right, also identified using Mxpost.
- **KS_6**. Subject and object of the verb, given by the Minipar parser [5].
- **KS_7**. Grammatical relations: verb-subject, verb-object, verb-modifier, subject-modifier, and object-modifier, as identified by Minipar.
- **KS_8**. The sense with the highest count of overlapping words in its dictionary definition and in the sentence containing the target verb, extracted from the bilingual dictionary *Password* [10], for the multilingual task, and from *Longman Dictionary* (LDOCE) [11], for the monolingual task.
- **KS_9**. Selectional restrictions of the verbs, defined in terms of the features required by its arguments, as extracted from LDOCE, e.g., the verb *come*, in the sense of *move toward*, requires an *animate* subject, and no object. If the restrictions imposed by the verb are not satisfied by its arguments, the features of synonyms and hyperonyms of these arguments – extracted from WordNet [4] – are also verified. A hierarchy of feature types is used to account for restrictions established by the verb that are more general than the features describing its arguments.

The following knowledge sources were designed for multilingual applications only:

- **KS_{10}**. Phrasal verbs potentially occurring in the sentence, identified using a list of phrasal verbs extracted from the same bilingual and monolingual dictionaries and simple heuristics to detect occurrences of separable and inseparable phrasal verbs

containing the verb under consideration.

- **KS_{11}**. Bag-of-words consisting of 5 Portuguese words to the right and left of the target verb in its sentence translation. This could be obtained using a machine translation system that would translate first the non-ambiguous words in the sentence. We extracted it using a parallel corpus.
- **KS_{12}**. Collocations consisting of 5 Portuguese words to the right and the left of the verb in its sentence translation.

Based on the examples, background knowledge and a series of settings specifying the predicate to be learned (i.e., the heads of the rules), the predicates that can be in the conditional part of the rules, how the arguments can be shared among different predicates and several other parameters, the inference engine produces a set of rules. Figure 1 shows examples of the rules induced for *come* in a multilingual task**.**

Models learned with ILP are symbolic and can be easily interpreted. Moreover, innovative knowledge about the problem can emerge from the rules learned by the system. For example, **Rule_1** states that the translation of the verb will be "chegar" (*arrive*) if it has a certain subject *B*, which occurs frequently with the word *today* as a bigram, and if the partially translated sentence contains the word "hoje" (the translation of *today*). **Rule_2** states that the translation will be "vir" (*move toward*) if the subject of the verb has the feature *animate* and there is no object, or if the verb has a subject *B* that is also a collocation C, in a position of a proper noun (*nnp*) or personal pronoun (*prp*).

```
Rule_1. sense(A, chegar) :-
  has_rel(A, subj, B), has_bigram(A, today, B),
  has_bag_trans(A, hoje).
Rule_2. sense(A, vir) :-
  satisfy_restriction(A, [animate], nil);
  (has rel(A, subj, B),
  (has collocation(A, C, B),
  (has pos(A, C, nnp); has pos(A, C, prp))).
```

Fig.1. Examples of rules produced for "come" in the multilingual task

4. Experiments and results

The model produced for each verb was tested by applying the rules in a *decision-list* like approach, i.e., retaining the order in which they were produced, using one rule at a time, removing all the examples covered by it from the test set, and backing off to the most frequent sense in the training set to classify cases that were not covered by the rules.

4.1 Multilingual task

For the first scenario, a corpus containing 5,000 sentences for 10 highly frequent and ambiguous verbs (500 for each verb) was extracted from corpora of different domains and genres, e.g., literary fiction and European Parliament proceedings. This corpus was semi-

automatically annotated with the translation of the verb using a tagging system based on parallel corpus, statistical information and translation dictionaries [20]. This tagging system outputs the most probable translation for each occurrence of the verb in the parallel corpus. It showed an average precision of approximately 82% in previous experiments, and thus we manually reviewed the automatic annotation. The sense repository of a verb was defined as the set of all the possible translations of that verb in the corpus. 80% of the corpus was used for training, and the remainder was retained for test. The verbs (and their number of senses in the corpus) are: ask (7), come (29), get (41), give (22), go (30), live (8), look (12), make (21), take (32) and tell (8).

The last column of Table 1 shows the accuracies (percentage of corpus instances which were correctly disambiguated) obtained by the ILP models. These are compared against the accuracy that would be obtained by using the most frequent translation in the training set to classify all the examples of the test set (*majority sense*). For comparison, we also experimented with three learning algorithms frequently used for WSD, which rely on knowledge represented by attribute-value vectors: C4.5 (decision-trees), Naive Bayes and Support Vector Machine (SVM). In an attempt to represent all knowledge sources in attribute-value vectors, KS_2, KS_7, KS_9 and KS_{10} were transformed into binary attributes. On average, the accuracy of the ILP approach is significantly better than the most frequent sense baseline and the other learning algorithms (paired t-test; $p < 0.05$). As expected, accuracy is generally higher for verbs with fewer possible translations.

Table 1. Accuracies in the multilingual task

Verb	Majority sense	C4.5	Naïve Bayes	SVM	Aleph
ask	0.68	0.68	0.82	0.88	**0.92**
come	0.46	0.57	0.61	0.68	**0.73**
get	0.03	0.25	0.46	0.47	**0.49**
give	0.72	0.71	**0.74**	**0.74**	**0.74**
go	0.49	0.61	**0.66**	**0.66**	**0.66**
live	0.71	0.72	0.64	0.73	**0.87**
look	0.48	0.69	0.81	0.83	**0.93**
make	0.64	0.62	0.60	0.64	**0.68**
take	0.14	0.41	0.50	0.51	**0.59**
tell	0.65	0.67	0.66	0.68	**0.82**
Average	**0.50**	**0.59**	**0.65**	**0.68**	**0.74**

The models produced by Aleph for all the verbs are very compact, containing 50 to 96 rules each. The various knowledge sources appear in different rules and therefore all of them seem to be useful for the disambiguation of verbs. Details about the experiments are presented in [18].

These results are very positive, particularly if we consider that: (1) the verbs are highly ambiguous; (2) the corpus was semi-automatically tagged, and sometimes distinct synonym translations were used to annotate different examples, but only one of these translations was considered to be correct for a given example; and (3) certain translations were very infrequent. It is likely that a less strict evaluation regime, such as one which takes account of synonym translations, would yield higher accuracies.

4.2 WSD for Machine Translation

Since the "senses" in the multilingual task are actually "translations", the quality of the models produced can be directly evaluated in any application involving translation, particularly Machine Translation (MT) itself. We investigated the contribution of the WSD models to Statistical Machine Translation (SMT), given the availability of such systems. Although it has been always thought that WSD can be useful for MT, only recently efforts have been made towards integrating both tasks to prove that this assumption is valid, particularly for SMT [3]. We propose a simple approach to efficiently integrate the use of rich contextual WSD features with standard SMT systems.

We used a phrase-based SMT system [12] in which candidate translations are scored according to a linear combination of feature functions. Our approach follows the n-best reranking technique proposed by [8], where a new feature (in this case, the WSD feature) is combined to the existing ones at translation time, as opposed to training time, to select the best scoring candidate translation from a list of n-best candidate sentences produced by the SMT system, the so called the *n-best list*. Given the procedure used to train standard SMT model parameters, using the n-best list reranking approach is considerably more feasible than adding the features at training time.

The original SMT system, which we call *baseline system*, has nine features, including the length of the translation, the probability of the translation given the source sentence, etc. The system was trained on a corpus of 700K English-Portuguese sentences extracted from several sources, mostly the European Parliament corpus. The estimation of WSD feature weight, as well as the re-estimation of the remaining feature weights, is performed using the n-best list of a 4K-sentence development set for which the sense annotation was available.

Based on the impact of the new feature in the SMT model and, as a consequence, the new global score produced by such model for each sentence in the n-best list, the candidate translations in that list can be reordered. For example, consider in Fig. 2 the top-2 candidate translations produced by the baseline SMT system for the sentence *s* (its reference translation being *r*) in the experiments with the translation of *ask*. The prediction given for this sentence by the WSD models is "perguntar" (*inquire*, *enquire*), but the top-scored sentence uses a different translation: "pedir" ("pediu-me") (*make a request*). The second candidate contains the correct prediction according to the WSD system, inflected as "perguntou". After the inclusion of the WSD feature, the second candidate becomes the top one.

s: He returned and *asked* me if I wanted anything else and whether I had enjoyed my meal.
r: Ele voltou, e **perguntou** se eu queria mais alguma coisa, se eu tinha gostado

Ele voltou, e **pediu-me** se eu queria mais alguma coisa e se eu tinha gostado.
Ele voltou, e **perguntou** se eu queria mais alguma coisa, se tinha gostado.

Fig. 2. Top-2 candidate translations for *s* as given by the SMT system

In order to assess the contribution of the WSD feature to the overall quality of the SMT system, we evaluate the system using an automatic evaluation metric, BLEU [9]. The score of the SMT system improved from 0.3248 to 0.34, which is statistically significant (paired t-test; $p < 0.05$). This improvement is comparable to that obtained by other approaches integrating WSD and SMT for other language pairs and datasets [3]. Details about the integration method can be found in [17].

4.3 Monolingual tasks – Senseval verbs and Semeval

For the monolingual scenario, we use the sense tagged corpus and sense repositories provided for verbs in Senseval-3 (www.senseval.org). There are 32 verbs with between 40 and 398 examples each. The number of senses varies between 3 and 10. The average accuracy obtained by Aleph 0.72, the same as the best performing system in the competition.

We also experimented with the monolingual dataset of SemEval-2007, which includes 35 verbs and 65 nouns. Results are detailed in [19]. Our system achieved an average accuracy of 85.1% and was ranked fourth place in that competition (out of 15 systems). An evaluation of the contribution of each knowledge source for the overall performance in this dataset can be found in [15].

Results are very encouraging for both datasets, considering that the system was not tuned for the monolingual task and, particularly, for the disambiguation of nouns.

4.4 ILP for feature construction

In [16] we present an alternative use of ILP for WSD. We examine the use of an ILP system as a method to construct a set of features from deep knowledge sources represented using first-order logic. The idea is to verify whether ILP systems could be used to improve the accuracy of WSD models induced from attribute-value representations. In essence, the predicates in the conditional part of rules learned by the ILP framework described in this paper are turned into binary features, filtered according to their coverage and accuracy, and then used by a common modeling technique (a support vector machine) to construct a classifier for predicting the sense of a word. Results are encouraging for monolingual and bilingual tasks: the ILP-assisted models show substantial improvements over those that simply use shallow features, and in some cases over the use of ILP as model learner. In addition, this procedure identifies smaller and better sets of features.

5. Conclusion

We have introduced a new hybrid approach to WSD which uses ILP to combine deep and shallow knowledge sources. ILP induces expressive disambiguation models which include relations between knowledge sources. It is an interesting approach to learning which had not yet been explored for WSD. Results from both multilingual and monolingual tasks demonstrate that the hypothesis put forward in this thesis, that ILP's ability to generate expressive rules which combine and integrate a wide range of knowledge sources is beneficial for WSD systems, is correct. Results for the multilingual task are validated in the experiments with the use of the WSD predictions in a machine translation system, yielding significant improvement in the translation accuracy.

By customizing the sense repository and knowledge sources, the proposed approach could be exploited for any other application requiring lexical disambiguation, particularly Information Extraction and Question Answering, both in monolingual and multilingual scenarios. Our goal for future work is to customize and integrate this approach to such applications.

References

1. Agirre, E., Rigau, G.: Word Sense Disambiguation using Conceptual Density. 15th Conference on Computational Linguistics, pp. 16--22, Copenhagen (1996)
2. Agirre, E., Marquez, L., Wicentowski, R.: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague (2007)
3. Carpuat, M., Wu, D.: Improving Statistical Machine Translation Using Word Sense Disambiguation. Empirical Methods in Natural Language Processing, pp. 61--72, Prague (2007)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Massachusetts (1998)
5. Lin, D.: Principle based parsing without overgeneration. 31st Meeting of the Association for Computational Linguistics, pp. 112--120, Columbus (1993)
6. Mihalcea, R, Chklovski, T. and Kilgarriff, A.: The Senseval-3 English Lexical Sample Task. Senseval-3: 3rd International Workshop on the Evaluation of Systems for Semantic Analysis of Text, pp. 25--28, Barcelona (2004)
7. Muggleton, S.: Inductive Logic Programming. New Generation Computing, 8(4):295--318 (1991)
8. Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., Radev, D.: A Smorgasbord of Features for Statistical Machine Translation. Human Language Technology, pp. 161--168, Boston (2004)
9. Papineni, K., Roukos, S., Ward, T. and Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation. 40th Annual Meeting of the Association for Computational Linguistics, pp. 311--318, Philadelphia (2002)
10. Parker, J., Stahel, M.: Password: English Dictionary for Speakers of Portuguese. Martins Fontes, São Paulo (1998)
11. Procter, P. (editor): Longman Dictionary of Contemporary English. Longman Group, Essex (1978)
12. Quirk, C., Menezes, A., Cherry, C.: Dependency Treelet Translation: Syntactically Informed Phrasal SMT. 43rd Annual Meeting of the Association for Computational Linguistics, pp. 271--279, Ann Arbor (2005)
13. Ratnaparkhi, A.: A Maximum Entropy Part-Of-Speech Tagger. Conference on Empirical Methods in Natural Language Processing, pp. 133--142, New Jersey (1996)
14. Schütze, H.: Automatic Word Sense Discrimination. Computational Linguistics, 24(1): 97--123 (1998)
15. Specia, L., Nunes, M.G.V., Stevenson, M.: Assessing the contribution of shallow and deep knowledge sources for word sense disambiguation. Journal of Language Resources and Evaluation, Springer Netherlands (2009)
16. Specia, L., Srinivasan, A., Ramakrishnan, G., Joshi, S., Nunes, M.G.V: An Investigation into Feature Construction to Assist Word Sense Disambiguation. Machine Learning, 76(1):109-136, Springer (2009)
17. Specia, L. Sankaran, B., Nunes, M.G.V.: n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. CICLing, pp. 399--410, Haifa (2008)
18. Specia, L., Nunes, M.G.V., Stevenson, M.: Learning Expressive Models for Word Sense Disambiguation. 45th Annual Meeting of the Association for Computational Linguistics, pp. 41--48, Prague (2007a)
19. Specia, L., Nunes, M.G.V., Srinivasan, A., Ramakrishnan, G.: USP-IBM-1 and USP-IBM-2: The ILP-based Systems for Lexical Sample WSD in SemEval-2007. 4th International Workshop on Semantic Evaluations, pp. 442--445, Prague (2007b)
20. Specia, L., Nunes, M.G.V., Stevenson, M.: Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. Conference on Recent Advances on Natural Language Processing, pp. 525--531, Borovets (2005)
21. Srinivasan, A.: The Aleph Manual. Technical Report. Computing Laboratory, Oxford University (2000)
22. Stevenson, M., Wilks, Y.: The Interaction of Knowledge Sources for Word Sense Disambiguation. Computational Linguistics, 27(3):321--349 (2001)