Translation Context Sensitive WSD

Lucia Specia[♦], Maria das Graças Volpe Nunes[♦], and Mark Stevenson[♣]

[♦] ICMC – University of São Paulo, Caixa Postal 668, São Carlos, 13560-970, SP, Brazil
[♣] DCS – University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK

{lspecia,gracan}@icmc.usp.br, M.Stevenson@dcs.shef.ac.uk

Abstract

While it is generally agreed that Word Sense Disambiguation (WSD) is an application-dependent task, the great majority of systems pursue application-independent approaches. We propose a strategy to support WSD for Machine Translation which is designed specifically for this application. It relies on the analysis of co-occurrences in the context that refer to words which have already been translated. Experiments on the English-Portuguese translation of 10 verbs using just this knowledge yielded an accuracy of 51%, which outperforms the baseline using the most frequent translation (37%). A less strict evaluation criterion considering the 10 best ranked translations proved the potential for this approach to be used as extra knowledge source for WSD: the correct translation was among the top 10 results in 92% of the cases.

1. Introduction

Word Sense Disambiguation (WSD) in Machine Translation (MT) is concerned with the choice of the most appropriate translation of an ambiguous word given its context. The question of whether WSD is useful for MT has recently been debated again. Vickrey et al. (2005), e.g., showed that a WSD module significantly improves the performance of their statistical MT system. Conversely, Carpuat and Wu (2005) claim that WSD does not yield significantly better translation quality than a statistical MT system alone. In this latter work, however, the WSD module was not specifically designed for MT.

In fact, although it has been agreed that multilingual WSD differs from monolingual WSD (Hutchins and Somers, 1992) and that WSD is ultimately a task which is only relevant in the context of a specific application (Wilks and Stevenson, 1998), little has been done on the development of WSD modules for particular applications. WSD models in general are application independent, and focus on monolingual contexts, particularly English. In the case of MT, the application we are dealing with, WSD approaches usually apply traditional sense repositories, such as the one provided by WordNet (Miller, 1990), to identify the monolingual senses, which are then mapped into the target language translations. However, mapping senses between languages is a very complex issue.

One of the reasons for this complexity is the difference in the sense inventories of the languages, as already discussed in (Hutchins and Somers, 1992) and recently evidenced by studies with certain pairs of languages. For example, Bentivogli et al. (2004) investigate the sense inventory discrepancies for English-Italian, Miháltz (2005), for English-Hungarian, Chatterjee et al. (2005), for English-Hindi, and Specia et al. (2006), for English-Portuguese. They show that there is not a one-to-one relation between the number of senses in the source language and their translations into another language. More specifically, they show that many source language senses are translated into a unique target language word, while some senses need to be split into different translations, conveying sense distinctions that only exist in the target language.

In addition to the differences in the sense inventory, the disambiguation process can vary according to the application. For instance, in monolingual WSD the main information is the context of the ambiguous word, that is, the surrounding words in a sentence or text. For MT purposes, however, the context may also include the translation in the target language, i.e., words in the text which have already been translated.

Although intuitively plausible, this strategy has not been explored specifically for WSD. On the other hand, some related approaches have exploited similar strategies for other purposes. For example, some approaches for MT, especially the statistics-based ap-

proaches, make use of the words which have already been translated as context, implicitly accomplishing basic WSD during the translation process (Egedi et al., 1994; Dagan and Itai, 1994). Some approaches for monolingual WSD use techniques which are similar to ours to gather co-occurrence evidence from bilingual corpus in order either to carry out WSD (Mihalcea and Moldovan, 1999), or to create monolingual sense tagged corpora (Fernández et al., 2004). Other monolingual related approaches somehow explore the already disambiguated or unambiguous words by taking into account the senses of the other words in the sentence in order to disambiguate a given word (Lesk, 1986; Hirst, 1987; Cowie et al., 1992).

In this paper we investigate the use of the translation context, that is, the surrounding words which have already been translated, as knowledge source for multilingual WSD. We present experiments on the disambiguation of 10 ambiguous verbs in English-Portuguese translation. The target language contextual information is applied by analysing the frequency of occurrence of the possible translations of the ambiguous word in text fragments found on the web by Google®, using n-grams and bags-of-words queries made of each translation along with other words in the target language context.

This investigation is part of an on-going project on WSD for MT, which makes use of resources and strategies specific to this application (Specia, 2005): parallel corpora (providing translations), bilingual dictionaries (as sense repositories), and this contextual information referring to target-words. The proposed approach also employs several knowledge sources exploring the source-language context, such as part-of-speech, syntactic relations and collocations. It learns a model in the form of a set of ordered rules from examples of translations, which are described by means of these knowledge sources. Knowledge provided by the strategy presented in this paper will be used to reinforce the available evidence.

In what follows we present our experimental settings (Section 2), the experiments carried out (Section 3) and the results achieved (Section 4).

2. Experimental setting

We limited our experiments to 10 verbs: seven highly frequent and ambiguous verbs previously identified as problematic for English-Portuguese MT systems: "to come", "to get", "to give", "to look", "to make", and "to take", along with other three also frequent but not so ambiguous verbs: "to ask", "to live", and "to tell". This set of verbs was also used in other experiments on WSD (Specia et al., 2005; Specia et al., 2006).

In order to experiment with these verbs we could use an MT system to produce the translation context. However, that would require adapting an MT system to our purposes. Since our intention is to investigate the feasibility of the strategy, we chose to use sentence aligned parallel corpora containing the ambiguous words to supply the translation context. This choice makes the experiment independent of translation methods and systems; in particular, it allows an evaluation which is not biased by the accuracy of an MT system. We assume the translations in our parallel corpus to be correct, since they were produced by human translators.

The parallel corpus consists of 100 English sentences for each of the 10 verbs (a total of 1,000 sentences) extracted from the corpus Compara (Frankenberg-Garcia and Santos, 2003), which is comprised of fiction books. In order to make the evaluation automatic, we use a version of this parallel corpus in which the translation of the ambiguous verbs has already been (automatically) annotated (Specia et al., 2005). For each occurrence of a certain verb, this corpus contains the English sentence, annotated with such translation, and the corresponding Portuguese sentence, as the example shown in Figure 1.

The strategy requires a list with all the possible translations of each verb. This list was extracted from bilingual dictionaries (e.g.: DIC Pratico Michaelis® and Password®), amounting to the numbers shown in Table 1. Our sample corpus includes translations of both phrasal and non-phrasal usages of the verbs.

Verb	# translations	Verb	# translations
ask	16	live	15
come	226	look	63
get	242	make	239
give	128	take	331
go	197	tell	28

Table 1. Verbs and their possible senses in the corpus

3. Obtaining information from the web

Our experiments explore translation information through the analysis of co-occurrence frequencies of n-grams and bags-of-words, created from the Portuguese words which have already been translated from the English source sentence and one possible translation of the verb under consideration. This is obtained by querying the web (via Google's API) with such n-grams or bag-of-words. In essence, an *n*-gram (or collocation) is a sub-sequence of *n* words from a

| tomar#I will **take** my medicines tonight, as prescribed by the doctor. |
| Vou **tomar** meus remédios hoje à noite, conforme indicado pelo médico. |

Figure 1. Example of parallel sentence in our corpus

given sequence of words in the translation context, including one possible translation of the ambiguous word. A bag-of-words is a subset of *m* words from the set of words in the translation context, including one possible translation of the ambiguous word, regardless of the order these words appear in the sentence (cf. examples in Table 2). In both cases, the relevant information is the number of hits (retrieved documents) provided by Google.

The web was chosen to be used as corpus to provide the statistical co-occurrence information because it is potentially the most representative corpus for Portuguese. However, the strategy could be applied using any monolingual corpus of the target language.

Since we are using the parallel corpus to provide translation context, all words in the target sentence, except the ambiguous one, can be used as context in the query. However, as previously mentioned, the parallel corpus is used here to simulate the environment that would be provided by an MT system. If we consider the use the approach in an MT system, the formulation of queries will vary depending on the translation approach. For example, if the MT system translates word-by-word, in the order they appear in the source sentence, the translation context can be constituted by the translations of the words in preceding positions in the source language sentence. Alternatively, if the system first translates all the possible words but the ambiguous one, any subset of the already translated words can be used as context. Other variations are possible for MT approaches translating chunks, translating all the words simultaneously, and translating the sentence based on the identification of the its main structure, usually given by the verb, for example.

In our experiments we consider a hypothetical rule-based transfer MT system which first translates all the non-ambiguous words in the sentence and then each ambiguous word - in the order they appear in the sentence - using the WSD module. In order to make it possible to use all the words in the sentence (except the ambiguous verb) as context, we assume that all these words will have already been translated, remaining only the verb to be disambiguated. Thus, our translation context could consist either of non-ambiguous words, which would have been already straightforwardly translated, or previously ambiguous words, which

would have been already disambiguated. Although any combination of words in that context could be used as a query, we limited the types of queries to the followings, each including one of the possible translations of the ambiguous verb (examples in Table 2):

(a) bigrams with the first word to the right of the verb;

(b) trigrams with the first two words to the right of the verb;

(c) trigrams with the first word to the right and to the left of the verb;

(d) n-grams with the first two words to the right and the first word to the left of the verb;

(e) bags-of-words with all the content words already translated in the sentence, requiring all of them to be included in the results;

(f) bags-of-words with all the content words already translated in the sentence, requiring any subset of the words to be in the results.

Two types of context are covered by the query sentences: (1) local context, give by n-grams, which consider a small window of surrounding words; and (2) topic context, given by bags-of-words, which consist of all the content words in the sentence.

Given the parallel corpus as exemplified in Figure 1, the automatic procedure to determine the queries is explained as follows (taking sentence in Figure 1):

1) Identify the ambiguous target word in the Portuguese side of the parallel corpus. Eg.: "Eu vou **target=tomar** meus remédios hoje à noite, conforme indicado pelo médico".

2) For every sentence, compose a set of queries for each type of bag-of-words and n-gram. Each query will contain one possible translation of the ambiguous verb. Stop words and other non-content words were eliminated from bags-of-words. Assuming that the verb "to take" has only three possible translations, "tomar" (consume, ingest), "pegar" (buy, select), and "levar" (take someone to some place), e.g., the queries that will be built for the sentence in Figure 1 are shown in Table 2.

3) Search every set of queries for a given sentence in Google, extracting the numbers of hits from the information provided by the search engine. The number of hits for our example is presented in the last column in Table 2[1].

---

[1] Obtained in 23/01/06.

Rule type	Queries	# Hits
(a)	"**tomar** meus"	324
	"**pegar** meus"	595
	"**levar** meus"	758
(b)	"**tomar** meus remédios"	228
	"**pegar** meus remédios"	17
	"**levar** meus remédios"	8
(c)	"vou **tomar** meus"	67
	"vou **pegar** meus"	320
	"vou **levar** meus"	913
(d)	"vou **tomar** meus remédios"	30
	"vou **pegar** meus remédios"	6
	"vou **levar** meus remédios"	0
(e)	+vou +**tomar** +remédios +hoje +noite +conforme +indicado +médico	10,300
	+vou +**pegar** +remédios +hoje +noite +conforme +indicado +médico	593
	+vou +**levar** +remédios +hoje +noite +conforme +indicado +médico	10,400
(f)	(**tomar** AND (remédios OR hoje OR noite OR conforme OR indicado OR médico))	3,500,000
	(**pegar** AND (remédios OR hoje OR noite OR conforme OR indicado OR médico))	2,290,000
	(**levar** AND (remédios OR hoje OR noite OR conforme OR indicado OR médico))	1,770,000

Table 2. Example of queries and their respective number of hits

4) For each type of query, choose the translation in the set of queries which achieves the maximum number of hits. In the example, the translation is "tomar", as correctly identified by queries (b), (d), and (f).

The rationale for experimenting with different query formulations was to identify the most suitable types of query for our purposes. After experimenting with the six types of query for a subset of five sentences of each verb, we concluded that, regarding local context queries, the potentially more appropriate queries are those including not larger contexts, particularly with more than one word to the right of the ambiguous word. This seems to be reasonable, since the ambiguous word is a verb, and thus the arguments in the object usually position play an important role. In fact, queries of type (a) are too short to identify the sense of the verb, while queries of type (b) presented good results.

The other n-grams (types (c) and (d)), which contain words to the left of the verb, did not yield good results. This was mainly due to the use of infinitive form of the possible translations. Portuguese verbs have different forms for tense, number, and person variations, and thus the use of the infinitive can result in grammatically inappropriate queries. The ideal would be to automatically identify the form of the verb in the translated sentence, and then generate all the possible translations in the same form. However, since there is not a Portuguese morphological analyser able to identify the form of a verb with a satisfactory accu-

racy, we used the infinitive form. Although this can be a drawback for any type of n-gram, it proved to be particularly problematic for queries including words to the left of the verb (i.e., (c) and (d)). For example, in the sentence in Figure 2, referring to the verb "to look", the target verb "examinando" is in the gerund tense. Queries of the types (c) and (d) with the infinitive form "examinar" would be grammatically inadequate.

Topical context queries in the form of bag-of-words are, in general, very flexible. They can be employed with most of the MT approaches: any non-empty subset of the already translated sentence can be queried together with each possible translation. With our preliminary experiments, we concluded that the type of query allowing any number of words to appear in the results (i.e., type (f)) is more appropriate. Since our corpus contains very long sentences[2], requiring all the words in the bag-of-words to be in the results (query of the type (e)) constrains the strategy, which will yield zero hits in most of the cases.

In summary, in these preliminary experiments with all types of queries for five sentences of each verb, results were particularly promising for queries of the types (b) and (f), and thus we limited our experiments with the whole corpus to these types of queries.

4. Results

In Table 3 we first show the accuracy of the baseline using the most frequent translation in the set of 100 sentences per verb: 0.37, on average. We assumed the

---

[2] Many sentences contain more than 100 words.

| examinar# Nizar was **looking** at X-rays of my knee, holding them up before a lighted screen. |
| Nizar estava **examinando** os raios X do meu joelho, segurando-os contra a tela iluminada. |
| query type (c): "estava **examinar**" ("was look") |
| query type (d): "estava **examinar** os" ("was look the") |

Figure 2. Example of grammatically inadequate queries

most frequent translation to be the one given first by the dictionary Password. As an additional baseline, in order to show that the number of hits provided by Google was indeed due to the formulation of the queries, and not only to the existence of the verb itself in the query, we searched Google with each of the possible translations alone. With the resultant estimation of the frequency of the verb when used alone, we calculated the accuracy of a baseline strategy which would assign the most frequent translation given by Google to all the sentences. As shown in the third column of Table 3, this accuracy (0.06, on average) is very low, and significantly smaller then the accuracy of the first baseline.

In Table 3.a we present the accuracies obtained in our experiments using queries of the type (d). The first column shows the percentage of sentences for which the query with the maximum number of hits contained the actual translation of the verb in that sentence. The other two columns show the percentages of sentences for which the query with the actual translation was included among the top 3 and top 10 positions, respectively, according to number of hits. In Table 3.b we show the corresponding accuracies for queries of the type (f).

In general, the results are considerably better for queries of the type (b). The problem with queries of the type (f) is that the long sentences produce queries consisting of many words and these are too general to accurately identify the sense of the verb. On average, the accuracy of the strategy for the first choice with queries of the type (b) (0.51) outperforms both baselines considering the most frequent sense (0.37 and 0.06). This result shows the potential for the strategy to be used as isolate and unique evidence to support WSD in MT. However, as previously mentioned, our goal is to use the knowledge provided by this strategy as additional evidence to our WSD approach, which already uses many other knowledge sources. In this sense, results in the second and third columns of Tables 3.a and 3.b could be exploited in many ways. For example, a number of translations considered as the most probable (the top 10, e.g.) could be used to filter the initial set of possible translations. In this case, the level of ambiguity would be significantly reduced, making the disambiguation problem much less complex. For example, it would be decreased from 331 to 10 translations, for the verb "to take". Alternatively, the ordering of the translations according to their corresponding number of hits could be used to assign weights to the possible translations and these could be employed to bias the use of the other knowledge sources when generating the WSD model.

In general, the results obtained can be considered satisfactory, if we take into account that our verbs are of very general use, having a great number of possible translations, and thus a great number of queries searched for each sentence. The fact that we also consider as possible translations those referring to English phrasal verbs increases the variety of translations, which also makes it more complicate to find the actual translation. Another important feature is that we are not grouping synonyms as classes of possible translations, since the existence of the synonym relation depends on the use of the verb, which makes it impossible to be automatically identified. It is also worth emphasizing that in many cases "relevant" documents were not retrieved by some queries since they were grammatically inappropriate, due to the lack of morphological analysis.

5. Conclusions

We described a strategy to support WSD in MT which is specific to this application. We experimented with 1,000 sentences containing 10 highly ambiguous verbs, searching Google with all the possible translations of the ambiguous verbs, by means of queries formulated as bags-of-words and n-grams.

In general, results were very promising, taking into account our experimental setting: general use and highly ambiguous verbs, the use of phrasal verbs as possible translations, and the lack of processing resources for Portuguese. The best results were achieved using queries with n-grams including two words to the right of the verb. When used as the sole knowledge source for WSD, the strategy outperformed the most frequent translation baselines. However, we consider that the great potential of the strategy is in its use as extra evidence to carry out WSD, together with other knowledge sources, rather than as the sole source of information.

Verb	Acc. most freq. T in Password	Acc. most freq. in Google
ask	0.3	0.1
come	0.4	0.1
get	0	0
give	0.8	0.1
go	0.4	0
live	0.6	0
look	0.4	0.1
make	0.8	0
take	0.2	0
tell	0.3	0.2
Aver.	**0.37**	**0.06**

Table 3.

Acc. 1st choice	Acc. 1st to 3rd choice	Acc. 1st to 10th choice
0.5	0.8	1
0.4	0.4	0.8
0.4	0.7	0.7
0.5	0.7	0.8
0.5	0.9	1
0.6	0.9	0.9
0.6	0.9	1
0.3	0.6	1
0.8	1	1
0.5	0.9	1
0.51	**0.78**	**0.92**

a. Accuracy for query type (b)

Acc. 1st choice	Acc. 1st to 3rd choice	Acc. 1st to 10th choice
0	0.6	1
0.2	0.2	0.2
0	0	0.2
0.4	0.8	1
0	0	0
0	0	1
0.4	0.4	0.4
0	0	1
0	0.2	0.4
0.4	0.6	0.8
0.14	**0.28**	**0.6**

b. Accuracy for query type (f)

References

Bentivogli, L., Forner, P., and Pianta, E. (2004). Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus. COLING, Geneva , pp. 364-370.

Carpuat, M. and Wu, D. (2005). Word sense disambiguation vs. statistical machine translation. 43rd ACL, Ann Arbor, pp. 387–394.

Chatterjee, N., Goyal, S., and Naithani, A. (2005). Pattern Ambiguity and its Resolution in English to Hindi Translation. RANLP, Borovets, pp. 152-156.

Cowie, J., Guthrie, J.A., and Guthrie, L. (1992). Lexical Disambiguation Using Simulated Annealing. COLING, Nantes, pp. 359-365.

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

Egedi, D., Palmer, M., Park, H.S., and Joshi, A.K. (1994). Korean to English Translation Using Synchronous TAGs. 1st AMTA, Maryland, pp. 48-55.

Fernández, J., Castilho, M., Rigau, G., Atserias, J., and Turmo, J. (2004). Automatic Acquisition of Sense Examples using ExRetriever. LREC, Lisbon, pp. 25-28.

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.

Hutchins, W.J. and Somers H.L. (1992). An Introduction to Machine Translation. Academic Press, Great Britain.

Lesk, M. (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. SIGDOC, Toronto, pp. 24-26.

Mihalcea, R. and Moldovan, D.I. (1999). A Method for Word Sense Disambiguation of Unrestricted Text. 37th ACL, Maryland.

Miháltz, M.(2005). Towards A Hybrid Approach to Word-Sense Disambiguation in Machine Translation. Workshop Modern Approaches in Translation Technologies, RANLP, Borovets.

Miller, G.A., Beckwith, R.T., Fellbaum, C.D., Gross, D., and Miller, K. (1990). WordNet: An On-line Lexical Database. International Journal of Lexicography, 3(4):235-244.

Specia, L. (2005). A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation. 8th CLUK, Manchester, pp. 71-78.

Specia, L., Nunes, M.G.V., Stevenson, M. (2005). Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. RANLP, Borovets, pp. 525-531.

Specia, L., Castelo-Branco, G., Nunes, M.G.V., Stevenson, M. (2006). Multilingual versus Monolingual WSD. To appear in Workshop Making Sense of Sense, EACL, Trento.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. HLT/EMNLP, Vancouver.

Wilks, Y. and Stevenson, M. (1998). The Grammar of Sense: Using Part-of-speech Tags as a First Step in Semantic Disambiguation. Natural Language Engineering, 4(1):1-9.