Building a WSD Module within an MT system to enable interactive resolution in the user's source language

Constantin Orasan*, Ted Marshall[+], Robert Clark[+], Le An Ha[*], Ruslan Mitkov*

[*] Research Group in Computational Linguistics,
University of Wolverhampton, UK
[+] Translution, UK
C.Orasan@wlv.ac.uk, Ted.Marshall@translution.com, Rober.Clark@translution.com,
L.A.Ha@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract. Ambiguous words pose very serious problems to existing machine translation systems. The Translation Checker, a system part of Translution Central addresses this problem by allowing users to disambiguate words in their own language, with little or no knowledge of the target language. In order to achieve this, a multilingual dictionary is being developed using EuroWordNet. Languages are too ambiguous to feasibly present users with all the senses available for a word. To this end, a suite of language processing modules has been developed to reduce the ambiguity of words. The implemented modules and an evaluation of their influence on English, French, German, Italian and Spanish corpora are presented. The results of the evaluation show that the proposed approach dramatically reduces the ambiguity of the language.

1. Introduction

Ambiguous words pose a very serious problem to existing machine translation systems because in many situations the translation engines do not know how to handle these words. This problem can be particularly serious for organisations which heavily rely on machine translation for their everyday operation. The work presented in this paper is part of a larger project to develop technologies which will enable people and organisations to improve communications by removing language barriers.[1] The technology is based on automatically redirecting e-mails, web pages and electronic documents to a centrally based translation facility termed *Translution Central*. Users simply write emails in their own language in the normal way, press the Send button, and the recipients will receive the email automatically translated into their own language. Similarly, incoming emails will be translated into the user's own language.

The initial release of the product will support five European languages: English, French, German, Spanish and Italian. Translution has developed three different product suites, aimed at different sectors of the market, *Translution Light*, *Translution Pro* and *Translution Corporate*[2].

This paper presents the Translation Checker, a tool integrated in the Translation Central which enables users to specify the meaning of polysemous words without the need of knowing how they translate in the target language or defining the domain of the source document. The paper is structured as follows: Section 2 discusses the structure of the translation checker. An evaluation of the system is presented in Section 3. The paper finishes with conclusions. [3]

[1] This project was initiated by Translution in 2002.

[2] More information about these products can be found at http://www.translution.com

[3] This is a collaborative work between Translution and the University of Wolverhampton the main objective being to develop tools which allow users to improve Machine Translation quality designed with the non-linguist in mind.

2. Translation Checker: a tool to deal with ambiguous words in MT

Translation Checker is a product designed to help the translation process by allowing users to specify which meaning of an ambiguous word is to be used, without the need of knowing how to translate it in the target language or the need to define domain-dependent meanings. In order to work, this product needs to have access to a multilingual dictionary that allows a user to obtain a definition for any word in a text. On the basis of this definition and of the information present in the multilingual dictionary, the user can indicate which meaning is to be used, thereby producing its translation in the target language without the necessity of knowing the word in the target language. For example for the verb *to address* the following three definitions will be displayed:

- to speak to someone formally
- to put an address on something
- to deal with something particular

but the user will not need to know that in French the first sense in translated by *s'adresser à*, the second one by *mettre une adresse*, whilst the third one by *traiter*.

Because it was noticed that natural language is highly ambiguous[4], it is not feasible to require the marking of all words with their senses, and ways to automatically reduce the number of ambiguous words have to be identified. In this section, the main features of the multilingual dictionary and the tools necessary to reduce this ambiguity are presented.

2.1. Multilingual dictionary

The dictionary used by the Translation Checker is based on the EuroWordNet, a multilingual database with wordnets for several European languages developed between 1996 and 1999 with funding from the European Union (EuroWordNet). The wordnets are structured in the same way as the American WordNet (WordNet) in terms of synsets (sets of synonymous words) with basic semantic relations between them. In addition, these wordnets are linked to an Inter-Lingual-Index (ILI), based on the American WordNet. Via this index, the languages are interconnected so that it is possible to go from words in one language to similar words in any other language. In addition to the synonymy relations, the wordnets also contain a large number of other relations, such as hypernymy (i.e. more general concepts), hyponymy (i.e. more specific concepts), etc. Because these relations cannot be directly used in the translation process, it was decided not to include them in the multilingual dictionary. The reason for employing WordNet as a basis for the dictionary is because work has been either performed or is currently going on a wide number of languages offering scalability to the proposed method.

Even though the languages in the EuroWordNet are supposed to be linked via a language independent index, because this index is based on the American WordNet, it inherited all its weaknesses. For this reason, soon after we started this project it became obvious that work was required to maximise the usefulness of the dictionary. After investigating the ILI (i.e. the English WordNet) it became clear that it can be improved in the following ways.

- Remove senses which are too specific, too rare or too obscure (e.g. the verb *accept* with the meaning *be sexually responsive*, the adjective *dark* with the meaning *in a state of intellectual or social darkness*, the noun *account* with the meaning *turned her writing skills to good account*)
- Conflate senses which are too close for the everyday user (e.g. two senses of the verb *gather*: *believe to be the case* and *conclude from evidence* were conflated in one sense to *understand/believe something even though it has not been explicitly stated*)
- Add missing senses (e.g. for the noun *gate* there is no meaning for the place *where you board a plane at the airport*)

In addition to the work undertaken on the ILI, a further step which needed to be taken for all five languages was to provide definitions for the words. The English WordNet has a large number of glosses which were provided by the

[4] Using the English and French WordNets it was determined that for the top 1000 English words the averages number of senses per word is 7.83, whilst for the top 1000 French words is 4.62.

	English	French	German	Italian	Spanish
Synsets	91803	22417	10284	14967	28066
Nouns	60647	17528	7594	11537	24047
Verbs	11597	4892	2688	1653	4019
Adjectives	16491	0	4	1573	0
Adverbs	3263	1	0	206	0

Table 1: Total number of synsets and words in WordNets

lexicographers in order to facilitate its creation, but in many cases these glosses are not appropriate as definitions (e.g. for the verb *accumulate* the gloss is *Journals are accumulating in my office* and therefore it was replaced with the definition *to (be) collect(ed) or gather(ed) over a period of time*). The rest of the WordNets do not have definitions attached to words, and therefore have to be introduced from scratch.

The third type of work which needed to be performed on the WordNet is to enrich the WordNets for languages other than English. Investigation of these wordnets revealed that the quantity of information varies enormously from one language to another. Table 1 presents the number of synsets present in different languages. As a result of this finding, it became evident that in order to have a high quality resource, it is necessary to have a similar number of synsets across languages. In addition it was necessary to add adjectives and adverbs for French, German and Spanish.

The work necessary to improve the multilingual dictionary was undertaken at the University of Wolverhampton and it involved all the steps described above. For each language, native speakers were employed to perform the described steps. In order to speed up the production of definitions in languages other than English, the English definitions provided by our English expert were automatically translated and presented to the other language experts. This approach had limited success because only in a few cases the definitions were correctly or nearly correctly translated. There are several explanations for this. First of all, many of the words in the definitions are ambiguous which makes the translation quite difficult. In addition, many of the definitions are not grammatical sentences, making them difficult to translate even for humans.

2.2. Implementation of the language processors

As aforementioned, presenting all the alternative meanings is not a practical solution to deal with ambiguous words because of the high ambiguity some words exhibit. This problem was addressed by implementing several language processing filters which reduce the ambiguity. At present the filters implemented in the system are:

- Part-of-speech taggers
- Named entity recognisers
- Identification of multiword units
- Cross-lingual references
- Document and domain sense selection

Each of these filters is presented in detail in the remaining of the section.

2.2.1. Part-of-speech taggers

Part-of-speech tagging is the process of assigning labels to words which indicate their grammatical category. In the context of the WSD project, this information is important for two reasons:

- The use of part-of-speech enables us to reduce the number of senses possible for a word. For example, the word *bank* has 10 senses as a noun and 8 senses as a verb. In a sentence like *He banks the money*, when using POS tagging, the word *bank* will be identified as a verb, and the number of meanings the users have to choose from is reduced by **10**.[5]
- In computational linguistics, part-of-speech information is considered basic information which is widely used to improve the performance or make possible other tasks such

[5] The numbers reported in this and next section use EuroWordNet and not the enriched dictionaries.

as named entity recognition and identification of multiword units.

The part-of-speech taggers implemented in the Translation Checker are based on Hidden Markov Models (HMM) which confers them language independence. On the basis of the error analysis, a set of rules which correct frequent errors of the part-of-speech tagger has been written for each language. Examples of rules are:

```
THE ADJ X OF → X is tagged as NOUN
ADJ COMMA X AND ADJ → X is tagged as ADJ
```

2.2.2. Named entity recognition

Named entity recognition is the task which identifies whether sequences of words refer to entities that have special meaning (e.g. names of people, locations, organisations, etc.) This information is important for this project for two reasons:

- Named entities contain words which have several senses, but their senses should not be shown to the user because they do not need disambiguation in this context. In a sentence such as *Bill Gates is the youngest multi-billionaire in the history of the United States*, if we can identify *Bill Gates* and *United States* as Named Entities, we can eliminate 13 senses of *bill*, 7 senses of *gate*, 8 senses of *united* and 11 senses of *states* (this eliminates **39** senses in total).
- The identification of named entities is important for the machine translation process because they either should not be translated, or when they are translated, this has to be done using special approaches such as table lookup (Babych and Hartley, 2003)

In the context of this project, it is not necessary to perform complete named entity recognition. It is enough to identify them, without determining their type. Sometimes this task is referred to as *normalisation*. A language independent named entity engine has been implemented in order to facilitate the identification of named entities in different languages. This engine relies on language specific gazetteers and language specific rules.

2.2.3. Multiword units

Multiword units are sequences of tokens which have a different meaning than the individual parts which constitute them. Identification of the multiword units can reduce the number of choices a user of the Translation Checker is making. For example, if the system can identify multiword units such as *prime minister*, *earnings per share*, there is no need to disambiguate component words (*prime*, *minister*, *earnings*, *share* eliminating **30** senses in total). At present only the multiword units which are nouns are identified and dealt with, but in the future it is intended to tackle other types of multiword units.

2.2.4. Cross-language reference

After analysing the data, it was noticed that there are cases where all the senses of a word in a language are translated using the same word in the target language (for example all the senses of the word *opponent* can be translated to French using the same word *adversaire*). In this context, even if the word is ambiguous, it is not necessary to ask the user of the system to disambiguate it, because the word which translates it can be accurately determined. This process is referred to as *cross-lingual reference*. The approach can be extended when there are several target languages.

2.2.5. Document and domain sense selection

The filters presented in Sections 2.2.1–2.2.4 try to reduce the number of senses which are presented to a user for a word. In addition to these filters, two others which do not reduce the number of senses which are presented to the user, but which influence the way they are displayed to the user, and therefore help the decision process, were tried.

In the *one sense per discourse* setting, the Translation Checker assumes that the user uses only one sense in the document, and when users choose a specific sense for a word, all the appearances of the same word in the text after the marked one will be automatically considered with the same sense. Given that this is not always the case the users can override this sense.

The subject of the text is another way to help the user in the disambiguation process. The *same subject prioritisation* determines which of the senses will be displayed higher in the list of

	English	French	German	Italian	Spanish
Synsets	9260	6459	7299	5410	5541
Nouns	4579	3219	3632	2757	2816
Verbs	2897	2043	2331	1705	1733
Adjectives	1678	1042	1150	826	884
Adverbs	206	155	186	122	108

Table 2: Total number of synsets in multilingual dictionary

	English	French	German	Italian	Spanish
Nouns	3816	3036	3329	3388	3350
Verbs	2706	2069	1951	2336	2183
Adjectives	1671	1258	965	1242	1077
Adverbs	233	193	177	221	166

Table 3: The number of words in the multilingual dictionary

senses on the basis of senses selected for other words. This process relies on an annotated version of the English WordNet, where all the synsets were annotated with a subject label (Magnini and Speranza, 2002). The ontology of senses used in the annotation is quite large, but we decided to keep only the 51 more general subjects.

3. Evaluation of the reduction techniques

The main aim of the evaluation was to automatically determine the effectiveness of elimination techniques implemented in the TranslationChecker. In addition, a small scale user-focused evaluation was performed on English to French translations. The experiments were performed on an improved version of the multilingual dictionary. Tables 2 and 3 present the number of words and synsets of the multilingual dictionaries. As can be seen the number of words and synsets is much lower than the ones which were in the original dictionary. The reason for this is that in the current dictionary we included only the words which have been checker. However, the words to be included in the dictionary were selected in such a way that they are the most frequent ones and ensure over 80% coverage of texts. In future we plan to continue working on the dictionary to enlarge it.

3.1. Automatic evaluation

In order to assess the influence of the filters on the number of senses available to a user, the system was run several times, each time switching on an additional filter. For the experiments corpora of 100,000 words per language have been extracted from the Multilingual Corpora for Co-operation (MLCC) distributed by ELDA. These corpora contain newswire texts in English, German, French, Spanish and Italian. Table 4 presents the number of words from the corpora which appear in dictionary and total number meanings for these words when no filtering is considered. As can be seen, for all the languages but German, more than 40% of the words can be processed by the TranslationChecker. The lower value for German will have to be investigated further. The remaining words are either closed class words such as prepositions, articles, conjunctions, or are unknown words and have to be discarded.

Language	# ambiguous	# meanings
English	41,745	185,086
French	40,631	157,278
German	23,841	56,821
Spanish	42,959	158,522
Italian	41,489	142,850

Table 4: Number of ambiguous words

The reduction of average number of senses per dictionary word (i.e. word which appears in our dictionary) is presented in Table 5. The columns of the table correspond to the different languages which can be processed by the TranslationChecker, whilst the rows correspond to different filters: *Nothing* when no filtering is applied, *+POS* when the part-of-speech tagger is used, *+NE* when the named entity is switched

	English	French	German	Spanish	Italian
Nothing	4.43	3.87	2.38	3.69	3.44
+POS	2.82	2.72	2.03	2.69	2.07
+POS+NE	2.80	2.72	1.97	2.47	2.03
+POS+NE+MWU	2.79	2.72	1.97	2.44	2.01
+POS+NE+MWU+ST	2.77	2.67	1.87	2.38	1.97

Table 5: The average number of senses per dictionary word

	EN	FR	DE	ES	IT
EN	-	2.55	2.68	2.59	2.53
FR	2.50	-	2.63	2.56	2.52
DE	1.67	1.80	-	1.75	1.75
ES	2.25	2.31	2.35	-	2.26
IT	1.78	1.81	1.94	1.84	-

Table 6: The average number of sense when the *Same translation* module is used

on, *+MWU* when the multiword units are considered, and *+ST* when the same translation module is turned on. Each of the modules is applied on the top of the other.[6]

For the same translation the results reported in Table 5 correspond to the situation when for each of the other four languages is possible to find one word which can be used in the translation for all the senses. A bigger reduction in the average number of senses as a result of the *Same translation module* is obtained when only one target language is considered. The results in this situation are presented in Table 6. The rows of correspond to the source language, whilst the columns to the target language. As can be seen the results vary a lot from one pair of languages to another.

Investigation of Table 3 reveals that part-of-speech tagging leads to a massive reduction in the average number of senses. The named entity recogniser has quite a small influence, but closer investigation of the corpora indicated that they do not contain a large number of named entities. Identification of multiword units proved more beneficial than named entity recognition, a result which was not expected, but which can be justified by the nature of the corpora. A small but useful reduction was achieved by the same translation module. As seen in Table 6 this reduction is larger when only one target language is used.

3.2. User-focused evaluation

In order to see how users find the Translation Checker, a user-focused experiment was conducted. In this experiment, the user was asked to use the Translation Checker from English to French with different settings on several small texts.[7] The main purpose of this experiment was not to record the time necessary to annotate the text with senses, but to get feedback about how the user feels while using the program.

The best combination determined empirically by the user was *POS+NE+MWU+One sense per discourse*. The same domain prioritisation did not prove useful because, as a result of constantly changing the place of a definition in the list according to the current domain, the user was confused. The *One sense per discourse* module did not prove as accurate as expected (i.e. there were quite a few texts where the same word was used with more than one sense), but overall, it reduced the time necessary to process texts.

[6] Actually it is not possible to apply the named entity recogniser without running the part-of-speech tagger. For this reason it was not possible to report the influence of individual modules on the reduction in the number of senses displayed.

[7] Actually for this experiment we did not use the TranslationChecker, but a tool which replicates its functionality, but it is not integrated in the TranslationCentral.

4. Conclusions and future plans

This paper has addressed the problem of polysemous words in machine translations by proposing the Translation Checker, a tool which relies on a multilingual dictionary and a series of natural language filters to help users disambiguate such words. Evaluation conducted on English, French, German, Spanish and Italian has revealed that each of the proposed filters help reducing the ambiguity.

For future we plan to continue enriching the dictionary in order to include more words. We also plan to continue the evaluation in several directions. The first one will focus on the evaluation of each individual module included in the TranslationChecker in order to find out its influence on the overall success of the system. In addition, evaluation of the impact of the TranslationChecker on the quality of the translation will also be investigated. Given that the TranslationChecker is part of a commercial product will enable us to conduct evaluations from the point of view of the user of the system.

5. References

BABYCH, B; HARTLEY, A. (2003) Improving Machine Translation quality with automatic Named Entity recognition. In: EACL 2003, 10th Conference of the European Chapter. Proceedings of the 7th International EAMT workshop on MT and other language technology tools. April 13th 2003, Budapest, Hungary. Pp. 1-8

EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

MAGNINI B., SPERANZA M. (2002), Merging Global and Specialized Linguistic Ontologies, ITC-irst, June 2002, 6 pp. Published in Simov K. (ed.), Proceedings of Ontolex 2002 Ontologies and Lexical Knowledge Bases, Workshop held in conjunction with LREC-2002, Las Palmas, Canary Islands, Spain, May 27-31, 2002, pp. 43-48.

WordNet: http://wordnet.princeton.edu/w3wn.html