WSD evaluation and the looking-glass

Elisabeth Aimelet, Veronika Lux, Corinne Jean and Frédérique Segond

Xerox Research Centre Europe
6, chemin de Maupertuis,
38240 Meylan France

1. Introduction

Human beings use natural language to communicate with their pairs. They want to use it to communicate with machines as well. But because natural language is ambiguous by nature, Word Sense Disambiguation (WSD) is a crucial research topic for Human Language Technologies (see (Ide and Véronis 98)). WSD is necessary in most natural language applications (e.g. information retrieval, machine translation) and essential in any language understanding application.

Within XRCE, we have developped a multilingual comprehension system based on the Oxford Hachette French-English dictionary (OUP-H))[1]. It integrates a component called SDL (Semantic Dictionary Look-up) that uses dictionary information to perform automatic WSD. SDL works on all words of the OUP-H dictionary, with a methodology reusable for any language with existing on-line dictionaries.

SDL has been evaluated, for French, within the Romanseval exercice [2]. Here, we present two evaluations of the SDL for French: the one of Romanseval and an internal bilingual evaluation. We conclude with a more general discussion on some methodological issues about the evaluation of WSD systems.

2. What do we evaluate?

2.1. Semantic dictionary look-up: goal, architecture and components

SDL is an all word desambiguation system that attempts to select the most appropriate translation of a word appearing in a given context. The dictionary entry associated to this word is reordered so that the prefered translation appears first. SDL is built on top of Locolex[3], an intelligent dictionary look-up which achieves some WSD using word's context (part-of speech and MultiWord Expression (MWEs)[4]. recognition). However, Locolex choices remain syntactic.

SDL [5] goes one step further towards semantic disambiguation by using information from the

[1] See (Oxford 94).

[2] See http://www.lpl.univ-aix.fr/projects/romanseval

[3] See (Bauer et al. 95).

[4] Multiword expressions range from compounds (e.g. salle de bain) and fixed phrases (e.g. a priori) to idiomatic expressions (e.g. to sweep something under the rug).

[5] A full description of the SDL can be found in (Segond et al. 98).

OUP-H that is:
- information about *subcategorization*. At a syntactic level, subcategorization frames encode the prototypical complements of predicates (e.g. the verb *présenter* allows a pronominal construction, it is therefore associated in the OUP-H, with the subcategorization frame *vpr*).
- information about *collocates*. At a semantic level, collocate are used to encode prototypical subjects and/or objects of predicates (e.g. in its pronominal subcategorization frame, the verb *présenter* has *difficulté* as subject collocate). In the OUP-H collocates are usually given as a list of words, sometimes as concepts.

SDL also relies on information (mainly functional) given by the XIFSP [6] shallow parser. XIFSP adds syntactic information at sentence level in an incremental way, depending on the contextual information available at a given stage. In particular, it allows automatic recognition and extraction of subject and object dependency relations. For instance, in the sentence: *Des difficultés se* **présentent** *lorsque l'entreprise d'assurance n'exerce ses activités qu'en régime de libre prestation de services.*[7] . XIFSP provides SDL with the following information: *difficulté* is the subject of the reflexive usage of the verb *présenter*.

Information extracted by XIFSP is then matched against the information encoded in the OUP-H. If matches are found (e.g. both for the *vpr.* subcategorization frame and for the collocate *difficulté*), SDL reorders theOUP-H entry for *présenter* and first proposes the translation *to arise, to present itself*[8].By default, if there no match is found, SDL returns the first translation/sense of the OUP-H. This is done for each word of any input sentence,

2.2. Rationale for the evaluation exercises

We evaluated SDL, for French, within the Romanseval competition. This monolingual evaluation has been achieved on 60 words (20 verbs, 20 nouns and 20 adjectives). The gold standard was a corpus semantically tagged by humans.
In addition, we performed an internal bilingual evaluation on the 20 verbs chosen in Romanseval[9]. In this case, the gold standard was a semi-automatically aligned bilingual corpus developped in the MULTEX project [10].

The first goal of the evaluation was for us to understand the SDL's strengths and weaknesses and to identify possible improvements. The second goal was to gradually build up a methodology for evaluating WSD systems.

3. How do we evaluate?

3.1. Gold standards

Within Romanseval, as described in (Segond 98), six human informants were asked to semantically tag a corpus (made out of excerpts of the Official Journal of the European Community) in order to create the gold standard. The semantic tags were the senses distinctions of the monolingual French dictionary Le Petit Larousse ((Larousse 95)) for the 60 words chosen for the evaluation. Each of these words (20 verbs, 20 nouns and 20 adjectives) appeared in 50

[6]The French Incremental Finite State Parser is developed at our Research Centre.
[7]Difficulties *arise* when the insurance company exercises its functions as...
[8]In case of information conflict between subcategorisation and collocates, priority is given to collocates.
[9]This evaluation was on verbs exclusively since OUP-H above all provides information for verbs
[10]This alignment had been built within the Arcade project (See http://www.lpl.univ-aix.fr/projects/arcade)

different contexts[11] which yielded to 3000 contexts to be manually sense-tagged. Annotators were instructed to chose either zero, one, or several senses for each word in each context [12].

In the second evaluation, the *gold standard* was a word aligned bilingual corpus (French-English) established by two human annotators starting from a bilingual paragraph alignment[13].

3.2. Evaluation process and scoring

In Romanseval, for each test item, the sense tag (of the Petit Larousse) that SDL had selected is checked against the gold standard, as described in (Segond 98). Roughly the adopted metrics were as follows:
- *Agree* which counts agreement when matches at least one human sense, weighted by the number of proposed senses: $\frac{(human \cap system)}{system}$
- *Kappa* which is the same, corrected for chance agreement In the second exercise as there was no reference dictionary involved in the tagging phase, we kept SDL as it were, that is, using the OUP-H bilingual dictionary. For each test item, we checked the English translation selected by SDL (in OUP-H) against the gold standard.

3.3. Results of the evaluations

In Romanseval (see (Segond 98)), precision and recall for the SDL system were the following:

POS	Precision	Recall
Adjective	0.49	0.56
Noun	0.43	0.44
Verb	0.29	0.32

A close study of verbs results shows that low results are mainly due to dictionary mapping and recognition of MWEs (see 4.1 below). SDL tagged 715 verbs out of 1502 verb occurrencies. Among these 715 tagged verbs, 400 were tagged using MWEs' information[14]. They were hence considered *wrong* answers in the Romanseval evaluation though, among the 400 verbs tagged as MWEs, 279 were properly recognized.

For the bilingual evaluation, in 341 cases among 1102, SDL was considered *right*, either because it provided the same translation as the one of the alignement (212 cases), or because it provided a translation for a correctly identified MWE [15].

Both evaluations show that, in conformity with its specification, the system succeeds when it recognizes: - an expression encoded as MWE. Identification of MWEs account for 55% of the correct answers,
- a subcategorization frame that allows or helps disambiguation. Identification of a subcategorization frame alone accounts for around 20% of the correct answers. - a *collocate*. Identification of both a subcategorization frame and a collocate accounts for around 20%. The remaining

[11] a context is a paragraph of one or several sentences

[12] Question mark were used when none of the senses matched the given context. They were treated as an additional sense for each word, grouping all meanings that were not found in the dictionary.

[13] In this alignement, French was not always the source language.

[14] For example, the verb *exercer* in *la Commission peut-elle dire si elle entend* exercer *des pressions sur les autorités grecques?* is tagged as part of the MWE *exercer une pression sur quelqu'un* and translated with *to put pressure on sb/sth*

[15] For exemple, for the verb *entrer* in the sentence *Le nouveau rgime transitoire de TVA doit entrer en vigueur dans la Communaut le 1er janvier 1993.*, the alignement gives *to enter* and the system *to come into force*

5% are cases where the default solution happened to be a correct translation.

4. Discussion on the evaluation methodology

These results are especially interesting because they throw lights on a many issues related to the evaluation exercice.

4.1. Was are the consequences of sense mapping in Romanseval?

Because SDL uses a bilingual dictionary, participating to Romanseval meant to *map* sense distinctions of the two dictionaries (the Petit Larousse and the OUP-H) for the 60 chosen words. *Mapping* (a task similar to corpus sense taging) consisted in assigning a Petit Larousse sense tag to a OUP-H sense (that is usually illustrated by a bunch of examples). This task raised a number of issues as sense distinctions are, of course, different in both dictionaries. First, both dictionaries usually do not distinguish the same number of senses for each word considered.[16]. Clearly, the less senses in the initial lexical resource used by the WSD system, the easier the mapping. Second, the two dictionaries do not distinguish the same senses.

Such differences show up between any two dictionaries and make dictionary mapping difficult if not impossible (see(Atkins et Levin 91)). In this case they were especially important since Petit Larousse is a monolingual traditional dictionary with a clear encyclopedic bias while OUP-H is a bilingual, corpus and frequency based dictionary.
Being monolingual and intended for French native speakers, Petit Larousse provides a rather sophisticated hierarchy of senses. Being bilingual and intended for non native speakers, the OUP-H provides a rather flat set of senses. For the same reason, Petit Larousse gives priority to semantic and provides only indicative syntactic information, while OUP-H explicitely mentions all the most common syntactic constructions and distinguishes one sense for each of them [17].

Sense mapping has clearly been an additional source of discrepancy with the *gold* standard. For example, while SDL provides one answer, the mapping phase led SDL to output a disjunction of tags (when one sense of the OUP-H mapped with several senses of Petit Larousse) or a question mark (when one sense of the OUP-H did not map with any sense of Petit Larousse, or when the human mapper did not know). MWEs are also a challenging issue for sense mapping. While Petit Larousse usually includes MWEs in a given word sense, OUP-H systematically lists them at the end of an entry with no link to any of the other senses. OUP-H distinguishes one sense for each MWE.
Following the OUP-H philosophy, we choose not to attach any of the Larousse senses to the OUP-H MWEs. When SDL identified a (OUP-H) MWE, its output was a translation and not a sense tag of Petit Larousse. As a consequence, all MWEs that were correctly identified by SDL (about 18% of the verbs occurrences) were computed as wrong answers in the evaluation. Paradoxally, one of the SDL's strength turned out to be a drawback within the Romanseval exercise.

[16]On average, the OUP-H distinguishes more senses than the Petit Larousse for verbs (15.5 for OUP-H, 12.66 for Petit Larousse) and less for nouns and adjectives (for nouns: 5.6 in OUP-H, 7.6 in Petit Larousse, for adjectives: 4.8 in OUP-H, 6.3 in Petit Larousse

[17]For example, for the verb *poursuivre* there is only a transitive construction according to Petit Larousse while OUP-H distinguishes a sense for the pronominal *se poursuivre* to account for occurrences such as *L'aide* **se poursuit** *dans le cadre du programme spécifique actuel pour les TI adopté le 8 juillet 1991*.

4.2. Are gold standards worthy of the name?

In Romanseval, building a *gold standard* for evaluation was a challenge in its own. The intrinsic difficulty of sense-tagging even for human beings, was reflected by a low inter-tagger agreement: inter-tagger agreement on the French corpus was below 50% (see (Véronis 98) for detailed results and analysis). As inter-tagger agreement defines the upper-bound for how well a system can perform (ie. if two human taggers agree on a sense-tag 80% of the time, than a system cannot be said to achieve more than 80% accuracy), this sheds a doubt on the very possibility of evaluating WSD systems within such a frame.

In the second evaluation, in contrast with Romanseval, the agreement between the human judges who checked the bilingual alignement is very high. But then the *gold standard* that had been relatively easy to establish proved to be difficult to use. First, because the *gold standard* is derived from text alignement, the reference translation provided for each test word occurrence can be empty. Therefore we cleaned the *gold standard* by removing all items with no target unit [18]. Second, because the *gold standard* is derived from text alignement, it includes translations that one cannot expect a dictionary-based WSD system to provide. For instance, in the *gold standard*, one occurrence of *comprend* is translated by *with*) since *leur famille* **comprend** *des enfants mineurs.* was aligned with ... *families* **with** *young children*..

In such cases, it was a real challenge to score the system *right* or *wrong*: our decision to SDL's answer *right* if and only if it provided the same translation as the alignement (212 cases), or it provided a translation for a correctly identified MWE is questionable.

In this evaluation, SDL which is integrated in a comprehension aid, was judged on its capacity to provide the desired translation rather than the right sense [19]. Furthermore, evaluation of SDL was mixed with evaluation of its resource, that is, with evaluation of the dictionary (OUP-H) itself.

5. Conclusions: further development, further evaluation

As for the SDL, we believe that the encouraging results obtained for verbs can be improved by using more of the functional relations provided by the XIFSP and richer dictionary information. For instance, we could use relations such as subject of the relative clause and indirect object. Furthermore, we plan to combine the dictionary based method described in this paper with the example-driven method described in [Dini et al. 99].

As for evaluation methodology, these exercises point out the difficulty of evaluating WSD *in vitro*. We therefore set up an *in vivo* evaluation of the SDL, to see how it helps improving the overall performance of the application in which it is integrated. Within the comprehension aid application, using senses distinctions of a general bilingual resource (OUP-H), the system will be evaluated on an all word disambiguation task.

Aknowledgements We are grateful to Marie-Hélène Corréard, Caroline Brun, Laurent Griot, Gregory Greffenstette, Irene Maxwell and Pierre Isabelle for their comments on this work. Our

[18] 320 items among 1502

[19] For instance, there are cases where the SDL selected the proper sense of the word appearing in the context, for instance, for a verb, the sense attached to its pronominal interpretation. But if within this sense it did not select any specific translation, the case is counted as a mistake in the evaluation.

thanks also go to Claude Roux and Hervé Poirier for their help with the integration of SDL within XeLDA.

References

S. Ait-Mokhtar, J-P. Chanod. 1997. Subject and Object Dependency Extraction Using Finite-State Transducers. In *Proceedings of Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications, ACL*, Madrid, Spain.

B. Atkins,B. Levin 1991. Admitting impediments In *Lexical acquisition: exploiting online resources to build a lexicon* Lawrence Erlbaum Asociates, Hillsdale NJ.

D. Bauer, F. Segond, A. Zaenen. 1995. LOCOLEX: the translation rolls off your tongue. In *Proceedings of ACH-ALLC*Santa-Barbara, USA.

L. Breidt, G. Valetto, F. Segond. 1996. Multiword lexemes and their automatic recognition in texts. In *Proceedings of COMPLEX*, Budapest, Hungaria.

L. Breidt, G. Valetto, F. Segond. 1996. Formal description of Multi-word Lexemes with the Finite State formalism: IDAREX In *Proceedings of COLING*, Copenhagen, Danmark.

L. Dini, V. Di Tomaso, F. Segond. 1998. GINGER II: an example-driven word sense disambiguator. In *Computer and the Humanities*, Same issue.

N. Ide, J. Véronis 1998. Introduction to the special issue on word sense disambiguation: the state of the art. In *Computational Linguistics - special issue on word sense disambiguation*, nb.1 vol.24, March 1998.

A. Kilgarriff. 1999. Gold standard datasets for evaluating word sense disambiguation programs. In *Computer and the Humanities*, same issue.

Larousse 1995. *Le petit Larousse illustré - dictionnaire encyclopédique*. Edited by P. Maubourguet, Larousse, Paris.

Oxford-Hachette 1994. *The Oxford Hachette French Dictionary*. Edited by M-H Corréard and V. Grundy, Oxford University Press-Hachette.

F. Segond, E. Aimelet, L. Griot. 1998. "All you can use!" or how to perform Word Sense Disambiguation with available resources In *Second Workshop on Lexical Semantic System*, Pisa, Italy.

F. Segond. 1998. *Framework and results for French*. to appear in a special issue of *Computer and the Humanities*.

F. Segond, E. Aimelet, C. Jean, V. Lux. 1999. "Dictionary-driven semantic look-up" to appear in a special issue of *Computer and the Humanities*.

J. Véronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and Advanced Papers of Senseval Workshop*, Herstmonceux Castle, UK.

Y. Wilks, M. Stevenson. 1998. Word Sense Disambiguation using Optimised Combinations of Knowledge Sources. In *Proceedings of COLING/ACL*, Montreal, Canada.