Computers and the Humanities **34:** 193–197, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Dictionary-Driven Semantic Look-up

FRÉDÉRIQUE SEGOND[1], ELISABETH AIMELET[1], VERONIKA LUX[1] and CORINNE JEAN[2]
[1]*Xerox Research Centre Europe, Meylan, France;* [2]*Université de Provence and Xerox Research Centre Europe*

1. Introduction

The French Semantic Dictionary Look-up (SDL) uses dictionary information about subcategorization and collocates to perform Word Sense Disambiguation (WSD). The SDL is fully integrated in a multilingual comprehension system which uses the Oxford Hachette French-English bilingual dictionary (OUP-H). Although the SDL works on all words both for French and English, Romanseval results are relevant for French verbs only because subcategorisation and collocate information is richer for this part of speech in the OUP-H. The SDL uses dictionaries as semantically tagged corpora of different languages, making the methodology reusable for any language with existing on-line dictionaries.

This paper first describes the system architecture as well as its components and resources. Second, it presents the work we did within Romanseval, namely sense mapping and results analysis.

2. Semantic Dictionary Look-Up: Goal, Architecture and Components

The SDL selects the most appropriate translation of a word appearing in a given context. It reorders dictionary entries making use of dictionary information. It is built on top of Locolex,[1] an intelligent dictionary look-up device which achieves some word sense disambiguation using the word's context: part-of speech and multiword expression (MWEs)[2] recognition. However, Locolex choices remain syntactic. Using the OUP-H information about subcategorization and collocates the SDL goes one step further towards semantic disambiguation.

To reorder dictionary entries the SDL uses the following components:

- the Xerox Linguistic Development Architecture (XeLDA),
- the Oxford University Press-Hachette bilingual French-English, English-French dictionary (OUP-H),
- the French Xerox Incremental Finite State Parser (XIFSP).

XeLDa is a linguistic development framework designed to provide developers and researchers with a common architecture for the integration of linguistic services. The OUP-H dictionary look-up and the French XIFSP are both integrated into XeLDA. The OUP-H (French-English),[3] an SGML-tagged dictionary, is designed to be used for production, translation, or comprehension, by native speakers of either English or French. The SDL uses OUP subcategorization and collocate tags. Collocate tags encode the kind of subject and/or object a predicate expects. Most of the time, they are given as a list of words, sometimes as a concept.

To extract functional information from input text in order to match it against OUP-H information, we use the French XIFSP. XIFSP adds syntactic information at sentence level in an incremental way, depending on the contextual information available at a given stage. Of particular interest to us is the fact that shallow parsing allows fast automatic recognition and extraction of subject and object dependency relations from large corpora, using a cascade of finite-state transducers. The extraction of syntactic relations does not use subcategorisation information and relies on part of speech information only.

For instance, suppose the task is to disambiguate the verb *présenter* in the sentence:

> *Des difficultés se* **présentent** *lorsque l'entreprise d'assurance n'exerce ses activités qu'en régime de libre prestation de services et s'en tient à la couverture de risques industriels.*

The SDL first calls the XIFSP which parses the sentence and extracts syntactic relations, among which: *SUBJREFLEX (difficulté,présenter)*. This relation encodes that *difficulté* is the subject of the reflexive usage of the verb *présenter*. This information is then matched against collocates information in the OUP-H for the verb *présenter*. Because matches are found (reflexive usage and collocate), the SDL reorders the dictionary entry and first proposes the translation "to arise, to present itself".

If no dictionary information matches the context of the input sentence, it returns, by default, the first sense of the OUP-H.[4]

In case of information conflict between subcategorisation and collocates, priority is given to collocates.[5]

3. Sense Mapping

Sense mapping is an additional source of discrepancy with the *gold standard* which has an influence on the evaluation of WSD systems. Mapping, in our case, consists of assigning a Larousse sense tag not to an example but to a sense that is usually illustrated by a number of examples in the OUP-H. We map two different sets of senses which usually do not have the same number of elements. On average, the OUP-H distinguished more senses than Le Larousse for verbs (15.5 for OUP-H, 12.66 for Larousse) and less for nouns and adjectives (for nouns: 5.6 in OUP-H, 7.6 in Larousse; for adjectives: 4.8 in OUP-H, 6.3 in Larousse).[6] Clearly, the

fewer senses in the initial lexical resource used by the WSD system, the easier the mapping.

These differences show up between any two dictionaries, but in this case they are especially important because of two additional factors: first, the Petit Larousse is monolingual while OUP-H is bilingual. Second, the Petit Larousse is a traditional dictionary with a clear encyclopedic bias while the OUP-H is corpus and frequency based.

Being monolingual and intended for French native speakers, the Petit Larousse provides a sophisticated hierarchy of senses. Being bilingual and intended for non-native speakers, the OUP-H provides a flat set of senses. For the same reason, Larousse gives priority to semantics and provides only indicative syntactic information, while OUP-H explicitly mentions all the most common syntactic constructions and distinguishes one sense for each of them.

Because of the mapping phase, the output of the SDL can be a disjunction of tags (one sense of the OUP-H maps to several senses of the Petit Larousse) or a question mark (one sense of the OUP-H does not map to any sense of Le Larousse, or, the human mapper did not know).

Another challenging issue for sense mapping concerns MWEs. While Larousse often 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 did not attach any of the Larousse senses to the OUP-H MWEs. When the SDL identifies a (OUP-H) MWE, its output is a translation and not a sense tag of the Larousse. As a consequence, all MWEs that were correctly identified by SDL (about 18% of the verb occurences) were computed as wrong answers in the evaluation. Paradoxically, one of the SDL's strength turns out to be a drawback within the ROMANSEVAL exercise.

4. Evaluation and Conclusion

For complete results and for a comparative analysis of these results with other systems, see Segond (this volume).

One of the strengths of the ROMANSEVAL exercise has been to make us understand in greater details the different factors that influence the evaluation of WSD systems. They include, for instance, the granularity of dictionaries used by the system (definition dictionaries, bilingual dictionaries, ontologies), how MWEs are handled as well as what is the goal of a given WSD system. Because what we are interested in is to see how much semantic disambiguation the SDL actually achieves according to our own dictionary (OUP-H) within our own application (comprehension aid), we computed another evaluation for the 20 verbs.[7] In this evaluation, we obtain 70% precision and 33% recall. Precision is the number of verbs correctly tagged divided by number of verbs tagged. By tagged verbs we mean verbs for which dictionary information has been used by the SDL to select

a meaning. Recall is the number of verbs correctly tagged divided by the total number of verbs. It gives an indication of how many times information needed is encoded in the dictionary.

A study of the results shows that the system tagged 715 verbs out of 1502 verbs occurences. Among these 715 tagged verbs 400 were tagged using MWEs' information and 315 using subcategorization and/or collocates information. Among the 400 tagged as MWEs, 279 were properly recognized. Wrong MWEs were recognized because of a too generous encoding of the possible variations of MWEs.[8] Among the 315 senses selected using subcategorisation and collocates information, 225 were correctly selected. Incorrect ones are mainly due to the two following factors:

- subject/object extraction error by the shallow parser,
- false prepositional phrase attachment.[9]

We see that MWEs recognition achieves about 18% of the verb semantic disambiguation while subcategorization and collocates achieve about 14%.

In this evaluation we did not take into account cases where we found the right tag using the first OUP-H sense by default. Two reasons guided this decision: first, we wanted to see how well the SDL performed when it actually performed a choice; second, as long as the first sense of the OUP-H usually does not map with the first sense of the Larousse, this information is difficult to interpret.

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 clauses, indirect object.

We are now working on combining the SDL with the semantic example-driven tagger developed with CELI.[10] The resulting semantic disambiguation module, a dictionary-based semantic tagger, will use a rule database encoding all together information about subcategorization, collocates and examples. Indeed, looking back at the overall evaluation exercise, we believe that the future of WSD lies not only in combining WSD methods, but also in creating WSD systems attached to a particular lexical resource which has been designed with a given goal. For instance, a WSD system attached to a general bilingual dictionary will perform better than a general ontology containing few senses distinctions in helping in understanding English texts from general newspapers.

Notes

[1] See Bauer et al. (1995).

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

[3] See Oxford (1994).

[4] Note that because of the encyclopedic vs corpus frequency based difference between OUP-H and Larousse, the first sense of Larousse often does not match the first sense of OUP-H.
[5] A full description of the SDL can be found in (Segond et al., 1998).
[6] Individual cases differ considerably from the average. Two particular verbs such as *comprendre* and *parvenir* which respectively have 11 and 3 senses in the OUP-H, both have 4 senses in the Larousse.
[7] No collocate information is attached to nouns in the OUP-H and for the adjectives chosen, very little collocate information was provided. When information is not present in the dictionary there is no way for us to perform any disambiguation.
[8] Using local grammar rules, Locolex encodes morpho-syntactic variations of MWEs in the OUP-H. In some cases this encoding has been too generous leading to the over-recognition of such expressions.
[9] For instance in the sentence “une aide destinée á couvrir les dettes des éleveurs” (help which is designed to cover debts of breeders), the shallow parser analyzes “des éleveurs” as a VMODOBJ of “couvrir” instead of as a complement of the NP “les dettes”. This is because in equivalent syntactic construction such as “couvrir les gens d’or”, “d’or” is VMODOBJ of “couvrir”.
[10] See Dini et al (this volume).

References

Ait-Mokhtar, S. and J-P. Chanod. “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 (1997).

Bauer, D., F. Segond and A. Zaenen. “LOCOLEX : The Translation Rolls Off Your Tongue”. In *Proceedings of ACH-ALLC*, Santa-Barbara, USA (1995).

Breidt, L., G. Valetto and F. Segond. “Multiword Lexemes and Their Automatic Recognition in Texts”. In *Proceedings of COMPLEX*, Budapest, Hungaria (1996a).

Breidt, L., G. Valetto and F. Segond. “Formal Description of Multi-word Lexemes with the Finite State formalism: IDAREX”. In *Proceedings of COLING*, Copenhagen, Danmark (1996b).

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

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

Segond, F., E. Aimelet and L. Griot. “ ‘All You Can Use!’ Or How to Perform Word Sense Disambiguation with Available Resources”. In *Second Workshop on Lexical Semantic System*, Pisa, Italy, 1998.

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