A new semantically annotated corpus for Word Sense Disambiguation evaluation

Myriam Rakho, Éric Laporte, Mathieu Constant

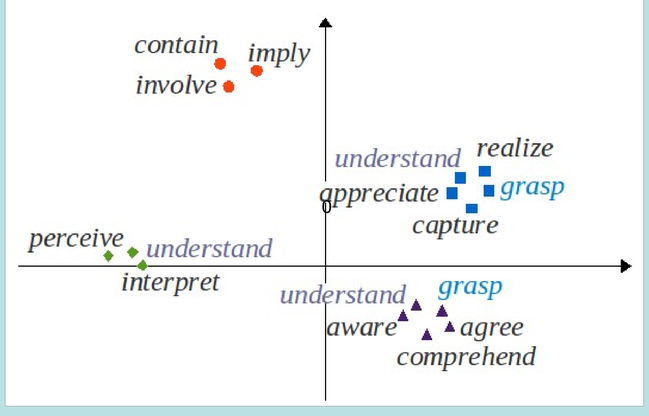
Université Paris-Est Marne-la-Vallée, LIGM
{rakho,ericlaporte,matthieuconstant}@univ-mlv.fr

1. Introduction

Word Sense Disambiguation (WSD) is the task of assigning one of several possible senses to particular occurrences of words in texts. WSD is an NLP intermediate task, so the inventory of word senses between which WSD systems are to disambiguate depend on the final application.

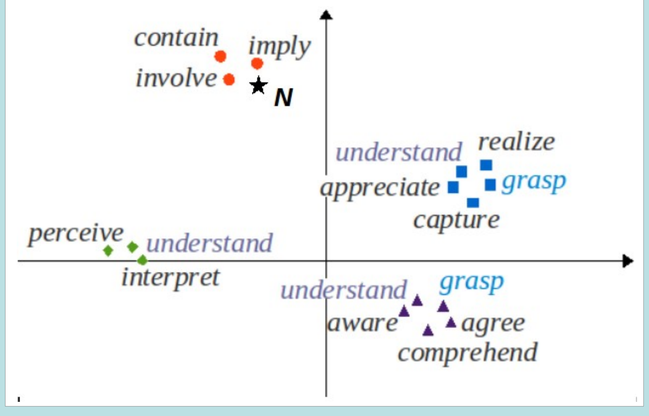
The Word Sense Induction approach (WSI) allows us to overcome the need for a predefined sense inventory:

1. CLUSTERING THE TRANSLATIONS



2. CLASSIFICATION OF NEW INSTANCES

N: *Le principe de précaution ne* ***comprend*** *pas la politisation de la science. The precautionary principle does not* ***involve*** *the politicisation of science.*



The advantage of WSI is that the semantic ambiguities of the translations are implicitly resolved:

(1) all synonymous translations of the same source language (SL) sense belong to the same cluster:

(2) each translation can belong to as many clusters as its corresponding SL senses:

(3) translations of the same SL sense that represent valid sense distinctions only in the TL are identified

2. Aims

WSI systems are evaluated by using a Gold Standard (GS) corpus in which each word occurrence has its reference translation(s).

We built a GS corpus for the evaluation of WSI for 20 French polysemous verbs. In this corpus, the translations are clustered on the basis of the entries of the verbs in a dictionary in which sense distinctions are based on syntactic-semantic evidence: the Lexicon-Grammar tables (LG).

3. Data

THE LEXICAL SAMPLE

The corpus is constitued of sentences that contain occurrences of 20 French polysemous verbs. Those verbs were selected on the basis of their polysemy during the ARCADE campaign (2000) for the evaluation of multilingual word alignment systems.

THE CORPUS

All contexts which contain an occurrence of one of the 20 polysemous verbs in the French version of the EuroParl parallel corpus and their aligned translation in the English version.

4. The Lexicon-Grammar tables

The LG is represented as a taxonomy of syntactic-semantic classes. The underlying assumption is Harris' view on meaning:

Difference of meaning correlates with difference of distribution (Harris, 1970)

Each class in the LG groups items that share some core *defining syntactic properties*

Example: Tables 37M1 and 37M6 group lexical items which select a transitive locative complement and a prepositional complement that differs semantically depending on the table:

couvrir [cover]:
37M1. Max a **couvert** Luc d'un paletot.
Max **wrapped** Luc in a jacket
37M6. Max a **couvert** son désarroi d'un empressement subit
Max **masked** his distress with sudden eagerness

Each class is represented as a table in which:

→ each entry describes one of the possible environments of a lexical item belonging to this class
→ with a selection of features (in columns)

Each verb has as many entries in the LG tables (sometimes in the same table) as it has different senses in the SL.

The verb part of the LG describes:

- 16872 non-idiomatic items, 5738 lemmas, 67 tables, 552 features
- 39628 idiomatic items, 38658 lemmas, 69 tables, 276 features

Example: *comprendre* [understand, comprise] (*table_entry: illustrative example*)
6_73: *Max a* ***compris*** *qu'Ida était coupable à cet indice.* Max understood that Ida was guilty from this indication.
12_17: *Max* ***comprend*** *qu'Ida ne vienne pas.* Max understands that Ida does not come.
38R_54: *Max a* ***compris*** *cette remarque comme une plaisanterie.* Max understood this remark as a joke.
10_46: *Faire ce travail* ***comprend*** *pour Max qu'il nettoie tout.* To make this work includes for Max that he cleans everything.

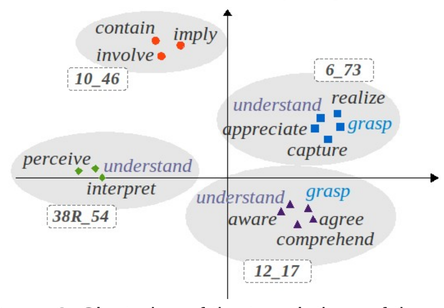


Figure 1. Clustering of the translations of the verb *comprendre* according to its entries in the LG

5. Data annotation

TRANSLATIONS

Sense inventory. The translations of the verbs in the English version of the EuroParl parallel corpus

Data annotation. (1) lexical alignment of the French and English versions of EuroParl with Giza++ and (2) manual validation of the alignment of the polysemous verbs with their correct translation.

Sense tags. *translation_pos-tag*

Example:
Fr: Les personnes commettant des fraudes doivent **comprendre** qu'elles seront poursuivies.
En: People need to **understand** that if they commit fraud they will be prosecuted.
Sense label: understand_vv

LEXICON-GRAMMAR ENTRIES

Sense inventory. All entries of the verbs in the LG tables (simple and idiomatic) found in the corpus.

Data annotation. Each context of a verb was manually assigned the LG entry that corresponds to its linguistic sense.

Sense tags. {*V*|*C*}*_table_entry*
C (*V*): table of idiomatic expressions (or not); *table*: name of a table; *entry*: unique identifier of the lexical Item in the table

Example:
Sense label: V_12_17

FINE-GRAINED SENSES

Sense inventory. Automatic concatenation, by context, of the translation and the LG entry.

Sense tags. {*V*|*C*}*_table_entry#translation_pos-tag*

Example:
Sense label: V_12_17#understand_vv

CLUSTERED FINE-GRAINED SENSES

Sense inventory. The fine-grained senses are grouped into clusters based on their LG part.

Sense tags. Clusters of fine-grained senses

Example:
Sense label: {V_12_17#understand#VV, V_12_17#appreciate_vv, v_12_17#grasp, v_12_17#realise_vv, Etc.}

6. Corpus statistics

	Nb samples	Nb TRSL	Nb LG	Nb LG#TRSL
arrêter	2033	12	150	242
comprendre	**8240**	8	183	308
conclure	3488	5	79	122
conduire	2114	10	96	145
connaître	5786	14	158	238
couvrir	2183	16	85	128
entrer	2325	6	107	189
exercer	**1851**	4	86	120
importer	2778	5	**71**	**101**
ouvrir	2656	17	127	253
parvenir	7469	**3**	152	185
porter	3301	**20**	**219**	**495**
poursuivre	5354	5	154	219
rendre	6731	14	177	347
tirer	2163	19	102	160
venir	7369	12	120	306
Overall mean	**3837**	**11**	**129**	**222**

Table 1. The polysemous verbs, their sample size (Nb saples) and the number of senses in each of their sense inventories: translations (TRSL), LG entries (LG) and fine-grained senses (LG#TRSL)

7. Conclusion

This corpus will be used for many purposes:

→ evaluation of WSI systems
→ optimization of their parameters
→ comparison of their results with different sense inventories
→ comparison of their results with those of supervised WSD systems

First WSD results

We trained a supervised SVM based WSD system on this corpus. We evaluated it with a best score assignment strategy: only the first sense label was assigned to each new instance. The system achieved 98.2% accuracy when assigning LG senses and 85.7% when assigning translations and fine-grained senses.

This corpus will be made freely available for the research community.

NLP tools and resources used: EuroParl parallel corpus, Unitex for compouds identification and lemmatisation, Giza++ for the lexical alignment
The Lexicon-Grammar tables are freely available at this adress: http://infolingu.univ-mlv.fr/english/