Evaluating the impact of some linguistic information on the performances of a similarity-based and translation-oriented Word-Sense disambiguation method

Myriam Rakho, Matthieu Constant
Université Paris-Est, LIGM
E-mail: rakho@univ-mlv.fr, mconstan@univ-mlv.fr

Abstract

In this article, we present an experiment of linguistic parameter tuning in the representation of the semantic space of polysemous words. We evaluate quantitatively the influence of some basic linguistic knowledge (lemmas, multi-word expressions, grammatical tags and syntactic relations) on the performances of a similarity-based Word-Sense disambiguation method. The question we try to answer, by this experiment, is which kinds of linguistic knowledge are most useful for the semantic disambiguation of polysemous words, in a multilingual framework. The experiment is about 20 French polysemous words (16 nouns and 4 verbs) and we make use of the French-English part of the sentence-aligned EuroParl Corpus for training and testing. Our results show a strong correlation between the system accuracy and the degree of precision of the linguistic features used, particularly the syntactic dependency relations. Furthermore, the lemma-based approach absolutely outperforms the word form-based approach. The best accuracy achieved by our system amounts to 90%.

1. Introduction

In word sense disambiguation (WSD) task, multiple experiments of parameter tuning in the representation of the semantic spaces have been carried out (Pancardo-Rodguez & al., 2005; Crestan & al., 2003). In this paper, we present the results of a similar experiment, looking, among all the cooccurrents of a polysemous word, for the best candidates to be employed as dimensions of its semantic space : those which are discriminative enough to give a WSD method the ability for distinguishing its different senses. For this purpose, we varied the combinations of some basic linguistic knowledge in the semantic spaces (lemmas, multi-word expressions, grammatical tags and syntactic relations).

The semantic spaces, built from pre-classified instances of the ambiguous words in their occurring contexts, are used in examplar-based classification methods like *k* Nearest Neighbors algorithm (*k*NN) (Veenstra & al., 2000), Support Vector Machines (SVM, Keok Lee & al., 2004), Semantic Classification Trees (SCT, Loupy & al., 2000) and other methods that use context-similarity measures (Apidianaki, 2009). The classes associated with the training instances in the semantic space of a word can be its senses as they are defined in traditional lexical resources (dictionaries, thesauri) (Gale & al., 1993). In a multilingual framework, the classes can also be its translation equivalents in one (Kaji & Morimoto, 2002) or more (Crego & al., 2009) other languages. For our experiment, we chose the machine translation (MT) oriented and similarity-based method described in (Apidianaki, 2009).

Our corpus for training and testing was the French (SL) and English (TL) aligned version of the sentence-aligned EuroParl corpus (Koehn, 2003), and we evaluated the performances of our WSD method for the disambiguation of 20 polysemous words (16 nouns and 4 verbs).

In the first section of this article, we give a description of the WSD method we used for this experiment. The second section is our definition of the semantic spaces. And in the third section, we propose an analysis for the results of our experiment.

2. Description of our WSD method

We used for this experiment the MT oriented method described in (Apidianaki, 2009).

The training instances, for a given word, are the vectors of cooccurrents representing the SL segments (***SL_segment***) in the aligned corpus in which this word occurs. The classes are its translation equivalents (***EQV***) in the corresponding TL parts.

(Apidianaki, 2009) takes things in two steps. First, the EQVs of the word are grouped into clusters representing its various senses, and the new instance is assigned the most suitable cluster. And secondly, the most probable EQV is chosen among those in this cluster. We evaluated the first step of this method.

Clustering the EQVS. First, every EQV is associated with a unique vector (EQV_i*_segment*) which contains the union of the components of all the *SL_segments* with which it is associated. Every cooccurrent (*j*) is assigned as many relative weights (rw_{ij}) as classes (*i*) with which it is associated. The value of rw_{ij} is the discriminating potential of *j* between EQV *i* and the other EQVs of the word (see Apidianaki, 2009). We then build a similarity matrix of EQVs using the weighted Jaccard coefficient (WJ, Grefenstette, 1994). While all the cooccurrents are taken into account when computing the relative weights, the computation of EQVs similarity can be made either taking into account all the cooccurrents or only the syntactic cooccurrents and the neighbors (parameter *sim*, described in *section 3*).

Clusters of semantically similar EQVs are built, in which the similarity between all EQVs is equal or higher than the average of all the similarities in the matrix. Every cluster is represented by a vector containing all the cooccurrents that appear in all the EQV_i*_segment* of its components at once.

The **decision function** used for determining the cluster of the new instance (w_{new}) is defined as follows :

- The similarity between the context vector of w_{new} and the vector of every cluster is calculated on the basis of their intersection : it is the ratio between, on one hand, the sum of the relative weights of the cooccurrents that appear in both vectors, and, on the other hand, the product of the sizes of the two vectors.
- And the most similar cluster is assigned to w_{new}.

3. Our definition of the semantic space(s)

Representing the dimensions of the semantic spaces have been done using various kinds of linguistic knowledge. *Table 1* describes the corresponding parameters and their modalities.

The linguistic preprocessing of every vector corresponding to a SL_segment was done in three steps :

- **step 1 :** *type* and *comp* parameters are applied to the whole vector.
- **step 2 :** *ctxt_type* parameter is applied. The output of this step is a vector in which the components that belong to the same kind of context (thematic, neighbours or syntactic cooccurrents) are assigned the same ***absolute weight*** (*aw*). Three kinds of vectors are obtained, corresponding to the three combinations of contexts we have tested :
  - a ***neighbours and thematic*** vector is a vector in which two kinds of contexts are represented : the neighbours of the ambiguous word and the other (thematic) cooccurrents of the word. The first ones are distinguished by a strong absolute weight (*aw*=2) while *aw* is 1 for the other (thematic) cooccurrents ;
  - a ***syntactic and thematic*** vector is a vector in which the direct syntactic cooccurrents are distinguished from the other cooccurrents by the same weighting procedure as in the preceding kind of vector ;
  - a ***thematic vector*** is a vector in which all the (thematic) components of the original vector are assigned the same absolute weight (*aw*=1).
- **step 3 :** *ctxt_component* is applied differently to every kind of context represented in the vector.

For a given cooccurrent *j*, all its absolute weights in the context vectors that are associated with a given EQV *i* are summed. Its weight relatively to this EQV (rw_{ij}, *section 2*) is then multiplied by this sum. Thus, the neighbours and the syntactic cooccurrents are favoured when computing EQVs similarity.

We give in *table 2* below an example, using, as a context of the word '*article*', the SL sentence : "*Selon l'article 22 du règlement, vous voulez que les membres dressent un compte-rendu détaillé de leurs activités*". This example illustrates the 'neighbour (in bold) and thematic' context. In the 'vectorial representation' row, we put in brackets the value of *aw* for each cooccurrent.

NLP tools. The NLP tools we used for preprocessing the corpus are Unitex (Paumier, 2008), for multi-word expressions extraction, TreeTagger (Schmidt, 1995), for lemmatization and grammatical tagging, and the Xip parser online demo (Xerox Incremental Parser, Aït-Mokhtar & al., 2002), for the detection of syntactic relations.

4. Evaluation

We evaluate the parameter combinations defined in *section 2* above, in order to find the more relevant one for our WSD method, in terms of representativeness for the different senses of a word. We use for that the French-English part of the sentence-aligned EuroParl corpus, which consists of about one million sentences, that is 30 million words for each language version.

We evaluate the disambiguation of 20 French polysemous words, 16 of which are nouns (*article*,

Linguistic knowledge	Correponding parameter	Parameter modalities	Signification	
Nature of a cooccurrent	***ctxt_type***	*neighbour*	a unit appearing in a window of size 1 before and after the word	
		syntactic	a unit linked to the word by a direct syntactic relation (*Subject*, *Object*, …)	
		thematic	a unit that is neither a neighbor nor a syntactic cooccurrent	
Type of a cooccurrent	***type***	*form*	a form only (word or lemma, depending on the modality of *ctxt_comp* parameter described below)	
		form#tag	a form and its grammatical tag (*word#tag* or *lemma#tag*)	
Lexical form	***comp***	*no*	the context vectors are composed of simple word-token units	
		yes	multi-word expressions are considered as single units	
Grammatical form and grammatical category	***ctxt_comp***	*1*	the cooccurrents of the ambiguous word	All the components of the vector
		2		Only the nouns, verbs and adjectives are used
		3	lemmas of the cooccurrents	All the components of the vector
		4		Only the nouns, verbs and adjectives are used
EQVs similarity	***sim***	*all*	WJ coefficient using all the components of the two vectors	
		strong	WJ coefficient using only the strong components in the two vectors	

Table 1: Parameters for the representation of the semantic spaces of the words

The ambiguous word : article

SL_segment :

*Selon l'**article** 22 du règlement, vous voulez que les members dressent un compte rendu détaillé de leurs activités*

TL_segment :

*Under **rule** 22 of the rules of procedure, you want us, as members, to give you an exact account of what we do, at what time.*

Parameters combination :

Step 1 : *type=form#tag ; comp=yes*
Step 2 : *ctxt_type = neighbours and thematic*
Step 3 : *ctxt*=1+3 (1 for *neighbours* ; 3 for *thematic cooccurrents*)

Vectorial representation :

article#nom(4) - *l*#det:art(2) – *22*#card(2) - *selon*#prp(1) – *du*#prp(1) – *règlement*#nom(1) – *vous*#pro :per(1) – *vouloir*#ver :pres(1) – *que*#kon(1) – *le*#det :art(1) – *membre*#nom(1) – *dresser*#ver :pres(1) – *un*#det :art(1) – *compte-rendu*#nom(1) – *détailler*#ver :pper(1) – *de*#prp(1) – *leur*#det :pos(1) – *activité*#nom(1)

Table 2: An illustration for the neighbours and thematic context.

barrage, *cadre*, *compte*, *conclusion*, *culture*, *matière*, *passage*, *produit*, *raison*, *rapport*, *reserve*, *société*, *traitement* and *vol*) while 4 are verbs (*lever*, *monter*, *porter* and *saisir*).

4.1 Training and testing data

We first manually built a bilingual lexicon in which each SL word is associated with its various TL translations (EQV) in the aligned corpus. Then, for every word, we extracted a corpus, consisting of sub-corpora for its EQVs : one sub-corpus consists of the SL part (*SL_segment*) of all the aligned segments in the corpus in which the word is translated by the EQV concerned. Every sub-corpus contains around one thousand *SL_segments*, 80% of which are used for training and 20% for testing.

Table 3 describes the words of our task : it summarizes the mean polysemy from monolingual and multilingual points of view (the number of usages of the words according to the French (Larousse, 2009) dictionary and the number of TL EQVs in our corpus, respectively), the size of the training and testing sets for each word and the size of the sub-corpora for each EQV (minimal and maximal size).

In the bilingual lexicon, each word is associated with all the TL EQVs that are used to translate it the aligned corpus. *Table 4* gives the lexicon entries for the word *compte* and their part-of-speech (POS). *Figure 1* is an illustration of the extraction of the semantic space for the word *article*.

4.2 Evaluation metrics

The metrics used for the evaluation are :

- **recall** : the ratio between the number of correct predictions and the number of reference instances
- **enriched precision** : the ratio between the number of correct predictions and the number of predictions
- ***f*-score** : (2* (*recall * precision*)) / (*recall * precision*)

A prediction is considered as *correct* if the selected cluster for w_{new} contains the EQV used as its reference translation in the aligned segment of the SL part which contains w_{new}.

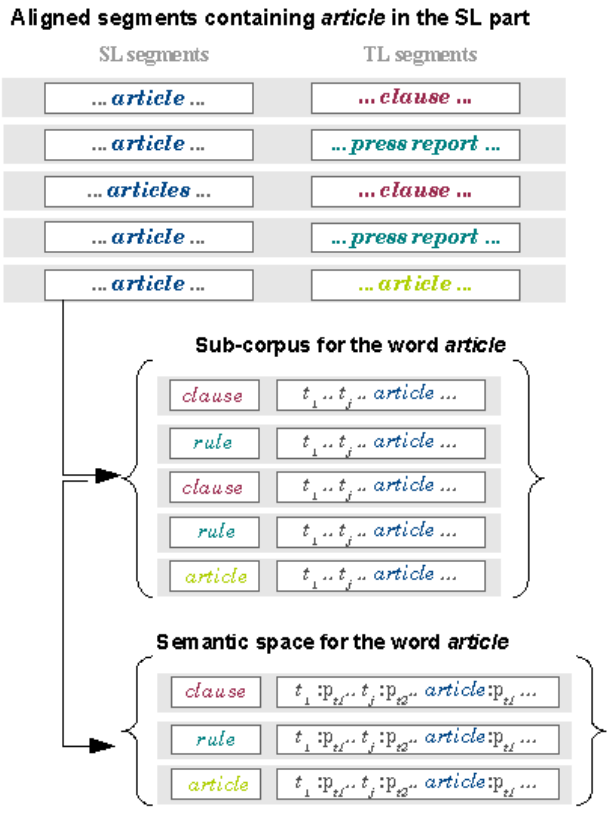

Figure 1: Building the semantic space of the word *article*

		NOUNS	VERBS
mean polysemy	#usages in SL	4.5	4.5
	#TL EQVs	9.5	8
#train (for each word)		72 to 2471	21 to 942
#test (for each word)		50 to 691	14 to 178
#examples for each EQV		1 to 1 000	1 to 514

Table 3: Description of the words and their sub-corpora

SL linguistic form.POS : {TL EQVs}
compte.N : {*account*}
en fin de compte.Adv. Loc. : {*in the end, ultimately*}
se rendre compte de.V : {*realise*}
rendre des comptes.V : {*be accountable*}
tenir compte de.V : {t*ake into consideration, take into account, take account of, give consideration to, be aware of, consider*}
compte tenu de.Prep. Loc. : {b*ecause (of), considering*}
Cour des Comptes.N : {*Court of Auditors*}

Table 4: Lexicon entries for the word *compte*

4.3 Best scores obtained

The best score obtained for our WSD method amounts to 90.5%. This score is equally obtained with the 'neighbours and thematic' context or with the 'syntactic and thematic' context, both defined with the following parameters combination : *type* is *form#tag*, *comp* is *yes*, *ctxt* is 3 for the thematic cooccurrents and any value for the neighbours or the syntactic cooccurrents, and *sim* is *strong*.

4.4 Parameters evaluation

4.4.1. Quantitative evaluation

Due to lack of space, we cannot show in this extended abstract all the results we have obtained. The diagrams in *figure 1* represent the evolution of the *f*-score depending on *ctxt* parameter, when type is *form#tag* and *comp* is *yes*. For parameters *type* and *comp*, we give the most significant scores only.

4.4.2. Global tendencies

In this sub-section, we propose several conclusions concerning the linguistic parameters we have drawn from the results of the experiment.

Table 5 illustrates quantitatively the interaction between the linguistic informations. Each cell of the table contains the highest score obtained with all the parameter combinations in which both the row entry and the column entry modalities of the two parameters concerned are activated. Concerning *ctxt_comp* parameter, only its application to the thematic context is represented, since we found that it have no influence for the neighbors and the syntactic coocurrents. We can draw, from this table, the global tendency for every parameter. For example, in the row and column that represent type parameter, the values in the *form#tag* (in dark gray in the table) part are always higher than the ones in the *form* part (in light gray), whatever the parameter with which *type* is combined.

4.4.3. Our findings

In this sub-section, we propose several conclusions concerning the linguistic parameters we have drawn from the results of the experiment.

Parameter sim. The best score (90.5%) falls to 73.9% when *sim* is *all* : the similarity between two EQVs is computed using all their cooccurrents. Representing the different usages of a polysemous word is then more precise when using 'syntactic patterns'. However, the thematic context cannot be ignored, since the best score fell to 74% when only the syntactic cooccurrents and the neighbors were considered in the computation of the relative weights of the cooccurrents (rw_{ij}).

Parameter *ctxt_type*. We observe that the *f*-scores are higher when neighbors and syntactic cooccurrents are used in the semantic spaces.

Parameter *type*. The representation of the semantic spaces was more precise with *form#tag* value for this parameter. With the optimal combination, we observed a strong decrease when *type* is *form* (81.4%). Then, we can say that the morpho-syntactic component plays a significant role in the representation of the linguistic context of the words.

Parameter *comp*. The *f*-scores were better when multi-word expressions were considered as single units : the best-score falls to 81.8% when *comp* is *no*. This is explained by the fact that the sense, for this kind of expressions, cannot be induced by a semantically compositional process.

Linguistic parameters and their modalities		**sim**		**type**		**comp**		**ctxt_comp (1 vs 3)**	
		all	**strong**	**form**	**form#tag**	**no**	**yes**	**words**	**lemmas**
type	**form**	81.6	81.6						
	form#tag	81.6	**90.5**						
comp	**no**	81.3	81.8	81.6	81.8				
	yes	81.3	**90.5**	81.4	**90.5**				
ctxt_comp (1 vs 3)	**words**	81.6	82.7	80.8	82.7	80.9	82.7		
	lemmas	81.3	**90.5**	81.7	**90.5**	81.1	**90.5**		
ctxt_comp (2 vs 4)	**filtered**	80	84.6	75.9	84.9	76	84.5	75.8	84.5
	all	81.6	**90.5**	81.6	**90.5**	85.7	**90.5**	82.7	**90.5**

Table 5: Global tendencies of the linguistic parameters and their interaction with each other

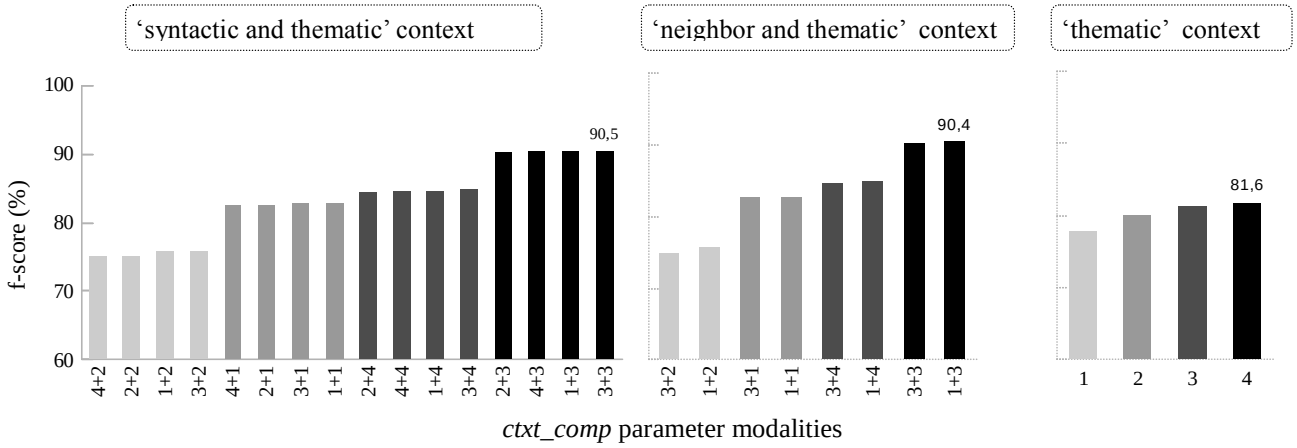

Figure 1: Impact of lemmatisation and grammatical filtering : *f*-score depending on *ctxt_comp* parameter (*comp* is *yes* and *type* is *form#tag*) (reminder : 1:*words* ; 2:*filtered words* ; 3:*lemmas* ; 4:*filtered lemmas*)

So, inserting the sense of the lexical units composing them in the semantic space of a word is literally incorrect.

Parameter *ctxt_comp*. This parameter was very influential for the performances of our WSD method. The grammatical filtering was absolutely bad. And the influence of the lemmatization varies depending on the kind of context. It was good for the thematic context, but had no influence for the neighbors and for the syntactic cooccurrents.

Conclusion. Making use of distributional hypothesis in order to describe the semantic space of the words cannot be done by considering all the co-textual elements in a homogenous way. Then, each word can be caracterized by a multifaceted representation of its local and global contexts of usage in which each kind of co-textual element (neighbors as left and right lexical and grammatical context, syntactic cooccurrents, first and second order thematic cooccurrents, predicates and arguments, semantic roles, and so on) has to be taken into account. Moreover, each kind of co-textual element has then to be favoured depending on the goal of the task. For example, (Baroni & Bisi, 2004) used narrow windows of size 2 and 5 (immediate lexical neighbors) to discover synonymy relations.

4.4.4. Analysis of the evolution of the scores

Figure 1 is a diagram representation of the evolution of the scores following *ctxt_comp* parameter. The evolution of the labels of the Y-axis shows us consistent relation between the scores and *ctxt_comp* parameter. In fact, in the two first diagrams, the scores are formed of four groups, corresponding to the fours modalities of *ctxt_comp* applied to the thematic context (*b* part in the *a+b* labels). This four groups are in the following ascending order : [2 1 4 3], which provides two rankings in the scores :

- a first order in terms of words and lemmas : (2,1) < (4,3), so words < lemmas ;
- and a second order in terms of grammatical filtering : 2<1 and 4<3, so filtering < no filtering.

In the first diagram, within this first ranking, a second ranking is observed that follows the modality of *ctxt_comp* applied to the syntactic cooccurrents (*a* part in the *a+b* labels). This second ranking is a corroboration of the order we observed in the first ranking, relating to the grammatical filtering : in the four groups, we effectively observe that (4,2) < (3,1). And in the second diagram, we observe an identical second ranking relating to the lemmatization and that corroborate the fact that lemmas are better than words : 3<1 in the four groups.

The third diagram represents the scores obtained when no distinction is made between the neighbors and syntactic cooccurrents on one hand, and the other (thematic) cooccurrents. Once again, we observe the same ranking concerning the grammatical filtering : (2,4)<(1,3). But the lemmas are better than words (80 vs. 77.7%, respectively) when this filtering is applied, while the words are (slightly) better than lemmas (81.6 vs. 81.3%) when no grammatical filtering is applied.

5. Conclusion and future work

We have evaluated the impact of some linguistic knowledge on WSD performances using a classification method based on the kNN algorithm. The best scores were obtained with five different parameter combinations, what corroborate (Habert & al., 1997)'s conclusion according to which the best linguistic model does not exist, theoretically. Every kind of linguistic knowledge has fluctuating effects depending particularly on the other kind of linguistic knowledge it is combined with and on the NLP application for which the WSD method concerned will be a sub-task.

Various improving factors should be considered, like combining both neighbors and syntactic cooccurrents, or using neighbors from windows of size higher than 1 and second order cooccurrences from every kind of context. Besides, we could define the semantic spaces differently according to the grammatical tag of the ambiguous word as suggested by (Habert & al., 1997) (adjectival and adverbial cooccurrents are certainly more semantically informative for nouns than for verbs, for example). We could also extend the training corpora and define more fine-grained entries in the bilingual lexicon by using more than one TL.

Finally, we should do the same experiment using WSD methods based on other learning techniques, like SVM and SCT, and complete our evaluation with a comparative one, both in monolingual and multilingual frames.

6. References

Aït-Mokhtar, S., Chanod, JP., Roux, C. (2002). Robustness beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8(2/3):121-144, Cambridge University Press.

Apidianaki, M. (2009). Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference on European Chapter of the ACL (EACL)*, pp. 77-85, Athènes, Grèce.

Baroni, M., Bisi, S. (2004) Using cooccurrence statistics and the Web to discover synonyms in technical language. In *Proceedings of the International Conference on Language Resources and Evaluation 2004*.

Carpuat, M., Wu, D. (2007) Word Sense disambiguation vs. statistical machine translation. In *Proceedings of the 43th Annual Meeting of the ACL*, p. 387-394, Ann Arbor, Michigan.

Crego, J.M., Max, A., Yvon, F. (2009) Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical. In *TALN'09* (à paraître).

Gale, W., Yarowsky, D. (1993) A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415-439.

Gavrilidou, M., Labropoulou, P., Desipri,, E., Antonopoulos, V. Piperidis, S. (2004) Building parallel corpora for eContent professionals. In *Proceedings of the Workshop on Multilingual Linguistic resources*, 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland, pp. 90-93.

Habert, B., Nazarenko, A., Salem, A. (1997) *Les linguisitques de corpus*. Armand Colin/Masson, Paris.

Kaji, H., Morimoto, Y. (2002) Unsupervised word sense disambiguation using bilingual corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pp. 1-17.

Keok Lee, Y., Tou Ng, H., Kiah Chia, T. (2004) Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of SENSEVAL-3*, pp.137-140, Barcelona, Spain.

Koehn, P. (2005) *EuroParl : A parallel corpus for Statistical Machine Translation*. MT Summit.

De Loupy, C., El-Bèze, M., Marteau, P.-F. (2000) Using Semantic Classification Trees for WSD. *Computers and the Humanities*, 34:187-192.

Veenstra, J., van den Bosch, A., Buchholz, A., Daelemans, W., Zavrel, J. (2000) Memory-based Word Sense disambiguation. *Computers and the Humanities*, 34:171-177.

Paumier, S. (2008) *Unitex 2.0 User Manual*.

Schmidt, H. (1994) Probabilistic part-of-speech tagging using Decision Trees. In *Actes de la Conférence Internationale sur les Nouvelles Méthodes en Traitement du Langage*, septembre.