Disambiguation of Partial Cognates

Oana Frunza and Diana Inkpen
E-mail: {ofrunza,diana@site.uottawa.ca}
School of Information Technology and Engineering; University of Ottawa;
Ottawa, ON, Canada, K1N 6N5

1. Introduction

Cognates – words that have similar spelling and meaning in two or more languages – can accelerate vocabulary acquisition and facilitate the reading comprehension task. A student has to pay attention to the pairs of words that look and sound similar but have different meanings – false-friend pairs, and especially to pairs of words that share meanings in some but not all contexts – partial cognates.

Our goal is to present a method to disambiguate French words that are partial cognate to English words. The task of disambiguating partial cognates for French and English can be seen as coarse-grain cross-language Word-Sense Disambiguation (WSD) task. A lot of work has been done on monolingual WSD systems that use supervised and unsupervised methods and report good results on Senseval data, but there is less work on disambiguating words across two cross-languages.

Although French and English belong to different branches of the Indo-European family of languages, their vocabularies share a great number of similarities due to the geographical, historical, and cultural contact between the two countries over many centuries. Most of these borrowings have changed their orthography and most likely their meaning as well.

Second language learners of French, native speakers of English, can be assisted by a partial-cognate disambiguation system during the learning process. Claims that false friends can be a hindrance in second language learning are supported by Carroll (1992). She suggested that a cognate pairing process between two words that look alike happens faster in the learner's mind than a false-friend pairing. Experiments with second language learners of different stages conducted by Heuven et al. (1998) suggest that missing false-friend recognition can be corrected when cross-language activation is used.

Besides second language learning, Machine Translation (MT) systems can also benefit from extra information when translating a certain word in context. Knowing if a French word is a cognate or a false friend with an English word can improve translation results. Cross-Language Information Retrieval systems can also use the knowledge of the sense of certain words in a query.

We describe a supervised and a semi-supervised method to discriminate the senses of a partial cognate in a French text (according to its English cognate or false-friend sense). The methods are based on Machine Learning (ML) techniques. The semi-supervised method uses a monolingual and bilingual bootstrapping technique. We use parallel corpora to automatically create training data / seeds for the bootstrapping techniques. Our methods are independent of the language pair at hand; they can be applied to any pair of languages for which a parallel corpus and two monolingual text collections are available.

2. Related Work

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons (Simard et al., 1992). Brew and McKelvie (1996) extracted French-English cognates and false friends from aligned bitexts using simple orthographic similarity measures. Kondrak (2001) identified cognates between various pairs of languages, paying attention to phonetic aspects, especially for **genetic cognates** – words in related languages that derive directly from the same word in the ancestor (proto)-language.

For French and English, substantial work on cognate detection was done manually. LeBlanc and Séguin (1996) concluded that cognates appear to make up over 30% of the French vocabulary. Inkpen et al. (2005) have looked at different combinations of orthographic similarity measures using automatic techniques (ML) to identify cognates and false friends between French and English.

From the wealth of publications on WSD we have chosen to briefly discuss only the ones that are related to our work. Determining the sense of an ambiguous word, using bootstrapping and texts from a different language was done by Yarowsky (1995), Hearst (1991), Diab and Resnik (2002), and Li and Li (2004). Yarowsky (1995) has used a few seeds and untagged sentences in a bootstrapping algorithm based on decision lists. He added two constrains – words tend to have one sense per discourse and one sense per collocation. The monolingual bootstrapping approach was used by Hearst (1991), who used a small set of hand-labeled data to bootstrap from a larger corpus for training a noun disambiguation system for English. Diab and Resnik (2002) used cross-language lexicalization for an English monolingual unsupervised WSD system.

The difference between our approach and the ones mentioned above, is that our technique uses the whole sentences from the parallel text, not only the target words (the translation of certain English words) like (Diab and Resnik, 2002); our focus is not only on nouns as in (Hearst, 1991), and we look at

words that are difficult to disambiguate even for humans, not only at words with distinct senses as in (Li and Li, 2004) and (Yarowsky, 1995).

Our task, disambiguating partial cognates between two languages is a new task, different than the Word Translation Disambiguation task because we do not see each translation as a different sense of a target word (two or more possible translation can have the same meaning). We perform a coarse-grained cross-lingual disambiguation into two senses: cognate and false friend. We use automatically-collected training data, eliminating the costly effort of the manual annotation; off-the-shelf ML and MT tools; and existing parallel corpora.

3. Data for Partial Cognate Disambiguation

We performed experiments with ten pairs of partial cognates (a not untypical size of test data in the WSD literature). For a French partial cognate we list its English cognate and several false-friend words in English. Often the French partial cognate has two senses (one for cognate, one for false friend), but sometimes it has more than two senses: one for cognate and several for false friends (nonetheless, we treat the false friends senses together). For example, the false-friend words for *note* include one sense for *grade*, *mark*, and one for *bill*, *check*, *account*. We selected ten partial cognates (for which we had enough parallel sentences), from a list of 64 partial cognates[1]. These ten partial cognates are words frequently used in the language. The words we worked with are:

French partial cognate (PC); English cognate (COG); English false friends (FF):

1. blanc; blank; white, livid

2. circulation; circulation; traffic

3. client; client; customer, patron, patient, spectator, user, shopper

4. corps; corps; body, corpse

5. détail; detail; retail

6. mode; mode; fashion, trend, style, vogue

7. note; note; mark, grade, bill, check, account

8. police; police; policy, insurance, font, face

9. responsable; responsible; in charge, responsible party, official, representative, person in charge, executive, officer

10. route; route; road, roadside

Both the supervised and the semi-supervised methods use a set of seeds. The seeds are parallel sentences, French and English, which contain the partial cognate. For each partial-cognate word, a part of the set contains parallel sentences with the cognate sense and the other part the false-friend sense.

[1] http://french.about.com/library/fauxamis/blfauxam_a.htm

Table I. Number of parallel sentences used as seeds.

Partial Cognates	Train COG	Train FF	Test COG	Test FF
Blanc	54	78	28	39
Circulation	213	75	107	38
Client	105	88	53	45
Corps	88	82	44	42
Dètail	120	80	60	41
Mode	76	104	126	53
Note	250	138	126	68
Police	154	94	78	48
Responsable	200	162	100	81
Route	69	90	35	46

The seed sentences are not hand-tagged with the sense (the cognate or false-friend), they are automatically annotated by the way we collect them. To collect the set of seed sentences we use parallel corpora from Hansard[2], EuroParl[3], and the manually aligned BAF corpus[4]. The cognate sense sentences were created by extracting parallel sentences that had on the French side the French cognate and on the English side the English cognate. The same approach was used to extract sentences with the false-friend sense of the partial cognate, only this time we used the false-friend English words. Here are examples of sentences from parallel corpus:

Fr (PC:COG) *Je note, par exemple, que l'accusé a fait une autre déclaration très incriminante à Hall environ deux mois plus tard.*

En (COG) *I note, for instance, that he made another highly incriminating statement to Hall two months later.*

Fr (PC:FF) *S' il gèle les gens ne sont pas capables de régler leur note de chauffage.*

En (FF) *If there is a hard frost, people are unable to pay their bills.*

We used 2/3 of the sentences for training (seeds) and 1/3 for testing when applying both the supervised and semi-supervised approach. In Table I we present the number of seeds used for training and testing.

Because our goal is to disambiguate partial cognates in general, not only in the particular domain of Hansard and EuroParl we created another set of automatically extracted and labeled sentences from a 1.5 million words multi-domain parallel corpus of magazine articles, modern fiction, texts from international organizations and academic textbooks (we will call this corpus MDC[5]). The number of extracted parallel sentences for the two senses varied from none to a maximum of 288. We used this corpus to perform further experiments.

[2] http://www.isi.edu/natural-language/download/hansard/; http://www.tsrali.com/

[3] http://people.csail.mit.edu/koehn/publications/europarl/

[4] http://rali.iro.umontreal.ca/Ressources/BAF/

[5] The MDC corpus was provided by Prof. Raphael Salkie, Brighton University, UK

4. Methods

In this section we describe our supervised and semi-supervised method. The goal is to determine which of the two senses (the cognate or the false-friend sense) of a partial-cognate word is present in a test sentence. Therefore the classes in which we classify a sentence are: COG (cognate) and FF (false-friend).

Supervised Method For both the supervised and semi-supervised method we used the bag-of-words (BOW) approach of modeling context, with binary values for the features. The features are words from the training corpus that appeared at least 3 times after removing the stopwords[6]. We ran experiments when we kept the stopwords as features but the results did not improve.

As a baseline for the experiments that we present we used the ZeroR classifier from WEKA[7], which predicts the class that is the most frequent in the training corpus. The classifier for which we report results is Naïve Bayes with a kernel estimator (NB-K). We performed experiments with other classifiers as well, with no better results. The supervised method consists in training the classifiers on the automatically-collected training seed sentences, for each partial cognate, and then test their performance on the test set.

Semi-Supervised Method For the semi-supervised method we add unlabeled examples, an average of 200 sentences for each of the senses, from monolingual corpora: the French newspaper Le Monde[8] 1994, 1995 (LM), and the BNC[9] corpus; these are different domain corpora than the seeds. The procedure of adding and using this unlabeled data is described below.

Monolingual Bootstrapping The monolingual bootstrapping algorithm that we used for experiments on French sentences (MB-F) and on English sentences (MB-E) is:

For each pair of partial cognates (PC):
1. Train a classifier on the training seeds.
2. Apply the classifier on unlabeled data, sentences that contain the PC word, extracted from Le Monde (MB-F) or from BNC (MB-E).
3. Take the first few newly classified sentences, both from the COG and FF class and add them to the training seeds.
4. Rerun the experiments training on the new training set.

For the first step of the algorithm we used NB-K classifier because it was the classifier that consistently performed better. We chose to perform attribute selection on the features after we tried the method without attribute selection. We obtained better results when using attribute selection. This sub-step was performed with the WEKA tool, the Chi-Square attribute selection was cho-

[6] http://www.site.uottawa.ca/ diana/csi5180/StopWords
[7] http://www.cs.waikato.ac.nz/ ml/weka/
[8] http://www.lemonde.fr/
[9] http://www.natcorp.ox.ac.uk/

sen because is commonly used for text processing tasks. In the second step of the MB algorithm the classifier that was trained on the training seeds was then used to classify the unlabeled data that was collected from the two additional resources. For the MB algorithm on the French side we trained the classifier on the French side of the training seeds and then we applied the classifier to classify the sentences that were extracted from Le Monde and contained the partial cognate. The same approach was used for the MB on the English side only this time we were using the English side of the training seeds for training the classifier and the BNC corpus to extract new examples. In fact, the MB-E step is needed only for the BB method. Only the sentences that were classified with a probability greater than 0.85 (experimentally chosen value) were selected for later use in the bootstrapping algorithm.

Bilingual Bootstrapping The algorithm for bilingual bootstrapping that we used in our experiments is:
1. Translate the English sentences that were collected in the MB-E step into French using an online MT[10] *tool and add them to the French training data.*
2. Execute the MB-F step (in order to re-train the classifier on the new labeled data and the original seeds).

The BB algorithm uses as a new source of knowledge sentences that were selected in the MB-E experiments. It has been shown (Li and Li, 2004) that two languages are more informative than one and since that task that we need to solve is similar to a cross-language word sense disambiguation the idea of using knowledge from English was straightforward.

5. Evaluation and Results

In this section we present the results that we obtained with our methods. Given limited space, we can only show a representative sample of the results. Table II contains the results for the supervised method, and for the MB and BB algorithms, on the French side. In the last line of the table (AVERAGE_MDC), we show the averaged results obtained when using as test set the multi-domain corpus.

Discussion of the Results The results of the experiments and the methods that we propose show that we can successfully use unlabeled data to learn from, and that the noise that is introduced due to the seed set collection is tolerable by the ML techniques that we use. The supervised method improves over the baseline with 20% for the test set and 15% for the MDC corpus.

The BB method improved the results on the NB-K classifier with 3.24%, compared with the supervised method (no bootstrapping), when we tested only on the test set, the one that represents 1/3 of the initially-collected parallel sentences. BB with NB-K brought an improvement of 1.95% from no

[10] http://www.freetranslation.com/free/web.asp

Table II. Results for the Supervised Method (SM), Monolingual Bootstrapping (MB), and Bilingual Bootstrapping (BB) methods on the initial test set data and on the multi-domain corpus.

		SM	MB	BB
PC	Baseline	NB-K	NB-K	NB-K
Blanc	58.00%	95.52%	97.01%	95.52%
Circulation	74.00%	91.03%	90.34%	92.41%
Client	54.08%	67.34%	77.55%	70.40%
Corps	51.16%	62.00%	78.00%	83.00%
Dètail	59.40%	85.14%	88.11%	91.08%
Mode	58.24%	89.01%	89.01%	87.91%
Note	64.94%	89.17%	85.05%	85.56%
Police	61.41%	79.52%	71.65%	80.31%
Responsable	55.24%	85.08%	87.29%	87.84%
Route	56.79%	54.32%	51.85%	60.49%
AVERAGE	59.33%	80.17%	**80.96%**	**83.41%**
AVERAGE_MDC	67.00%	71.97%	67.03%	**73.92%**

bootstrapping, when we tested on the multi-domain corpus, the line for AVERAGE_MDC. According to a t-test this improvement is statistically significant.

For some experiments MB did better, for others BB was the method that improved the performance; nonetheless for some combinations of experiments (we performed additional experiments when we used the multi-domain corpus in the training data set as well and experiments when we combined the two semi-supervised methods) MB together with BB was the method that worked the best. Improvements over the supervised method were always obtained using the semi-supervised methods. This observation is also valid in experiments with different combinations of training and testing data sets that we conducted for our task.

Another positive aspect that we want to emphasize throughout the experiments that we performed is that the number of features that were extracted from the seeds was more than double at each MB and BB experiment, showing that even though we started with seeds from a restricted domain, the method is able to capture knowledge form different domains as well. Besides the change in the number of features, the domain of the features has also changed form the parliamentary one to others, more general, showing that the method will be able to disambiguate sentences where the partial cognates cover different types of context.

Unlike previous work that has been done with monolingual or bilingual bootstrapping, we tried to disambiguate not only words that have senses that are very different, e.g., *plant* with a sense of *biological plant* or with the sense

of *factory*. In our set of partial cognates the French word *route* is a difficult word to disambiguate even for humans: it has a cognate sense when it refers to a *maritime* or *trade route* and a false-friend sense when it is used as *road*. The same observation applies to *client* (the cognate sense is *client*, and the false-friend sense is *customer*, *patron*, or *patient*) and to *circulation* (cognate in *air* or *blood circulation*, false friend in *street traffic*).

6. Conclusion and Future Work

We showed that with simple methods and using available tools we can achieve good results in the task of partial cognate disambiguation. The accuracy might be increased by using dependency relations, lemmatization, part-of-speech tagging (to extract sentences where the partial cognate has the same POS in both languages), and other types of data representation combined with other semantic tools. In future work we plan to try different representations of the data, to use knowledge of the relations that exists between the partial cognate and the context words, and to run experiments when we iterate the MB and BB steps more than once.

References

Brew, C. and D. McKelvie: 1996, 'Word-pair extraction for lexicography'. In: *Procs. of 2nd Intl. Conf. on New Methods in Language Processing*. Ankara, Turkey, pp. 45–55.

Carroll, S.: 1992, 'On Cognates'. Technical report, Second Language Research.

Diab, M. and P. Resnik: 2002, 'An unsupervised method for word sense tagging using parallel corpora'. *ACL-2002* pp. 255–262.

Hearst, M.: 1991, 'Noun homograph disambiguation using local context in large corpora'. *Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research* pp. 1–19.

Heuven, W. V., A. Dijkstra, and J. Grainger: 1998, 'Orthographic neighborhood effects in bilingual word recognition'. *Journal of Memory and Language* **39**, 458–483.

Inkpen, D., O. Frunza, and G. Kondrak: 2005, 'Automatic identification of Cognates and False Friends in French and English'. In: *In RANLP-2005*. Bulgaria, pp. 251–257.

Kondrak, G.: 2001, 'Identifying Cognates by Phonetic and Semantic Similarity'. In: *Proceedings of the 2nd Meeting of the NAACL*. pp. 103–110.

LeBlanc, R. and H. Séguin: 1996, *Les congénères homographes et parographes anglais-français*. Twenty-Five Years of Second Language Teaching at the Univ. of Ottawa.

Li, H. and C. Li: 2004, 'Word translation disambiguation using bilingual bootstraping'. *Computational Linguistics* **30**(1), 1–22.

Simard, M., G. F. Foster, and P. Isabelle: 1992, 'Using Cognates to Align Sentences in Bilingual Corpora'. In: *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal, Canada, pp. 67–81.

Yarowsky, D.: 1995, 'Unsupervised Word Sense Disambiguation Rivaling Supervised Methods'. *In Proc. of the 33th Annual Meeting of the ACL-95* pp. 189–196.