Semi-Supervised Learning of Partial Cognates using Bilingual Bootstrapping

Oana Frunza and **Diana Inkpen**

School of Information Technology and Engineering
University of Ottawa
Ottawa, ON, Canada, K1N 6N5
{ofrunza,diana}@site.uottawa.ca

Abstract

Partial cognates are pairs of words in two languages that have the same meaning in some, but not all contexts. Detecting the actual meaning of a partial cognate in context can be useful for Machine Translation tools and for Computer-Assisted Language Learning tools. In this paper we propose a supervised and a semi-supervised method to disambiguate partial cognates between two languages: French and English. The methods use only automatically-labeled data; therefore they can be applied for other pairs of languages as well. We also show that our methods perform well when using corpora from different domains.

1 Introduction

When learning a second language, a student can benefit from knowledge in his / her first language (Gass, 1987), (Ringbom, 1987), (LeBlanc *et al.* 1989). Cognates – words that have similar spelling and meaning – can accelerate vocabulary acquisition and facilitate the reading comprehension task. On the other hand, a student has to pay attention to the pairs of words that look and sound similar but have different meanings – false friends pairs, and especially to pairs of words that share meaning in some but not all contexts – the partial cognates.

Carroll (1992) claims that false friends can be a hindrance in second language learning. She suggests 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 Van et al. (1998) suggest that missing false-friend recognition can be corrected when cross-language activation is used – sounds, pictures, additional explanation, feedback.

Machine Translation (MT) systems can benefit from extra information when translating a certain word in context. Knowing if a word in the source language is a cognate or a false friend with a word in the target language can improve the translation results. Cross-Language Information Retrieval systems can use the knowledge of the sense of certain words in a query in order to retrieve desired documents in the target language.

Our task, disambiguating partial cognates, is in a way equivalent to coarse grain cross-language Word-Sense Discrimination. Our focus is disambiguating French partial cognates in context: deciding if they are used as cognates with an English word, or if they are used as false friends.

There is a lot of work done on monolingual Word Sense Disambiguation (WSD) systems that use supervised and unsupervised methods and report good results on Senseval data, but there is less work done to disambiguate cross-language words. The results of this process can be useful in many NLP tasks.

Although French and English belong to different branches of the Indo-European family of languages, their vocabulary share a great number of similarities. Some are words of Latin and Greek origin: e.g., *education* and *theory*. A small number of very old, "genetic" cognates go back all the way to Proto-Indo-European, e.g., *mére* - *mother* and *pied* - *foot*. The majority of these pairs of words penetrated the French and English language due to the geographical, historical, and cultural contact between the two countries over

many centuries (borrowings). Most of the borrowings have changed their orthography, following different orthographic rules (LeBlanc and Seguin, 1996) and most likely their meaning as well. Some of the adopted words replaced the original word in the language, while others were used together but with slightly or completely different meanings.

In this paper we describe a supervised and also a semi-supervised method to discriminate the senses of partial cognates between French and English. In the following sections we present some definitions, the way we collected the data, the methods that we used, and evaluation experiments with results for both methods.

2 Definitions

We adopt the following definitions. The definitions are language-independent, but the examples are pairs of French and English words, respectively.

Cognates, or True Friends (Vrais Amis), are pairs of words that are perceived as similar and are mutual translations. The spelling can be identical or not, e.g., *nature* - *nature*, *reconnaissance* - *recognition*.

False Friends (Faux Amis) are pairs of words in two languages that are perceived as similar but have different meanings, e.g., *main* (= *hand*) - *main* (= *principal* or *essential*), *blesser* (= to *injure*) - *bless* (= bénir).

Partial Cognates are pairs of words that have the same meaning in both languages in some but not all contexts. They behave as cognates or as false friends, depending on the sense that is used in each context. For example, in French, *facteur* means not only *factor*, but also *mailman*, while *étiquette* can also mean *label* or *sticker*, in addition to the cognate sense.

Genetic Cognates are word pairs in related languages that derive directly from the same word in the ancestor (proto-)language. Because of gradual phonetic and semantic changes over long periods of time, genetic cognates often differ in form and/or meaning, e.g., *père* - *father*, *chef* - *head*. This category excludes lexical borrowings, i.e., words transferred from one language to another at some point of time, such as *concierge*.

3 Related Work

As far as we know there is no work done to disambiguate partial cognates between two languages.

Ide (2000) has shown on a small scale that cross-lingual lexicalization can be used to define and structure sense distinctions. Tufis et al. (2005) used cross-lingual lexicalization, wordnets alignment for several languages, and a clustering algorithm to perform WSD on a set of polysemous English words. They report an accuracy of 74%.

One of the most active researchers in identifying cognates between pairs of languages is Kondrak (2001; 2004). His work is more related to the phonetic aspect of cognate identification. He used in his work algorithms that combine different orthographic and phonetic measures, recurrent sound correspondences, and some semantic similarity based on glosses overlap. Guy (1994) identified letter correspondence between words and estimates the likelihood of relatedness. No semantic component is present in the system, the words are assumed to be already matched by their meanings. Hewson (1993), Lowe and Mazadon (1994) used systematic sound correspondences to determine proto-projections for identifying cognate sets.

WSD is a task that has attracted researchers since 1950 and it is still a topic of high interest. Determining the sense of an ambiguous word, using bootstrapping and texts from a different language was done by Yarowsky (1995), Hearst (1991), Diab (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. He reported high accuracy scores for a set of 10 words. The monolingual bootstrapping approach was also 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. Unlike Yarowsky (1995), we use automatic collection of seeds. Besides our monolingual bootstrapping technique, we also use bilingual bootstrapping.

Diab (2002) has shown that unsupervised WSD systems that use parallel corpora can achieve results that are close to the results of a supervised approach**.** She used parallel corpora in French, English, and Spanish, automatically-produced with MT tools to determine cross-language lexicalization sets of target words. The major goal of her work was to perform monolingual English WSD. Evaluation was performed on the nouns from the English all words data in Senseval2. Additional knowledge was added to the system

from WordNet in order to improve the results. In our experiments we use the parallel data in a different way: we use words from parallel sentences as features for Machine Learning (ML). Li and Li (2004) have shown that word translation and bilingual bootstrapping is a good combination for disambiguation. They were using a set of 7 pairs of Chinese and English words. The two senses of the words were highly distinctive: *e.g. bass* as *fish* or *music; palm* as *tree* or *hand.*

Our work described in this paper shows that monolingual and bilingual bootstrapping can be successfully used to disambiguate partial cognates between two languages. Our approach differs from the ones we mentioned before not only from the point of human effort needed to annotate data – we require almost none, and from the way we use the parallel data to automatically collect training examples for machine learning, but also by the fact that we use only off-the-shelf tools and resources: free MT and ML tools, and parallel corpora. We show that a combination of these resources can be used with success in a task that would otherwise require a lot of time and human effort.

4 Data for Partial Cognates

We performed experiments with ten pairs of partial cognates. We list them in Table 1. For a French partial cognate we list its English cognate and several false friends 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 them together). For example, the false friend words for *note* have one sense for *grades* and one for *bills*.

The partial cognate (PC), the cognate (COG) and false-friend (FF) words were collected from a web resource[1]. The resource contained a list of 400 false-friends with 64 partial cognates. All partial cognates are words frequently used in the language. We selected ten partial cognates presented in Table 1 according to the number of extracted sentences (a balance between the two meanings), to evaluate and experiment our proposed methods.

The human effort that we required for our methods was to add more false-friend English words, than the ones we found in the web resource. We wanted to be able to distinguish the senses of cognate and false-friends for a wider variety of senses. This task was done using a bilingual dictionary[2].

Table 1. The ten pairs of partial cognates.

French partial cognate	**English cognate**	**English false friends**
blanc	blank	white, livid
circulation	circulation	traffic
client	client	customer, patron, patient, spectator, user, shopper
corps	corps	body, corpse
détail	detail	retail
mode	mode	fashion, trend, style, vogue
note	note	mark, grade, bill, check, account
police	police	policy, insurance, font, face
responsable	responsible	in charge, responsible party, official, representative, person in charge, executive, officer
route	route	road, roadside

4.1 Seed Set Collection

Both the supervised and the semi-supervised method that we will describe in Section 5 are using 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 the cognate sense and another part the false-friend sense.

As we mentioned in Section 3, the seed sentences that we use are not hand-tagged with the sense (the cognate sense or the false-friend sense); they are automatically annotated by the way we collect them. To collect the set of seed sentences we use parallel corpora from Hansard[3], and EuroParl[4], and the, manually aligned BAF corpus.[5]

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. See the upper part of Table 2 for an example.

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. See lower the part of Table 2.

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

[2] http://www.wordreference.com

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

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

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

Table 2. Example 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*.

To keep the methods simple and language-independent, no lemmatization was used. We took only sentences that had the exact form of the French and English word as described in Table 1. Some improvement might be achieved when using lemmatization. We wanted to see how well we can do by using sentences as they are extracted from the parallel corpus, with no additional pre-processing and without removing any noise that might be introduced during the collection process.

From the extracted sentences, 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 3 we present the number of seeds used for training and testing.

We will show in Section 6, that even though we started with a small amount of seeds from a certain domain – the nature of the parallel corpus that we had, an improvement can be obtained in discriminating the senses of partial cognates using free text from other domains.

Table 3. Number of parallel sentences used as seeds.

Partial Cognates	Train CG	Train FF	Test CG	Test FF
Blanc	54	78	28	39
Circulation	213	75	107	38
Client	105	88	53	45
Corps	88	82	44	42
Détail	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
AVERAGE	132.9	99.1	66.9	50.1

5 Methods

In this section we describe the supervised and the semi-supervised methods that we use in our experiments. We will also describe the data sets that we used for the monolingual and bilingual bootstrapping technique.

For both methods we have the same goal: 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. The classes in which we classify a sentence that contains a partial cognate are: COG (cognate) and FF (false-friend).

5.1 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 were words from the training corpus that appeared at least 3 times in the training sentences. We removed the stopwords from the features. A list of stopwords for English and one for French was used. We ran experiments when we kept the stopwords as features but the results did not improve.

Since we wanted to learn the contexts in which a partial cognate has a cognate sense and the contexts in which it has a false-friend sense, the cognate and false friend words were not taken into account as features. Leaving them in would mean to indicate the classes, when applying the methods for the English sentences since all the sentences with the cognate sense contain the cognate word and all the false-friend sentences do not contain it. For the French side all collected sentences contain the partial cognate word, the same for both senses.

As a baseline for the experiments that we present we used the ZeroR classifier from WEKA[6], which predicts the class that is the most frequent in the training corpus. The classifiers for which we report results are: Naïve Bayes with a kernel estimator, Decision Trees - J48, and a Support Vector Machine implementation - SMO. All the classifiers can be found in the WEKA package. We used these classifiers because we wanted to have a probabilistic, a decision-based and a functional classifier. The decision tree classifier allows us to see which features are most discriminative.

Experiments were performed with other classifiers and with different levels of tuning, on a 10-fold cross validation approach as well; the classifiers we mentioned above were consistently the ones that obtained the best accuracy results.

The supervised method used in our experiments consists in training the classifiers on the

[6] http://www.cs.waikato.ac.nz/ml/weka/

automatically-collected training seed sentences, for each partial cognate, and then test their performance on the testing set. Results for this method are presented later, in Table 5.

5.2 Semi-Supervised Method

For the semi-supervised method we add unlabelled examples from monolingual corpora: the French newspaper LeMonde[7] 1994, 1995 (LM), and the BNC[8] corpus, different domain corpora than the seeds. The procedure of adding and using this unlabeled data is described in the Monolingual Bootstrapping (MB) and Bilingual Bootstrapping (BB) sections.

5.2.1 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 – using the BOW approach and a NB-K classifier with attribute selection on the features.
2. Apply the classifier on unlabeled data – sentences that contain the PC word, extracted from LeMonde (MB-F) or from BNC (MB-E)
*3. Take the first **k** newly classified sentences, both from the COG and FF class and add them to the training seeds (the most confident ones – the prediction accuracy greater or equal than a threshold =0.85)*
4. Rerun the experiments training on the new training set
*5. Repeat steps 2 and 3 for **t** times*
endFor

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 chosen.

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 LeMonde 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 were selected for later use in the bootstrapping algorithm.

The number of sentences that were chosen from the new corpora and used in the first step of the MB and BB are presented in Table 4.

Table 4. Number of sentences selected from the LeMonde and BNC corpus.

PC	LM COG	LM FF	BNC COG	BNC FF
Blanc	45	250	0	241
Circulation	250	250	70	180
Client	250	250	77	250
Corps	250	250	131	188
Détail	250	163	158	136
Mode	151	250	176	262
Note	250	250	178	281
Police	250	250	186	200
Responsable	250	250	177	225
Route	250	250	217	118

For the partial-cognate *Blanc* with the cognate sense, the number of sentences that had a probability distribution greater or equal with the threshold was low. For the rest of partial cognates the number of selected sentences was limited by the value of parameter ***k*** in the algorithm.

5.2.2 Bilingual Bootstrapping

The algorithm for bilingual bootstrapping that we propose and tried in our experiments is:

1. Translate the English sentences that were collected in the MB-E step into French using an online MT[9] tool and add them to the French seed training data.
*2. Repeat the MB-F and MB-E steps for **T** times.*

For the both monolingual and bilingual bootstrapping techniques the value of the parameters **t** and **T** is 1 in our experiments.

[7] http://www.lemonde.fr/
[8] http://www.natcorp.ox.ac.uk/
[9] http://www.freetranslation.com/free/web.asp

6 Evaluation and Results

In this section we present the results that we obtained with the supervised and semi-supervised methods that we applied to disambiguate partial cognates.

Due to space issue we show results only for testing on the testing sets and not for the 10-fold cross validation experiments on the training data. For the same reason, we present the results that we obtained only with the French side of the parallel corpus, even though we trained classifiers on the English sentences as well. The results for the 10-fold cross validation and for the English sentences are not much different than the ones from Table 5 that describe the supervised method results on French sentences.

Table 5. Results for the Supervised Method.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58%	95.52%	98.5%	98.5%
Circulation	74%	91.03%	80%	89.65%
Client	54.08%	67.34%	66.32%	61.22%
Corps	51.16%	62%	61.62%	69.76%
Détail	59.4%	85.14%	85.14%	87.12%
Mode	58.24%	89.01%	89.01%	90%
Note	64.94%	89.17%	77.83%	85.05%
Police	61.41%	79.52%	93.7%	94.48%
Responsable	55.24%	85.08%	70.71%	75.69%
Route	56.79%	54.32%	56.79%	56.79%
AVERAGE	59.33%	**80.17%**	77.96%	80.59%

Table 6 and Table 7 present results for the MB and BB. More experiments that combined MB and BB techniques were also performed. The results are presented in Table 9.

Our goal is to disambiguate partial cognates in general, not only in the particular domain of Hansard and EuroParl. For this reason we used another set of automatically determined sentences from a multi-domain parallel corpus.

The set of new sentences (multi-domain) was extracted in the same manner as the seeds from Hansard and EuroParl. The new parallel corpus is a small one, approximately 1.5 million words, but contains texts from different domains: magazine articles, modern fiction, texts from international organizations and academic textbooks. We are using this set of sentences in our experiments to show that our methods perform well on multi-domain corpora and also because our aim is to be able to disambiguate PC in different domains. From this parallel corpus we were able to extract the number of sentences shown in Table 8.

With this new set of sentences we performed different experiments both for MB and BB. All results are described in Table 9. Due to space issue we report the results only on the average that we obtained for all the 10 pairs of partial cognates.

The symbols that we use in Table 9 represent:

S – the seed training corpus, TS – the seed test set, BNC and LM – sentences extracted from LeMonde and BNC (Table 4), and NC – the sentences that were extracted from the multi-domain new corpus. When we use the + symbol we put together all the sentences extracted from the respective corpora.

Table 6. Monolingual Bootstrapping on the French side.

PC	ZeroR	NB-K	Dec.Tree	SMO
Blanc	58.20%	97.01%	97.01%	98.5%
Circulation	73.79%	90.34%	70.34%	84.13%
Client	54.08%	71.42%	54.08%	64.28%
Corps	51.16%	78%	56.97%	69.76%
Détail	59.4%	88.11%	85.14%	82.17%
Mode	58.24%	89.01%	90.10%	85%
Note	64.94%	85.05%	71.64%	80.41%
Police	61.41%	71.65%	92.91%	71.65%
Responsable	55.24%	87.29%	77.34%	81.76%
Route	56.79%	51.85%	56.79%	56.79%
AVERAGE	59.33%	**80.96%**	75.23%	77.41%

Table 7. Bilingual Bootstrapping.

PC	ZeroR	NB-K	Dec.Tree	SMO
Blanc	58.2%	95.52%	97.01%	98.50%
Circulation	73.79%	92.41%	63.44%	87.58%
Client	45.91%	70.4%	45.91%	63.26%
Corps	48.83%	83%	67.44%	82.55%
Détail	59%	91.08%	85.14%	86.13%
Mode	58.24%	87.91%	90.1%	87%
Note	64.94%	85.56%	77.31%	79.38%
Police	61.41%	80.31%	96.06%	96.06%
Responsable	44.75%	87.84%	74.03%	79.55%
Route	43.2%	60.49%	45.67%	64.19%
AVERAGE	55.87%	**83.41%**	74.21%	82.4%

Table 8. New Corpus (NC) sentences.

PC	COG	FF
Blanc	18	222
Circulation	26	10
Client	70	44
Corps	4	288
Détail	50	0
Mode	166	12
Note	214	20
Police	216	6
Responsable	104	66
Route	6	100

6.1 Discussion of the Results

The results of the experiments and the methods that we propose show that we can use with success 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.

Some results of the experiments we present in Table 9 are not as good as others. What is important to notice is that every time we used MB or BB or both, there was an improvement. For some experiments MB did better, for others BB was the method that improved the performance; nonetheless for some combinations MB together with BB was the method that worked best.

In Tables 5 and 7 we show that BB 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 (TS), the one that represents 1/3 of the initially-collected parallel sentences. This improvement is not statistically significant, according to a t-test.

In Table 9 we show that our proposed methods bring improvements for different combinations of training and testing sets. Table 9, lines 1 and 2 show that BB with NB-K brought an improvement of 1.95% from no bootstrapping, when we tested on the multi-domain corpus NC. For the same setting, there was an improvement of 1.55% when we tested on TS (Table 9, lines 6 and 8). When we tested on the combination TS+NC, again BB brought an improvement of 2.63% from no bootstrapping (Table 9, lines 10 and 12). The difference between MB and BB with this setting is 6.86% (Table 9, lines 11 and 12). According to a t-test the 1.95% and 6.86% improvements are statistically significant.

Table 9. Results for different experiments with monolingual and bilingual bootstrapping (MB and BB).

Train	Test	ZeroR	NB-K	Trees	SMO
S (no bootstrapping)	NC	67%	***71.97%***	73.75%	76.75%
S+BNC (BB)	NC	64%	**73.92%**	60.49%	74.80%
S+LM (MB)	NC	67.85%	67.03%	64.65%	65.57%
S +LM+BNC (MB+BB)	NC	64.19%	70.57%	57.03%	66.84%
S+LM+BNC (MB+BB)	TS	55.87%	81.98%	74.37%	78.76%
S+NC (no bootstr.)	TS	57.44%	***82.03%***	76.91%	80.71%
S+NC+LM (MB)	TS	57.44%	82.02%	73.78%	77.03%
S+NC+BNC (BB)	TS	56.63%	**83.58%**	68.36%	82.34%
S+NC+LM+ BNC(MB+BB)	TS	58%	83.10%	75.61%	79.05%
S (no bootstrapping)	TS+NC	62.70%	***77.20%***	77.23%	79.26%
S+LM (MB)	TS+NC	62.70%	**72.97%**	70.33%	71.97%
S+BNC (BB)	TS+NC	61.27%	**79.83%**	67.06%	78.80%
S+LM+BNC (MB+BB)	TS+NC	61.27%	77.28%	65.75%	73.87%

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 language 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 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)*.

7 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 dependencies relations, lemmatization, part-of-speech tagging – extract sentences where the partial cognate has the same POS, and other types of data representation combined with different semantic tools (e.g. decision lists, rule based systems).

In our experiments we use a machine language representation – binary feature values, and we show that nonetheless machines are capable of learning from new information, using an iterative approach, similar to the learning process of humans. New information was collected and extracted by classifiers when additional corpora were used for training.

In addition to the applications that we mentioned in Section 1, partial cognates can also be useful in Computer-Assisted Language Learning (CALL) tools. Search engines for E-Learning can find useful a partial cognate annotator. A teacher that prepares a test to be integrated into a CALL tool can save time by using our methods to automatically disambiguate partial cognates, even though the automatic classifications need to be checked by the teacher.

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

Susane Carroll 1992. On Cognates. *Second Language Research*, 8(2):93-119

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, pp. 255-262.

S. M. Gass. 1987. The use and acquisition of the second language lexicon (Special issue). *Studies in Second Language Acquisition*, 9 (2).

Jacques B. M. Guy. 1994. An algorithm for identifying cognates in bilingual word lists and its applicability to machine translation. *Journal of Quantitative Linguistics*, 1(1):35-42.

Marty Hearst 1991. Noun homograph disambiguation using local context in large text corpora. *7th Annual Conference of the University of Waterloo Center for the new OED and Text Research*, Oxford.

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

John Hewson 1993. A Computer-Generated Dictionary of Proto-Algonquian. Ottawa: Canadian Museum of Civilization.

Nancy Ide. 2000 Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34:1-2, *Special Issue on the Proceedings of the SIGLEX SENSEVAL Workshop*, pp.223-234.

Grzegorz Kondrak. 2004. Combining Evidence in Cognate Identification. *Proceedings of Canadian AI 2004: 17th Conference of the Canadian Society for Computational Studies of Intelligence*, pp.44-59.

Grzegorz Kondrak. 2001. Identifying Cognates by Phonetic and Semantic Similarity. *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pp.103-110.

Raymond LeBlanc and Hubert Séguin. 1996. Les congénères homographes et parographes anglais-français. *Twenty-Five Years of Second Language Teaching at the University of Ottawa*, pp.69-91.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrap. *Computational Linguistics*, 30(1):1-22.

John B. Lowe and Martine Mauzaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20:381-417.

Hakan Ringbom. 1987. *The Role of the First Language in Foreign Language Learning*. Multilingual Matters Ltd., Clevedon, England.

Dan Tufis, Ion Radu, Nancy Ide 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned WordNets. *Proceedings of the 20th International Conference on Computational Linguistics*, COLING 2004, Geneva, pp. 1312-1318.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp 189-196.