Macrothink Institute™

International Journal of Linguistics
ISSN 1948-5425
2009, Vol. 1, No. 1: E2

Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques

Oana Frunza and Diana Inkpen

School of Information Technology and Engineering

University of Ottawa

Ottawa, ON, K1N 6N5, Canada

Tel: 1-613-562-5800 E-mail:{ofrunza,diana}@site.uottawa.ca

Abstract

Cognates are words in different languages that have similar spelling and meaning. They can help a second-language learner on the tasks of vocabulary expansion and reading comprehension. The learner needs to pay attention to pairs of words that appear similar but are in fact false friends, have different meanings. 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 article we present a method to automatically classify a pair of words as cognates or false friends. We use several measures of orthographic similarity as features for classification. We study the impact of selecting different features, averaging them, and combining them through machine learning techniques. We also present a supervised and a semi-supervised method to disambiguate partial cognates between two languages. The methods applied for the partial cognate disambiguation task use only automatically-labeled data therefore they can be applied to other pairs of languages as well. We also show that our methods perform well when using corpora from different domains. We applied all our methods to French and English.

Keywords: Cognates, False friends, Partial cognates, Machine learning

1. Introduction

When learning a second language a student can benefit from knowledge in his/her first language (Gass, 1987, Ringbom, 1987, LeBlanc, 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 meaning, false friend pairs, and especially to pairs of words that share meaning in some but not all contexts, partial cognates.

A French second language learner has to be able to distinguish if the French partial cognate word is used with the same meaning with the English cognate word, orthographically similar and with similar meaning, or with a different meaning, a false friend. For example in the sentence: *L'avocat a accepté et lui a conseillé de ne rien dire a la police. (The advocate accepted and consulted him not to tell anything to the police.)* the French partial cognate *police* has the same meaning, institution, with the English word *police* but in the following sentence: *Il s'agit ni plus ni moins d'une police d'assurance. (It is nothing more than an insurance policy.)* the same French partial cognate has a different meaning, insurance policy, which is different than the English word *police*.

Although French and English belong to different branches of the Indo-European family of languages their vocabularies 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, and here we talk about borrowings. Other cognates can be traced to the conquest of Gaul by Germanic tribes after the collapse of the Roman Empire and by the period of French domination of England after the Norman conquest.

Most of the borrowings have changed their orthography following different orthographic rules (LeBlanc & Séguin, 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.

Cognates have been employed in natural language processing. The applications include sentence alignment (Simard, Foster, & Isabelle 1992, Melamed 1999), inducing translation lexicons Mann and Yarowsky (2001), improving statistical machine translation models Marcu, Kondrak, & Knight (2003), and identification of confusable drug names Kondrak (2004). All these applications depend on an effective method of identifying cognates by computing a numeric score that reflects the likelihood that the two words are cognates.

Machine Translation (MT) systems can 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 could improve the translation results. Cross-Language Information Retrieval systems can also use knowledge of the sense of certain words in a query in order to retrieve desired documents in a target language.

We focus on automatic identification of cognates and false friends for the purpose of preparing lists of them for inclusion in dictionaries and other learning aids. Special cognate lists exist only for very few language pairs. Moreover, it takes a lot of time for humans to prepare such lists manually, while programs can do this very quickly.

The definitions that we adopt are the following:

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*. Some researchers refer to cognates as being pairs of words that are orthographically identical and to near-cognates as the ones that have slightly different spelling. In our work, we adopt the cognate definition for both;

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* (meaning *principal* or *essential*), *blesser* (=*to injure*) - *bless* (that is translated with *bénir* in French);

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*.

Our approach of identifying cognates and false friends is based on several orthographic similarity measures that we use as features for machine learning classification algorithms. We test each feature separately and we also test for each pair of words the average value of all the features. We explore various ways of combining the features by applying several Machine Learning (ML) techniques from the Weka package (Witten & Frank, 2005). The two classes for the automatic classification task are: Cognates/False friends and Unrelated. Cognates and false friends can be distinguished on the basis of an additional "translation" feature. If the two words are translations of each other in a bilingual dictionary they are classified as Cognates. Otherwise, they are assumed to be false friends.

The task of disambiguating partial cognates can be seen as a coarse grain cross-language word-sense discrimination task. 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 competitions, but there is less work on disambiguating cross-language words. The results of this process can be useful for different Natural Language Processing (NLP) tasks and applications.

Our goal is to disambiguate French partial cognate words to help second language learners of French, who are native English speakers, in a reading comprehension task. The same approach that we propose in this article can be followed to help second language learners of English, who are native French speakers, in a reading comprehension task. The only difference is the partial cognate list of pairs that will be used while the methods will be similar. The definition for partial cognates that we adopt is the following:

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.

Our proposed methods are bilingual, can be applied to any pair of languages for which a parallel corpus is available and two monolingual collections of text. Our main focus for our methods is to disambiguate a French partial cognate looking at its English cognate and false friend senses.

To be able to use our methods to disambiguate partial cognates in different domains we

combined and used corpora of different domains. Evaluation and experiments with the semi-supervised method using 10 pairs of French-English partial cognates are also presented and discussed.

Besides the theoretical research that we did, we also implemented a practical application that uses cognates and false friends. We implemented a Computer-Assisted Language Learning (CALL) tool that is capable of annotating cognates and false friends in French texts.

Section 2 presents related work, section 3 is dedicated to the cognate and false friend identification tasks, section 4 to the partial cognate disambiguation(Note1) task, section 5 describes the practical application tool that we implemented, a CALL tool, and section 6 gathers conclusions and future work.

2. Related Work

2.1 Cognates and False Friends Identification

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons. Simard et al. (1992) use cognates to align sentences in bitexts. They employ a very simple test: French-English word pairs are assumed to be cognates if their first four characters are identical.

Brew & McKelvie (1996) extract French-English cognates and false friends from aligned bitexts using a variety of orthographic similarity measures based on DICE's coefficient measure. They look only at pairs of verbs in French and English pairs that were automatically extracted from the aligned corpus.

Guy (1994) identifies 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) and Lowe & Mauzaudon (1994) use systematic sound correspondences to determine proto-projections for identifying cognate sets.

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 especially genetic cognates. He uses in his work algorithms that combine different orthographic and phonetic measures, recurrent sound correspondences, and some semantic similarity based on gloss overlap. In Kondrak (2004) he looks directly at the vocabularies of related languages to determine cognates between languages. Kondrak (2004) reports that a simple average of several orthographic similarity measures outperforms all individual measures for the task of identifying drug names.

Complex sound correspondence was used by Kondrak (2003) to help the process of identifying cognates between languages. The algorithm was initially designed for extracting non-compositional compounds from bitexts and was shown to be capable of determining complex sound correspondences in bilingual word lists. He reports 90% results for precision and recall for cognate identification.

For French and English substantial work on cognate detection was done manually. LeBlanc & Séguin (1996) collected 23,160 French-English cognate pairs from two general-purpose dictionaries: Robert-Collins Collins (1987) and Larousse-Saturne. 6,447 of these cognates

had identical spelling, disregarding diacritics. Since the two dictionaries contain approximately 70,000 entries, cognates appear to make up over 30% of the vocabulary.

There is a considerable amount of work done on cognate identification but not as much for false friends identification. In fact, we could not point out work that is focusing on false friends identification between English and French. Work on cognates and false friends identification for German and English was done by Friel & Kennison (2001).

Barker & Sutcliffe (2000) propose a semi-automatic method of identifing false cognates between English and Polish. The method uses a set of morphological transformation rules to convert an English word into a number of candidate Polish words. The resulted Polish words are classified based on human judgments as being: false cognates, true cognates, unrelated, and nonexistent.

Claims that false friends can be a hindrance in second language learning are supported by Carroll (1992). 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 Heuven, Dijkstra, & Grainger (1998) suggest that missing false friend recognition can be corrected when cross-language activation is used: sounds, pictures, additional explanation, feedback.

MT and Word Alignment tools use with success cognate pairs between two languages. Marcu et al. (2003) report results for experiments aimed to improve translation quality by incorporating the cognate information into translation models. The results confirm that the cognate identification approach can improve the quality of word alignment in bitexts without the need for extra resources.

Cross-Language Information Retrieval systems can use knowledge about the sense of certain words in a query, in order to retrieve desired documents in a target language. So far, we are not aware of work that has been done on cross-language information retrieval system that uses cognate knowledge

Identifying cognates and false friends has been an attractive research area not only for researchers in the NLP domain but also for linguists and psycholinguists.

Compared with the above mentioned research the work that we present in this paper is different and improved by the way we identify cognates and false friends, use various orthographic and phonetic similarity measures in combination with various learning algorithms, and by the way we use the available resources e.g. dictionaries, corpora, etc.

WSD is an NLP task that is attracting 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 is done by (Yarowsky 1995, Hearst 1991, Diab & 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. 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.

Diab & Resnik (2002) used cross-language lexicalization for an English monolingual unsupervised WSD system. Besides the parallel data and MT tools they also used additional knowledge from WordNet in order to improve the results. Their task and technique are different from our task and our methods. The difference is that our technique uses the whole sentence from the parallel text, while Diab & Resnik (2002) are using only the target words (the translation of certain English words.)

Li & 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*.

Vickrey, Biewald, Teyssier, & Koller (2005) propose a method that determines the correct translation of a word in context, a task that they consider as a different formulation of the word-sense disambiguation task. They used the European Parliament English French parallel corpus as a training data for the logistic regression model in order to determine the correct translation in context for a set of ambiguous words that they have chosen. They report a 95% recall for translating English ambiguous words into the correct French translation.

Our task, disambiguating partial cognates between two languages, is a new task different from the above mentioned tasks. It is different than the Word Translation Disambiguation task because we do not see each translation as a different sense of a target word (most of the times it is not true, two or more possible translation can have the same meaning). We make the distinction between the meanings of a target word (French partial cognate) into the English possible translations. We perform a coarse-grained Cross-Lingual Word Sense Disambiguation into two senses: cognate and false friend. We will show later on in the article that the method that we propose is different than the methods used before. Our method is based on a supervised and also a semi-supervised technique that uses bootstrapping, to discriminate the senses of a partial cognate between French and English. In addition to all the methods that use bootstrapping and parallel text, we bootstrap our method with corpora from different domains. Hansard, the French-English parallel text, is one of the biggest and well-known parallel corpora Our method uses a small set of seeds from Hansard, but additional knowledge from different domains is added using a bootstrapping technique.

Our work described in this article shows that monolingual and bilingual bootstrapping can be successfully used in disambiguating partial cognates between two languages. The strength of our methods is that are using automatically-collected training data, eliminating the costly effort of the manual annotation, and in the fact that we use only off-the-shelf tools and resources: free Machine Translation and Machine Learning tools, and parallel corpora.

3. Cognates and False Friends Identification

This section presents the methods and data sets that we use to experiment and evaluate the methods that we propose for cognates and false friend identification and partial cognate disambiguation. We also present evaluation results for all methods.

3.1 Data Sets for Cognate and False Friend Identification

The training data set that we use consists of 1454 pairs of French and English words (see Table 1). None of the pairs that we work with contain multi-word expressions. They are extracted from the following sources:

1. An on-line(Note 2) *bilingual list of 1047 basic words and expressions. (After excluding multi-word expressions, we manually classified 203 pairs as Cognates and 527 pairs as Unrelated.)*

2. A manually word-aligned bitext (Melamed, 1998). (We manually identified 258 Cognate pairs among the aligned word pairs.)

3. A set of exercises for Anglophone learners of French (Tréville, 1990) (152 Cognate pairs).

4. An on-line(Note3) *list of "French-English false cognates" (314 false friends).*

A separate test set is composed of 1040 pairs (see Table 1), extracted from the following sources:

1. A random sample of 1000 word pairs from an automatically generated translation lexicon. We manually classified 603 pairs as Cognates and 343 pairs as Unrelated.

2. The above-mentioned on-line list of "French-English false cognates", 94 additional false friends not used for training.

Table 1: Data sets. The numbers in brackets are counts of word pairs that are identical (ignoring accents).

	Training set	Test set
Cognates	613 (73)	603 (178)
False Friends	314 (135)	94 (46)
Unrelated	527 (0)	343 (0)
Total	1454	1040

In order to avoid any overlap between the two sets, we removed from the test set all pairs that happened to be already included in the training set.

The data set has a 2:1 imbalance in favor of the class Cognates/False-Friends; this is not a

problem for the classification algorithms (the precision and recall values are similar for both classes in the experiments presented in Evaluation section 3.4). All the Unrelated pairs in our data sets are translation pairs. It would have been easy to add more pairs that are not translations but we wanted to preserve the natural proportion of cognates in the sample translation lexicons.

From the whole data set, 73 cognates and 135 false friends in the training data set have identical spelling in both languages. When counting identical words we ignore the accents in the French words. The number of identical pairs without ignoring diacritics is 58 cognates and 121 false friends.

This is the data on which we evaluate our proposed methods.

3.2 Orthographic Similarity Measures

The measures that we use as features for the cognate and false friend classification task are described in details in this section.

Many different orthographic similarity measures have been proposed in the literature and their goal is to quantify human perception of similarity, which is often quite subjective. Each measure that we use returns a real value (between 0 and 1, inclusive) that describes the similarity between two words. In the following part of the section we explain each method followed by some examples.

- IDENT is a measure that we use as a baseline. The measure returns 1 if the words are identical and 0 otherwise.
- PREFIX is a simple measure that returns the length of the common prefix divided by the length of the longer string(Note4). E. g., the common prefix for *factory* and *fabrique* has length 2 (the first two letters) which, divided by the length string 8, yields 0.25.
- DICE (Adamson & Boreham, 1974) is calculated by dividing twice the number of shared letter bigrams by the total number of bigrams of both words:

$$DICE\,(X,Y) = \frac{2|bigrams\ (x) \cap bigrams\ (y)|}{|bigrams\ (x)| + |bigrams\ (y)|}$$

  where bigrams(x) is a multi-set of character bigrams in word x. E. g., DICE(*colour, couleur*) = 6/11 = 0.55 (the shared bigrams are *co*, *ou*, *ur*).
- TRIGRAM is defined in the same way as DICE, but employs trigrams instead of bigrams.
- XDICE (Brew & McKelvie, 1996) is also defined in the same way as DICE, but employs "extended bigrams", which are trigrams without the middle letter.

- XXDICE (Brew & McKelvie, 1996) is an extension of the XDICE measure that takes into account the positions of bigrams. Each pair of shared bigrams is weighted by the factor:

$$\frac{1}{1+(pos(a)-pos(b))^2}$$

where pos(a) is the string position of the bigram a (Note5).

- LCSR (Melamed, 1999) stands for the Longest Common Subsequence Ratio, and is computed by dividing the length of the longest common subsequence by the length of the longer string. E. g., LCSR (*colour, couleur*) = 5/7 = 0.71.

- NED is a normalized edit distance. The edit distance (Wagner & Fischer, 1974) is calculated by counts up the minimum number of edit operations necessary to transform one word into another. In the standard definition, the edit operations are substitutions, insertions, and deletions, all with the cost of 1. A normalized edit distance is obtained by dividing the total edit cost by the length of the longer string.

- SOUNDEX (Hall & Dowling, 1980) is an approximation to phonetic name matching. SOUNDEX transforms all but the first letter to numeric codes and after removing zeros truncates the resulting string to 4 characters. For the purposes of comparison, our implementation of SOUNDEX returns the edit distance between the corresponding codes.

- BI-SIM, TRI-SIM, BI-DIST, and TRI-DIST belong to a family of n-gram measures (Kondrak & Dorr, 2004) that generalize LCSR and NED measures. The difference lies in considering letter bigrams or trigrams instead of single letter (e.g., unigrams). For example, BI-SIM finds the longest common subsequence of bigrams, while TRI-DIST calculates the edit distance between sequences of trigrams. n-gram similarity is calculated by the formula:

$$s(x_1 \ldots x_n, y_1 \ldots y_n) = \frac{1}{n}\sum_{i=1}^{n} id(x_i, y_i)$$

where id(a, b) returns 1 if a and b are identical, and 0 otherwise.

3.3 Method

Our contribution to the task of identifying cognates and false friends between languages is the method itself by the way we approach the identification task: using ML techniques and a combination of various orthographic similarity measures. Other methods that have been proposed for cognate and false friend identification require intensive human knowledge (Barker & Sutcliffe, 2000), (Friel & Kennison, 2001).

As we described in the second chapter, ML techniques require the data to be in a certain format. Each instance that is going to be used by the technique has to be

transformed into a feature value representation.

We can associate the representation with a flat relational data base where each row is an instance and the columns are features used to represent the data.

From the Weka package (Witten & Frank, 2005) we used different supervised ML algorithms to best discriminate between the two classes that we have chosen: Cognates/False-Friends-orthographically similar and Unrelated - not orthographically similar.

3.3.1 Instances

Instances are small parts of the whole data that are used in a ML technique and have a label attached. The label is the class to which the instance belongs. By using ML techniques we want to create classifiers that are able to discriminate between different classes.

What an instance is and how it is represented are interesting aspects of ML techniques. A big role here is taken by the human knowledge and intuition. The choices of data representation differ from method to method and from task to task.

For our task, an instance is a pair of words containing a French word and an English word. The data that we use consists of different lists of pairs of words that will be described in detail in the next section.

3.3.2 Features and Feature Values

The features that we choose to use in our method are the 13 orthographic similarity measures that we described in Section 3.2. For different experiments we used different features: each orthographic measure as a separate all 13 features together, and the average of the results of all 13 orthographic measures as a single feature.

Another goal of our methods is to automatically identify threshold values for the orthographic measures that we use. In order to do so, we use the ML algorithms with data represented by a single feature, the one for which we want to identify the threshold. We also tested to see if the average of all measures performs better than each measure in part or all put together. For this experiment, we again used only one feature that has as value the average of all measures.

No matter what are the features that the method uses, the values of the features are real numbers between 0 and 1 (inclusively) that reflect the orthographic similarity between two words that belong to one instance, a pair of French-English words, in our experiments.

In addition to other research that has been done on cognate and false friend identification, we look at different orthographic measures combined by ML techniques to classify pairs of words as being cognates, false friends or unrelated pairs.

3.4 Evaluation on Training and Testing Data Sets

We present evaluation experiments using the two data sets described in Section 3.1, a training/development set and a test set. We classify the word pairs on the basis of similarity into two classes: Cognates/False-Friends and Unrelated. Cognates are later distinguished

from false friends by virtue of being mutual translations. We report the accuracy values for the classification task (the precision and recall values for the two classes are similar to the accuracy values). We test various feature combinations for our classification task. We test each orthographic similarity measure individually and we also average the values returned by all the 13 measures. Then, in order to combine the measures, we run several machine learning classifiers from the Weka package.

3.4.1 Results on the Training Data Set

Individual Orthographic Measures Table 2 presents the results of testing each of the 13 orthographic measures individually. For each measure, we need to choose a specific similarity threshold for separating Cognates/False-Friends from the Unrelated pairs. The separation has to be made such that all the pairs with similarity above or equal to the threshold are classified as Cognates/False-Friends and all the pairs with similarity below the threshold are classified as Unrelated.

For the IDENT measure, the threshold was set to 1 (identical in spelling ignoring accents). This threshold leads to 49.4% accuracy since the number of pairs with identical spelling in the training data is small (208 pairs out of 1454 that is 14.3% identical pairs, ignoring accents). We could also use the value 0 for the threshold, in this case all the pairs would be classified as Cognates/False-Friends since all scores are greater or equal to zero. This achieves 63.75% accuracy the same as the baseline obtained by always choosing the class that is the most frequent in the training set (reported in Table 3).

For the rest of the measures, we determine the best thresholds by running Decision Stump classifiers with a single feature. Decision Stumps are Decision Trees that have a single node containing the feature value that produces the best split. When we run the decision stump classifier for one feature (each measure in part), we obtain the best thresholds. An example for the XXDICE measure Decision Stump tree is presented in Fig1. The values of the thresholds obtained in this way are also included in Table 2.

```
XXDICE  <= 0.21710000000000002  :  UNREL
XXDICE  > 0.21710000000000002 : CG_FF
```

Figure 1. Example of Decision Stump Classifier

Combining the Measures. The training data set representation for the machine learning experiments consists in 13 features for each pair of words, the values of the 13 orthographic similarity measures. We traine several machine learning classifiers from the Weka package: OneRule (a shallow Decision Rule that considers only the best feature and several values for it), Naive Bayes, Decision Trees, Instance-based Learning (IBK), AdaBoost, Multi-layered Perceptron, and a light version of Support Vector Machine (SMO). Unlike some other machine learning algorithms, Decision Tree classifier has the

advantage of being relatively transparent. Figure 2 shows the Decision Tree obtained with the default Weka parameter settings. For each node, the numbers in round brackets show how many training examples were in each class.

Table 2: Results of each orthographic similarity measure individually, on the training data set. The last line presents a new measure which is the average of all measures for each pair of words.

Orthographic similarity measure	Threshold	Accuracy
IDENT	1	43.90%
PREFIX	0.03845	92.70%
DICE	0.29669	89.40%
LCSR	0.45800	92.91%
NED	0.34845	93.39%
SOUNDEX	0.62500	85.28%
TRI	0.0476	88.30%
XDICE	0.21825	92.84%
XXDICE	0.12915	91.74%
BI-SIM	0.37980	94.84%
BI-DIST	0.34165	94.84%
TRI-SIM	0.34845	95.66%
TRI-DIST	0.34845	95.11%
Average Measure	0.14770	93.83%

```
TRI-SIM <= 0.3333
| TRI-SIM <= 0.2083: UNREL (447/17)
| TRI-SIM > 0.2083
| | XDICE <= 0.2
| | | PREFIX <= 0: UNREL (74/11)
| | | PREFIX > 0
| | | | SOUNDEX <= 0.5
| | | | | BI-SIM <= 0.3
| | | | | | SOUNDEX <= 0.25: CG_FF (6/2)
| | | | | | SOUNDEX > 0.25
| | | | | | | LCSR <= 0.1818: UNREL (3)
| | | | | | | LCSR > 0.1818
| | | | | | | | TRI-DIST <= 0.29: CG_FF (2)
| | | | | | | | TRI-DIST > 0.29: UNREL (2)
```

```
| | | | | BI-SIM > 0.3: UNREL (7)
| | | | SOUNDEX > 0.5: CG_FF (3)
| | XDICE > 0.2
| | | BI-SIM <= 0.3: UNREL (3)
| | | BI-SIM > 0.3: CG_FF (9)

TRI-SIM > 0.3333
| BI-SIM <= 0.4545
| | LCSR <= 0.25: UNREL (5/1)
| | LCSR > 0.25
| | | BI-DIST <= 0.4091
| | | | TRI-DIST <= 0.3333
| | | | | XXDICE <= 0.1222: CG_FF (7)
| | | | | XXDICE > 0.1222: UNREL (2)
| | | | TRI-DIST > 0.3333: CG_FF (26)
| | | BI-DIST > 0.4091
| | | | TRI-DIST <= 0.4286
| | | | | XXDICE <= 0.2273: UNREL (7/1)
| | | | | XXDICE > 0.2273: CG_FF (4/1)
| | | | TRI-DIST > 0.4286: CG_FF (11/1)
| BI-SIM > 0.4545: CG_FF (836/3)
```

Figure 2. Example of Decision Tree classifier (default Weka parameters, CF=25%). The two classes are Cognates/False-Friends (CG FF) and Unrelated (UNREL). Decisions are based on values of the orthographic similarity measures. The numbers in parentheses show how many examples were classified under each node.

```
TRI-SIM <= 0.3333
| TRI-SIM <= 0.2083: UNREL (447.0/17.0)
| TRI-SIM > 0.2083
| | XDICE <= 0.2: UNREL (97.0/20.0)
| | XDICE > 0.2
| | | BI-SIM <= 0.3: UNREL (3.0)
| | | BI-SIM > 0.3: CG_FF (9.0)
TRI-SIM > 0.3333: CG_FF (898.0/17.0)
```

Figure 3. Example of Decision Tree classifier, heavily pruned (confidence threshold for pruning CF=16%).

Some of the nodes in the decision tree contain counter-intuitive decisions. For example, one of the leaves classifies an instance as Unrelated if the BISIM value is greater than 0.3. Since

all measures attempt to assign high values to similar pairs and low values to dissimilar pairs, the presence of such a node suggests overfitting. One possible remedy to this problem is more aggressive pruning. We kept lowering the confidence level threshold from the default CF = 0.25 until we obtained a tree without counter-intuitive decisions, at CF = 0.16 (Figure 3). Our hypothesis was that the latter tree would perform better on a test set.

The results presented in the rightmost column of Table 3 are obtained by 10-fold cross-validation on training data set (the data is randomly split in 10 parts, a classifier is trained on 9 parts and tested on the tenth part; the process is repeated for all 10 splits). We also report in the middle column the results of testing on the training set. These results are artificially high due to overfitting. The baseline algorithm in Table 3 always chooses the most frequent class in the data set which happened to be Cognates/False-Friends. The best classification accuracy (for cross-validation) is achieved by Decision Trees, OneRule, and AdaBoost (95.66%). The performance equals the one achieved by the TRI-SIM measure alone in Table 2.

Table 3: Results of several classifiers for the task of detecting cognates/False-Friends versus Unrelated pairs on the training data (cross- validation).

Classifier	Accuracy on training set	Accuracy for cross-validation
Baseline	63.75%	63.75%
OneRule	95.94%	95.66%
Naive Bayes	94.91%	94.84%
Decision Trees	97.45%	95.66%
DecTree (pruned)	96.28%	95.66%
IBK	99.10%	93.81%
AdaBoost	95.66%	95.66%
Perceptron	95.73%	95.11%
SVM (SMO)	95.66%	95.46%

Error Analysis: We examined the misclassified pairs for the classifiers built on the training data. There were many shared pairs among the 60–70 pairs misclassified by several of the best classifiers. Here are some examples from the decision tree classifier of false negatives (Cognates/False-Friends classified as Unrelated): *égale - equal, boisson - beverage, huit - eight, cinquante - fifty, cinq - five, fourchette - fork, quarante - forty, quatre - four, plein - full, coeur - heart, droit - right, jeune - young, faire - do, oreille - ear, oeuf - egg, chaud - hot*.

Most of the false negatives were genetic cognates that have different orthographic form due to changes of language over time (13 out of the 16 examples above). False positives on the other hand were mostly caused by accidental similarity: *arrêt - arm, habiter - waiter, peine - pear*. Several of the measures are particularly sensitive to the initial letter of the word which is a strong clue of cognation. Also, the presence of an identical prefix made some pairs look

similar but they are not cognates unless the word roots are related.

3.4.2 Results on the Test Set

The rightmost column of Table 4 shows the results obtained on the test set described in Section 3.1. The accuracy values are given for all orthographic similarity measures and for the machine learning classifiers that use all the orthographic measures as features. The classifiers are the ones built on the training set.

The ranking of measures on the test set differs from the ranking obtained on the training set. This may be caused by the absence of genetic cognates in the test set. Surprisingly, only the Naive Bayes classifier outperforms the simple average of orthographic measures. The pruned Decision Tree shown in Figure 3 achieves higher accuracy than the overtrained Decision Tree from Figure 2 but still below the simple average. Among the individual orthographic measures, XXDICE performs the best supporting the results on French-English cognates reported in (Brew & McKelvie, 1996). Overall, the measures that performed best on the training set achieve more than 93% on the test set. We conclude that our classifiers are generic enough since they perform very well on the test set.

3.4.3 Results for Three-Class Classification

Since all the examples of pairs of the class Unrelated in our training set were mutual translations we had to add Unrelated pairs that are not translations (otherwise all pairs with the translation feature equal to 0 would have been classified as false friends by the machine learning algorithms). We generated these extra pairs automatically by taking French and English words from the existing Unrelated pairs and pairing them with words other than their pairs. We manually checked to ensure that all these generated pairs were not translations of each other by chance.

As expected, this experiment achieved similar but slightly lower results than the ones from Table 2 when running on the same data set (cross-validation). Most of the machine learning algorithms (except the Decision Tree) did not perfectly separate the Cognate/False-Friends class. We conclude that it is better to do the two-way classification that we presented above, Cognates/False-Friends and Unrelated, and then split the first class into Cognates and False-Friends on the basis on the value of the translation feature. Nevertheless, the three-way classification could still be useful provided that the translation feature is assigned a meaningful score such as the probability that the two words occur as mutual translations in a bitext. The results for the three-way classification are presented in Table 5.

Table 4: Results of testing the classifiers built on the training set (individual measures and machine learning combinations). The rightmost column tests on the test set of 1040 pairs.

Classifier (measure or combination)	Accuracy on test set
IDENT	55.00%
PREFIX	90.97%
DICE	93.37%
LCSR	94.24%
NED	93.57%
SOUNDEX	84.54%
TRI	92.13%
XDICE	94.52%
XXDICE	95.39%
BI-SIM	93.95%
BI-DIST	94.04%
TRI-SIM	93.28%
TRI-DIST	93.85%
Average measure	94.14%
Baseline	66.98%
OneRule	92.89%
Naive Bayes	94.62%
Decision Trees	92.08%
DecTree (pruned)	93.18%
IBK	92.80%
AdaBoost	93.47%
Perceptron	91.55%
SVM (SMO)	93.76%

Table 5: Results of several classifiers for the task of detecting Cognates, False-Friends and Unrelated pairs using cross-validation.

Classifier	Accuracy on cross-validation
Baseline	39.51%
OneRule	71.18%
Naive Bayes	92.08%
Decision Trees	96%
DecTree (pruned)	96%
IBK	95.18%
AdaBoost	96%
Perceptron	95.75%
SVM (SMO)	95.4%

4 Partial Cognate Disambiguation

This section presents the data that we used and experimented with the two methods that we propose for disambiguating partial cognates between two languages.

4.1 Data for Partial Cognates

We performed experiments using our proposed methods with ten pairs of partial cognates. We list them in Table 6. For a French partial cognate we list its English cognate word 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 the false friend senses together). For example, the false friend words for the French partial cognate *note* include one sense for *grades* and one sense for *bills*. In our experiments, the false friend meaning will contain both senses.

The partial cognate (PC), the cognate (COG) and false friend (FF) words were collected from a Web(Note 6) resource. The resource contains a list of 400 false friends including 64 partial cognates. All partial cognates are words frequently used in the language. We selected the ten partial cognates according to the number of extracted sentences, have a balance between the two meanings of cognates and false friends.

Table 6: 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
route	Route	road, roadside

To show how frequent the ten pairs of partial cognates are, we run some experiments on the LeMonde(Note 7) corpus, a collection of French newspaper news from 1994 and 1995. We used this time frame for this corpus because we had it available and also because our lists of cognates, false friends, and partial cognates, contain words that are representative to both languages throughout various time periods. We counted the frequency of all content words that we found in the corpus; we did not use any lemmatization for the experiment. To filter out the stop words, we used a list of 463 stop French words from the web. The total number of content words from the corpus that remain after filtering out the stop words is 216,697. From all extracted content words we took into account only the ones that have a frequency greater or equal to 100, bellow 100 almost all words had the frequency 1 and very few had a frequency between 1 and 100, a total number of 13,656. If we compute the average frequency for the chosen content words, the value is 695.52. All 10 partial cognate had the absolute frequency above the average frequency.

With the ten pairs of partial cognates collected, the human effort that we require for our methods is 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(Note 8) After adding

more false friend words the final set of pairs for which we evaluate our methods is the one from Table 6.

4.1.1 Seed Set Collection

Both the supervised and the semi-supervised method that we describe in the next section 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.

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(Note 9) EuroParl(Note 10), and the manually aligned BAF(Note 11) corpus from University of Montreal.

The cognate sense sentences were created by extracting parallel sentences that have on the French side the French cognate and on the English side the English cognate. See the upper part of Table 7 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 the lower the part of Table 7.

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 6. 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 preprocessing and without removing any noise that might be introduced during the collection process.

Table 7: 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)	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.

Table 8: Number of parallel sentences used as seeds.

Partial Cognates	Train COG	Train FF	Test COG	Test FF	Fr Features	Eng Features
Blanc	54	78	28	39	83	76
Circulation	213	75	107	38	363	328
Client	105	88	53	45	229	187
Corps	88	82	44	42	198	163
Détail	120	80	60	41	195	178
Mode	76	104	126	53	163	156
Note	250	138	126	68	377	326
Police	154	94	78	48	373	329
Responsable	200	162	100	81	484	409
Route	69	90	35	46	150	127
AVERAGE	132.9	99.1	66.9	50.1	261.5	227.9

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 8 we present the number of seeds used for training and testing as well as the number of features selected from the training seed sets for each partial cognate.

We will show later on in the article 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.

4.2 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 are used for the monolingual and bilingual bootstrapping techniques.

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). Our goal is to determine the sense of a partial cognate in a French sentence, determine if the partial cognate is used with a cognate sense with the corresponding English word or with a false friend sense. Both the cognate and false friend English words are translated as the same French word.

4.2.1 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 stop words from the features. A list of stop words for French was used on the French sentences and one for English was used on the English parallel sentences. We ran some additional experiments when we kept the stop words as features but the results did not improve.

As a baseline for the experiments that we present we used the ZeroR classifier from WEKA, it predicts the class that is the most frequent in the training corpus. The classifiers for which we report results are: Naive 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.

The supervised method used in our experiments consists in training the chosen classifiers on the automatically-collected training seed sentences, separately for French and for English, for each partial cognate and then testing their performance on the test set. Results for this method are presented later in section 4.3.

4.2.2 Semi-Supervised Methods

Besides the supervised method that we propose to disambiguate a partial cognate we look at a semi-supervised method as well. For the semi-supervised method we add unlabeled examples, sentences that contain the partial cognate with one of the two senses, cognate or false friend, from monolingual French newspaper LeMonde 1994, 1995 (LM), and the BNC(Note 12) corpus. The domain of these additional corpora is different than the domain of the seeds. The procedure of adding and using this unlabeled data is described in the Monolingual Bootstrapping (MB) and Bilingual Bootstrapping (BB) algorithms.

Monolingual Bootstrapping The monolingual bootstrapping algorithm that we use for experiments on French sentences (MB-F) or 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 monolingual 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 use the NB-K classifier because it was the classifier that consistently performed better when we run experiments on the training data set using 10-fold cross validation technique with different classifiers and different levels of tuning. 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 algorithm was chosen. The attribute selection was performed only when we trained the classifiers to be used for labeling the unlabeled data from the additional resources.

In the second step of the MB algorithm the classifier that is trained on the training seeds is then used to classify the unlabeled data that is collected from the two additional resources, separately. For the MB algorithm on the French side we train the classifier on the French side of the training seeds and then we apply the classifier to classify sentences extracted from LeMonde and contain the partial cognate, as belonging to the COG class or the FF class. The same approach is used for the MB on the English side only this time we use the English side of the training seeds to train the classifier and the BNC corpus to extract new examples. In fact, the MB-E step is needed only for the BB method.

Only sentences classified with a probability greater than 0.85 are selected for later use in the bootstrapping algorithm. This value of the parameter is a heuristic value in our experiments. All results that will be described in Section 4.3 use the threshold of 0.85 for the probability distribution.

The number of sentences that are selected from the new corpora and used in the MB and BB algorithms is presented in Table 9.

Table 9: 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 have a

probability distribution greater or equal with the threshold is low. For the rest of partial cognates the number of selected sentences was limited by the value of the parameter k that was 250, in the algorithm.

4.2.3 Bilingual Bootstrapping

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

1. Translate the English sentences that were collected in the MB-E step into French using an online MT tool(Note 13) *and add them to the French seed training data.*

2. Repeat the MB-F and MB-E steps T times.

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

In the bilingual bootstrapping algorithm we take the English sentences that are extracted from the BNC corpus, as described in the MB-E algorithm, translate them into English using an on-line MT tool and then we add them to the French training corpus.

Our proposed methods are bilingual. They can be applied to any pair of languages for which a parallel corpus is available and two monolingual collections of text. Our main focus for our methods is to disambiguate a French partial cognate looking at its English cognate and false friend senses.

4.3 Evaluation and Results

This section is dedicated to the results that we obtain with the supervised and semi-supervised methods that we use to disambiguate partial cognates.

4.3.1 Evaluation Results for the Supervised Method

As we mentioned in section 4.2 subsection 4.2.1 we report results, for our supervised method to disambiguate partial cognates, only for three classifiers selected based on the results obtained by 10-fold cross-validation evaluation on the training data.

Table 10 presents the results obtained on the French data using our supervised technique. We used for training the 2/3 of the seed sets and for testing the other 1/3 of the seeds.

Table 10: Results for the Supervised Method on the French test set data.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.00%	95.52%	98.5%	98.5%
Circulation	74.00%	91.03%	80.00%	89.65%
Client	54.08%	67.34%	66.32%	61.22%
Corps	51.16%	62.00%	61.62%	69.76%
Détail	59.40%	85.14%	85.14%	87.12%
Mode	58.24%	89.01%	89.01%	90.00%
Note	64.94%	89.17%	77.83%	85.05%
Police	61.41%	79.52%	93.70%	94.40%
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%

4.3.2 Results for Semi-Supervised Methods

We want to disambiguate partial cognates not only in a parliamentary domain, the domain of our collected seeds but in different domains as well. To vary the domain of the training data and improve the classification results we proposed to algorithms MB and BB presented in Section 4.2.

Results that we obtain with these two algorithms are presented in Table 12 for the French MB (MB-F) and Table 14 for BB. For the MB experiments the training examples (training seeds) both for the French side of the parallel corpus and the English one are complemented with sentences extracted from LeMonde corpus for the French experiments and sentences extracted from BNC corpus for the English experiments. The training data and the number of features extracted after we add the new training data for the French MB experiments are presented in Table 11.

Table 13 presents the data and results obtained with the BB algorithm on the French side. To the French training seeds we added the translated sentences extracted from the BNC corpus and trained the classifier on them. The classifier performance is tested on the seed testing set, 1/3 of the collected seed sets. Results of this experiment are presented in Table 14.

We also combined MB and BB techniques and evaluated the classifiers for this combination. We trained the classifiers on the training seed sentences plus sentences from LeMonde plus sentences from BNC. For these experiments the best accuracy is 81.98% on the average of all 10 partial cognates using the Naive Bayes classifier. For all results reported until now we test the classifiers on the test set of the automatically collected seeds. These results are discussed in Section 4.3.2.

Table 12: Monolingual Bootstrapping results (accuracies) on the French test set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	97.01%	97.01%	98.5%
Circulation	73.79%	90.34%	70.34%	84.13%
Client	54.08%	77.55%	55.10%	70.40%
Corps	51.16%	78%	56.97%	69.76%
Détail	59.4%	88.11%	85.14%	82.17%
Mode	58.24%	89.01%	90.10%	85.00%
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 13: Data sets for Bilingual Bootstrapping on the French data.

PC	Train COG	Train FF	No. Features
Blanc	54	319	331
Circulation	283	255	686
Client	182	337	636
Corps	219	269	582
Détail	278	215	548
Mode	252	365	714
Note	428	419	922
Police	340	293	915
Responsable	377	386	1028
Route	286	207	533
AVERAGE	269.8	306.4	689.5

Table 14: Accuracies results for Bilingual Bootstrapping on the French test set data.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	95.52%	97.01%	98.50%
Circulation	73.79%	92.41%	63.44%	87.58%
Client	45.91%	70.40%	45.91%	63.26%
Corps	48.83%	83.00%	67.44%	82.55%
Détail	59.00%	91.08%	85.14%	86.13%
Mode	58.24%	87.91%	90.10%	87.00%
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	43.20%	60.49%	45.67%	64.19%
AVERAGE	55.87%	83.41%	74.21%	82.40%

Table 15: Accuracies for Monolingual Bootstrapping plus Bilingual Bootstrapping on the French test set data.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	98.50%	97.01%	98.50%
Circulation	73.79%	88.96%	69.65%	86.20%
Client	45.91%	71.42%	54.08%	70.40%
Corps	48.83%	80.00%	60.46%	76.74%
Détail	59.40%	90.09%	85.14%	87.12%
Mode	58.24%	86.81%	90.10%	82.00%
Note	64.94%	85.05%	70.61%	78.35%
Police	61.41%	74.80%	95.27%	70.86%
Responsable	44.75%	88.39%	75.69%	79.00%
Route	43.20%	55.55%	45.67%	58.02%
AVERAGE	55.87%	81.98%	74.37%	78.76%

Results for additional experiments with different data sets. Besides the experiments that we did with the semi-supervised method using unlabeled corpus from LeMonde and BNC we run additional experiments with another set of automatically collected 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. The corpus was provided to us by Brighton University, UK. We use 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 16.

Table 16: Number of sentences collected from the New Corpus (NC).

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

With this new set of sentences we performed different experiments both for the MB and BB the algorithms. All the results are described in Table 17. The results are reported for the average of the accuracies for the ten pairs of partial cognates.

The symbols that we use in Table 17 represent: S — the seed training corpus, TS — the seed test set, BNC and LM — sentences extracted from LeMonde and BNC (Table 9), 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 17: Results for different experiments with Monolingual and Bilingual Bootstrapping (MB and BB), when the New Corpus (NC) is used either in training or in testing.

Train	Test	ZeroR	NB-K	Dec. Trees	SMO
S (no bootstrapping)	NC	67%	71.97%	73.75%	76.75%
S+BNC (BB)	NC	64%	73.92%	60.49%	74.8%
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+NC (no bootstrapping)	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	62.7%	72.97%	70.33%	71.97%
S+BNC (BB)	TS+NC	61.27%	79.83%	67.06%	78.8%
S+LM+BNC (MB+BB)	TS+NC	61.27%	77.28%	65.75%	73.87%

Figure 4 presents in a graphical way the results obtained with the four methods, No Bootstrapping, Monolingual French Bootstrapping, Bilingual Bootstrapping and Monolingual plus Bilingual Bootstrapping on different sets of French sentences for the average over all 10 pairs of partial cognates. The sets of French sentences set that the method uses are shown on the X axis of the chart. The set used initially for training, no bootstrapping, is presented before the underscore line, and the set used for testing is presented after the underscore line.

Error Analysis: Most of the errors that the classifiers made were on the hard to disambiguate words, but still improvement was obtained even for these partial cognates. The errors that appear can also be caused by the noise that was introduced in the seed set collection process and in the bootstrapping process.

For example for the partial cognate *circulation*, on the seed testing set from a total of 145 testing instances the Naive Bayes classifier trained on a number of 288 instances made 13 mistakes. 6 mistakes were on the Cognate class, the actual class was the Cognate class but the classifier predicted the False Friend class and the rest of 7 mistakes were made on the False Friend class.

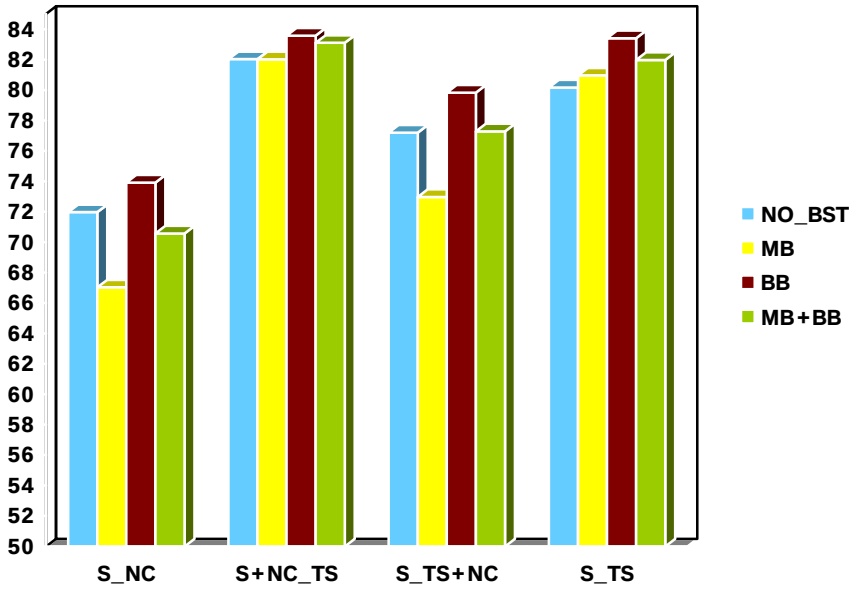

Figure 4. Results for the average of the PC set with different methods and data sets.

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 introduced due to the seed set collection is tolerable by the ML techniques that we use. The noise that can be introduced is due to the fact that we do not use a word-aligned corpus. For example a French sentence can contain the partial cognate and the English parallel sentence can contain the cognate English word but the meaning of the English sentence could be the false friend one and the cognate word could appear in another part of the sentence.

Some results of the experiments we present in Table 17 are not as good as others. What is important to notice is that every time we used MB or BB or both, there is 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.

The supervised method results, Table 10, were outperformed by the semi-supervised methods on the French set, Table 12 showing that unlabeled data can be used with success to boost the results of the classification task.

In Tables 10 and 14 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 17 we show that our proposed methods bring improvements for different combinations of training and testing sets. Table 17, lines 1 and 2 show that BB with NB-K brings an improvement of 1.95% from no bootstrapping when we tested on the multi-domain

corpus, NC. For the same setting, there is an improvement of 1.55% when we test on TS (Table 17, lines 6 and 8). When we test on the combination TS+NC, again BB brings an improvement of 2.63% from no bootstrapping (Table 17, lines 10 and 12). The difference between MB and BB with this setting is 6.86% (Table 17, lines 11 and 12). According to a t-test the 1.95% and 6.86% improvements are statistically significant.

The results presented in the above mentioned tables are only performed with French data, our goal was to be able to disambiguate a French partial cognate in a certain context.

Unlike previous work with monolingual or bilingual bootstrapping Diab & Resnik (2002), Li & Li (2004), 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, blood*, etc. circulation and false friend for *street traffic*).

We also show that our method is able to bootstrap the initial knowledge of the chosen classifiers, parliamentary domain knowledge, with information from different domains that was obtained in the monolingual and bilingual bootstrapping steps. The number of features that is extracted at each semi-supervised step more than doubled compared with the initial one extracted from the seeds. These additional features also belong to different domains.

5 The Tool for Cross-Language Pair Annotations

This section is devoted to our tool called Cross-Language Pair Annotator (CLPA) that is capable of automatically annotating cognates and false friends in a French text. The tool uses the Unstructured Information Management Architecture (UIMA)(Note 14) Software Development Kit (SDK) from IBM and Baseline Information Extraction (BaLIE)(Note 15), an open source Java project capable of extracting information from raw texts.

5.1 Tool Description

CLPA is a tool that has a Graphical User Interface (GUI) capability that makes it easy for the user to distinguish between different annotations of the text. We designed the tool as a Java open source downloadable kit that contains all the additional projects (Balie and UIMA) that are needed. It can be downloaded from the following address: CLPA(Note 16).

The tool is a practical follow up to the research that we did on cognates and false friends between French and English. Since one of our main goals is to be able to use the research that we did in a CALL tool can help second language learners of French, CLPA is intended to be the first version of such a tool. At this point, the tool uses as knowledge a list of 1,766 cognates and a list of 428 false friends. The list of false friends contains a French definition for the French word and an English definition for the English word of the pair. Both lists contain the cognates and false friend pairs that were used in the Machine Learning experiments for the cognate and false friend identification task described in Section 3. The

tool offers an easy management of the resources. If the user would like to adjust/use other lists of cognates and/or false friends he/she needs to add the new source files to the resource directory of the project. The directory can be found in the home project directory.

UIMA is an open platform for creating, integrating, and deploying unstructured information management solutions from a combination of semantic analysis and search components. It also has various GUI document analyzers that make it easy for the user to visualize the text annotations.

UIMA offers CLPA the GUI interface and an efficient management of the annotations that are done for a certain text. The user can select/deselect the cognate or false friend annotations. By default, both type of cross language pairs are annotated.

BaLIE is a trainable Java open source project that can perform: Language Identification, Sentence Boundary Detection, Tokenization, Part of Speech Tagging and Name Entity Recognition for English, French, German, Spanish and Romanian. BaLIE provided the tokenization and part-of-speech tagging tasks for the French texts. The tokenization is done using a rule based method and the part-of-speech by using a probabilistic part-of-speech tagger, QTag(Note 17).

In a single run, the tool can annotate not only one document but a directory that contains more than a single text document. UTF-8 is the character encoding chosen to represent a document. The reason why we chose this format is due to the French characters and also for a consistency with the other projects that are used by CLPA the BaLIE project.

Figure 5 provides a snapshot of the interface that the user will see after the annotation process is completed.

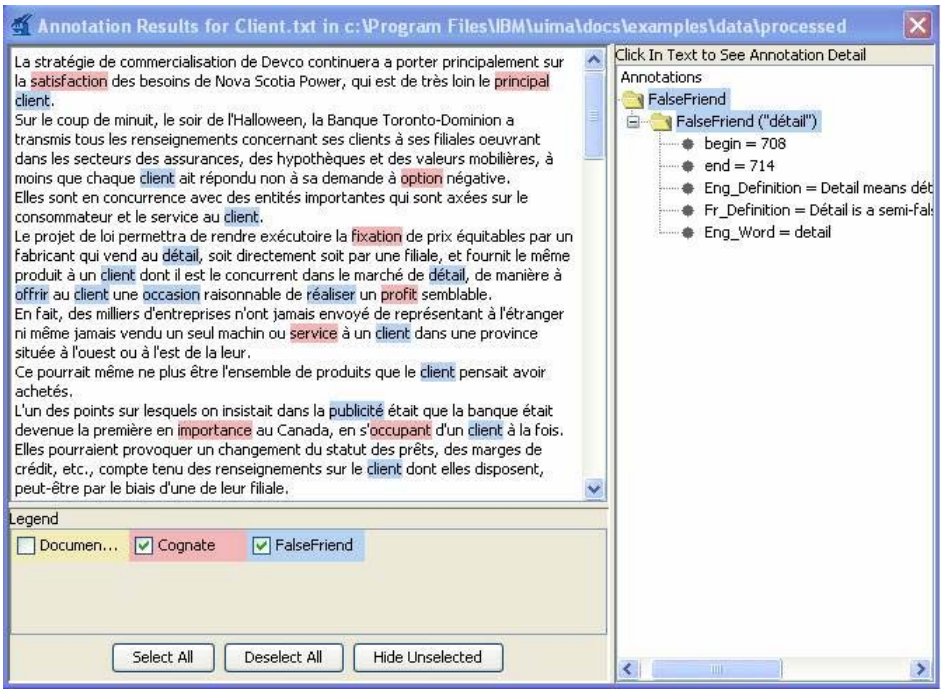

Figure 5: Cognate and false friend annotations.

The user has to click on one of the text annotations to obtain additional information about the chosen annotation, (e.g. at what position in the text does the chosen word starts, what position does it end, the French definition of the French false friend word, the English definition of the English false friend word, etc.).

5.2 Tool Capabilities

In its early stage of existence, first version of CLPA, can annotate cognates and false friends between French and English in a French text. The cognate and false friend knowledge that the tool has is provided by lists of pairs of cognates and false friends. For now we intended high accurate lists instead of automatically produced lists. Instead of the lists that we use in this first version, we can use the lists that we automatically produced and described in Section 3.

In addition to the colored annotations (cognates are annotated with one color and false friends with another color), the tool provides other useful information about the annotations. For the cognate words it provides the position in the text and the English cognate word. For the false friend words it provides: the position in the text, the English false friend word/words, the definition of the French word, and the definition of the English words. The definitions were collected from the same resource(Note 18) as the false friend word pairs.

The lists that the CLPA uses to annotate the French texts are free to download and can be used for future research. They are contained in the same package as the tool.

The annotations that the tool makes are only for French content words: nouns, adjectives, adverbs and verbs. We have chosen to annotate only the content words not to introduce some false alarms (e.g. the French word *pour* can be either adverb (*pro*), or preposition (*for; to*), and it is a false friend with the English word *pour* that is a verb), and also because they are of more interest for second language learners.

Since BaLIE can provide information regarding the part-of-speech tag for each token in the text, it was easy for us to make the distinction between the content and close class French words.

The tool does not lemmatize the text and for this reasons some words might not be annotated or some errors might be introduced. Some annotations might be missed because the words are not in the base form and some errors might be introduced because the inflected form corresponds to the base form for another word (e.g. the verb *être* has the singular third person *est* form that corresponds to the base form of the cardinal point *est* that is cognate with the English word *east*).

The annotation will be done only for the tokens in the text that have the same form as the pair of words in the lists, the base form of the French and English words. For the next version of the tool we will have the lemmatization step performed on the text before we do the annotation step.

Both UIMA and BaLIE are Java projects that can be easily downloaded and used with Eclipse(Note 19) SDK. In fact, UIMA has some of the features to be easily used with Eclipse. For both projects, documentation on how to install the projects is available from the

corresponding web pages. For CLPA, the web page will provide instructions on how to install and put all the resources together so they can be ready to run for French text annotations.

6 Conclusions and Future Work

This section is dedicated to bring to the reader some conclusions for both tasks that we tackled in this work and also puts into perspective some plans for future work to further develop our research.

6.1 Cognate and False Friend Identification

In this article we presented and evaluated a new method of identifying cognates and false friends between French and English. The method uses 13 orthographic similarity measures that are combined through different ML techniques. For each measure separately we also determined a threshold of orthographic similarity that can be used to identify new pairs of cognates and false friends. The novelty that we bring to this task is the way we use and combine different orthographic similarity measures by ML techniques. The results show that for French and English it is possible to achieve very good accuracy even without the training data by employing orthographic measures of word similarity.

As future work we want to apply the cognate and false friend identification task to other pairs of languages that lack this kind of resource (since the orthographic similarity measures are not language-dependent).

We want to increase the accuracy of the automatically generated lists of cognates and false friends by increasing the threshold used — we could obtain better precision but less recall for both classes. We could eliminate some falsely determined false friends by using other orthographic measures or the same measure with a higher threshold on the initial list determined with the same threshold for both cognates and false friends.

6.2 Partial Cognate Disambiguation

In our work we have shown that the task of partial cognate word disambiguation can be done with success using a supervised and more likely with a semi-supervised method that uses a bootstrapping technique. We proposed two semi-supervised algorithms that use unlabeled data from different languages, French and English, which can improve the accuracy results of a simple supervised method. We have also shown that classifiers, computer algorithms, are able to capture knowledge in an incremental process like humans and that are sensitive to knowledge from different domains. The unlabeled data that was extracted and added to the training data for the different algorithms was collected from different domains than the initial training seed data.

Simple methods and available tools have proved to be resources of great value to achieve good results in the task of partial cognate disambiguation. The accuracy results might be increased by using dependency relations, lemmatization, part-of-speech tagging — extract only 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. New information was collected and extracted by classifiers when additional corpora were used for training.

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.

6.3 A Tool for Cross-Language Pair Annotations

In the Section 5 we present and describe a CALL tool, CLPA, which can annotate cognates and false friends in a French text. In its first version, the tool uses the list of cognates and false friends that were used to experiment with the cognate and false friend identification technique in Section 3 but any other list can be easily integrated. CLPA has an easy to use GUI that allows users to choose between annotations e.g. only cognate annotation, only false friend annotation or both and also provides additional information to the users. This information can be useful to a second language learner similar to the feedback from a tutor.

For the future we want to continue to develop the tool, add other features, perform the lemmatization step, and also annotate partial cognates with the corresponding meaning in the texts.

We also want to use French second-language students to evaluate the usefulness of the tool. We want to see if the cognate and false friend annotations are helpful and more likely if the additional information that we provide helps students in the learning process.

Trying to develop the tool for other languages is also one of our future aims. In order to do this all we need is to plug in lists of cognates and false friends for the corresponding languages.

The overall contribution of our work consists in the new methods that we proposed, the evaluation experiments, and the new directions that we followed for cognate, false friend, and partial cognate words between French and English.

References

Adamson, G. W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval, 10, 253–260.

Barker, G., & Sutcliffe, R. F. E. (2000). An experiment in the semi-automatic identification of false-cognates between English and Polish. Tech. rep., Department of Languages and Cultural Studies, University of Limerick, Ireland.

Brew, C., & McKelvie, D. (1996). Word-pair extraction for lexicography. In Proceedings of the 2nd International Conference on New Methods in Language Processing, pp. 45–55, Ankara, Turkey.

Carroll, S. (1992). On cognates. Tech. rep., Second Language Research.

Diab, M., & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. Proc. of the Association for Computational Linguistics (ACL 2002), 255–262.

Friel, B. M., & Kennison, S. M. (2001). Identifying German-English cognates, false cognates, and non-cognates: Methodological issues and descriptive norms. Bilingualism: Language and Cognition, 4, 249–274.

Frunza, O., & Inkpen, D. (2006). Semi-supervised learning of partial cognates using bilingual bootstrapping. In Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, COLING-ACL 2006, pp. 433–440, Sydney, Australia.

Frunza, O., & Inkpen, D. (2008). Disambiguation of partial cognates. Language Resources and Evaluation, Volume 42, Number 3, September, 2008.

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

Guy, J. B. M. (1994). The use and acquisition of the second language lexicon (special issue). *9* (2), 35–42.

Hall, P. A. V., & Dowling, G. R. (1980). Approximate string matching. Computing Surveys, *12* (4), 381–402.

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, 1–19.

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

Hewson, J. (1993). A computer-generated dictionary of proto-algonquian. Tech. rep., Ottawa:Canadian Museum of Civilization.

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

Kondrak, G. (2001). Identifying Cognates by Phonetic and Semantic Similarity. In Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 103–110.

Kondrak, G. (2003). Identifying complex sound correspondences in bilingual wordlists. In Proceedings of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003), pp. 432–443, Mexico City.

Kondrak, G. (2004). Combining evidence in cognate identification. In Proc. of Canadian AI 2004: 17th Conference Canadian Society for Computational Studies of Intelligence, pp. 44–59.

Kondrak, G., & Dorr, B. J. (2004). Identification of confusable drug names: A new approach and evaluation methodology. In Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004), pp. 952–958, Geneva, Switzerland.

LeBlanc, R. (1989). L'enseignement des langues secondes aux adultes: recherches et pratiques. Les Presses de l'Universit´e d'Ottawa.

LeBlanc, R., & Séguin, H. (1996). Les congénères homographes et parographes anglais-français, Vol. Twenty-Five Years of Second Language Teaching at the University of Ottawa.

Li, H., & Li, C. (2004). Word translation disambiguation using bilingual bootstrap. Computational Linguistics, *30* (1), 1–22.

Lowe, J. B., & Mauzaudon, M. (1994). The reconstruction engine: a computer implementation of the comparative method. Computational Linguistics, 20, 381–417.

Mann, G. S., & Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In Proceedings of 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001), pp. 151–158.

Marcu, D., Kondrak, G., & Knight, K. (2003). Cognates can improve statistical translation models. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), 46–48.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The Blinker project. Tech. rep., University of Pennsylvania.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. Computational Linguistics, 25, 107–130.

Ringbom, H. (1987). The Role of the First Language in Foreign Language Learning. Multilingual Matters Ltd., Clevedon, England.

Robert-Collins (1987). Robert-Collins French-English English-French Dictionary. Collins, London.

Simard, M., Foster, G. F., & Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 67–81, Montreal, Canada.

Tréville, M.-C. (1990). Rôle des congénères interlinguaux dans le dévelopment du vocabulair réceptif. Ph.D. thesis, Universit´e de Montreal.

Vickrey, D., Biewald, L., Teyssier, M., & Koller, D. (2005). Word-sense disambiguation for machine translation. Conference on Empirical Methods in Natural Language Processing (EMNLP), 779–786.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. JACM, 21, 168–173.

Witten, I. H., & Frank, E. (2005). Data Mining - Practical Machine Learning Tools and Techniques (Second edition). Morgan Kaufmann.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL-95), 189–196.

Notes

Note 1. Parts of this article were presented in (Inkpen, Frunza, & Kondrak, 2005) and (Frunza & Inkpen, 2006) and (Frunza & Inkpen, 2008)

Note 2. http://mypage.bluewin.ch/a−z/cusipage/basicfrench.html

Note 3. http://french.about.com/library/fauxamis/blfauxam.htm

Note 4. The PREFIX measure can be seen as a generalization of the approach of Simard et al. (1992).

Note 5. The original definition of XXDICE does not specify which bigrams should be matched if they are not unique within a word. In our implementation, we match non-unique bigrams in the order of decreasing positions, starting from the end of the word.

Note 6. http://french.about.com/library/fauxamis/blfauxam_a.htm

Note 7. http://www.lemonde.fr

Note 8. http://www.wordreference.com

Note 9. http://www.isi.edu/natural-language/download/hansard/ and http://www.tsrali.com/

Note 10. http://people.csail.mit.edu/koehn/publications/europarl/

Note 11. http://rali.iro.umontreal.ca/Ressources/BAF/

Note 12. http://www.natcorp.ox.ac.uk/

Note 13. http://www.freetranslation.com/free/web.asp

Note 14. http://www.research.ibm.com/UIMA/

Note 15. http://balie.sourceforge.net/

Note 16. www.site.uottawa.ca/ ofrunza/CLPA.html

Note 17. http://www.english.bham.ac.uk/staff/omason/software/qtag.html

Note 18. http://french.about.com/library/fauxamis/blfauxam_a.htm

Note 19. http://www.eclipse.org/