Multilingual lexical resources to detect cognates in non-aligned texts

Haoxing Wang
Queensland University of Technology
2 George St, Brisbane, 4000
haoxing.wang@hdr.qut.edu.au

Laurianne Sitbon
Queensland University of Technology
2 George St, Brisbane, 4000
l.sitbon@qut.edu.au

Abstract

The identification of cognates between two distinct languages has recently started to attract the attention of NLP research, but there has been little research into using semantic evidence to detect cognates. The approach presented in this paper aims to detect English-French cognates within monolingual texts (texts that are not accompanied by aligned translated equivalents), by integrating word shape similarity approaches with word sense disambiguation techniques in order to account for context. Our implementation is based on BabelNet, a semantic network that incorporates a multilingual encyclopedic dictionary. Our approach is evaluated on two manually annotated datasets. The first one shows that across different types of natural text, our method can identify the cognates with an overall accuracy of 80%. The second one, consisting of control sentences with semi-cognates acting as either true cognates or false friends, shows that our method can identify 80% of semi-cognates acting as cognates but also identifies 75% of the semi-cognates acting as false friends.

1 Introduction

Estimating the difficulty of a text for non-native speakers, or learners of a second language, is gaining interest in natural language processing, information retrieval and education communities. So far measures and models to predict readability in a cross-lingual context have used mainly text features used in monolingual readability contexts (such as word shapes, grammatical features and individual word frequency features). These features have been tested when estimating readability levels for K-12 (primary to high school) readers. As word frequency can be partially estimated by word length, it remains the principal feature for estimation second language learners with the assumption that less frequent words are less likely to have been encountered and therefore accessible in the memory of the learnt language by readers. However such features are not entirely adapted to existing reading abilities of learners with a different language background. In particular, for a number of reasons, many languages have identical words in their vocabulary, which opens a secondary access to the meaning of such words as an alternative to memory in the learnt language. For example, English and French languages belong to different branches of the Indo-European family of languages, and additionally European history of mixed cultures has led their vocabularies to share a great number of similar and identical words. These words are called cognates.

Cognates have often slightly changed their orthography (especially in derived forms), and quite often meaning as well in the years and centuries following the transfer. Because of these changes, cognates are generally of one of three different types. First of all, true cognates are English-French word pairs that are viewed as similar and are mutual translations. The spelling could be identical or not, e.g., *prison* and *prison*, *ceramic* and *céramique*. False cognates are pairs of words that have similar spellings but have totally unrelated meanings. For example, *main* in French means *hand*, while in English it means *principal* or *essential*. Lastly, semi-cognates are pairs of words that have similar spellings but only share meanings in some contexts. One way to understand it in a practical setting is that they behave as true cognates or false cognates depending on their sense in a given context.

In this paper, we present a method to identify the words in a text in a given target language and that could acceptably be translated by a true cognate in a given source language (native language of a reader learning the target language). Accept-

Haoxing Wang and Laurianne Sitbon. 2014. Multilingual lexical resources to detect cognates in non-aligned texts. In *Proceedings of Australasian Language Technology Association Workshop*, pages 14−22.

ability in this context does not necessarily mean that a translator (human or automatic) would chose the true cognate as the preferred translation, but rather that the true cognate is indeed a synonym of the preferred translation. The method we present takes into account both characteristics of true cognates, which are similar spelling and similar meaning.

Most of the previous work in cognate identification has been operating with bilingual (aligned) corpora by using orthographic and phonetic measurements only. In such settings, the similarity of meaning is measured by the alignment of sentences in parallel texts. Basically, all the words in a parallel sentence become candidates that are then evaluated for orthographic similarity.

In the absence of aligned linguistic context, we propose that candidates with similar meaning can be proposed by a disambiguation system coupled with multilingual sense based lexicon where each word is associated to a set of senses, and senses are shared by all languages. A multilingual version of WordNet is an example of such lexicons. In this paper, we use BabelNet, which is an open resource and freely accessible semantic network that connects concepts and named entities in a very large network of semantic relations built from WordNet, Wikipedia and some other thesauri. Furthermore, it is also a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms in different languages. In order to disambiguate the word sense, BabelNet provides an independent Java tool that is called Babelfy. It employs a unified approach connecting Entity Linking (EL) and Word Sense Disambiguation (WSD) together. Moro et al. (2014) believe that the lexicographic knowledge used in WSD is useful for tackling EL task, and vice versa, that the encyclopedic information utilized in EL helps disambiguate nominal mentions in a WSD setting. Given an English sentence, Babelfy can disambiguate the meaning of each named entity or concept. For example, "You will get more <volume when beating egg whites if you first bring them to room <temperature." The words with bracket in front are cognates. *Volume* has been disambiguated as "The amount of 3-dimensional space occupied by an object", and *temperature* refers to "The degree of hotness or coldness of a body or environment". After the English word has been processed, it will search the words in other languages that contain this particular sense as candidates. The English word in the source is then compared to the candidates in the target language to establish orthographic/phonetic similarity. Formula 1 shows how we measure the cognateness *C* of an English word W based on the word shape similarity WSS of all its possible translations CW, and will motivated further in sections 3 and 4.

$$C(W) \approx Max_{CW}\big(WSS(CW)\big) \quad (1)$$

Because there are several types of orthographic and phonetic similarities used in the literature, we first establish which is most discriminative of cognates. We then evaluate a threshold-based approach and machine learning based approach to leverage orthographic/phonetic similarity to discriminate cognates from non cognates.

The first evaluation focuses on the performance of our method on a cognate detection task in natural data. The natural dataset contains 6 different genres of text. A second evaluation focuses specifically on semi-cognates classification in controlled sentences, where 20 semi-cognates were each presented in a sentence where they would translate as a cognate and a sentence where they would not.

The paper is organized as follows. Section 2 presents related research on cognate identification and introduces word sense disambiguation with Babelfy. Section 3 describes a general approach to tackle the cognate identification work, while section 4 specifically presents our implementation process. Finally, section 5 focuses on the evaluation and experiment results. Discussion, conclusion and future work are presented in section 6 and 7 respectively.

2 Related Work

2.1 Identifying cognates using orthographic/phonetic similarity

The most well-known approach to measuring how similar two words look to a reader is to measure the Edit Distance (ED) (Levenshtein, 1966). The ED returns a value corresponding to the minimum number of deletions, insertions and substitutions needed to transform the source language word into the target language word. The Dice coefficient measurement (Brew and McKelvie, 1996) is defined as the ratio of the number of n-grams that are shared by two strings and the total number of n-grams in both strings. The Dice coefficient with bi-grams (DICE) is a particularly popular word similarity measure. In their work, Brew and McKelvie looked only at pairs of verbs in English and French, pairs that are extracted from aligned sentences in a parallel corpus. Melamed (1999) used another popular

technique, the Longest Common Subsequence Ratio (LCSR), that is the ratio of the length of the longest (not necessarily contiguous) common subsequence (LCS) and the length of the longer word. Simard, Foster and Isabelle (1992) use cognates to align sentences in bi-texts. They only employed the first four characters of the English-French word pairs to determine whether the word pairs are cognates or not.

ALINE (Kondrak, 2000), is an example of a phonetic approach. It was originally designed to align phonetic sequences, but since it chooses the optimal alignment based on the similarity score, it could also be used for computing word shape similarity between word pairs. Kondrak believed that ALINE provides a more accurate result than a pure orthographic method. Kondrak and Dorr (2004) reported that a simple average of several orthographic similarity measures outperforms all the measures on the task of the identification of cognates for drug names. Kondrak proposed the n-gram method (Kondrak, 2005) a year later. In this work, he developed a notion of n-gram similarity and distance, which revealed that original Levenshtein distance and LCSR are special cases of n-gram distance and similarity respectively. He successfully evaluated his new measurement on deciding if pairs of given words were genetic cognates, translational cognates or drug names cognates respectively. The results indicated that Bi gram distance and similarity are more effective than Tri gram methods. Bi gram methods outperform Levenshtein, LCSR and Dice coefficient as well. Rama (2014) combines subsequence feature with the system developed by Hauer and Kondrak, which employs a number of word shape similarity scores as features to train a SVM model. Rama stated, "The subsequences generated from his formula weigh the similarity between two words based on the number of dropped characters and combine vowels and consonants seamlessly". He concludes that using the Hauer and Kondrak's system with a sequence length of 2 could maximize the accuracy. However, none of the work mentioned above has taken the word context to account.

2.2 Identifying cognates using semantic similarity

Kondrak (2001) proposed COGIT, a cognate-identification system that combines ALINE with semantic similarity. Given two vocabulary lists (L1, L2) in distinct languages, his system first calculates the phonetic similarities between each pair of entries (i, j) ∈ (L1 × L2). The semantic similarity of each pair of word is calculated based on the glosses information between a pair of words. The glosses are available in English for all words in both lists. The overall similarity is a linear combination of phonetic and semantic similarity, with different importance assigned to them respectively. The final outcome of this system is a list vocabulary-entry pair, sorted according to the estimated likelihood of their cognateness. Although their evaluation suggested that their methods employing semantic information from glosses perform better than methods based on word shape (phonetic and orthographic), they only focus on finding cognates between different Native American languages.

Frunza (2006) focuses on different machine learning techniques to classify word pairs as true cognates, false cognates or unrelated. She designed two classes called "orthographically similar" and "not orthographically similar" to separate these three types of cognates. However, since the cognate and false cognate are likely to have a high orthographical similarity, their features also include one form of semantic similarity that is whether the words are translations of each other. As a result, this third class - "translation of each other" allows the classifiers to make a decision when a false cognate has a high orthographical similarity. Similar to Kondrak who uses Wordnet and European Wordnet to fetch the glosses, Frunza employs bilingual dictionaries to retrieve the translations.

The method proposed by Mulloni, Pekar, Mitkov and Blagoev (2007) also combines orthographic similarity and semantic similarity. They first extract candidate cognate pairs from comparable bilingual corpora using LCSR, followed by the refinement process using corpus evidence about their semantic similarity. In terms of the semantic similarity, they believe that if two words have similar meanings – and are therefore cognates – they should be semantically close to roughly the same set of words in both (or more) languages. For example, for English *article* and French *article*, their method first finds a set of ten most similar words in the representative language respectively. Then, the method uses a bilingual dictionary to find the correspondence between the two sets of words. Thirdly, a collision set is created between two sets of neighbors, saving words that have at least one translation in the counterpart set. Lastly, The Dice coefficient is used to determine the similarity of the two sets which becomes the semantic similarity of the two original words.

2.3 Word Sense Disambiguation

Unlike all the previous methods which take semantic similarity into consideration, our proposed approach is based on word sense disambiguation (WSD) within monolingual texts, as we aim to use the sense of words as a pivot to identify candidate cognates. There are two mainstream approaches to word sense disambiguation. One is supervised WSD, which uses machine learning methods to learn a classifier for all target words from labeled training sets. Navigli (2012) asserts that memory-based learning and SVM approaches proved to be most effective. The other approach is Knowledge-based WSD, which exploits knowledge resources such as semantic networks to determine the senses of words in context. Such approaches use network features to identify which interpretation of the words in a sentence leads to the most connected representation with the words (as a semantic graph). The application we employ in this paper is called Babelfy which is powered by BabelNet.

2.4 BabelNet

BabelNet follows the structure of a traditional lexical knowledge base and accordingly consists of a labeled directed graph where nodes represent concepts and named entities, while edges express semantic relations between them. The network contains data available in WordNet and also incorporates new nodes and relationships extracted from the Wikipedia (Navigli and Ponzetto, 2012).

Each node in BabelNet is called a Babel synsets. Navigli (2013) explains that each Babel synset represents a given meaning and contains all the synonyms that express that meaning in a range of different languages. More precisely, a Babel synset contains (1) a synset ID; (2) the source of the synset (such as WIKI or WordNet); (3) the corresponding WordNet synset offset; (4) the number of senses in all languages and their full list; (5) the number of translations of the sense and their full list. For example, when the English word *bank* means a financial institution, its translations in other languages as (German *bank*), (Italian *banca*), and (French *banque*); (6) the number of semantic pointers such as relations to other Babel synsets and their full list; (7) its corresponding glosses (possibly available in many languages).

The early version of BabelNet can disambiguate verbs, nouns, adverbs and adjectives, but only provides French synonyms for nouns. The newly released Babelfy tool, which is fully supported by BabelNet, can disambiguate all nominal and named entity mentions within a text, and access French synonyms for all types of words (verbs, nouns, adverbs and adjectives). According to the description from Moro et al. (2014), the WSD and entity linking is achieved in three steps: (1) associating each vertex such as concept or named entity to generate a semantic signature of a given lexicalized semantic network. The semantic network refers to Babelnet; (2) extracting all linkable fragments from a given text and list possible meanings based on semantic network for each of them; (3) creating a graph-based semantic interpretation of the whole input text by linking candidate meanings of extracted fragments using the previously-generated semantic signature, followed by a dense sub-graph of this representation to select the best candidate meaning of each fragment.

Babelfy has been evaluated on various datasets and compared with different systems, and has been shown to achieve a better disambiguating performance among all the participating systems by using the selected datasets.

3 General Framework

We first present a general framework for cognate detection supported by disambiguation and multilingual resources. This framework provides a score of "cognateness" with regards to a source language for every word in text written a target language. Such a score can be interpreted as the likelihood that a reader learns the target language would be assisted by their native language (the source language) to understand the meaning of the word.

To calculate the score of a word W, the main steps in our framework are as follows:

- Identify the likelihood of each possible sense of W (semantic similarity score (SS))
- For each sense, all the translations of the sense in the source language become candidate cognates CW
- For each candidate cognate (CW), calculate its word shape similarity score (WSS), its orthographic similarity with W.
- Determine the cognateness of W as the maximum combined SS and WSS score, that is:

$$C(W) = Max_{CW}\big(x.SS(CW) + (1-x).WSS(CW)\big)$$

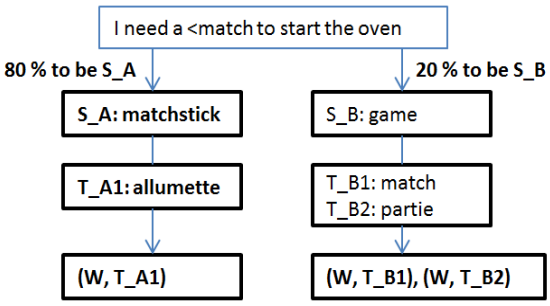

Figure 1 Process of general framework

For example, there are two possible meanings for the word *match* after disambiguation, which are S_A *matchstick* and S_B *game*, with 80% and 20% of sense likelihood respectively (this then becomes the (SS) score between the possible translations of each sense and the initial word). In the second step, all the possible translations of each sense in French will be retrieved according the multilingual resource. Finally, the retrieved translation will be paired with the word *match*. As shown in the figure 1, the final pairs under sense *A* would be (W, T_A1), similarly, pairs generated under sense *B*, which are (W, T_B1), (W, T_B2) and so on. For each of the candidate pair, the possible translation leads to the WSS score by applying orthographic/phonetic distance between the translation and the initial word (e.g., between *match* and *allumette*, *match* and *partie*). We then determine the cognateness of the word *match* by using the maximum combined SS and WSS score.

4 Methodology

The general approach presented in section 3 would be suited to the early version of BabelNet (version 1.0.1). Babelfy has a much higher accuracy for disambiguating; it does not provide sense likelihood for several candidate senses but only a single candidate sense. This is taken into account in our implementation by providing a simplified approach that does not use sense similarity. Indeed, in this paper we are assuming that the semantic similarity score is a static value which is always 1 and leave the combined formula for future work. As a result, the cognateness of a word W is now estimated by:

$$C(W) \approx Max_{CW}\big(WSS(CW)\big)$$

While scores are suited to applications that will not directly need a decision (such as used as a feature in readability measure, or used with graded colours in a visual interface), many will require a binary interpretation, including our evaluation framework. The binary decision is whether a word is a potential cognate of at least one of its likely translations. Such a decision can be based on a threshold for the cognate score presented above, or can be modeled using a panel of scores with a machine learning approach.

The implementation process is depicted in Figure 2 and the steps described as follows.

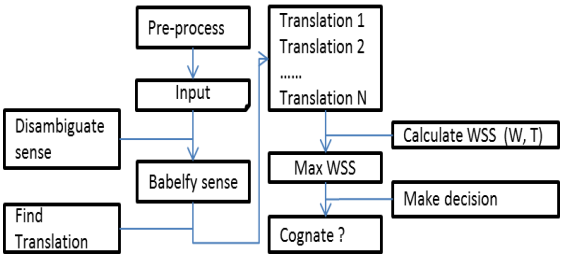

Figure 2 Process of implementation

Pre-Processing: The document is split in sentences; then each word within the sentence is stemmed using Krovetz Stemmer algorithm.

Disambiguate Sense: The stemmed sentence is disambiguated by Babelfy, with a specified "Exact" matching level. This matching level was empirically found to provide more accurate results than other options. For each word W in the sentence, we obtain a Babel Sense S.

Find Translations: query BabelNet based on the Babel Sense id of S. This provides a list of translations [T1, T2, …] of sense S.

Calculate WSS Score: Several measures can be used to calculate WSS score between word W and its translations. For example, we could get WSS1 as the score between (W, T1) and WSS2 for (W, T2) by using DICE. In the end, the Max [WSS1, WSS2] is selected as the final WSS score under sense S for word W with DICE.

Make Decision: We propose two approaches to decide whether or not a word W is cognate. The threshold approach states that true cognates are likely to have higher scores than non-cognates (Mulloni et Al., 2007). As a result, we build a training and a testing set for both cognate pairs (known cognates) and non-cognate pairs (random word pairs), and estimate the threshold that best separates cognates from non-cognates in the training set. The second approach proposes that several orthographic similarity measures can be retained, and the decision can be made using ma-

chine learning. A model is learnt from all W and all similarity measures in a training set of natural annotated data. For example, if a word W has 2 translations [T1, T2]; list_a which is [WSS1a, WSS2a, WSS3a, WSS4a, WSS5a] would be the WSS scores of T1, similarly, list_b ([WSS1b, WSS2b, WSS3b, WSS4b, WSS5b]) for T2. The 5 WSS scores in each list are calculated from Levenshtein, Bi Distance, LCSR, Dice and Soundex respectively. By finding the biggest value of (WSS1a, WSS1b), (WSS2a, WSS2b), (WSS3a, WSS3b) and so on, we generate a best value list which is Max [WSS1, WSS2, WSS3, WSS4, WSS5] for a word W.

5 Evaluation

5.1 Tuning the decision models

For the threshold approach, the training set contains 600 English French true cognate pairs and 600 English French non cognate pairs. The testing set contains 300 English French true cognate pairs and 300 English French non-cognate pairs. True cognate pairs were collected from various online resources. Non-cognate pairs were compiled by randomly selecting English words and French words from news websites[1].

While most cognate pairs on existing lists are of exactly identical words, this does not reflect the reality so we purposely included one third of non-identical cognate pairs in the training set. We have compared Bi Distance, Dice coefficient, Soundex, Levenshtein, and LCSR.

Measure	Threshold	Accuracy
BI distance	0.4	0.911
LCSR	0.444	0.898
Dice Coefficient	0.235	0.895
Levenshtein	0.428	0.882
Soundex	0.675	0.871

Table 1 Threshold and Accuracy of each orthographic measure.

Table 1 shows the accuracy of each measure used as a threshold on the testing set. In future evaluation, we will use the BI Distance to generate WSS score, but also Soundex. The reason we still employ Soundex despite its lowest accuracy is that it is a popular phonetic measure, so it is interesting to make comparisons with the BI distance.

For the machine learning approach, two models are trained from different training corpus described in the following section, one using Support Vector Machines (SVM) and Naïve Bayes (NB).

5.2 Cognate detection in natural data

The first experiment aims to evaluate our approach on natural language texts.

A corpus has been collected from web sources based on 6 different genres: cooking recipes (cook), political news (politics), sports news (sport), technical documentation (tech), novel (novel) and subtitles (sub). For each genre, we have collected 5 documents of roughly 500 words, resulting in a total 30 documents. A bilingual English/French speaker has manually annotated this training corpus to identify the true cognates[1]. Recent borrowings (such as *croissant* in English or *weekend* in French) were also annotated as cognates. The annotator reported that while some true cognates are very obvious as they have exactly the same spelling, there were cases where the words in French and English obviously shared some etymology and had some similarity (i.e. *juice* vs. *jus* or *spice* vs. *épice*), but it was difficult to decide if they would support a reader's understanding. Some words had a different spelling but very similar pronunciation and were therefore considered as cognates in this annotation process.

Table 2 lists the total numbers of cognates (C), non-cognates (N), stop words (S) and non-word characters (NW) for both the testing and training set. In brackets we show the number of cognates and non cognates that are actually processed by Balelfy and considered in the evaluation.

	Training	Testing
S	5,503	6,711
NW	585	752
C	1,623 (1,441)	2,138 (1,978)
N	3,368 (2896)	3,736 (2,008)
Total	11,709 (4,337)	13,337 (3,986)

Table 2 Natural Data corpus characteristics.

When testing our approaches on this dataset, we are interested in the overall accuracy, but also more specifically in the capacity of our system to identify the cognates and only the cognates. We

[1]Datasets presented in this paper are all available here: https://sourceforge.net/projects/cognates/files/?source=navbar

therefore use 3 measures of evaluation, Namely Accuracy (A), Recall (R) and Precision (P).

	BI Distance			Soundex		
	A	P	R	A	P	R
cook	0.81	0.73	0.81	0.79	0.71	0.8
politics	0.80	0.82	0.76	0.77	0.78	0.77
tech	0.80	0.78	0.78	0.8	0.77	0.8
sport	0.79	0.68	0.64	0.74	0.64	0.67
novel	0.81	0.56	0.76	0.8	0.54	0.77
sub	0.81	0.51	0.78	0.81	0.49	0.76
avg	0.80	0.72	0.75	0.78	0.69	0.77

Table 3 Results from decisions made by the thresholds approach with BI and Soundex

Table 3 shows the results of the threshold method, using either the BI distance or the Soundex similarity, for each genre and the average (avg). These results show that BI Distance has a higher overall detecting accuracy than Soundex, with average 0.8 compared with 0.78. It is interesting to observe that Soundex has a better recall rate than BI, which is to be expected given our definition of cognates as being words supported via a source language, rather than purely orthographically similar. There are no major differences across genres between Soundex and BI Distance. Both measures have higher precision and recall rate in cooking recipe (cook), political news (politics) and technology (tech), but lower results in sport news (sport), novel (novel) and subtitles (sub).

Table 4 shows the results for the two trained models. NB improves the precision across all genres but reduce the recall rate compared with SVM, which provides a completely reversed trend. The largest difference is observed for the sport news, novels and subtitles. NB dramatically improves their precision and still provides acceptable recall values, while SVM has lower precision but similar recall rate. The results also suggest that in addition to having an overall higher accuracy, NB is more robust across genres as there are smaller variations in precision and comparable variations in recall. For example, the precision range of SVM is between [0.47, 0.82] but [0.63, 0.85] for NB. If we compare the results between machine learning and threshold approaches, the BI distance, which is the best threshold approach exhibits variations of a similar order and range as those from SVM across the genres. As a result, the NB model is more likely to provide a balanced precision, recall and overall accuracy rate.

	SVM			NB		
	A	P	R	A	P	R
cook	0.82	0.75	0.81	0.8	0.77	0.73
politics	0.81	0.82	0.78	0.79	0.85	0.70
tech	0.82	0.78	0.82	0.83	0.84	0.76
sport	0.76	0.65	0.74	0.78	0.73	0.62
novel	0.79	0.53	0.78	0.85	0.65	0.74
sub	0.78	0.47	0.77	0.87	0.63	0.74
avg	0.80	0.70	0.78	0.82	0.77	0.71

Table 4 Results from decisions made by the machine learning approach

Finally, we establish 2 baselines to situate our results. The first baseline model (BL1) assumes that all words in the testing set are non cognates. To establish the second baseline (BL2), we employ an English/French cognate word list provided by Frunza (2006), and apply a simple decision rule that every word in the text that is present in the list should be returned as a cognate.

The results from two baselines are in the table 5. Because novel and subtitles contain less cognates, this results in the overall accuracy of BL1 and BL2 on these two genres being almost as good as the rates calculated from SVM. Precision and recall are not applied to BL1, and there is a huge variation between precision and recall values in BL2 across all the genres. This highlights the limits of a list-based approach.

	BL1	BL2		
	A	A	P	R
cook	0.58	0.64	0.95	0.14
politics	0.44	0.52	0.92	0.15
tech	0.52	0.58	0.86	0.14
sport	0.59	0.64	0.88	0.12
novel	0.77	0.78	0.61	0.13
sub	0.8	0.83	0.69	0.24
avg	0.6	0.65	0.85	0.15

Table 5 Results from decisions made by a naïve (BL1) and a list-based (BL2) baseline.

5.3 Testing semi-cognates in controlled sentences

Our second evaluation aims to test the robustness of our approach specifically for semi-cognates. For example, the English word *address* is a true cognate when it means "mailing, email" or "deftness, skill, dexterity". However, it is a false

cognate when it refers to "discourse". This task is highly dependent on the quality of the disambiguation.

20 semi-cognates are used to create controlled sentences where they appear in either as a true cognate or a false cognate. For each semi-cognate, we created 2 sentences, one where it is a true cognate and one where it is a false cognate. Additionally, we ensured that the other words in the sentences were neither true nor false cognates. Using *address* as an example again, the sentences created were "What is your <address?" and "His keynote <address is very insightful."

In this evaluation we use the NB model to make decisions since it provided the best accuracy in the previous evaluation.

True Cognate		F. Cognate	
C	N	C	N
15	4	14	5

Table 6 Results from NB model.

Table 6 shows the confusion matrix when the model is applied to the 20 sentences containing true cognates and the 20 sentences containing false cognates. The confusion matrices first show that 2 semi-cognates fail to be annotated or that BabelNet did not contain translation for the disambiguated sense. On the 4 errors made on recognizing the true cognates, 2 of them are due to an error in disambiguation, and for the other 2 Babelfy fails to give provide the correct translations because the extracted text fragment is a combination of two words or more. For example, "I like action movie", the sense of word *action* is correct but mapped to *action_movie* instead of *action* itself. Of the 14 errors made on recognizing false cognates, 6 were due to errors in the disambiguation, 7 were due to erroneous translations of the sense, and only 2 were due to an error of the model (word *organ* and *orgue* were considered cognates). For example, the word *assume* in sentence "I <assume full duty in this matter" was disambiguated as "Take to be the case or to be true and accept without verification or proof." It has translations such as *assumer, supposer*. Since we will only take the translation that has the highest WSS score, the *assumer* is selected instead of *supposer*.

6 Discussion

While the performance of our approach show improvements over a baseline using a dictionary based approach, there are a number of errors that could be avoided by integrating a probabilistic disambiguation approach as proposed in section 3. The issue of the quality of the disambiguation system, even though we selected a system with high performances, has been highlighted in section 5.3 on the semi-cognate evaluation, but has also been observed on natural data.

Another issue that is that Babelfy is not able to process all the words that should be disambiguated. For example, "You can bring egg whites to room <temperature by setting the eggs out on the counter at least 30 <minutes in <<advance of your preparation", and the *advance* was ignored. Table 2 shows how many such missing words are occurring in the natural dataset. The number of missing words varies across genres, for example, subtitles may only have 9 missing words out of 2,000 while sport news may have 55. Non cognate words are more likely to be ignored compared with true cognates; especially the cooking recipes and political news may include lots of low frequency word and name entities.

Additionally, there are several cases where an identified sense does not have a French translation in BabelNet (although we verified that the language has some). For instance, "Place the egg whites in a <bowl in a pan of warm water", although Babelfy successfully disambiguates *bowl* as "A round vessel that is open at the top; mainly used for holding food or liquids", BabelNet simply does not have a French translation *bol* under this specific sense in its network. Furthermore, some errors come from erroneous translations provided by BabelNet, even though we filter the translation sources to only use open multilingual wordnet (omwn), wiki, wikidata, wiki translation (wikitr). For instance, *marshmallow* shows French translation *marshmallow* instead of *chamallow*, and *soccer* shows French translation *soccer* instead of *football*, thus impacting on the precision. Finally, annotations are sometime subjective for similar but non identical words, or close but non identical meanings.

7 Conclusion and Future work

We presented a methodology to identify potential cognates in English sentences for a French reader. The accuracy is around 80%, and it is high enough to successfully be used in sentence selection schemes to support learners to get better understanding before tackling the hard texts, which has been proposed an alternative learning method for English Learners (Uitdenbogerd, 2005).

As implied earlier, our proposed approach is highly dependent on the sources used; our future work will first try to develop a strategy to minimize the noise and analyze how much the performance can be improved with ideal settings. Future work will also focus on integrating the decision model, or directly the cognateness score, into readability measures. In a pilot study, we found that word level criteria such as frequency or length are indeed not applicable when the word is a cognate (that is, very difficult words such as word *disambiguation* can actually be very transparent and therefore easy in the context of a multilingual reader). Thirdly, more work is needed to more accurately detect usages of semi-cognates, and integrating the SS score with WSS score, so that the actual readability measure to balance the impact from semantic and word shape feature and possibly alleviate errors made by a disambiguation system.

References

Brew, C., and McKelvie, D. (1996). Word-pair extraction for lexicography. *Proceedings of the second International conference on new methods in language processing*, Ankara, Turkey, 45-55.

Frunza, O.M. (2006). Automatic identification of cognates, false friends, and partial cognates. *Masters Abstracts International, 45,* 2520.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics*, Seattle, Washington, 288-295.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, Pittsburgh, Pennsylvania, USA, 1-8. doi: 10.3115/1073336.1073350.

Kondrak, G. (2005). N-gram similarity and distance. *String processing and information retrieval, 3772*, 115-126, Springer Berlin Heidelberg.

Kondrak, G., and Dorr, B. (2004*).* Identification of confusable drug names: A new approach and evaluation methodology. *Proceedings of the 20th international conference on Computational Linguistics,* 952-958. doi: 10.3115/1220355.1220492.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady, 10,* 707.

Melamed, I.D. (1999). Bitext maps and alignment via pattern recognition. *Comput Linguist.*, *25*(1), 107-130.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics,* 2.

Mulloni, A., Pekar, V., Mitkov, R., & Blagoev, D. (2007). Semantic evidence for automatic identification of cognates. *Proceedings of the 2007 workshop of the RANLP: Acquisition and management of multilingual lexicons*, Borovets, Bulgaria , 49-54.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. *Proceedings of the 38th international conference on Current Trends in Theory and Practice of Computer Science*, Czech Republic, 115-129. doi: 10.1007/978-3-642-27660-6_10

Navigli, R. (2013). A quick tour of babelnet 1.1. Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing, Samos, Greece, I, 25-37. doi: 10.1007/978-3-642-37247-6_3

Navigli, R., and Ponzetto, S.P. (2012). Joining forces pays off: Multilingual joint word sense disambiguation. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 1399-1410.

Rama, T. (2014). Gap-weighted subsequences for automatic cognate identification and phylogenetic inference. arXiv: 1408.2359.

Simard, M., Foster, G.F., and Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing*, Toronto, Ontario, Canada, *2*.

Uitdenbogerd, S. (2005). Readability of French as a foreign language and its uses. *Proceedings of the Australian Document Computing Symposium*, 19-25.