Overview of the 2015 ALTA Shared Task:

Identifying French Cognates in English Text

Laurianne Sitbon
Queensland University of Technology (QUT)
l.sitbon@qut.edu.au

Diego Molla
Macquarie University, Sydney, Australia
diego.molla-aliod@mq.edu.au

Haoxing Wang
Queensland University of Technology (QUT)
haoxing.wang@hdr.qut.edu.au

Abstract

This paper presents an overview of the 6th ALTA shared task that ran in 2015. The task was to identify in English texts all the potential cognates from the perspective of the French language. In other words, identify all the words in the English text that would acceptably translate into a similar word in French. We present the motivations for the task, the description of the data and the results of the 4 participating teams. We discuss the results against a baseline and prior work.

1 Introduction

Because many languages have evolved from a shared source language (e.g. Indo-European languages), many words in their vocabularies are the same or are very similar. Additionally, global communications have facilitated the transfer of words from one language to another in modern languages. As a result, when learning a related language, a learner's native language can support the acquisition and understanding of vocabulary words that are identical or similar in both languages.

A vocabulary word is a spelling associated to a particular meaning. Such pairs of identical or similar words that also share meaning across two languages are referred to as cognates. Definitions can vary in the level of similarity (exact or similar spelling, exact or similar pronunciation, or both). So far, research on detecting cognates has focused on being able to identify pairs of cognates in lists of presented pairs of words.

In contrast, in this shared task we use the notion of potential cognate in a target language with reference to a source language: a word in the target language that could be translated by a similar word in the source language such that these words form a cognates pair. Being able to identify these potential cognates in texts could provide technologies to extract easy to understand sentences and could support measures of reading difficulty (Uitdenbogerd, 2005) which can in turn be embedded in ranking information retrieval results or in sentence selection for summarization.

In 2015, the sixth Australasian Language Technology Association (ALTA) shared task was set to identify in English texts all the potential cognates from the perspective of the French language. A total of 6 teams registered to the competition, with 4 teams submitting their results.

In this paper we present some background for the task, describe the dataset and contrast the results of the participants against baselines and previous work. Section 2 presents some background and prior work, Section 3 presents the task, the dataset and the evaluation measures. Section 4 provides the results of the participants. Section 5 discusses the results and future work.

2 Cognates identification and detection

Cognates are pairs of words that are similar in spelling/pronunciation as well as meaning in two languages. By extension, as mentioned above, we refer here to cognates as words in one language that would, in their context of use, acceptably translate into a word in the second language with which they would form a cognate pair.

We also refer here to true cognates as per this definition, as opposed to false cognates (also referred to as false friends) which appear in both languages' lexicons but bear different meanings

Laurianne Sitbon, Diego Mollá and Haoxing Wang. 2015. Overview of the 2015 ALTA Shared Task: Identifying French Cognates in English Text . In *Proceedings of Australasian Language Technology Association Workshop*, pages 134−137.

(such as *pain* in English and *pain* in French (bread)), and as opposed to semi-cognates, which, depending on their context of use, may be either true cognates or false cognates (such as *chair* in English that translates into French as *chaise* if one refers to furniture (false cognate) as *chaire* if one refers to a University position (true cognate), while *chair* in French means *flesh* in English (false cognate)).

The task of detecting potential cognates is in contrast to many experimental settings in the literature that focused on detecting pairs of cognates amongst pairs of words in both languages.

Early work investigated the use of single orthographic or phonetic similarity measures, such as Edit Distance (ED) (Levenshtein, 1966), Dice coefficient (Brew and McKelvie, 1996), Longest Common Subsequence Ratio (LCSR) (Melamed, 1999).

Kondrak and Dorr (2004) reported that a simple average of several orthographic similarity measures outperformed all the measures on the task of the identification of cognates for drug names. More recently, Rama (2014) combined the subsequence features and a number of word shape similarity scores as features to train a SVM model. Kondrak (2001) proposed COGIT, a cognate-identification system that combines phonetic similarity with semantic similarity, the latter being measured from a distance between glosses in a lexical handcrafted resource. Frunza (2006) explored a range of machine learning techniques for word shape similarity measures, and also investigated the use of bi-lingual dictionaries to detect if the words were likely translations of each other. Mulloni, Pekar, Mitkov and Blagoev (2007) also combined orthographic similarity and semantic similarity, the latter being measured based on lists of collocated words.

In previous work, Wang (2014) established an initial version of the dataset proposed in the shared task, and used it to evaluate a new approach. This approach uses word shape similarity measures on pairs selected using word sense disambiguation techniques in order to account for context when seeking possible translations. The implementation is based on BabelNet, a semantic network that incorporates a multilingual encyclopedic dictionary. This work explored a variety of ways to leverage several similarity measures, including thresholds and machine learning.

3 The 2015 ALTA Shared Task

The task of the 2015 ALTA Shared Task was to identify in English texts all the potential cognates from the perspective of the French language. In other words, identify all the words in the English text that would acceptably translate into a similar word in French.

3.1 Dataset

Participants were provided with a training set that is approximately the same size as the testing set. Each set was composed of 30 documents, 5 in each of the following genres: novel, subtitles, sports news, political news, technology news, and cooking recipes. While the separations between the documents was included in both the training and testing data, the categories of documents were not released for the task.

Because we focus on transparency for understanding, we consider similarity (not equality) in either spelling or pronunciation as supporting access to meaning. A single human annotator has identified the potential cognates accordingly.

Similarity: typically similarity is examined at the level of the lemma, so the expected level of similarity would ignore grammatical markers and language-specific suffixes and flexions (for example *negociating* and *negocier* would be considered cognates as the endings that differ respond to equivalent grammatical markers in the languages, similarly for *astrologer* and *astrologue,* or *immediately* and *immediatement*), accented letters are considered equivalent to those without accents and unpronounced letters are ignored (hence *chair* in the French sense *chaire* would be considered true cognate since the *e* at the end is not pronounced). In addition, weak phonetic differences (such as the use of *st* instead of *t* in words such as *arrest* vs. *arrêt,* some vowel substitutions such as *potatoes* vs. *patates)* tend to be ignored and there is more flexibility on long words than on short words.

Rules for proper names: people's names are never considered cognates. Names of companies and products are not considered cognates where the name is a unique word (eg. *Facebook)*, but the words are considered on an individual basis where the name is also a noun compound (eg. in Malaysian Airlines, where *Malaysian* is a cognate, but not *Airlines*). Names of places may be cognates depending to their level of similarity with their translation.

3.2 Task description

The data presented for the task was divided into document text and annotation files. Document text files were formatted with one word (with punctuation attached, if present) per line and each line starts with the line number followed by a space (see Fig.1.a). Document boundaries were indicated by a document id marker.

```
1   <docid 1>
1   Chewy
2   little
3   drops
4   of
5   chocolate
6   cookies,
7   covered
8   with
9   peanuts
```

Figure 1.a Document Text File

```
Eval_id,Cognates_id
1, 6 7
```

Figure 1.b Annotation File

Annotation files were in .csv format. Each line comprised a document number in the first column, and a space delimited list of cognate term indices in the second column.

For instance, to indicate that `chocolate' (index 6) and `cookies' (index 7) are cognates of French words, the annotation file will include the entry shown on Figure 1.b.

Participants were provided with a document text file and corresponding annotation file for the training set, and with a document text file and a sample annotation file (produced by the baseline system, see below) for the test set, and they had to submit their own corresponding annotation file.

3.3 Evaluation

The evaluation measure used for the competition is the mean f-score as defined by the “Kaggle in Class” [1]platform:

$$F1=2pr/(p+r)$$

$$\text{where } p=tp/(tp+fp),\ r=tp/(tp+fn)$$

Where precision (p) is the ratio of true positives (tp) to all predicted positives (tp + fp) and recall (r) is the ratio of true positives to all actual positives (tp + fn).

However we will discuss the results in terms of recall and precision as well.

3.4 Baselines

The baseline for the task was produced by using a list of 1,765 known English/French cognate words (also matching for singular forms of plurals). Each word in the document text that belonged to the list was deemed to be a cognate for the purpose of the task. As demonstrated in prior work, such baseline tends to yield a high precision but a very low recall.

In addition to the baseline, we ran the task against the system proposed by Wang (2014). The implementation uses BabelNet (Navigli and Ponzetto, 2012) for disambiguating and accessing candidate translations, and integrates 5 measures of similarity (Bi Distance, Dice coefficient, Soundex, Levenshtein, and LCSR) using a Naïve Bayes classifier to assign the cognates labels.

4 Results

The evaluation was performed via the “Kaggle in Class” platform. This platform supports the partition of the test data into a public and a private component. When a team submitted a run, the participants received instant feedback on the results of the public data set, and the results of the private data set was kept for the final ranking. We used the default 50-50 partition provided by Kaggle in Class. The results are reported in Table 1. The table also includes the results returned by the baseline and the system proposed by Wang (2014).

System	Public	Private
LookForward	**0.705**	**0.769**
LittleMonkey	0.671	0.714
Wang(2014)	0.63	0.669
MAC	0.599	0.669
toe_in	0.37	0.367
Baseline	0.229	0.342

Table 1: F1 measure results

In Table 2 are presented the results evaluated posterior to the task in terms of recall and precision.

System	Public	Private

[1] http://inclass.kaggle.com/c/alta-2015-challenge/

	R	P	R	P
LookForward	0.76	0.69	**0.79**	0.76
LittleMonkey	0.77	0.62	0.77	0.67
Wang(2014)	**0.81**	0.54	**0.79**	0.60
MAC	0.72	0.54	0.74	0.63
toe_in	0.27	0.62	0.27	0.63
Baseline	0.15	**0.72**	0.22	**0.91**

Table 2**:** recall (R) and precision (P) results

5 Discussion and future work

The rankings between the public and the private test sets are consistent, and therefore the team LookForward is a clear winner. Both LookForward and Little Monkey achieved better results than Wang (2014), and MAC lagged closely behind. The descriptions of the systems used by LookForward, MAC, and toe_in can be found in the proceedings of the ALTA 2015 workshop. Whereas in the teams LookForward and MAC the system used a distance metric that compared the original word with the translation provided by a machine translation system, in the team toe_in the system was based on the intersection of an English and a French lexicon after applying a set of lexical transformations.

As predicted, the baseline had a high precision, and in fact it was the highest of all runs. It is also interesting to observe that the Wang (2014) system is the next highest in recall, while a bit lower in precision. It is important to note that while similar, the annotations on the dataset used in the 2014 paper was slightly different to the one of the 2015 shared task, however the system has not been retrained. This explains a drop in f-measure compared to the results presented in the paper.

Because of a fairly subjective definition of cognates, the annotation of the data can strongly depend on the annotator's personal viewpoint. It would be very interesting to have the dataset re-annotated by 2 more annotators to be able to measure inter-annotator agreement. This would allow judging whether the performance of the best systems reaches the level of humans on the task.

However, in order to put some perspective on the results, it will be even more interesting to measure the impact of the f-measure levels on various tasks such as measuring readability, or selecting sentences or paragraphs in a computer supported language learning system. One could think that a system stronger in precision would be more appropriate to select easy-to-read sentences, while a system stronger in recall may lead to better estimates of reading difficulty.

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. (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., 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.

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

Navigli R. and Ponzetto S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network . *Artificial Intelligence*, 193, Elsevier, pp. 217-250

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

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

Wang, H., Sitbon, S. (2014), Multilingual lexical resources to detect cognates in non-aligned texts, *Proceedings of the Twelfth Annual Workshop of the Australasia Language Technology Association (ALTA)*, Melbourne, Australia, November 2014.