Cognate Identification using Machine Translation

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

Abstract

In this paper we describe an approach to automatic cognate identification in monolingual texts using machine translation. This system was used as our entry in the 2015 ALTA shared task, achieving an F1-score of 63% on the test set. Our proposed approach takes an input text in a source language and uses statistical machine translation to create a word-aligned parallel text in the target language. A robust measure of string distance, the Jaro-Winkler distance in this case, is then applied to the pairs of aligned words to detect potential cognates. Further extensions to improve the method are also discussed.

1 Introduction

Cognates are words in different languages that have similar forms and meanings, often due to a common linguistic origin from a shared ancestor language.

Cognates play an important role in Second Language Acquisition (SLA), particularly between related languages. However, although they can accelerate vocabulary acquisition, learners also have to be aware of false cognates and partial cognates. False cognates are similar words that have distinct, unrelated meanings. In other cases, there are *partial* cognates: similar words which have a common meaning only in some contexts. For example, the word *police* in French can translate to *police*, *policy* or *font*, depending on the context.

Cognates are a source of learner errors and the detection of their incorrect usage, coupled with correction and contextual feedback, can be of great use in computer-assisted language learning systems. Additionally, cognates are also useful for estimating the readability of a text for non-native readers.

The identification of such cognates have also been tackled by researchers in NLP. English and French are one such pair that have received much attention, potentially because it has been posited that up to 30% of the French vocabulary consists of cognates (LeBlanc and Séguin, 1996).

This paper describes our approach to cognate identification in monolingual texts, relying on statistical machine translation to create parallel texts. Using the English data from the shared task, the aim was to predict which words have French cognates. In §2 we describe some related work in this area, followed by a brief description of the data in §3. Our methodology is described in §4 and results are presented in §5.

2 Related Work

Much of the previous work in this area has relied on parallel corpora and aligned bilingual texts. Such approaches often rely on orthographic similarity between words to identify cognates. This similarity can be quantified using measures such as the edit distance or dice coefficient with n-grams. Brew and McKelvie (1996) applied such orthographic measures to extract English-French cognates from aligned texts.

Phonetic similarity has also been shown to be useful for this task. Kondrak (2001), for example, proposed an approach that also incorporates phonetic cues and applied it to various language pairs.

Semantic similarity information has been employed for this task as well; this can help identify false and partial cognates which can help improve accuracy. Frunza and Inkpen (2010) combine various measures of orthographic similarity using machine learning methods. They also use word senses to perform partial cognate between two languages. All of their methods were applied to English-French. Wang and Sitbon (2014) combined orthographic measures with word sense disambiguation information to consider context.

Shervin Malmasi and Mark Dras. 2015. Cognate Identification using Machine Translation . In *Proceedings of Australasian Language Technology Association Workshop*, pages 138−141.

Cognate information can also be used in other tasks. One example is Native Language Identification (NLI), the task of predicting an author's first language based only on their second language writing (Malmasi and Dras, 2015b; Malmasi and Dras, 2015a; Malmasi and Dras, 2014). Nicolai et al. (2013) developed new features for NLI based on cognate interference and spelling errors. They propose a new feature based on interference from cognates, positing that interference may cause a person to use a cognate from their native language or misspell a cognate under the influence of the L1 version. For each misspelled English word, the most probable intended word is determined using spell-checking software. The translations of this word are then looked up in bilingual English-L1 dictionaries for several of the L1 languages. If the spelling of any of these translations is sufficiently similar to the English version (as determined by the edit distance and a threshold value), then the word is considered to be a cognate from the language with the smallest edit distance. The authors state that although only applying to four of the 11 languages (French, Spanish, German, and Italian), the cognate interference feature improves performance by about 4%. Their best result on the test was 81.73%. While limited by the availability of dictionary resources for the target languages, this is a novel feature with potential for further use in NLI. An important issue to consider is that the authors' current approach is only applicable to languages that use the same script as the target L2, which is Latin and English in this case, and cannot be expanded to other scripts such as Arabic or Korean. The use of phonetic dictionaries may be one potential solution to this obstacle.

3 Data

The data used in this work was provided as part of the shared task. It consists of several English articles divided into an annotated training set (11k tokens) as well as a test set (13k tokens) used for evaluating the shared task.

4 Method

Our methodology is similar to those described in §2, attempting to combine word sense disambiguation with a measure of word similarity. Our proposed method analyzes a monolingual text in a *source* language and identifies potential cognates in a *target* language. The source and target languages in our work are English and French, respectively.

The underlying motivation of our approach is that many of the steps in this task, *e.g.* those required for WSD, are already performed by statistical machine translation systems and can thus be deferred to such a pre-existing component. This allows us to convert the text into an aligned translation followed by the application of word similarity measures for cognate identification. The three steps in our method are described below.

4.1 Sentence Translation

In the first step we translate each sentence in a document. This was done at the sentence-level to ensure that there is enough context information for effectively disambiguating the word senses.[1] It is also a requirement here that the translation include word alignments between the original input and translated text.

For the machine translation component, we employed the Microsoft Translator API.[2] The service is free to use[3] and can be accessed via an HTTP interface, which we found to be adequate for our needs. The Microsoft Translator API can also expose word alignment information for a translation.

We also requested access to the Google Translate API under the University Research Program, but our query went unanswered.

4.2 Word Alignment

After each source sentence has been translated, the alignment information returned by the API is used to create a mapping between the words in the two sentences. An example of such a mapping for a sentence from the test set is shown in Figure 1.

This example shows a number of interesting patterns to note. We see that multiple words in the source can be mapped to a single word in the translation, and vice versa. Additionally, some words in the translation may not be mapped to anything in the original input.

[1] We had initially considered doing this at the phrase-level, but decided against this.

[2] `http://www.microsoft.com/en-us/translator/translatorapi.aspx`

[3] For up to 2m characters of input text per month, which was sufficient for our needs.

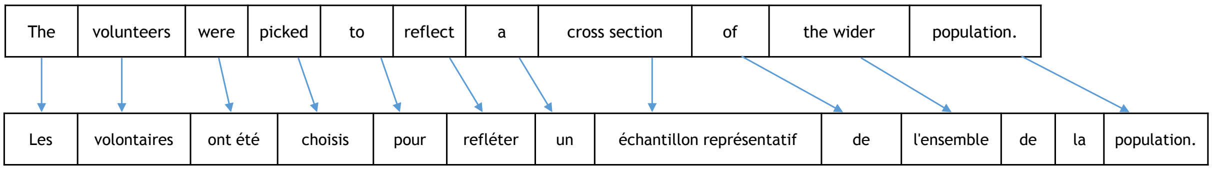

Figure 1: An example of word alignment between a source sentence from the test set (top) and its translation (bottom).

4.3 Word Similarity Comparison Using Jaro-Winkler Distance

In the final step, words in the alignment mappings are compared to identify potential cognates using the word forms.

For this task, we adopt the Jaro-Winkler distance which has been shown to work well for matching short strings (Cohen et al., 2003). This measure is a variation of the Jaro similarity metric (Jaro, 1989; Jaro, 1995) that makes it more robust for cases where the same characters in two strings are positioned within a short distance of each other, for example due to spelling variations. The measure computes a normalized score between 0 and 1 where 0 means no similarity and 1 denotes an exact match.

For each pair of aligned phrases, every word in the source phrases was compared against each word in the aligned phrase to calculate the Jaro-Winkler distance. A minimum treshold of 0.84 was set to detect matches; this value was chosen empirically.

Under some circumstances, such as appearing before a vowel, French articles and determiners may combine with the noun.[4] Accordingly, we added a rule to remove such prefixes (*d'*, *l'*) from translated French words prior to calculating the distance measure. Additionally, all accented letters (*e.g.* *é* and *è*) were replaced with their unaccented equivalents (*e.g.* *e*). We found that these modifications improved our accuracy.

4.4 Evaluation

Evaluation for this task was performed using the the mean F1 score, conducted on a per-token basis. This is a metric based on precision – the ratio of true positives (tp) to predicted positives (tp + fp) – and recall – the ratio of true positives to actual positives (tp + fn). The F1 metric is calculated as:

$$F1 = 2\frac{pr}{p+r} \text{ where } p = \frac{tp}{tp+fp}, \quad r = \frac{tp}{tp+fn}$$

Here p refers to precision and r is a measure of recall.[5] Results that maximize both will receive a higher score since this measure weights both recall and precision equally. It is also the case that average results on both precision and recall will score higher than exceedingly high performance on one measure but not the other.

5 Results and Discussion

Our results on the test set were submitted to the shared task, achieving an F1-score of 0.63 for detecting cognates.[6] The winning entry was 10% higher and scored 0.73.

The key shortcoming of our approach is that we only consider the best translation for detecting cognates. However, a word in the source language may translate to one or more words in the target language, one or more of which could be cognates. However, the cognate(s) may not be the preferred translation chosen by the translation system and therefore they would not be considered by our system.

This was not by design, but rather a technical limitation of the Microsoft Translator API. Although the API provides word alignment information, this is only available for the preferred translation.[7] A separate method is provided for retrieving the n-best translations which could contain relevant synonyms, but it is unable to provide word alignments.

[4] For example, *l'enfant* (the child).

[5] See Grossman (2004) for more information about these metrics.

[6] We obtained F1-scores of 0.67 and 0.59 on the private and public leaderboards, respectively.

[7] Details about this and other restrictions can be found at `https://msdn.microsoft.com/en-us/library/dn198370.aspx`

By using a different machine translation system, one capable of providing alignment information for the n-best translations, our approach could be extended to consider the top n translations. Given the good results using only the preferred translations, this can be considered a very promising direction for additional improvement and is left for future work.

We also noted that there were some idiosyncrasies in the annotation of the training data that were not explicitly outlined. One example is that proper nouns referring to locations, *e.g. Russia, Ukraine* and *Afghanistan*, were annotated whilst other proper nouns were not. Our system would require additional components to distinguish different classes of named entities to be able to implement this logic.

To conclude, we proposed an approach that takes an input text in a source language and uses statistical machine translation to create a word-aligned parallel text in the target language. A robust measure of string distance, the Jaro-Winkler distance in this case, was then applied to the pairs of aligned words to detect potential cognates. The results here are promising and could potentially be further improved using the extensions described in this section.

References

Chris Brew and David McKelvie. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55.

William Cohen, Pradeep Ravikumar, and Stephen Fienberg. 2003. A comparison of string metrics for matching names and records. In *KDD workshop on data cleaning and object consolidation*, volume 3, pages 73–78.

Oana Frunza and Diana Inkpen. 2010. Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, 1(1).

David A Grossman. 2004. *Information retrieval: Algorithms and heuristics*, volume 15. Springer.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Matthew A Jaro. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5-7):491–498.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

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

Shervin Malmasi and Mark Dras. 2014. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 95–99, Gothenburg, Sweden, April. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015a. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, pages 1403–1409, Denver, CO, USA, June. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015b. Multilingual Native Language Identification. In *Natural Language Engineering*.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and Misspelling Features for Natural Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145, Atlanta, Georgia, June. Association for Computational Linguistics.

Haoxing Wang and Laurianne Sitbon. 2014. Multilingual lexical resources to detect cognates in non-aligned texts. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, volume 12, pages 14–22.