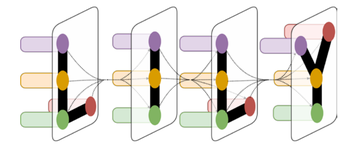
Word alignment with a pivot language

R. Thomas McCoy[1]

[1]Department of Linguistics, Yale University



Introduction

Many techniques in NLP crucially depend upon having large quantities of training data. For mainstream languages, such as English and French, this is not a problem, since large quantities of data exist for these languages; but for many other languages, termed low-resource languages, there simply is not enough data available to train high-quality NLP tools. One NLP task that relies heavily on a large training corpus is word alignment, which is an important step in the training of a machine translation system. I propose a technique for compensating for data shortages for the task of word alignment. This technique works by inferring knowledge about the language for which data is lacking by using a high-resource relative of the low-resource language as a pivot language that can act as an intermediary between its low-resource relative and the high-resource language with which it is being aligned.

Materials and Methods

The basic structure of my approach is illustrated in Figure 1. To align a sentence in a low-resource language with a sentence in a high-resource language, first a high-resource relative of the low-resource language is chosen as a pivot language. Edit distance is then used to interface between the low-resource language and the pivot language, while the IBM word alignment models are used to interface between the pivot language and the high-resource language.

In my experiments, I use Spanish as an artificially low-resource language being aligned with the high-resource language English. I test several different edit distance algorithms: Levenshtein edit distance, edit distance based on phonological features, character-embedding-based edit distance, and edit distance trained on cross-lingual cognates. I also test each successive IBM Model from Model 1 to Model 4. Finally, I test 7 different pivot languages of varying relatedness to Spanish, namely Finnish, Danish, German, Italian, French, Portuguese, and Spanish itself

(1) a. Ikárámu umukoôbwa a-ra-andik-a ſbárúwa nziza ná yo.
pen girl she-PRES-write-ASP letter good with it
'The pen, the girl is writing a nice letter with it.'

b. Ikárámu umukoôbwa araandika ſbárúwa nziza ná yo

The pen, the girl is writing a nice letter with it

Figure 1. Word alignment example

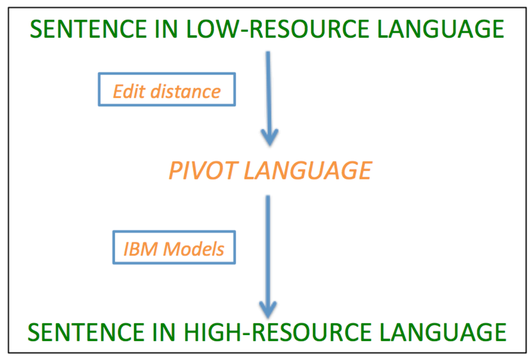


Figure 2. Approach.

- **Spanish:**
  en el caso de turquìa la cuestión es diferente .
- **English:**
  in the case of turkey it is different .
- **Guessed English:**
  i like case of turkey de justice mr different .
- **Guessed English (refined):**
  in the case of turkey the question is different .

Figure 3. Example translations.

Edit distance algorithm	Alignment model	Pivot language	AER
-	Random	-	0.948
-	Diagonal	-	0.819
Levenshtein	IBM M1	Portuguese	0.673
VC	IBM M1	Portuguese	0.670
Feature-based	IBM M1	Portuguese	0.674
char2vec	IBM M1	Portuguese	0.672
Cognate-based	IBM M1	Portuguese	0.663
Cognate-based	IBM M1	Portuguese	0.554
Cognate-based	HMM	Portuguese	0.504
Cognate-based	IBM M3	Portuguese	0.548
Cognate-based	IBM M4	Portuguese	**0.488**
Levenshtein	IBM M4	Portuguese	0.500
Levenshtein	IBM M4	Italian	0.571
Levenshtein	IBM M4	French	0.613
Levenshtein	IBM M4	German	0.722
Levenshtein	IBM M4	Danish	0.732
Levenshtein	IBM M4	Finnish	0.776
Levenshtein	IBM M4	Spanish	0.360
-	fast-align	-	0.288

Table 1. Test results.

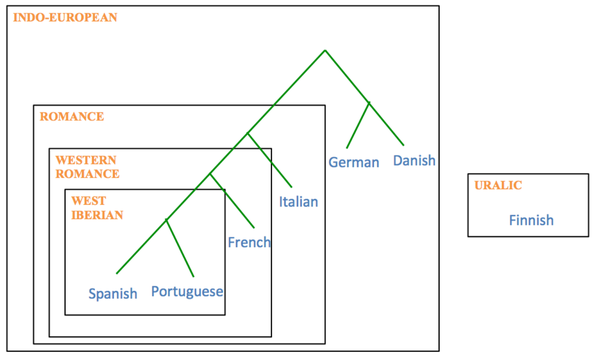


Figure 5. Linguistic family tree.

Results

Test results are reported in Table 1. It was found that all edit distance algorithms performed similarly well to each other, but the cognate-based edit distance was the best-performing by a slight margin. The best-performing alignment model was IBM Model 4. Finally, the best-performing pivot language (other than Spanish itself) was Portuguese. In general, the performance of the word aligner with different pivot languages had levels of success corresponding to how closely related the pivot language was to Spanish as illustrated in Figure 5.

The best-performing word alignment model was then used in conjunction with the Moses translation toolkit to create a preliminary Spanish to English translation program, trained only on Portuguese data. Figure 3 shows an example translation with this program. The intial program performed quite poorly, but the manual addition of a small quantity of Spanish information (namely, the meanings of 65 Spanish function words) significantly refined the translations that the program outputted, as show by the refined guess below the initial guess in Figure 3.

Conclusion

These results show that the use of a pivot language is a viable approach to rough machine translation of low-resource languages. The importance of the relatedness of the pivot language to the source language was verified, and a novel edit distance algorithm was introduced. Although the translation programs created with this approach are never likely to attain state-of-the-art performance, they may be useful in scenarios (such as disaster relief) in which conveying the main gist of a message is the most improtant goal.

Acknowledgement

Thank you to Bob Frank for the many hours spent discussing the topics in this presentation. Thank you also to Drago Radev for allowing me to present with the LILY group.