CS697: Unsupervised Cross Lingual Alignment Mid-Term Report

Pranjal Singh 10327511
Indian Institute of Technology,Kanpur

April 10, 2014

The problem of word-alignment is to find correspondence between words and phrases in parallel texts. Suppose we have two languages L1 and L2, then the task in hand is to identify which word token in L1 corresponds to which word token in L2.
For simplicity, we will take two languages, i.e., English and Hindi for describing examples. This can be extended to any pair of two languages.
The alignment of punctuation is also one of the important tasks but for now, the main focus is not the punctuations. The difficulty is that we don't have a parallel corpora and is generally the most common scenario. Most of the word in cross-lingual alignment which has been done by today is mainly on a parallel corpora using tools such as GIZA++.

1 Data

The data for this type of work can be any data which is focussed on only one topic so that there is a a large correspondence between the tokens in both the languages.
For example, the data which was used in my previous work in CS498.
For evaluation, Gold Standard Word Aligned Data can be used with each annotator adding Certain and Probable tags to each aligned pair.

2 Methodology

We have in hand two corpus which are basically focussed on a single topic Z.
We have used Latent Topic Clustering and Unsupervised Syntax Induction Model of ADIOS to derive corresponding clusters in both the languages which have almost similar tokens in both the languages. These can serve as a ID to the corresponding lines in the two languages.
These can be identifier to complex expressions which cannot be aligned word to word. So they are handled in pre-processing step (Smadja et al., 1996; Ahrenberg et al., 1998; Tiedemann, 1999).

We plan to use the following three features of tokens of both languages:

1. Direct Mapping using bilingual dictionaries for same meaning tokens
2. Predicted Mapping using Clusters derived above
3. Semantic Mapping such as synonymy using Word-Net resource and POS tags in both the languages.

The problem reduces to a matrix solving problem which is basically a Bag-of-Words Model and further can be converted into a *tf-idf* model. Rows contain sentences in one language and column contains those in the other language. Each row element is itself a n-dimensional vector, where n is the number of tokens in that sentence.

We need to reorder rows and columns such that entries in the cells of the matrix maximize everywhere.

3 Syntax

The final word-aligned file will be a description of each word-to-word mapping with the corresponding line number. For example,
18 1 1
18 2 2

This means that word numbers 1 & 2 in line numbers 18 of both the languages align with each other.
In addition, we can have two fields called Certain(C) and Probable(P) denoting the type of alignment. Certainty can vary between [0-1] adding another field which we will call as Accuracy(A).

- Items which are separated by a white space or a punctuation will be called a word. They are the ones which need to be aligned.
- Any extra tokens which have no corresponding tokens in the other language will be substituted by COPY token which means, no subtraction in score for that.
- Sometimes two lines may mean the same thing or a part of the line may mean the same, so the two sub-parts will be said as aligned.

4 Evaluation

Evaluation will be done using three different measures- Precision(P), Recall(R) and F-Score(F).
We have alignment A derived by our algorithm and G is the Gold Standard Alignment.

$$P_X = \frac{|A_X \cap G_X|}{A_X}$$
$$R_X = \frac{|A_X \cap G_X|}{G_X}$$
$$F_X = \frac{2P_X R_X}{P_X + R_X}$$

where X =C,P, C-Certain, P-Probable

References

[1] Marianna Apidianaki. Unsupervised cross-lingual lexical substitution. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 13–23, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Jörg Tiedemann. Word to word alignment strategies. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[3] Paola Virga and Sanjeev Khudanpur. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, MultiNER '03, pages 57–64, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.