

# CS697: Unsupervised Cross Lingual Alignment

## Report

Pranjal Singh 10327511  
*Indian Institute of Technology, Kanpur*  
Advisor: Prof. Satyadev Nandakumar

April 2, 2014

### Abstract

The world today is full of large volumes of text resources spreading over internet, journals and digital libraries. These resources are present in almost all major languages. The main problem here is that texts may differ in two languages and there is a need that these should be aligned in all languages. This can be used to detect inconsistencies in the two texts. We present a framework which understands this problem of cross lingual alignment and tries to provide a solution for a very special set of problems where data is focused on only one topic. The data used is a manually built corpus in English and Hindi which is focused on only one topic, i.e., coal scam.

## 1 Introduction

With an ever increasing volume of text, there is a need that these should be connected in some way. This need can be easily understood by the example of Wikipedia. Wikipedia has very well grown into a multilingual resource containing information from every field. English has the largest wikipedia among all languages and the amount of information present in English wikipedia exceeds information present in any single wikipedia. This is mainly due to cultural and regional differences present across all languages, sometimes also called as cultural or regional bias. The need is to enrich each and every wikipedia so that every group of people come to know more about every other thing.

Same is the case with any electronic dictionary in medical domain, which are not properly connected. The word cross-lingual alignment in both these cases essentially means that corresponding lexicons in, say for example, two languages, English and Hindi, are directly connected or atleast have some semantic connection. For the ease of understanding, we will take only two languages in this work and in the coming sections we will discuss only these two languages, namely English and Hindi.

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, as told above, 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 are not the punctuations. The difficulty is that we don't have a parallel corpora and is generally the most common scenario. Most of the work in cross-lingual alignment which has been done by today is mainly on a parallel corpora using tools such as GIZA++.

The rest of the text is organized as follows. Section 2 discusses some of the previous works which have been done in similar area. Section 3 discusses about the data which we will be using. Section 4 discusses about the Methodology which we have adopted. Section 5 discusses about the syntax of the final output and how can it

be interpreted. Section 6 discusses about the Evaluation parameters which describe the extent of goodness of an alignment. Section 7 describes my previous work which is a basis of this work. Section 8 further discusses the limits of this work and what can be done in future. Section 9 acknowledges my advisor and Section 10 are the references for this work.

## 2 Previous Work

Most of the previous work in this area has been supervised and there is as such no currently existing work which looks into cross lingual alignment in great detail, mainly on a large corpus which contains all sought of information related to a topic.

There is a work called DbPedia (Auer et al., 2008) which is an ongoing, large project and it automatically extracts information from Wikipedia, normalizes the extracted information, links the information with other online data repositories and provides an interactive access.

It has been said that word alignment, like many other NLP tasks (Banko and Brill, 2001), highly improves if there is a large amount of training data available but then providing large amount of training data is also not trivial. This is due to the fact that different domains need different type of training data and different type of feature vectors. Currently, there is no general technique that can serve useful for different NLP tasks. However, there has been work in this direction where people are looking into general representations of tokens (Collobert and Weston, 2008) and (Bengio et. al, 2009) which can aid in almost all NLP tasks but still it lags in accuracy. They have been referred as embeddings.

The above work discusses unsupervised approach but nowadays there has been more focus on semi-supervised approaches where people are trying to improve existing baseline systems by introducing some unsupervision into them. Unsupervised word representations are pushed into existing systems and they try to improve upon their result (Turian, Retinov & Bengio, 2010).

Recent days, there also has been focus on Lexical Substitution which aims to substitute target words (or phrases) in a context, which could prove very useful in paraphrasing and textual entailment. Apidianaki et al. has worked on Cross-Lingual lexical substitution where he has developed a framework that can substitute a target word in a context with several alternative substitute words in another language. The work is unsupervised and exploits the results of a cross-lingual word sense induction method that derives the sense of the words by clustering their translations using their semantic similarity. The output of this work may prove to be an input to cross-lingual Information Retrieval and Machine Translation Systems (Sinha et al., 2009) and (Mihalcea et al., 2010).

There also have been various statistical techniques and alignment heuristics (Brown et al., 1993; Vogel et al., 1996; Melamed, 2000) which have looked into automatic alignment of words and phrases. These word alignment techniques generally contain links between multi-word units or phrases, where some of the lexical items cannot be separated into words which match in another language.

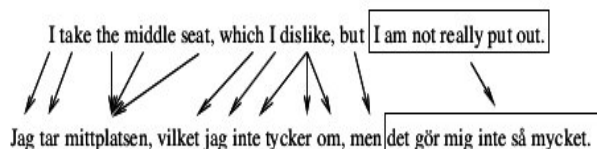


Figure 1: Example of word alignment from Saul Bellow "To Jerusalem and back: a personal account" (Bellow, 1976) and its Swedish translation (Bellow, 1977) (the *Bellow Corpus*)

Here, it can be easily seen that most of the words are properly connected to that in other language but the English unit "I am not really put out" doesn't have any proper way of connecting it to words in Swedish expression. So the whole expression is linked to the corresponding Swedish expression.

In some approaches, few complex expressions are separated in pre-processing step and later they are handled as complex units in the same ways as single words. This is very similar to the way ADIOS (Zach Solan, 2005) handles complex phrases and later treats them as single units. ADIOS can detect hidden syntactic units present

in a large document which, in general, cannot be analysed just by looking into the document. It also summarizes the documents as few production rules which may serve useful in grammar induction.

Having discussed some previous work in this area, the next few sections will now describe how the current problem is formulated and how we can try to solve the problem.

### 3 Data

The data for this type of work can be any data which is focused on only one topic so that there is a large correspondence between the tokens in both the languages.

For example, the data which was used in my previous work in CS498.

It is a manually built corpus on *Coal Scam* in English and Hindi collected from various online Hindi and English newspaper and news channel websites such as *Dainik Jagran*, *Hindustan*, *Aaj Tak*, etc. Hindi corpus has more than 35,000 tokens whereas English corpus has over 50,000 tokens. Stemming is to be ignored in this work to preserve syntax.

For evaluation, Gold Standard Word Aligned Data can be used with each annotator adding Certain and Probable tags to each aligned pair.

## 4 Methodology

### 4.1 PLSA: Probabilistic Latent Semantic Analysis

PLSA models co-occurrence information in a probabilistic framework which can discover the underlying semantic structure of the data. We have our corpus in term-document matrix form as input to this PLSA framework which extracts topics as clusters from it. PLSA considers the following three variables:

- Documents:  $d \in D = \{d_1, d_2, \dots, d_N\}$  are the observed/visible variables where  $N$  is the total number of documents given by the size of our corpus.
- Words:  $w \in W = \{w_1, w_2, \dots, w_M\}$  are again the observed variables where  $M$  is the number of distinct words in the corpus.
- Topics:  $z \in Z = \{z_1, z_2, \dots, z_K\}$  are the latent/hidden variables where  $K$  is to be assigned in advance.

PLSA models the following generative process:

- Select a document  $d_n$  with probability  $P(d)$
- For every word  $w_i, i \in 1, \dots, N_w$  present in the document  $d_n$ :
  - Choose a topic  $z_i$  with probability  $P(z|d_n)$  from a multinomial conditioned on the given document
  - Choose a word  $w_i$  with probability  $P(w|z_i)$  from a multinomial conditioned on the topic chosen previously

The predictive probability of PLSA model will be denoted by  $P(w|d)$ , and hence the objective function is given by:

$$L = \prod_{d,w} P(w|d) = \prod_{d \in D} \prod_{w \in W} P(w|d)^{n(d,w)}$$

where  $n(d, w)$  represents the frequencies observed, as in, the number of times word  $w$  appears in document  $d$ . Since, this turns out to be a non-convex optimization problem, it can be solved using EM(Expectation Maximization) algorithm.

PLSA uses EM-algorithm for following log-likelihood problem:

$$M = \log L = \prod_{d \in D} \prod_{w \in W} n(d, w) \cdot \log \sum_{z \in Z} P(w|z)P(z|d)$$

## 4.2 Formulation

We have in hand two corpus which are basically focused 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 also serve as 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 for 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 scores for each of the cell of the matrix maximizes.

A general representation for word-to-word alignment  $L$  for a given cross-lingual text with  $N$  words( $a_1, a_2, \dots, a_N$ ) from one language and  $M$  words( $b_1, b_2, \dots, b_M$ ) from other language can be written as:

$$L = L_1, L_2, \dots, L_p, L_p = [a_{x_1}, b_{x_2}], x_1 \in \{1, \dots, N\}, x_2$$

Direct mapping is certainly the best alignment and should be given the maximum score. Since predicted mapping take into account both semantics and syntax, its score will be less than direct mapping score. Semantic mapping score will also be less than direct mapping and will depend on the amount of semantic similarity present.

For semantic mapping, words can be extracted from English and Hindi WordNet. It can provide Synonym relation among words. Although this can be time consuming, but if each word in the corpus has an associated vector containing all *\*-nymy* relations in it from the very beginning, we can definitely save some time.

## 5 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.
- Alignment scores will also be given when two numbers are aligned in both languages, say for example, when dates are aligned

The Gold standard alignment will also be scored against our current system so that we can assign proper scores.

## 6 Evaluation

Evaluation will be done using three different measures- Precision(P), Recall(R) and F-Score(F). Precision and recall have turned out to be quite useful in various NLP tasks, so we will be using them as our evaluation parameter.

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

Precision(also called positive predictive value) is the fraction of retrieved instances that are relevant.

Recall(also known as sensitivity) is the fraction of relevant instances that are retrieved.

F-Score is the *harmonic mean* of precision and recall.

## 7 Base Results/Motivation

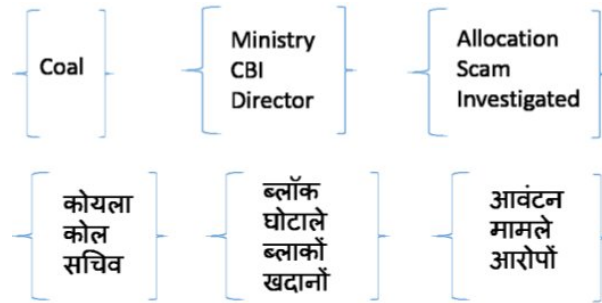


Figure 2: Cluster of size 3 obtained from PLSA technique which indicates cross lingual relationship between Hindi and English in our focused corpora on Coal Scam

[कोयला, आवन्टन्, कोल, जिदल, ब्लॉक, कंप्रियो, कंप्री, हिडाल्को, ब्लाक]  
['coal', 'cbi', 'the', 'allocation', 'birla', 'fir', 'block', 'hindalco', 'alleged']

[कोयला, घोडाले, सीबीआई, दर्ज, सरकार, सीबीआई, ब्लॉक, मामले, पूर्व]  
['minister', 'coal', 'prime', 'bjp', 'the', 'scam', 'government', 'party', 'issue']

[जांच,कोर्ट,कोयला,सरकार,रिपोर्ट,सीबीआई,सुप्रीम,प्रधानमंत्री,मंत्री,घोडाले]  
['cbi', 'coal', 'the', 'court', 'ministry', 'report', 'probe', 'files', 'agency', 'government']

Figure 3: An output from ADIOS which again indicates cross lingual relationship between entities in Hindi and English in our focused corpora on Coal Scam

## 8 Future Work

The quality of results can be easily improved if we use a large corpus though it slows down the rate of computing. We can also insert this unsupervised methodology into an existing supervised system and can improve its result. This would be more of a semi-supervised system.

We can also take into account phonemes for better alignment of tokens such as *Nouns*. This work mainly describes a system for a focused corpora but it can be extended to work for a very general corpora.

We could also have tried working with n-gram clusters which could have provided better alignment for complex expressions.

We can also choose alignment strategies to suit particular needs. We are re-ordering sentences in the corpus to align them in both languages but this may lead to loss of some semantic information because semantic information varies across various sentences.

Also, score of predicted mapping should vary depending on the tokens but our system cannot infer directly how to manage scoring for this case. In future work, we can experiment with different set of languages, in particular, generalize this approach for any two set of languages.

## 9 Acknowledgement

I would like to show my sincere gratitude towards Prof. Satyadev Nandakumar, Computer Science and Engineering Department, IIT Kanpur for his motivation to do this work and becoming my advisor for this work. I would also like to thank my all friends who have helped me throughout this project. I would also like to thank CSE Department, IIT Kanpur for providing such as courses as CS697 where we figure out completely new objectives and try to strengthen our concepts related to the topic by reading existing papers.

## 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] Zellig S Harris. Distributional structure. *Word*, 1954.
- [3] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
- [4] Heng Ji. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning.

- [5] Heng Ji. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, UMSLLS '09, pages 27–35, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [6] Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11629–11634, 2005.
- [7] Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 477–487, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [8] Oscar Tckstrm, Ryan T. McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *HLT-NAACL*, pages 477–487. The Association for Computational Linguistics, 2012.
- [9] 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.
- [10] 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.