

An Empirical Study of Cross-Lingual Unsupervised Alignment Based on Syntactic Features

Pranjal Singh(10511)

November 10, 2013

References

- [1] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

The authors have presented a novel statistical method for factor analysis of binary and count data which is closely related to a technique known as Latent Semantic Analysis. It analyses two-mode and co-occurrence data, which has many applications in NLP, information retrieval, etc.

LSA addresses the questions of polysems and synonyms which are important for lexical and semantic level distinction. LSA maps high-dimensional count vectors, such as those arising in vector space representation of text documents. The goal was to provide a mapping from syntactic level to semantic level but it remains unsatisfactory on many grounds.

PLSA takes as input a high-dimensional document by term matrix which has many intrinsic latent variables hidden in it. This vector has words which are unique excluding all the stop words. The words may be stemmed also because they are not looking into syntactical features much.

The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable with each observation. The parameters of this model are $P(term|k)$ and $P(k|doc)$. It is possible to derive the equations for computing these parameters by Maximum Likelihood. So authors have also used EM algorithm to derive the parameters.

$P(t|k)$ for all t and k , is a term by topic matrix

(gives which terms make up a topic)

$P(k|doc)$ for all k and doc , is a topic by document matrix

(gives which topics are in a document)

Lagrangian terms are added to ensure the constraints. Derivatives are taken w.r.t the parameters (one of them at a time) and are equated to zero. The resulting equations are solved to get fixed point equations which can be solved iteratively. This is the PLSA algorithm.

The authors have used two datasets to evaluate the perplexity performance:

a) a standard information retrieval test collection MED with 1033 document,

b) a dataset with noun-adjective pairs generated from a tagged version of the LOB corpus.

- [2] 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.

The authors have presented two approaches to automatically extract cross-lingual predicate clusters, based on bilingual parallel corpora and cross-lingual information extraction. They have proposed a new inductive learning framework to automatically augment background data for low- confidence events and then conduct global inference.

For each test document, a background document is constructed by gradually replacing the low-confidence events with the predicates in the same cluster. Then they have conducted cross-document inference technique as described in (Ji and Grishman, 2008) to improve the performance of event extraction which is basically concerned with ACE evaluations. They have worked with 33 distinct event types.

In the first approach, they have used 852 Chinese event trigger words in ACE05 training corpora as their anchor set. For each Chinese trigger, they have searched its automatically aligned English words from a Chinese-English parallel corpus including 50,000 sentence pairs. The word alignment was obtained by running Giza++.

They have used two cross-lingual IE techniques on TDT5 Chinese corpus to generate more clusters:

Chinese IE_MT

Apply Chinese IE on the Chinese texts to get a set of Chinese triggers *ch-trigger-set1*, and then use word alignments to translate (project) *ch-trigger-set1* into a set of English triggers *en-trigger-set1*;

MT_English IE

Translate Chinese texts into English, and then apply English IE on the translated texts to get a set of English triggers *en-trigger-set2*.

They have used the RWTH Aachen Chinese-to-English statistical phrase-based machine translation system. The derived cross-lingual predicate clusters consisted of 438 English predicate clusters and 543 Chinese predicate clusters.

The inductive learning algorithm they have proposed consists of background document generation. For each event mention in a test document, the baseline event tagger produces the following local confidence value:

$LConf(trigger, etype)$: The probability of a string trigger indicating an event mention with type *etype* in a context sentence *S*;

If $LConf(trigger, etype)$ is lower than a threshold, and it belongs to a predicate cluster *C*, they create an additional background document *BD* by:

For each predicate *i* belonging to *C*, they replace trigger with predicate *i* in *S* to generate new sentence *S'*, and add *S'* into *BD*.

They have applied the cross-document inference techniques as described in (Ji and Grishman, 2008) to improve trigger and argument labeling performance.

For both English and Chinese, the inductive learning approach using cross-lingual predicate clusters provided significant improvement over the baseline event extraction system (about 4% absolute improvement on trigger labeling and 2%-2.3% on argument labeling). The most significant gain was provided for the recall of trigger labeling 5.9% absolute improvement for English and 5.4% absolute improvement for Chinese.

- [3] 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.

The authors have presented an algorithm for inducing cross-lingual clusters. They have extended the work on improving prediction of linguistic structure for other languages. They provide an algorithm for inducing cross-lingual clusters and have shown that features derived from these clusters significantly improve the accuracy of cross-lingual structure prediction. They have mainly focused on semi-supervised and linguistic transfer methods for multilingual structure prediction.

Their algorithm generates cross-lingual word clusters which are consistent across all languages. This is achieved by means of a probabilistic model over large amounts of monolingual data in two languages, coupled with parallel data through which cross-lingual word-cluster constraints are enforced. They have worked on word clusters by using predictive class bigram model introduced by Uszkoreit and Brants (2008).

The authors have also evaluated the use monolingual word clusters (around 256 such clusters of 1 million most frequent word types) over thirteen languages for dependency parsing and NER. For NER experiments, they have studied three German languages. For all NER experiments, they have used a sequential first-order conditional random field (CRF) with a unit variance Normal prior, trained with L-BFGS until epsilon-convergence (epsilon = 0.0001, typically obtained after less than 400 iterations).

They have used delexicalized direct transfer method proposed by McDonald et al. (2011) based on work by Zeman and Resnik (2008) as their starting point. For a given training set, the learner ignores all lexical identities and only observes features over other characteristics. They have extended this approach to universal parsing by adding cross-lingual word cluster features. Their method for inducing cross-lingual clusters has two stages: First, it clusters a source language (S) as in the monolingual case, and then projects these clusters to a target language (T), using word alignments. Due to some drawbacks, they have modeled word sequences in each language by the monolingual language model with likelihood function so that clusterings are good in both the languages. They have coupled the clusterings defined by these individual models, by introducing additional factors based on word alignments, as proposed by Och (1999).

Cross-lingual cluster features are helpful across the board and give a relative error reduction ranging from 3% for DA to 13% for PT, with an average reduction of 6%, in terms of unlabeled attachment score (UAS). X-LINGUAL CLUSTERS provides roughly the same performance as PROJECTED CLUSTERS suggesting that even simple methods of cross-lingual clustering are sufficient for direct transfer dependency parsing. While the performance of the transfer systems is very poor when no word clusters are used, adding cross-lingual word clusters give substantial improvements across all languages. X-LINGUAL CLUSTERS reduce the error on test set by 22% in terms of F_1 and upto 26% for ES.