Cross-Lingual Sentiment Analysis

Sayali Borkar
Indian Institute of Technology, Bombay
sayalib@cse.iitb.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
pb@cse.iitb.ac.in

Abstract

Sentiment Analysis requires large sentiment labelled corpus for training. Many rare languages have insufficient labelled corpus. Such languages are called low resource languages. Training for such languages leads to insignificant results. In this paper, we study various approaches to perform sentiment analysis for low resource languages using high resource languages. We also look into word embedding approaches, different models to train cross-lingual word embeddings and their evaluation techniques.

1 Introduction

Sentiment Analysis has become one of the Natural Language Processing (NLP) tasks which has gained popularity over the past decade. The sentiment classification model requires large labelled corpus in order to learn the language patterns. Many rare languages do not have sufficient labelled corpora. This results into poor classification results. This lack of labelled corpus leads to the use of cross-lingual sentiment analysis.

Cross-lingual Sentiment Analysis (CLSA) is the use of resource rich languages to solve the problem of sentiment analysis for resource poor languages. The task is to train the sentiment model on language L_1 (source), for which a corpus is available, and to test it on another language L_2 (target) for which labeled data is unavailable. The biggest problem faced is to address the gap between the two languages. Many different strategies are used to solve this problem. Machine Translation was used by Wan (2009); Wei and Pal (2010) to solve the problem of CLSA by translating the source language corpus into target language. Lu et al. (2011); Meng et al. (2012) used parallel corpus in order to bridge the gap between two languages. Word embeddings (Luong et al., 2015; Singhal and Bhattacharyya, 2016; Barnes et al., 2018) have been used extensively to represent the source and target language corpus to train the cross-lingual model. Following section mentions different strategies to perform cross-lingual sentiment analysis.

Section 2 provides general approaches used for performing cross-lingual sentiment analysis. Section 3 explains cross-lingual word embeddings with examples. We discuss cross-lingual word embedding models in Section 4. Section 5 focuses on the evaluation methods used for measuring the quality of word embeddings. We prove our conclusions in Section 6.

2 Traditional Approaches for CLSA

Cross-lingual sentiment analysis is the task of using resource rich language to predict the sentiment of a resource poor language. The biggest problem is to address the gap between the two languages. This section focuses on different strategies used to solve this problem.

A lot of work in sentiment analysis has been done in English. Due to this, abundance of resources are available which can be used for training. These resources can be used to perform sentiment analysis on a low resource language. (Singhal and Bhattacharyya, 2016) used word embeddings and polar words in English to perform sentiment analysis for Hindi and Marathi as well as a few European Languages. Given the training data in one language, the text is converted into English using Google translate. These English words are then mapped to English pre-trained word embeddings. A CNN network is trained for sentiment classification.

Another approach (Singhal and Bhattacharyya, 2016) used to improve the correct prediction of sentiment and to learn the correct patterns in a

sentence, English polar words were appended to the training data. This addition allows the training model to predict the negative sentiment label correctly.

(Balamurali et al., 2012) proposed the use of WordNet senses as a replacement of words to train bilingual word embeddings. The WordNet of target language L_2 is created by adding words from corresponding synsets in source language L_1. So, the words in two languages having similar context will have same synset identifiers. (Balamurali et al., 2012) used WordNet to perform cross-lingual sentiment analysis for Hindi and Marathi. For a target language, the text in training and testing corpus is replaced by synset identifiers. Each word is annotated by senses using two methods - manually or using an automated system. A classifier is trained on this processed corpus.

3 Cross-Lingual Word Embeddings

Word Embeddings have been widely used in various NLP tasks like tagging, sentiment analysis, translation, etc. with successful results. The main focus has been trying to solve problems for a single language. Increase in multilingual NLP tasks has motivated the training of cross-lingual word embeddings.

Cross-lingual Word Embeddings are nothing but word representations of two or more languages into a common vector space. The words having similar meaning or similar context should be close to each other in the vector space. To illustrate: the English word happy and the Hindi word – (khush) should be close to each other in the vector space.

Following figure describes

Cross-lingual Word Embeddings can be trained using two approaches - online method or offline method. Online method is to jointly learn the word embeddings for multiple languages. Consider training bilingual word embeddings, then using a bilingual signal the word vectors can be trained in a common vector space. Offline method is to learn the monolingual embeddings separately and then project them into a common vector space. Online methods require strong bilingual signal whereas offline methods require weaker bilingual signal.

4 Cross-lingual Word Embedding Models

Recent study shows that it is important to use a good bilingual supervision signal in order to train bilingual word vector representations. This section focuses on the different techniques used to train bilingual word embeddings.

4.1 Using Word Alignments

The main focus while training cross-lingual word embeddings is to improve the bilingual quality as well as to preserve the monolingual quality of the word embeddings. This section explains a joint model to learn bilingual word embeddings that learn the context co-occurrence information through monolingual component and meaning equivalent signals from the bilingual constraint. It is an extension to skip-gram model (Mikolov et al., 2013a) in multilinguality domain. The bilingual constraint used to bridge the gap between two languages is word alignments.

Word Alignments

The main objective of word alignments is, given a source sentence s and a target sentence t, to find the correspondence between the words in s and t. The link between the words in s and t signify the translations of each other. One of the main approaches used to produce alignments is using parallel text. Parallel text includes same data translated into multiple languages. A model is trained on this corpus to get alignments.

Bilingual Skipgram (BiSkip) Model

The next step is to train the skipgram for two languages using the word alignments that are generated using the parallel text. The BiSkip model (Luong et al., 2015) is an updated skipgram model so that it predicts words crosslingually. Given an alignment link (s_i, t_j), the word s_i is used to predict the neighbours of word t_j and vice versa. This is equivalent to training a single skipgram model which considers the two words s_i and t_j same. So along with training the skipgram model monolingually, alignment links are used to train the model bilingually. This is equivalent to training four skipgram models which predict words between different pairs of languages. Consider two languages L_1 and L_2. The different skipgram models will learn to predict words between language pairs $L_1 \rightarrow L_1$, $L_1 \rightarrow L_2$, $L_2 \rightarrow L_1$ and $L_2 \rightarrow L_2$.

4.2 Using Bilingual Dictionary

This approach (Ammar et al., 2016) learns bilingual word embeddings using a bilingual dictio-

nary and monolingual data. The monolingual corpus is used to learn the semantic similarity between words of same language and the bilingual dictionary is used for cross-lingual similarity. We find clusters from bilingual dictionary by allocating the same cluster to translationally equivalent words in both the languages. Once the clusters are formed, each cluster is represented using a cluster ID and the monolingual corpora for both languages is used to estimate the multilingual word embeddings. The words in the monolingual corpora are replaced by the cluster IDs and all the monolingual datasets available are concatenated to train the model based on the clusters. By doing this, all the words in the cluster will have same word embedding thus creating anchor points in the vector space to bridge the two languages. Once the dataset of all languages is pre-processed, any monolingual embedding model can be used to learn the word vectors.

4.3 Minimizing Euclidean Distance for Mapping

This approach (Mikolov et al., 2013b) is an offline approach which uses independent vector spaces for two languages and projects one vector space into the second. The mapping is carried out by minimizing the distance between the vector representations of two words in a bilingual dictionary.

Consider X and Z to be the word embedding matrices of a bilingual dictionary for two languages such that X_i and Z_i are vector representations of i^{th} entry in the dictionary. To find the linear transformation W, such that $XW \approx Z$, we minimize the sum of squared Euclidean distance.

$$W = argmin \sum_{i} \|X_i W - Z_i\|^2 \qquad (1)$$

W is the least square solution for the equation $XW = Z$.

An improvement over this approach, is to preserve the monolingual quality of the word embeddings after mapping. This can be achieved if W is an orthogonal matrix. So the exact solution with this constraint is $W = VU^T$ where, $U\Sigma V^T$ is the SVD factorization of $Z^T X$.

4.4 Using Phrase Translations for Mapping

This approach (Zhao et al., 2015) finds the translation from source vector space to target vector space by using phrase translations which have similar continuous representations. The main focus of the approach is to find translation rules for phrases in a sentence. The relative positions of words in the vector space are preserved between languages. The task is to find a linear mapping between the vector representations of source phrases and target phrases. This transform is called Global Linear Projection as a single mapping is used to project every source phrase.

The source and target phrases are denoted by f and e respectively. Their vector representations are denoted by $f \in R^{1\times d}$ and $e \in R^{1\times d}$ where d is the dimension size. We need to find a linear transformation matrix $W \in R^{d\times d}$ with the help of phrase translations $(f_1, e_1), (f_2, e_2), (f_3, e_3)...(f_n, e_n)$ where n is the number of translations available. Let F and E be two matrices such that $F = [f_1^T, f_2^T, ...f_n^T]$ and $E = [e_1^T, e_2^T, ...e_n^T]$. We calculate W using -

$$FW = E \qquad (2)$$

This can be solved by -

$$W \approx (F^T F)^{-1} F^T E \qquad (3)$$

Given an unlabeled source phrase s, we can find the target vector representation by $t = sW$. This target representation will be close to the real translation phrases. Experiments have been conducted on these phrase representations by performing Urdu-English translation. The classification accuracy achieved is approximately 27%.

4.5 Maximizing Correlation for Mapping

In this approach (Faruqui and Dyer, 2014), we use pre-trained word embeddings of individual languages which are trained on large unlabeled corpus. This approach uses independent vector spaces of the two languages and projects them into a shared vector space. It uses the Canonical Correlation Analysis (CCA) (Hotelling, 1936) to measure the linear relationship between two multidimensional variables. Given two vectors, CCA will find two projection directions such that the new projected vectors will have a maximum correlation. The dimension of these direction vectors is equivalent to the smaller dimension of the two vector spaces. A translation is required from one vector space to another to find the direction vectors. A bilingual dictionary acts as a resource to find the points, in the independent vector spaces, which should overlap.

Let the word embeddings for two languages be represented by $\Sigma \in R^{v \times d_1}$ and $\Omega \in R^{v \times d_2}$ respectively. d_1 and d_2 are dimensions of vectors for two languages respectively and v is the number of words in the bilingual dictionary.

Given two vector matrices Σ and Ω, CCA returns two projection matrices.

$$P, Q = CCA\,(\Sigma, \Omega) \tag{4}$$

where, $P \in R^{d_1 \times d}$, $Q \in R^{d_2 \times d}$ and d is the dimension of new vector space.

For corresponding two vectors x and y in the matrices, it finds direction vectors p and q using

$$p, q = argmax\ \rho\,(xp, yq) \tag{5}$$

where, ρ is the correlation between two vectors.

Once the projection directions are determined, the vocabulary of both languages can be projected into a third vector space.

$$\Sigma^* = \Sigma P \quad \Omega^* = \Omega Q \tag{6}$$

where, Σ^* and Ω^* are the word embeddings for two languages in the new vector space.

5 Evaluation Methods

The quality of word embeddings can be evaluated using a number of tasks to check whether the word vector representations capture the syntactic and semantic relations.

5.1 Word Similarity

Word Similarity is used to measure the semantic quality of word representations using various word similarity datasets. This can be performed monolingually as well as cross-lingually. The similarity between two words is measured by calculating cosine similarity between their vector representations. There are four different benchmark datasets that are widely used.

1. WS-353 dataset (Finkelstein et al., 2001) contains 353 pairs of English words which are labelled with similarity ratings by humans.

2. RG-65 dataset (Rubenstein and Goodenough, 1965) contains 65 pairs of Nouns assigned with similarity ratings from 0 to 4.

3. MC-30 (A. Miller and G. Charles, 1991) dataset contains 30 pairs of nouns that are a subset of RG-65.

4. MTurk-287 (Radinsky et al., 2011) dataset contains 287 pairs of words assigned with similarity ratings using crowd-sourcing.

To evaluate word embeddings across languages, we can translate these benchmark datasets into respective languages and then calculate the Spearman's correlation coefficient (L. Myers and Well, 2003) with the human similarity ratings.

5.2 Cross-lingual Dictionary Induction

The task of cross-lingual dictionary induction (Vulić and Moens, 2013) is to measure how good the word embeddings are in finding out semantically similar word pairs across languages. Given a word in source language L_1, we can find top-k similar words in target language L_2 using cosine similarity. A gold bilingual dictionary is required to measure the accuracy of the trained word embeddings. For each pair (w_1, w_2) in the dictionary, we find if w_2 is a part of top-k similar words for w_1. The accuracy is the fraction of entries in the dictionary which fulfill the above condition.

6 Conclusion

Cross-lingual sentiment analysis is one of the important NLP tasks with a lot of scope in multilinguality as well as in exploiting various cross-lingual signals across languages. We discussed various approaches used for cross-lingual sentiment analysis. We have also discussed the role of word embeddings in the improvement in performance of cross-lingual sentiment analysis. Different approaches to train bilingual word embeddings are described. We explain the evaluation methods for the trained word embeddings as well as the datasets available.

References

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*.

R Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *COLING*, pages 73–82.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 462–471.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. volume 20, pages 406–414.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

J L. Myers and Arnold Well. 2003. Research design statistical analysis. XVII.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 320–330, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Herbert Rubenstein and John Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8:627–633.

Prerana Singhal and Pushpak Bhattacharyya. 2016. Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In *COLING*.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia. Association for Computational Linguistics.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bin Wei and Christopher Pal. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 258–262, Uppsala, Sweden. Association for Computational Linguistics.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning translation models from monolingual continuous representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1527–1536, Denver, Colorado. Association for Computational Linguistics.