**Cross Bilingual Embeddings**

Saumitra Yadav(201507647)

Sahil Chelaramani(20162051)

Pranav Dhakras(20162303)

1. Introduction
2. Methods
   1. Neural network
   2. Co-occurrence matrix
   3. GloVe matrix
   4. Sentence Alignments & Word Translation
3. Sample outputs
4. Future work
   1. Bootstrapping Sentence Alignments
   2. Named Entity Recognition
   3. Zero shot sentence alignment
5. Conclusion
6. References

**Introduction**

Hinton[4] gave a hint towards learning distributed representations for symbolic input which was explored in Bengio[5] paper to create neural network based probabilistic language model, this distributed representation led to word2vec[6]. These developments led to usage of vectors as representations for textual data. Word2vec, phrase2vec, doc2vec were all the resultant tools. In simple form word2vec model gives vector representations of words depending on their context in data (on which model is trained). Now this model creates a sort of complex multidimensional space for all vectors (words) of particular language. These are said to be feature embeddings of words in specific language. Now if we have some sort of vector space in which there are vector representations of two different languages. Than that is said to cross bilingual embeddings. One thing to note is that these embedding are not found directly by say training word2vec model on data containing data from both languages. Bilingual embeddings mean that feature vectors of words for both languages will exist in a certain meaningful way. Examples for bilingual embeddings :

* Words like **run** and **दौड़ना** in bilingual embedding setting will have vector representations which will be very closed or these two vectors can be mapped to each other through some transformation function.
* Word like **book** can have vector in some place in between vector points of **किताब** and **दर्ज** in this bilingual settings.

**Methods**

**Neural Network**

Neural Network was used to find a transformation vector reach from Language L1 vector space to Language L2 vector space. The basic premise is get vector space for parallel corpus. Now find alignment between these two parallel text using Giza++ or fast align from CDec[7]. Take most aligned words (W1 from language L1 and W2 from language L2) across parallel corpus and get their word vectors (v1 for W1 and v2 for W2), which are in turned used as input and output to learn weights of multi layered neural network (with gradient descent). Than use these learned weights to predict vector v4 for an input vector v3 and get nearest word W4 to point v4. This approach had many sub approaches, (all on some assumptions?)

1. Consider full vocabulary of L1 as input and corresponding most aligned words of L2 as output and train neural network. This approach yielded poor results. May be complex transformations between languages used in experiment ( English and Hindi) weren’t captured by neural network.
2. Next experiment attempted to find just alignment between verbs of two languages (English and Hindi). And use the setting of neural network as mentioned earlier with vectors of verb of L1 as input and vectors of verbs of language L2 which are correspondingly aligned to them as output. This experiment also yielded poor results.
3. Previous experiments were done on languages which have different word orderings (English and Hindi). So we repeated experiment **b** with Punjabi and Hindi languages with assumption that transformation between aligned verbs of these two similar structured language might give us transformation among them. But this experiment also yielded poor results.

**Co-occurrence matrix**

As a very basic method we are computing co-occurrence matrices between languages. A co-occurrence matrix notes down how often pairs of words occur with each other in the same context. For this specific project a sentence is considered as the context. Each row of the matrix is word vector from source language to target language while each column is a word vector from target to source language. Further, each cell of the matrix represents a (normalized) count of how often a word from source language occurred with some word from the target language.

We start with a sentence aligned parallel corpus for English-Hindi language pair. We extract all the unique words(unique tokens, all converted to lowercase) for both English and Hindi. We then pass over each pair of aligned sentences <Ei,Hi> and for each word (ei) in a sentence in English we increment the count in the row for all the Hindi words (hj) which occur in the corresponding Hindi sentence. This approach should intuitively yield high values in cells which represent the true translation of a word from source to destination language.

**GloVe matrix**

In this approach we first build GloVe vectors for both source and target languages[1]. Now, for every word in in the source language, we add the GloVe vectors of the words in the target sentence together. We repeat this across every word in the source corpus, and hence get a semantic representation for every word in the source language, in terms of the words that co-occur with that word in the target language. This representation will be the summation of GloVe vectors of the words.

**Word Translation & Sentence Alignments**

After building suitable vector representations of the words in the semantic space, we attempt to translate the word to the target language. To find the translation, for the case of co-occurrence based vectors, we can simply check the target vectors for the words with the highest co-occurrence with the source word.

This is a very simplistic model, and results in noisy translations since other words also co-occur with the source word. Think about a case where one word translates to multiple words in the source language or the case where we have function words which co-occur just as frequently as the translated words themselves!! However, it was observed that including high frequency words often skewed the counts in their favour. To limit this undesirable effect we took the approach of normalizing each column of the matrix with the sum of all the values in that column. This is effectively normalizing the co-occurrence count for each word in target language by the total occurrences of that word.

Additionally, we adopt the pointwise mutual informationsimilarity metric for comparing vectors in the source and target languages. To compare similarity of words, we need a method to do build two matrices, one from source to target, telling us about which words in the target language co-occur with words in the target language. Another, to tell us which words in the target language co-occur with itself. Given these two matrices, we have a comprehensive description of how each word is related to another both across languages, and within each language.

The case of GloVe vectors is similar to the above mentioned, except that we no longer need to explicitly normalize the vectors by the weights, we simply convert them to unit vectors, and then apply the similarity metric.

Another approach we tried was to align sentences using the trained vectors. If “similar” words can be obtained using the trained vectors, our intuition tells us, that a simple summation of word vectors of words in both the source and target languages should give us similar vectors. Now, these vectors could be checked for similarity to obtain the sentences across languages, hence giving rise to a novel algorithm for sentence alignments.

**Sample outputs**

**Sentence alignments**

**Some success cases:**

* Some famous beaches of Goa are Dona Paula , Calangute , Anjuna , Arambol , Colva , Miramar , Vagator , Agonda etc .
* गोआ के कुछ प्रसिद्ध ' बीच ' दोला पाउला , कैलेंगुट , अंजुना , आरामबोल , कोलवा , मीरामार , वागाटोर , अगोंडा आदि हैं ।
* For honeymoon of the newly married also Goa is a good place .
* नवविवाहितों के हनीमून के लिए भी गोआ एक बढ़िया स्थान है ।
* On Baisakhi celebrated in 1999 AD , 300 years of the Khalsa Panth got completed .
* सन् 1999 में मनाये गये बैसाखी पर खालसा पंथ के 300 साल पूरे हुए ।
* Golf competitions of the national level are organized here every year .
* यहाँ हर साल राष्ट्रीय स्तर की गोल्फ प्रतियोगिताएँ आयोजित की जाती हैं ।

**Some failure cases:**

* There are several shops of Bal Mithai and Singori on the Mall Road itself .
* कटरा के बस स्टैंड का विस्तार किया गया है , बाज़ार और पार्क बनाए गए हैं जिसके परिणामस्वरूप यह क्षेत्र और समृद्ध दिखाई देने लगा है ।
* Almora Angora is famous for the clothes made of the wool of rabbit .
* लेह से 240 किलोमीटर की दूरी पर संरक्षित क्षेत्र के अंतर्गत , सुविस्तृत पर्वत श्रृंखलाओं और विविध पक्षियों के बसेरे के मध्य दुनिया के एक सिरे पर स्थित सो-मोरीरी झील का सौंदर्य अद्‍भुत है ।
* From the sea level the height of Binsar is 2412 kilometres .
* नौकुचियाताल समुद्रतल से 1 , 220 मीटर की ऊँचाई पर भीमताल से 4 किलोमीटर एवं नैनीताल से 26 किलोमीटर की ऊँचाई पर स्थित है ।

**Word translations:**

Below are some of the examples of the word translations of words that GloVe captures accurately, along with the similarity measures. These measures are taken such that if the correct prediction appears anywhere in the top 5, we consider it as a correct match.

* आलू: potato : 0.94245625315
* कुफ़री: kufri : 0.994272913746
* भूपेंद्र: bhupendra : 1.0
* चैल: chail : 0.976168885497
* होम: home : 0.907562282141
* विश्वविद्यालय: university : 0.968271346425
* चाडविक: chadwick : 1.0
* पुस्तकों: books : 0.962314263953
* Some interesting cases of similarity where the translations failed were :-
* कीं: height 0.906067325874
* नहा: defend : 0.933498210903
* रिसने: pandals : 0.809153225993
* कृशकाय: vertical : 0.825471334778

As can be easily observed, Some simple cases fail while relatively complex named entities are easily captured with an extremely high accuracy.

**4. Future work**

**4.1 Bootstrapping Sentence Alignments**

An extension to the work done is that after finding an “aligned” sentence pair from our testing we can add this sentence pair back into the our training set and retrain our sentence alignment model. This bootstrapping approach will have to be performed smartly and efficiently since retraining the model can be computationally expensive. Overall it can be a promising approach to perform semi supervised learning for sentence alignment.

**4.2 Named Entity Recognition**

Another interesting observation made while working on the word to word translation approach is that the translation accuracy seemed particularly higher for named entities in the corpus. We, in future, wish to explore techniques that can leverage this observation, supplemented by other grammatical observations about named entities, to create a novel approach to identify named entities from a given parallel corpus.

**4.3 Zero shot sentence alignment**

The most interesting extension that we propose is a method to align/translate words across two languages in which we do not have a corpus to directly train on. We call this a zero shot method for aligning sentences. The idea is that, we have a source language Si and a some other language Sj. We build the vectors for the source language in terms of some target space(With respect to some intermediate language Sk). Now we build the vectors for Sj in terms of Sk as well. Now, since both Si and Sj both exist in the same target space, we can simply find translations between words in S*i* and Sj, without having an explicit training corpus for those two languages!!!

**Conclusion**

As we have shown, relatively simple methods such as counting co-occurrence and GloVe vectors form extremely robust representations for words. In this paper, we have leveraged this robustness to demonstrate their usefulness in the context of sentence alignments. Additionally, we have further proved that these representations can easily be bootstrapped to continuously improve word representations over time, or identify named entities with a high degree of confidence and finally, we have also presented a novel method to perform zero-shot word alignments without explicitly having an aligned corpus between the source and target language.

**References**

[1] GloVe: Global Vectors for Word Representation -- Jeffrey Pennington, Richard Socher, Christopher D. Manning <http://nlp.stanford.edu/projects/glove/>

[2] Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification -- Aditya Mogadala, Achim Rettinger <https://www.aclweb.org/anthology/N/N16/N16-1083.pdf>

[3] Bilingual Word Embeddings for Phrase-Based Machine Translation -- Will Y. Zou† , Richard Socher, Daniel Cer, Christopher D. Manning <http://ai.stanford.edu/~wzou/emnlp2013_ZouSocherCerManning.pdf>

[4] Hinton, Geoffrey E. "Learning distributed representations of concepts." *Proceedings of the eighth annual conference of the cognitive science society*. Vol. 1. 1986.

<http://www.cogsci.ucsd.edu/~ajyu/Teaching/Cogs202_sp13/Readings/hinton86.pdf>

[5] Bengio, Yoshua, et al. "A neural probabilistic language model." *journal of machine learning research* 3.Feb (2003): 1137-1155.

<http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>

[6] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

<https://arxiv.org/abs/1301.3781>

[7] Dyer, Chris, et al. "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models." *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, 2010.

<http://cs.jhu.edu/~jonny/pub/P10-4002.pdf>