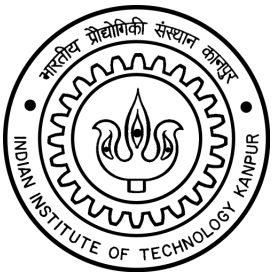
Cross Lingual Plagiarism Detection

Utsav Sinha, 12775 Md Enayat Ullah, 12407

14th November 2015

CS671: Natural language Processing
Guide: Amitabha Mukerjee



Contents

List of Figures

List of Tables

1 Motivation

Plagiarism, as Wikipedia defines it, is the wrongful appropriation and stealing and publication of another author's language, thoughts, ideas, or expressions and the representation of them as one's own original work. A lot of emphasis has been attributed to automatic detection of text reuse in the research community [Alz12]. But most of the work is focused on monolingual comparison (mostly english to english) and multilingual domain is yet largely unexplored.

For example, a hindi novel by Premchand adapted to english with little modification can be published as an original work. Identifying it is difficult even for humans as it involves comprehension of both the languages. This form of paraphrasing often involves some translation model which in itself is a growing research field in natural language processing.

In this project, we have attempted to build a multilingual word embedding space and use it for cross lingual plagiarism detection.

2 Introduction

Most of the work in cross lingual plagiarism detection relies on syntactic similarities between the language pair involved or on sentence aligned parallel corpora. See Related Work. But many languages neither have huge amounts of digital text, nor sentence aligned corpora. Moreover, aligning sentences requires translation dictionaries which itself are not entirely accurate. Therefore to overcome these shortcomings, we have used comparable corpus in our work. A comparable corpus is topic aligned instead of sentence aligned - the same article in two different languages. The two articles are generally not translations of each other.

Taking inspiration from [VM15], a unified word vector space for two languages is trained using word2vec from pseudo-bilingual documents. This step merges the two languages as if they were a single language. Therefore, monolingual plagiarism detection techniques can be applied to this unified model. To verify the robustness of the vectors learnt, they are tested on Bilingual Lexical Extraction (BLE) and Suggested Word Translation in Context (SWTC) tasks.

With extensive learning on large corpus, both the languages learn the context of each other. These vectors are then fed as input to a recursive auto-encoder (RAE) which generates phrase vectors for sentences. These phrase vectors are trained and tested on paraphrase labeled dataset.

3 Dataset

English German comparable corpus was available at [Too] which was built from Wikipedia articles. Comparable Hindi-English articles were not readily avail-

able, so they were extracted using Interwiki SQL dump which mapped Hindi and English Wikipedia articles. 41001 topic aligned articles were obtained which had 227MB of Hindi and 422MB of English data. This has been made available at [Wik]. Articles of comparable length (within a document length ratio of two) were only used for training word vectors. This was to avoid learning too much of monolingual context. So after pruning, only 9474 articles of Hindi-English were used.

For the BLE task, words in English were translated to Hindi and vice-verse by hand to obtain the gold truth against which our proposed model would be tested with. Similarly, test cases were created for English-German language pair. For the SWTC task, **50 English-Hindi sentences** and their correct translation in context has been created. All tagged data has been made available at [Sin].

Language Pair	POS	Number of test cases
EN-HI	Noun	135
EN-HI	Verb	100
EN-HI	Adjective	100
EN-HI	Others	65
EN-DE	All	100

Table 1: Labeled Dataset for BLE Task

For paraphrase detection, MSR Paraphrase Corpus [Res] is used which contains 5801 English to English labeled sentences. These sentences are translated into Hindi using Google Translate to get the data for plagiarism task.

4 Related Work

Traditional methods involve sentence aligned parallel corpus along with a translation thesaurus [BCRAL10] to translate one language to the other and compute their similarity. Further extension using machine translation has also been explored but generally leads to poorer results due to limited accuracy of translation. [PBCSR11] discusses CL-CNG (Cross Lingual Character N-Gram) which performs relatively better for syntactically similar languages despite its simplicity. [Alz12] summarizes existing methods which employ clustering techniques, grammatical constructs, fuzzy logic based approaches and stylometric features (used in author identification).

With increased computational power, focus has shifted towards learning dense word embeddings such as word2vec [MCCD13] and GloVe amongst others. [SLLS15] uses PMI matrix co-factorization to learn bilingual word vectors from a parallel corpora. [VM15] goes beyond sentence aligned parallel corpora and applies SGNS (Skip Gram with Negative Sampling) on comparable articles

(Wikipedia) to obtain a unified multilingual word representation space. Its results are outlined in table 2. [FD14] further discusses other methods for learning distributional representation of words.

[SHP+11] deals with monolingual plagiarism detection. It generates similarity matrix of two phrases after learning phrase representations using recursive autoencoders from a large corpus. This similarity matrix is later trained on a classifier based on labeled paraphrase corpus.

Languages	Accuracy	Model
ES-EN	70.1%	BWESG Length-Ratio
NL-EN	39.7%	BWESG Length-Ratio
IT-EN	69.2%	BWESG Merge and Shuffle

Table 2: Results from [VM15]

5 Approach

5.1 Implementation Overview

Pre-processing was done on comparable wiki corpus for sanitization and stemming. Then it was fed into word2vec model after applying the shuffling strategy. The word vectors learnt were fed as initialization parameter to the Recursive Autoencoder (RAE) and sentence representations were learnt as described in [SHP+11]. Then a classifier, namely SVM and logistic regression were applied to the labeled paraphrase corpus and this was later used for detecting cross lingual as well as monolingual paraphrase. Figure 1 presents the overall overview of the implementation.

5.2 Unified Multilingual Word Embeddings

Our approach is inspired from [VM15] where pseudo-bilingual documents were created from comparable articles. These were later fed into the word2vec model for training.

To get a pseudo-bilingual document, we merge two comparable articles together based on a shuffling strategy. Shuffling strategy can be:

Length Ratio
: Documents are merged based on their length ratio by preserving the ordering of words in both the documents. For example, if EN article has 400 words and HI article has 200 words, then after every two words of EN article, one word of HI article is inserted.

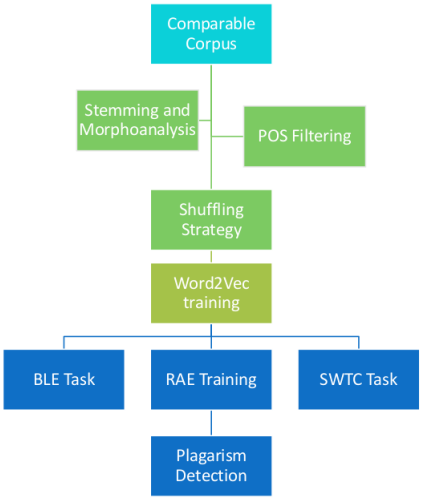


Figure 1: Implementation Overview

Random Shuffle

The documents are randomly shuffled together after merging without any regard to ordering of words in both the documents.

Order Preserved Random Merging

The documents with lower word count is inserted at random positions into the larger document and the word ordering of both the documents are maintained.

An example 2 would make it more clear.

English

gangubai hangal gangubai hangal was an indian singer of the "khyal" genre of hindustani classical music, who was known for her deep and powerful voice.

Hindi

गंगूबाई हंगल गंगूबाई हंगल (कन्नड़: "ಗಂಗೂಬಾಯಿ ಹಾನಗಲ್") हिंदुस्तानी शास्त्रीय संगीत की प्रख्यात गायिका थीं उन्होने स्वतंत्र भारत में खयाल गायिकी की पहचान बनाने में महती भूमिका निभाई |

गंगूबाई gangubai हंगल hangal गंगूबाई gangubai हंगल (कन्नड़: hangal "ಗಂಗೂಬಾಯಿ march ಹಾನಗಲ್") was an हिंदुस्तानी शास्त्रीय संगीत की indian प्रख्यात singer गायिका of थीं the उन्होने "khyal" genre स्वतंत्र भारत में खयाल of गायिकी की पहचान hindustani बनाने में महती classical भूमिका music, who निभाई .

Figure 2: Pseudo Bilingual Document

The intuition behind this merging of documents is that articles about the same topic would contain words from both the languages which can co-occur together

semantically. Moreover, by putting them next to each other, even though they are not translations, we are learning the appropriate context in our own language as well as in the other language. So to make this realizable, the context window is increased to 30 and 48 in word2vec training to capture the cross-lingual context. There is also a threshold on the ratio of the sizes (word count) of the two articles which are merged: this is to ensure that a small article in one language is not merged with its large comparable article in other language. This has been done to prevent bias towards monolingual learning of word embeddings.

[VM15] considers only nouns for learning the unified word space while we have not restricted ourselves to such filtering. We have explored lemmatization of entire documents using NLTK library [BKL09] for English and Shallow Parser [Hyd] for Hindi.

5.3 Learning Phrase Representation

Feature Extraction using RAE

A recursive autoencoder(RAE) is a deep-learning framework wherein simple autoencoders are recursively applied to get a series of low-dimensional representations of the data. Figure3 shows an instance of RAE applied to a parse tree, which produces multi-word(phrase) vectors of the sentence. An unsupervised training is performed on the RAE using all the sentences of English and Hindi Wikipedia. The non-convex objective function is solved by an L-BFGS solver with mini-batch training.

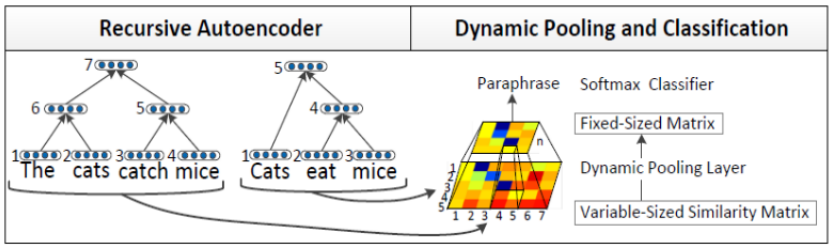


Figure 3: Recursive Autoencoder with Dynamic Pooling [SHP+11]

Dynamic Pooling

A similarity matrix is then constructed using the features vectors of two sentences. However, the size of the similarity matrix is dependent of the lengths of two sentences, and cannot be used as input to a classifier. Dynamic pooling is a technique which maps a variable size matrix to a fixed size representation. Non-overlapping windows slide over the similarity matrix and a min pooling is performed in each of the windows to get a fixed size pooled matrix.

Currently a single layer RAE is used due to lack of computational power, but it can be extended to a deep RAE having multiple encoding layers at each node

in the tree, and may capture finer abstractions within the data.
The size of the pooled matrix is taken as 15×15, which is the average length of the sentence in the MSR paraphrase corpus.

Why min-pooling?
Other statistics, like averaging or max pooling, are also possible. But average pooling will obscure the presence of similar phrases, and max pooling will overshoot in case of word-level matches.
The pooled matrix is then normalized to have zero mean and unit variance.

Classification
The normalized pooled matrix serve as the feature vectors representing a pair of sentences, and standard supervised classification techniques like soft-max and SVMs are employed, with the hyperparameters being further tuned.

Hand-engineered features
A set of heuristic features on numbers based on domain knowledge [SHP+11]:
f1 – 1, if the two sentences have same set of numbers.
f2 – 1, if the two sentences have a common number.
f3 – 1, if number set of one sentence is a subset of the number set of other.
f4 – Difference between sentence length.
f5 – Fraction of words of one sentence present in other.
The implementation code which was available at [Soc] was tuned for our experimental settings.

6 Results

The tasks which are considered for testing the unified language models are Bilingual Lexicon Extraction (BLE) and Suggested Word Translation in Context (SWTC) as outlined in [VM15].

The BLE task finds out the ten nearest neighbours of a test word and sees if its translation is amongst them. If the languages are successfully able to learn each others' context, then the nearest neighbours will contain some of the possible translations. Various parameters like the word vector dimensions, context size and shuffling strategies were tuned and the outcomes are in figure 4. The best performing model is when the number of dimensions in 200 and context window is of length 48 with order preserved random merge strategy. This model's accuracies are only further used for reporting the performances of SWTC and paraphrase detection task.

The accuracy of cross lingual nearest neighbours in case of English-German was 22%. For Hindi-English, the performance for BLE task of the best performing bilingual model is in table 5. Some examples of successfully found neighbours and some failure cases are in table 3

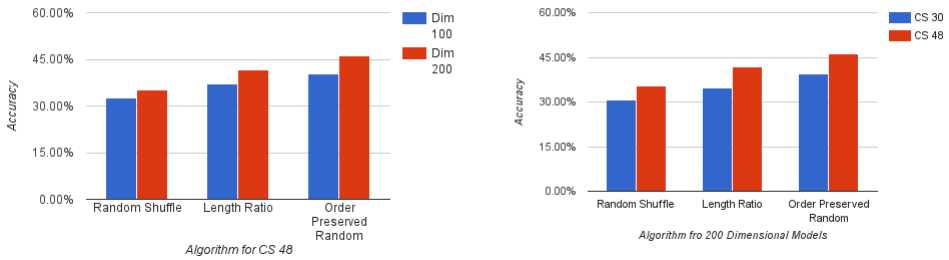


Figure 4: Performance of word2vec with tuned parameters

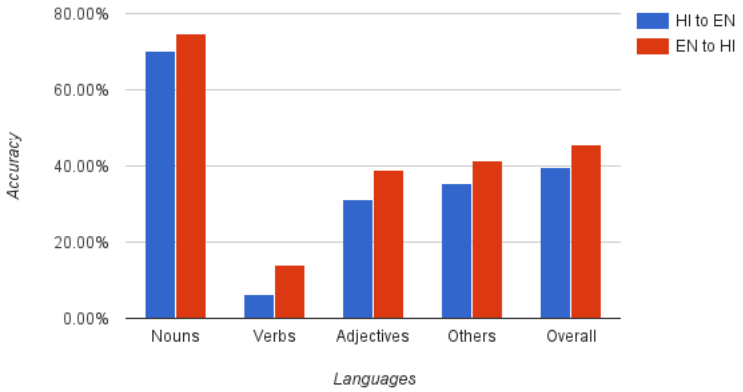


Figure 5: BLE Task performance of different parts of speech

Further to investigate the effect of morphological constructs in Hindi to English, stemming was performed and its results are in table 4

In SWTC task, the best translation in context is identified. To find the context vector, the interpolation of target word's vector and sum of all word vectors in the sentence is taken. This vector's similarity measure is found with all possible translations of the target word. If the one with highest similarity is the correct translation in context as per the gold truth, then it is counted as positive detection. This task is somewhat similar to the Word Sense Disambiguation (WSD) but in cross lingual domain. The accuracies of this task was 30%[1]. Some examples of the testing is in table 5

[1]The SWTC test corpus was increased in size, so the accuracy decreased from that reported while poster presentation

Word	Nearest Neighbours in Sorted Order
knowledge	श्रीमद्भगवद्गीता doer vedas ignorance ज्ञान
father	पिता mother wife माता child
बाज़ार	price शेयर बाजार market markets
बेहतर	better अच्छी अगर improve निर्धारित
sudden	cardiac पूर्णहृदरोध defibrillation ऊष्माघात अतिताप
in	a the में to की
भेजना	भेजने ईमेल प्रेषण email bes
पीना	रिसेप्टरों मूत्रवर्धक सेवन drinks पेय
run	spielen away travel fahrt athleten
vater	**father** bruder wife son eltern
send	भेजने protocol BDR SSH NSSA

Table 3: BLE Task Examples

Languages	BLE Task	SWTC Task
EN to HI	32	24
HI to EN	28	-

Table 4: Accuracy after Stemming

Sentence	Possible Translations	Gold Truth	Model's Prediction
He is engaged to that foreign actress.	सगाई व्यस्त बंधना संलग्न लगना	सगाई	सगाई
Why are you engaging me in useless conversation?	सगाई व्यस्त बंधना संलग्न लगना	संलग्न लगना	लगना
The match was well played.	मैच दियासलाई विवाह जोड़ा मुक़ाबला होड़ मेल खेल प्रतियोगिता माचिस	मैच मुक़ाबला खेल	मेल
The couple looks like a perfect match.	मैच दियासलाई विवाह जोड़ा मुक़ाबला होड़ मेल खेल प्रतियोगिता माचिस	विवाह जोड़ा मेल	मुक़ाबला

Table 5: SWTC Task Examples

The best performing model was fed as initialization to the the RAE and the performance of the SVM and Softmax classifiers on the paraphrase detection task is in table 6. The results with and without using features have also been reported.

Languages	Softmax	Soft-max+features	Linear SVM+features	RBF SVM+features
EN to EN	66.21	68.14	68.81	70.15
EN to HI	65.53	66.55	63.23	66.55
HI to EN	64.64	65.98	63.86	67.21
HI to HI	60.78	60.34	62.45	64.67

Table 6: Accuracy of Paraphrase Detection

Algorithm	Reference	Accuracy
EN to HI	This model	66.55
Explicit Semantic Space	Hassan (2011)	67
Explicit Semantic Space	Hassan (2011)	67
EN to EN	This model	70.15
JCN WordNet Similarity with Matrix	Fernando and Stevenson (2008)	74.1
RAE with dynamic pooling	Socher et al. (2011)	76.1
Matrix Factorization and Supervised Reweighing	Ji and Eisenstein (2013)	80.4

Table 7: Accuracy of Paraphrase Detection on MSR Corpus

Some sentences successfully classified by our project and their ground truth along with other cases which were mis-classified are in table 8 and table 9.

Statement	Paraphrase	Gold Truth
They had published an advertisement on the Internet on June 10, offering the cargo for sale	वे बिक्री के लिए माल की पेशकश, 10 जून को इंटरनेट पर एक विज्ञापन प्रकाशित किया था	YES
The initial report was made to New York Police department.	आरोप दिसंबर को किए गए कुछ पुलिस रिपोर्ट की वजह से उपजी	NO
बेटों एंथनी और केली, बेटियों लिंडा आशा और नोरा सोमर्स - और चार पोते वह अपने चार बच्चों के रूप में करते उसे जीवित रहते हैं।	Hope is survived by his wife; sons Anthony and Kelly; daughters Linda and Nora Somers; and four grandchildren.	YES
In response to sluggish sales cisco pared spending.	सिस्को सुस्त बिक्री के लिए क्षतिपूर्ति की तिमाही के दौरान खर्च मुकाबले।	NO

Table 8: Paraphrases detected successfully

Statement	Paraphrase	Gold Truth	Our Verdict
”Americans don’t cut and run, we have to see this misadventure through,” she said.	She also pledged to bring peace to Iraq: ”Americans don’t cut and run, we have to see this misadventure through.”	NO	YES
नेटवर्क भी यह शुक्रवार की रात "डेटलाइन" संस्करण है गिरती है।	नेटवर्क "डेटलाइन," अपने समाचार पत्रिका मताधिकार का एक संस्करण छोड़ देंगे।	YES	NO
टेक्सास इंस्ट्रूमेंट्स कल $19.25 के लिए $1.37 पर चढ़ गए और Novellus सिस्टम्स इंक $36.31 के लिए $1.76 उन्नत।	Texas Instruments climbed $US1.37 to $US19.25 and Novellus Systems advanced $1.76 to $US36.31, each having been raised to ”overweight” by Lehman.	NO	YES
He and John believed that only a new board would have had the credibility to restore el paso to health.	वह और जॉन केवल एक नए बोर्ड स्वास्थ्य के लिए एल पासो बहाल करने की साख पड़ता था कि माना जाता है।।	YES	NO

Table 9: Paraphrases detection failures

7 Conclusions

- Order preserved random strategy outperforms deterministic length ratio and purely random shuffling strategies. So preserving the ordering of the original text leads to better performances as only the random shuffling strategy garbled the the articles randomly.
- Increasing the context window size in word2vec training lead to improved performance on bilingual comparable corpora
- Reason for lower accuracy for our model than 2 is because Hindi and English are syntactically different whereas Italian, Spanish are closer to English. Moreover, we trained our model for all Parts of Speech rather than just nouns as done in that paper. For nouns, our accuracy is also 72%.
- Adding hand-engineered features in the classifier gives better results. But still, many sentences which have most words as common but are not paraphrases are detected to be so by our model 9.
- The procedure is language independent and doesn’t require any aligned corpus or translation
- SVM with RBF kernel outperforms the others classification methods

- Performance of EN to HI and HI to EN was also at par with EN to EN in paraphrase detection task. Our results were also not very far behind from other works done for EN to EN on MSR Paraphrase Corpus. 7. So the technique of learning a unified word embedding space from comparable corpus is not only learning cross lingual context, but it is also appropriate for learning monolingual word vectors.

7.1 Analysis of Stemming

Due to lack of faster execution of shallow parser, only around 100 articles were used for building the word2vec model while stemming. Therefore, the results are also relatively poor.

Qualitatively, we saw that neighbours in case of adjectives improved for HI to EN. This was due to the fact that Hindi is a gendered language, therefore without stemming all different forms like छोटा छोटे छोटी were separately coming among the top nearest neighbours of *small*, but after stemming, only the root form छोटा was present.

But due to this lemmantization, many nouns and verbs were combined in Hindi like दौड़ना to दौड़, सोचना to सोच which led to merging of their corresponding nearest neighbours too. Therefore, many previously correct top neighbours slipped out of the list. For example, for the verb *run*, the neighbour became *walk,* रन*, play, race,* खेल all of which signify *race* instead of the verb *run*. Due to this, the overall performance with stemming also remained almost the same as before.

8 Future Work

- Since the sizes of the articles in EN-HI comparable corpus had huge disparity with English articles being five to ten times longer in many cases, so only 9474 articles with comparable sizes for both the languages were used for training. So expanding the corpus can lead to better learning.
- As in [VM15], POS specific learning of word vectors can be done.
- This method can be further used for learning multilingual word embeddings for many Indian Languages like Hindi and Bengali, Tamil and Telugu etc which are structurally more similar than Hindi and English.
- Deep recursive autoencoders can be used instead for better phrase vectors representations as done in the paper [SHP+11].
- More intuitive features can be engineered for the paraphrase classifier.

References

[Alz12] Naomie & Abraham Ajith Alzahrani, Salha M & Salim. Understanding plagiarism linguistic patterns, textual features, and detection methods. *Systems, Man and Cybernetics Part C:Applications and Reviews, IEEE Transactions on*, 2012.

[BCRAL10] Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. Plagiarism detection across distant language pairs. 2010.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[FD14] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics, 2014.

[Hyd] IIIT Hyderabad. Shallow parser for indian languages.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[PBCSR11] Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-language plagiarism detection. *Language Resources and Evaluation*, 2011.

[Res] Microsoft Research. Msr paraphrase corpus.

[SHP+11] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. 2011.

[Sin] Utsav Sinha. Translations with and without context for english-hindi, https://github.com/utsavsinha/cross-lingual-paraphrase/.

[SLLS15] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. *Volume 2: Short Papers*, page 567, 2015.

[Soc] Richard Socher. Rae for paraphrase detection, http://www.socher.org/index.php/main/dynamicpoolingandunfoldingrecursiveautoencodersforpa

[Too] Lingua Tools. Lingua tools, www.linguatools.org/tools/corpora/wikipedia-comparable-corpora/.

[VM15] Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data. *arXiv preprint arXiv:1509.07308*, 2015.

[Wik] Wikipedia. Hindi english comparable corpus, https://www.drive.google.com/file/d/0byhk0jhidfugrjnwehbcde5yce0/view?usp=sharing.