Cross Lingual Plagiarism Detection

Utsav Sinha, 12775 Md Enayat Ullah, 12407
Guided by Professor Amitabha Mukerjee
CSE Department, IIT Kanpur

Motivation

Plagiarism ⇒ wrongful appropriation, stealing and publication of another author's language, thoughts and ideas, as one's own original work [Wiki]

Monolingual plagiarism, especially English with English has received a lot of effort in the research community [Alz12], but multilingual domain is yet largely unexplored.

Hindi Novel → (Translated / Modification) → English Work → (Difficult to Detect)

This paraphrasing often involves some translation model, an already growing NLP task.

Introduction

We learn word embeddings in unified multilingual distribution space from freely available comparable wiki articles.

These word vectors are then trained on translations of MSR Paraphrase Corpus which is used for cross lingual plagiarism detection.

Data Gathering

Interwiki SQL dump → Hindi ID to English title map → Search in English corpus

41001 Hindi-English comparable articles extracted

2GB of DE-EN comparable data obtained from [LIN]

100 most used words in different POS in English, German and Hindi and their translations were created by hand for BLE and STWC Task

Used Google Translate on MSR Paraphrase Corpus to get 4000 HI-EN training and 1000 test sentences

Related Work

[BCRA10] uses sentence aligned parallel corpus , [PBCSR11] employs grammatical and syntactic structures and other approaches using machine translation and stlyometric techniques explored

Multilingual word vector learning using PMI matrix co-factorization [SLLS15] on parallel data and using word2vec on comparable data [VM15]Their results:

Language Pair	Accuracy
ES-EN	70.1%
NL-EN	39.7%

Plagiarism Detection Results:

Method	Accuracy
RAE + Dynamic pooling	76.8%
Matrix factorization with supervised reweighting (State of the art)	80.4%

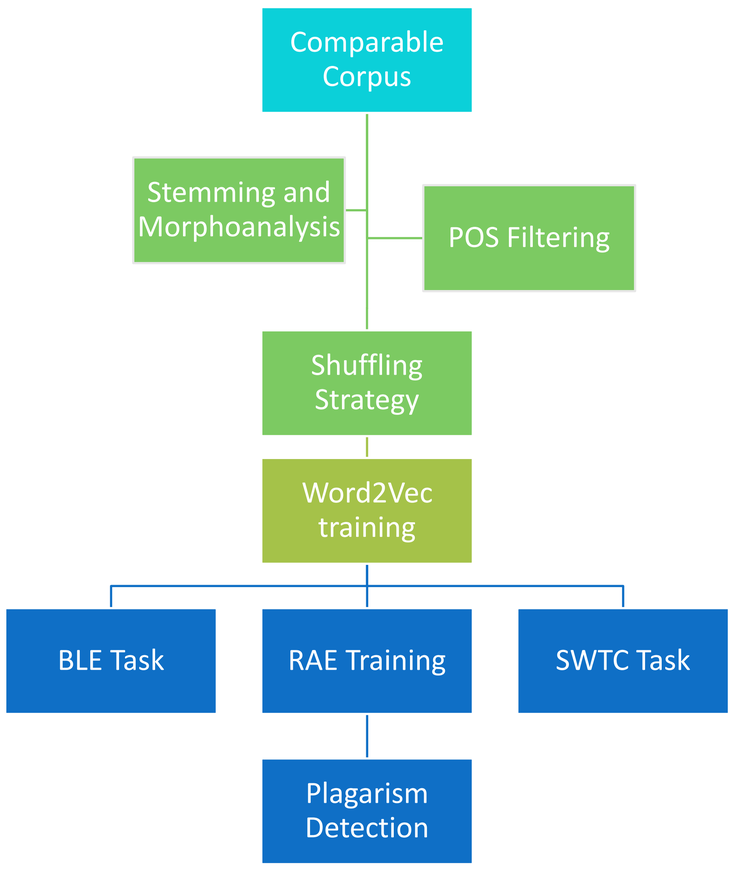
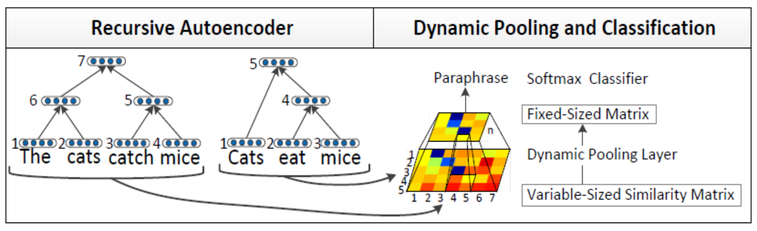


Figure 1. Implementation Flow Chart

Implementation

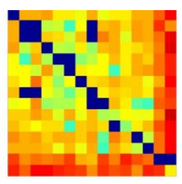
Taking inspiration from [VM15], we generate pseudo-bilingual documents using deterministic and random shuffling strategies.

Hindi and English sentences are then used to train Recursive Auto Encoder(RAE) , which outputs phrase vectors for a given sentence. [SHP+11]



Dynamic pooling (non-overlapping min pooling) on the similarity matrices constructs fixed size representations. A supervised learning is performed on the fixed size representations using Logistic Regression as well as SVM, whose parameters are tuned using grid search
Another set of features is added to the classifier:
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

Positive:
S1: They had published an advertisement on the Internet on June 10, offering the cargo for sale
S2: वे बिक्री के लिए माल की पेशकश , 10 जून को इंटरनेट पर एक विज्ञापन प्रकाशित किया था



Negative:
S1:The initial report was made to New York Police department.
S2:आरोप दिसंबर को किए गए कुछ पुलिस रिपोर्ट की वजह से उपजी

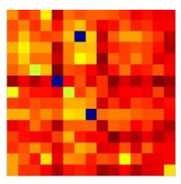


Figure 2. Sentence pairs along with their pooled similarity matrix

Results

Cross-lingual nearest neighbour task(BLE) performs better when tested on non-nouns
Accuracies:

Language Pair	BLE	SWTC
EN-HI	28%	39%
EN-DE	22%	--

Language Pair	RAE+ Softmax	RAE + features+ softmax	RAE+ features+ SVM (Linear)	RAE+ features+ SVM (RBF)
EN-EN	66.21	68.14	68.81	70.15
HI-EN	65.53	66.55	63.23	66.55
HI-HI	60.78	60.34	62.45	64.67

Table 1. Accuracies in Paraphrase Task

माँ	she	Vater	run
माता	he	**father**	walk
mother	**वह**	eltern	play
husband	it	bruder	race
सौतेली	her	wife	रन

Table 2. Top four neighbours (BLE Task)

English	Hindi
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.	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 3. Pseudo-bilingual document

Conclusion

- Length ratio random strategy outperforms deterministic length ratio and purely random shuffling strategies
- Increasing the context window size in word2vec training ⇒ improved performance on bilingual comparable corpora
- Adding hand-engineered features in the classifier gives better results
- The procedure is language independent and doesn't require any aligned corpus or translation
- SVM with RBF kernel outperforms the others

Future Improvements

- Deep RAE can be used for better phrase vectors
- Increased labelled paraphrase for classifier
- Extension for other Indian languages
- More intuitive features for classifier
- POS specific word2vec training for BLE and SWTC
- Max/Aggregate pooling on overlapping regions can be performed in the dynamic pooling layer

References

[VM15] Ivan Vulic and Marie-Francine Moens: Bilingual distributed word representations from document-aligned comparable data, 2015
[Alz12] Naomie & Abraham A. Alzahrani, Salha M & Salim: Understanding plagiarism linguistic patterns, textual features and detection methods, IEEE Transactions 2012
[BCRAL10] Alberto Barron-Cedeno, Paolo Rosso, Eneko Agirre and Gorka Labaka: Plagiarism detection across distant language pairs, 2010
[PBCSR11] Martin Potthast Alberto Barron-Cedeno, Benno Stein & Paolo Rosso: Cross-language plagiarism detection, Language Resources and Evaluation, 2011
[SLLS15] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun: Learning cross-lingual word embeddings via matrix co-factorization, 2015
[SHP+11] Richard Socher, E HHuang, J Pennington, Andrew Y. Ng, C D Manning: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, 2011
[LIN] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/