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 [AI12], but multilingual domain is yet largely unexplored.

Hindi Novel Translated 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

[B CRA10] uses sentence aligned parallel corpus , [PBCSR11] employs grammatical and syntactic structures and other approaches using machine translation and stylometric techniques explored

Multilingual word vector learning using PMI matrix co-factorization [SLS15] 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%

Positive:
S1: They had published an advertisement on the Internet on June 10, offering the cargo for sale.
S2: वे विज्ञान के लिए मल की रचना, 10 जून की इंटरनेट पर एक ज्वाइन्पुर फ्लाक लिखी थीया

Negative:
S1: The initial report was made to New York Police department.
S2:अप्रैल दिवस को की एंग्रु कु ल प्रतिबन्ध निर्देश की जब दे उखाणी

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

Figure 1. Implementation Flow Chart

Figure 2. Sentence pairs along with their pooled similarity matrix

Table 1. Accuracies in Paraphrase Task

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>BLE</th>
<th>SWTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-EN</td>
<td>66.55</td>
<td>66.55</td>
</tr>
<tr>
<td>HI-EN</td>
<td>60.78</td>
<td>60.34</td>
</tr>
<tr>
<td>HI-HI</td>
<td>62.45</td>
<td>64.67</td>
</tr>
</tbody>
</table>

Table 2. Top four neighbours (BLE Task)

<table>
<thead>
<tr>
<th>neighbour</th>
<th>Hindi</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>गांवी</td>
<td>he</td>
<td>father</td>
</tr>
<tr>
<td>पत्र</td>
<td>he</td>
<td>walk</td>
</tr>
<tr>
<td>परी</td>
<td>it</td>
<td>bruder</td>
</tr>
<tr>
<td>बहन</td>
<td>her</td>
<td>wife</td>
</tr>
</tbody>
</table>

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

Cross-lingual Plagiarism Detection

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

References
[VM15] Ivan Vulic and Marie-Francine Moens: Bilingual distributed word representations from document-aligned comparable data, 2015