Cross Lingual Plagiarism Detection

Guide: Prof. Amitabha Mukerjee
October 16th, 2015

Group 7
Utsav Sinha
Enayat Ullah

Plagiarism Example

Once upon a time, there was a race between a hare and a tortoise. What transpired is that the hare slept midway in the race and the tortoise eventually won the race. The moral of the story is: slow and steady wins the race.

एक घमंडी और तेज लड़के ने एक बार एक मेहनती लड़के से दौड़ जीतने कि शर्त लगाई । घमंडी बीच मे ही सो गया और अपनी मेहनत से दूसरा लड़का धीरे दौड़ने के बावजूद विजय हुया|

Related Work

- Sentence aligned corpus [BCRAL10]

  Ishaan is playing on the mobile *ईशान मोबाइल पर खेल रहा है*

  Problem: parallel corpus, aligning text

- Machine Translation

  Bottleneck: translation accuracy, loss of context in literary works

- Stylometric Construct - Author identification
- Grammatical and Syntactic features [PBCSR11]

  Problem: syntactically far languages - English and Hindi

Bilingual word embeddings

PMI Matrix Co-factorization from parallel Corpus[SLLS15]

Cross lingual document classification task

en ⟶ de 92.7% en ⟶ de 80.2%

Document-Aligned Comparable Data [VM15]

Bilingual Lexicon Extraction Task

es-en 70.1% nl-en 39.7%

Approach

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 .

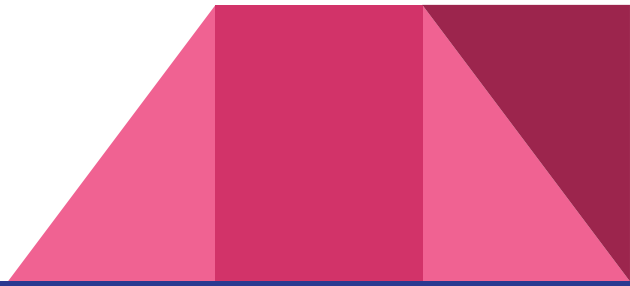
Approach

Shuffling and merging documents related to same topic

Train word vector model on these assuming it to be monolingual - word2vec used with wider context window

Heuristic: the domain of articles would be same, cross lingual context would be captured

Advantage: No sentence aligned corpus required !



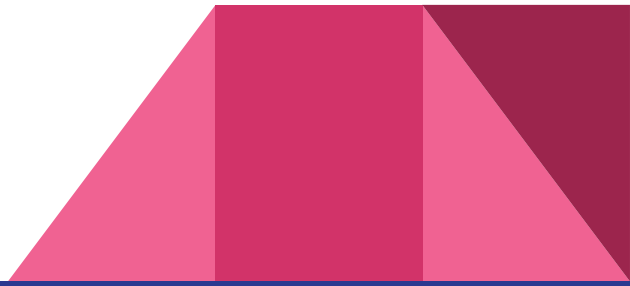
Approach

But what shuffling strategy to use ?

- Random
- Length ratio maintenance (deterministic)
- Length ratio random

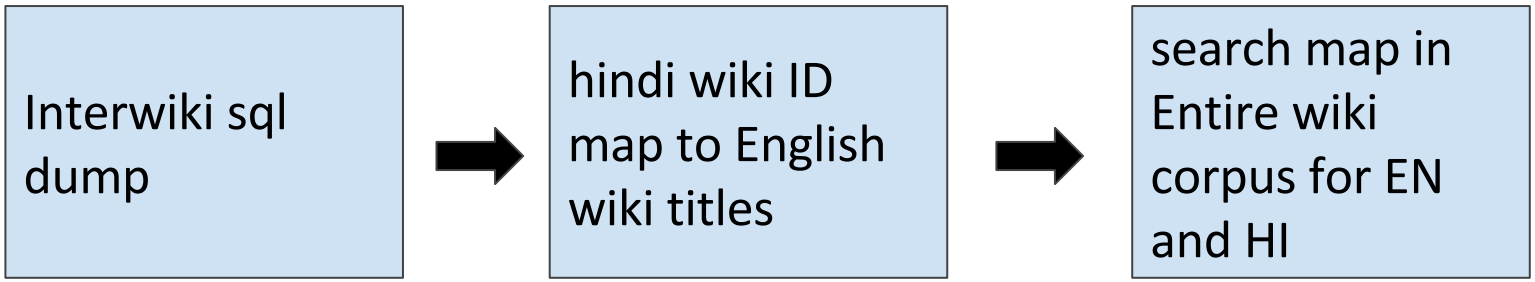
Preprocessing:

Lemmatization, remove punctuations, shuffle documents and merge



Data Collection

It consumed 90% of our put in efforts !



Result: 30,000 Hindi-English aligned articles extracted and still counting !

Meanwhile, experiment continued for available English German Dataset

Experiments

Task I: Bilingual Lexicon Extraction (BLE) The top 5 nearest neighbours

माँ	**she**	**Vater**	**run**
सौतेली	cameron	Father	spielen
बहन	winslet	Bruder	away
husband	नाइटली	Elder	travel
mother	पोर्टमैन	Wife	fahrt
भीष्म	she's	Eltern	athleten

Experiments

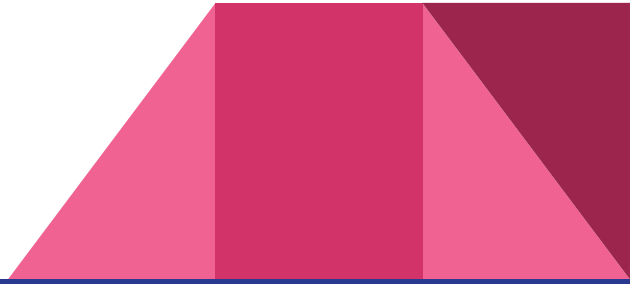
Task II: Suggested Word Translation in Context (SWTE)

play	खेल	नाटक	क्रीड़ा	अभिनय	जुआ	बजाना	आसान	सरल

Sentence	Context translations and similarity score		
That was a child's play:	खेल 0.47	नाटक 0.34	**आसान** 0.29
They play football:	**खेल** 0.43	नाटक 0.27	**क्रीड़ा** 0.19
I am a play character in Hamlet:			
	खेल 0.37	**नाटक** 0.32	आसान 0.26

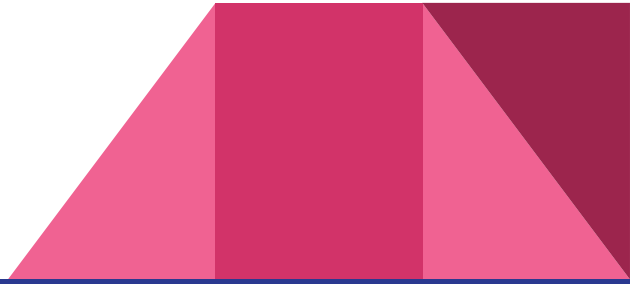
Observations

- Words like **माँ** correctly get mother as its top 5 nearest neighbours
- Moreover context sensitive words like बहन, husband were also found
- limited training done - nearest neighbours of **she** turned out to be names of actresses in the wiki articles !
- The SWTE is not working well - limited training size



Future Work

- Perform preprocessing before sending input to word2vec
- changing context window and vector dimension of word2vec
- Use the big data corpus extracted to train models
- Employ new shuffling heuristics
- Can we use some translation dictionaries while creating pseudo-documents ?
- Adopt the existing bilingual word embeddings for plagiarism task - the ultimate goal of the project !



References

[BCRAL10]	Alberto Barron-Cedeno et. al: *Plagiarism detection across distant language pairs*, 2010
[PBCSR11]	Martin Potthast et al: *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
[VM15]	Ivan Vulic and Marie-Francine Moens. *Bilingual distributed word representations from document-aligned comparable data*, 2015