# Cross Lingual Plagiarism Detection
# Guide: Prof. Amitabha Mukerjee

Enayat Ullah 12407
Utsav Sinha 12775

CS671 Project Proposal
October 4, 2015

## 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 (that is 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. This is difficult to achieve 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. We propose to discover multilingual word embeddings and use them for cross lingual plagiarism detection, thereby automating the entire process in an unsupervised way.

## 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 cofactorization 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. [FD14] further discusses other methods for learning distributional representation of words.

## Proposed Approach

Taking inspiration from [VM15], we plan to extend it for Hindi-English language pair. This will involve joint learning of word vectors in unified multilingual distributional space using word2vec. New shuffling heuristics will be explored to create better pseudo-bilingual documents. This will help to capture semantic context more effectively.

2 subtasks [VM15] will be employed on the learnt vector space:
Bilingual Lexicon Extraction (BLE)
Suggested Word Translation in Context (SWTC)

BLE lists out the semantically nearest neighbouring words for a given word while SWTC lists the most appropriate sense of the used word in the given context. Both these tasks helps to estimate the robustness of the multilingual word space.

These accuracies would be improved before applying it on the plagiarism detection task. Since it is a unified space, standard monolingual detection techniques can be used here. [SHP$^+$11] trains an unsupervised model using dynamic pooling and autoencoders to generate sentence vector representations which can be fed as input to the paraphrase detection module.

## Dataset

We plan to use document aligned Wikipedia articles of Hindi and English.

## 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.

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

[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.

[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.

[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.

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