Using Phonemes for better Cross-lingual unsupervised alignment using unsupervised methods

In my previous work, I used PLSA and ADIOS for aligning sentences from two languages, i.e., Hindi and English.

For this, I built a corpus manually in Hind and English containing roughly around 35000 and 50000 tokens respectively.

The results were quite good as far as appearance of similar tokens were concerned. I got 4-5 clusters using PLSA, each containing similar tokens in both languages but the problem was of alignment. They were not properly aligned.

One approach which I have thought is of using phonemes for improving the alignment and I am positive that this will surely increase the alignment and most importantly everything is unsupervised.

References

- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. *In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACLHLT '12, pages 477–487, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics

- Heng Ji. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, UMSLLS '09, pages 27–35, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics

- Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15* (MultiNER '03), Vol. 15. Association for Computational Linguistics, Stroudsburg, PA, USA, 57-64.