Gold standard for English/Hindi

Version: 1.0, Date: 15/04/11

NOTE: Check for a newer version of the data at
www.ims.uni-stuttgart.de/~sajjad/resources.html

 We request a citation of the following paper if the data is used in a publication.

Sajjad, Hassan; Fraser, Alexander; Schmid, Helmut (2011). An Algorithm for Unsupervised Transliteration Mining with an Application to Word Alignment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11).

The word pairs in the gold standard are extracted from a word aligned parallel corpus of English/Hindi made available for the shared task on word alignment, organized as part of the ACL 2005 Workshop on Building and Using Parallel Texts [1]. The data is available at www.cse.unt.edu/~rada/wpt05/

The English/Hindi gold standard is in a single line format, where each line contains a word pair and its tag (indicating whether a word is a transliteration). The words and tags are separated by a tab like the following:

english hindi tag

There are three kinds of tags.

1. All transliteration pairs have tag ”ti”.

The non-transliterations are divided in two categories.

2. There are a few word pairs which are close transliterations. They differ by one or two characters to consider as a transliteration pair. We tag them as ”tm”. Example: OFFICERS अफ़सर/officer

3. All non-transliterations other then ”tm” are tagged ”ma”.

Note: In the paper, we have not analyzed the results using different non-transliteration categories. We have merged them under non-transliteration pairs. However we hope that they are useful for the analysis of future transliteration mining methods.

Reference:

[1] Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In ParaText '05: Proceedings of the ACL Workshop on Building and Using Parallel Texts, Morristown, NJ, USA. Association for Computational Linguistics.