RC25527 (WAT1503-097) March 31, 2015
Other

IBM Research Report

Labeled Multi-Lingual Data for Cognate Detection

Jiri Navratil
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
USA

Nanyun Peng
Johns Hopkins University
3400 North Charles Street
Baltimore, MD
USA

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

Labeled Multi-Lingual Data for Cognate Detection

Jiri Navratil (jiri@us.ibm.com) IBM Watson, 1101 Kitchawan Rd., Yorktown Heights, NY
Nanyun Peng (npeng1@jhu.edu), Johns Hopkins University, 3400 North Charles Street, Baltimore, MD

The attachments to this document contain files with data sets that may be used to develop, evaluate, and compare methods for automatic detection of cognates. Two language pairs are included: English-French (filenames with "FR-EN"), and English-Russian (filenames with "RU-EN").

The data sources have been reviewed by a human annotator (in our case, a person native in the respective foreign language and proficient in English) for accuracy. The data format follows that in [1], in which word pairs (from the two languages) are labeled as positive (i.e., cognate) or negative (not cognate).

To speed up the annotation process, candidate lists were created for the annotators wherein one word in a language is accompanied by five meaningful candidates in the other language, some of which might be cognate to the given word. Annotators are asked to label which of the five candidates are cognates to the given, if any. The candidates are generated by running a bi-text machine translation system and extracting five most frequent translations for a given word. This was done for two language pairs: English-French, and English-Russian. For each pair, the top 10,000 most frequent words were considered on the non-English side with corresponding English words as candidates.

The review process took place in two rounds to improve correctness.

For each language pair, one main file (*.data) and one additional TAB-separated file (*.tsv) file is provided, the latter defining our development subset. In our experiments, this subset served the purpose of training and tuning various parameters in our automatic cognate detection system (described elsewhere). The format of the tsv files is as follows: 1st column has index referring to the corresponding line in the data (*.data) file from which the entry was taken; 2nd column contains tokenized word in the foreign language; 3rd column contains tokenized word in English. Note that the TAB character is used to separate the columns.

Files accompanying this document:

```
cognate_FR-EN.data
cognate_FR-EN.dev.tsv
cognate_RU-EN.data
cognate_RU-EN.dev.tsv
```

References:

[1] Shane Bergsma, Grzegorz Kondrak, "Alignment-Based Discriminative String Similarity," In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (June 2007), pp. 656-663