NLPCC-ICCPOL: 2016 Shared Task Guideline:

Chinese Lexical Similarity Computation

1. Task

Lexical similarity computation is a fundamental task for natural language processing. This task provides a dataset of Chinese word similarity, including 500 word pairs with their similarity scores. We try to provide a benchmark dataset to evaluate and compare different lexical similarity methods. All kinds of strategies are welcome, including the traditional corpus-based distributional similarity, dictionary-based similarity computation, as well as the recently developed word embedding methods and deep learning strategies. Also, the participating system is encouraged to use any additional resources.

2. Data

In this task, we will only provide a test data. Some examples in test data are given below:

word 1	word 2	similarity
没戏	没辙	4.9
只管	尽管	4
GDP	生产力	6.5

When selecting words, we take into consideration the following factors.

- **Domain**. We select words mainly from news articles and Weibo text, in order to capture word usages both in formal written documents and causal short texts.
- **Frequency.** We first make statistics about word frequency on our corpus, and then select words with high frequency (about 30%), middle frequency (50%) and low frequency (20%), respectively.
- **POS Tags.** The selected words consist of nouns, verbs and adjectives, as well as some functional words like adverbs and conjunctions.
- **Word length.** The selected words consist of one-character words, two-character words, three-character words and four-character words (idioms).
- **Senses.** We pick up some ambiguous words with multiple senses.

Word pairs are constructed by an expert in computational linguistics, sometimes referring to Tongyici Cilin.

Twenty post-graduate students mayor in linguistics are asked to give a score 1-

10 for each word pair according to their semantic similarity, and we calculate the average score of twenty humans as the final similarity score of each pair.

3. Evaluation

We use Spearman's rank correlation coefficient to evaluate the statistical dependence between automatic computing results and the golden human labelled data:

$$r_R = 1 - \frac{6\sum_{i=1}^{n} (R_{Xi} - R_{Yi})^2}{n(n^2 - 1)}$$

Where n is the number of observations, R_{Xi} and R_{Yi} are the standard deviations of the rank variables.