**Instruction for phrase similarity data construction**

* How to determine the phrase type

According to the Chinese tree bank (chtb), the most frequent Chinese phrase types are noun-noun, cardinal number-quantifier and adjective-noun phrase. We do not consider cardinal number-quantifier for it is hard to evaluate the similarity. We include one more phrase types, which is verb-noun phrase. Verb-noun phrase is common in sentences and often co-occur with an adjective in the middle of verb and noun, so it didn’t achieved high frequency in the Chinese tree bank.

* How to extract candidate phrases

To extract candidate phrase, we use Sougou corpus and Baidu corpus (13G in total). We first do segmentation and part of speech (pos) tagging. Then we use the pos tag to extract noun-noun phrase and adjective-noun phrase. We use rules to extract verb-noun phrase according to Chinese usage. Finally, we get 12958552 noun-noun phrases, 61038730 verb-noun phrases, and 2379905 adjective-noun phrases.

* How to select phrase for human annotation

Firstly, to reduce the size of phrase pairs, we hypothesis that sentence pairs, which are high frequency in the corpus after exchanging the one word, are similar to each other. Next, we hold the most similar phrase to each phrase. We use Chinese Tongyi Cilin dictionary similarity (which is defined as the overlap of the words label) to calculate similarity score for the left phrases. We select the 200 most similar phrase pairs and keep those occur in Renmin Paper (which are annotated with pos tag by human) to exclude annotation error. Leaving 56 noun-noun phrases, 86 verb-noun phrases and 52 adjective phrases. At last, we delete the fake phrase (not linguistically legal phrase) manually and keep the 40 most high frequency phrase pairs for each phrase type.

* how to determine the final evaluation pairs

To make the final phrase dataset to reflect high level, middle level and low level similarity, we recombine the 40 phrase pairs and calculate similarity score by Tongyi Cilin similarity. We use the original phrase pairs as the high level similarity pairs. We select the 40 most similar phrase pairs in the recombination pairs as middle level similarity pairs and 40 phrases with most low frequency pairs as the low level similarity pairs. In this way, we get 120 phrase pairs for each phrase type.

**Human annotation data collection:**

The questionnaire in the experiment is collected with <http://wj.qq.com/index.html>. There are about 150 people involved in the experiment and was paid. We collected 178 valid questionnaires. Specifically, 40 noun-noun phrases, 54 verb-noun phrases and 42 adjective-noun phrases. After preprocessing (checking for low correlation subject), leaving 39 for noun phrase, 46 for verb phrase and 40 for adjective phrase.

In order to distinguish between relatedness and similarity, we explain the two concepts and give examples in the experimental instruction. We put 9 phrases in a group and display in one page to make annotation easier with other phrase as reference.

**Data analysis：**

We use a series of Kruskal-Wallis rank sum tests to examine the relationship between our similarity bands and the elicited similarity ratings. In each phrase type, subjects rating is significantly different(p<.01）. the table below show mean value, standard deviation and standard error of different phrase type on the three band. From the mean value (High>Meduim>Low), we can see the usefulness of the three band from Tongyi Cilin.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Adjective-noun | | | Noun-noun | | | Verb-object | | |
|  | M | SD | SE | M | SD | SE | M | SD | SE |
| High | 4.46 | 1.859 | 0.046 | 4.02 | 1.738 | 0.044 | 3.94 | 1.902 | 0.044 |
| Medium | 2.77 | 1.672 | 0.042 | 2.79 | 1.758 | 0.045 | 2.60 | 1.570 | 0.037 |
| Low | 1.48 | 0.830 | 0.021 | 1.87 | 1.241 | 0.031 | 1.53 | 0.895 | 0.021 |

We also examine how well participants agreed in their similarity judgments for each phrase type. Intersubject agreement is an upper bound for the task and allows us to interpret how well our models are donging in relation to humans. To calculate intersubject agreement, we used leave one-out resampling. For each subject group we divided the set of the subjects’ responses with size m into a set of size m-1 (we average the m-1 human data) and a set of size one. We then correlated the ratings of the former set with the ratings of the latter using Spearman’s correlation

coefficient ρ. This was repeated m times. We get the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| Spearman correlation | Noun-Noun | Verb-Noun | Adjective-Noun |
| Mean | 0.7836 | 0.817707 | 0.836137 |
| Max | 0.904774 | 0.915898 | 0.91593 |
| Min | 0.589749 | 0.608979 | 0.562192 |
| SD | 0.073536 | 0.066712 | 0.065086 |

The result show that the inter correlation between subjects is high even though they think it is a hard task. Noun-noun phrase type is harder and adjective-noun phrase type is easier. The reason is probably that noun-noun phrase contain the meaning of two nouns. Comparing with adjective-noun phrase which the adjective is a modifier of noun and the phrase meaning is determined by the noun, the noun-noun phrase needs more cognitive resources.