Semantic Clustering of Morphologically Related Chinese Words

Chia-Ling Lee[1], Ya-Ning Chang[2], Chao-Lin Liu[3], Chia-Ying Lee[2], Jane Yung-jen Hsu[1]
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan[1]
Institute of Linguistics, Academia Sinica, Taipei, Taiwan[2]
Department of Computer Science, National Chengchi University, Taipei, Taiwan[3]
http://www.csie.ntu.edu.tw/~r00922072/aaai14stu.html
r00922072@ntu.edu.tw

Abstract

A Chinese character embedded in different compound words may carry different meanings. In this paper, we aim at semantic clustering of a given family of morphologically related Chinese words. In Experiment 1, we employed linguistic features at the word, syntactic, semantic, and contextual levels in aggregated computational linguistics methods to handle the clustering task. In Experiment 2, we recruited adults and children to perform the clustering task. Experimental results indicate that our computational model achieved a similar level of performance as children.

Introduction

Morphological awareness, defined as "children's conscious awareness of the morphemic structure of words and their ability to reflect on and manipulate that structure", is associated with children's reading ability and comprehension (Carlisle and Feldman 1995).

A Chinese character embedded in different words may carry different meanings. For example, "商代(Shang Period)", "商朝(Shang Dynasty)", "商店(store)", and "商品(commodity)" can form two clusters: {"商代", "商朝"} and {"商店", "商品"}. In terms of meaning of the character "商/shang1/", the former subgroup conveys concepts about a Chinese dynasty, and the latter carries information about commerce. Differentiating the meanings of the shared character in such morphologically related words can facilitate Chinese word sense disambiguation, improve Chinese word segmentation, and contribute to Chinese learning.

There are numerous semantic similarity measures proposed in the literature. Several knowledge-based approaches proposed by exploiting WordNet[1] (Pedersen, Patwardhan, and Michelizzi 2004). For corpus-based approaches, perhaps the commonest one is the LSA (Landauer, Foltz, and Laham 1998).

In this research, we employed techniques of computational linguistics to differentiate the meanings of a shared character. We applied different methods which took diverse factors into account, such as syntax, semantics, and context. We also aggregated all methods and built a better ensemble model. To contrast these results, we conducted another experiment in which we asked adults and children to do the same clustering task. Experimental results indicate that our model can achieve a similar level of performance as children in the clustering task.

[1]http://wordnet.princeton.edu

Experiments

Experiment 1

Word-to-word Semantic Similarity In Chinese compounds, a constituent character provides some clues to the semantic of a compound. We adopted four different methods to estimate word-to-word semantic similarities. Harris (1954) proposed a hypothesis that "words that occur in similar contexts tend to have similar meanings." In each method, a target word w was represented by a feature vector. The similarity was determined by the cosine similarity.

- **LSA**: LSA assumes that words with closer meaning will occur in similar documents. Using our corpus which has 5 million words, we constructed a matrix with the LSA method. Each row was a feature vector which corresponded to a target word. Each value of the feature vector was mapped to a so-called latent topic.
- **Document**: We would like to capture the document level context of a word by counting the frequency of a target word that occurred in specific document types. We had totally 90 genres, styles, and topics in our corpus, so the feature vectors had 90 dimensions.
- **Relation**: We would also like to utilize the semantic relation between words in a sentence. We parsed sentences with the Stanford Parser[2] to obtain dependency relationships among words. A typed dependency is a triplet: name of the relation, governor and dependent. We counted the frequency of a target word playing the role of a dependent in each kind of relation, and used the frequencies to build the feature vector for a target word.
- **POS**: Part-Of-Speech (POS) is an important vehicle for text processing. Based on our corpus and using the Stanford parser, we counted the frequencies of a target word serving different POS types to build the feature vector of

[2]http://nlp.stanford.edu/software/lex-parser.shtml

the word. We normalized the vectors to build a word-POS matrix.

Ensemble and Clustering For each method mentioned above, we generated one word-to-word similarity matrix. We then aggregated the similarity matrices by accumulating them with the different weights. Based on heuristic, the weight of each method were determined based on its rank of individual performance (e.g., 1.0, 1.2, 1.3, 1.4).

Since the number of clusters was not determined in advance and we only had pairwise similarities between words, we employed hierarchical agglomerative clustering (HAC) algorithm (Manning and Schütze 1999) in our work. To compute the similarity between two clusters, the average link method was adopted. We applied HAC on the ensemble matrix to cluster words in a family.

Experiment 2

In Experiment 2, we recruited two groups of participants, 14 adults and 9 children, to perform the clustering task. We did not limit the number of clusters in our experiments. Each subject in the adult group completed a questionnaire. Subjects in the child group were given word cards for grouping. During the task, if children did not know a target word, we would ask them to guess or put it to another group named "I do not know." On average, 8% were not recognized among 285 target words. When evaluating the groupings, we would view each word in that group as a single cluster.

Results and Discussion

We used the Academia Sinica Balanced Corpus[3] as the reference corpus. Our test data and ground truth were provided by three psycholinguistic researchers (actually there may be no absolute right answer). There are 11 morphological families, including 285 target words. To evaluate our performance, we first used F1 and normalized mutual information (NMI). However, we found some trends that we did not expect. In terms of F1, when the threshold of HAC increased, meaning that a larger number of clusters were generated, F1 became worse. When evaluated with NMI, the performance improved as the number of clusters increased. To prevent the number of clusters dominating our performance, we therefore introduced a new metric named F-NMI by combining *F1* and *normalized mutual information* (NMI) (Manning and Schütze 1999). F-NMI is defined as $\alpha \times F1 + (1-\alpha) \times \text{NMI}$ where α is set to 0.5 in the current experiments.

We averaged the performances observed in experiments for 11 families of each method and each human group. Table 1 summarizes the results of Experiments 1 and 2.

The ensemble method achieved the best performance (50.84%) since it considered various linguistic factors. In addition, among other four computational methods, POS method was the best performer. This was because that POS tags provided relatively more specific clues to how a word functions in a sentence. Even though other methods did not have distinguished performances, all of them contributed useful information to the ensemble method.

[3]http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

Table 1: F-NMI of Experiment 1 (computational methods) and Experiment 2 (human clustering result)

Experiment 1					
Random	LSA	Document	Relation	POS	Ensemble
25.44%	36.68%	29.28%	34.54%	47.45%	**50.84%**
Experiment 2					
Adult Group			Child Group		
76.80%			55.52%		

In Experiment 2, the adult group achieved 76.80% of F-NMI on average, showing a high agreement with our ground truth. The child group reached 55.52%. As illustrated in Figure 1 with standard error bars, the ensemble method accomplished a similar performance level with the child group. There is a big room for us to improve our methods before we can compete with human performances.

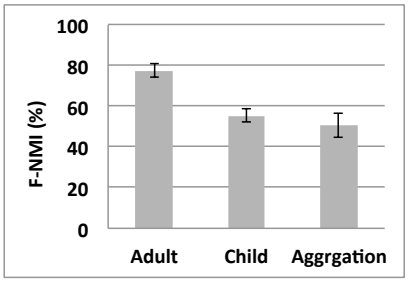

Figure 1: The ensemble method accomplished a similar performance level with the child group.

Acknowledgments

This work was supported in part by the grants NSC 99-2221-E-002-139-MY3, NSC 101-2627-E-002-002, NSC101-2221-E-004-018, and NSC 102-2420-H-001-006-MY2 from the National Science Council, Taiwan, and NTU 102R890864 from National Taiwan University, Taiwan.

References

Carlisle, J. F., and Feldman, L. 1995. Morphological awareness and early reading achievement. In *Morphological aspects of language processing*. Psychology Press. 189–209.

Harris, Z. S. 1954. Distributional structure. *Word* 10:146–162.

Landauer, T. K.; Foltz, P. W.; and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25(2-3):259–284.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. WordNet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 38–41. San Jose, CA: Association for Computational Linguistics.