Sense Tagging and Sense Discrimination of the Meaning of Chinese Polysemous Free Morphemes in Compound Words

Bruno GALMAR
Institute of Education
National Cheng Kung University
Tainan, Taiwan
Email: hsuyueshan@gmail.com

Jenn-Yeu CHEN
Institute of Cognitive Sciences
National Cheng Kung University
Tainan, Taiwan
Email: jennyeu.chen@gmail.com

I. Introduction

Among the most famous unsolved scientific problems, the problem of language acquisition is one which drives much scientific effort from an heterogenous community of researchers: linguists, cognitive scientists and neuroscientists. One of the first steps towards a full understanding of language acquisition is the study of how the morphology of words is acquired. Morphology deals firstly with the knowledge of how words are formed from minimal morphological units called morphemes and secondly with the mechanisms by which morphemes and words contact the semantic system to form their meaning. For the study of morphology, computational models relying upon machine learning approaches are useful in two ways: firstly by providing insights to cognitive scientists about how a system could learn the morphology of a human language and secondly by creating programs that could be empowering tools for linguists in analyzing corpora or preparing experimental materials [1]. The world languages differ greatly in the nature of their morphology: to one extreme, Chinese morphology is mainly based on the compounding of pairs of monomorphemic words and to another extreme Finnish is a highly inflectional language. If for certain European languages, the computational study of morphology is already advanced [2], for the Chinese language, computational research of morphology acquisition is lagging because much necessary effort is directed towards the difficult task of Chinese Word Segmentation (CWS) of raw texts. For our study pertaining to Chinese morphology, we bypass the CWS problem by using a Chinese corpus already parsed.

II. Corpus-based approaches to the study of both Chinese morphology and semantics

A. Background

Computational corpus-based approaches to the study of Chinese morphology have been initiated a decade ago mainly by [3] and be applied essentialy to CWS and Chinese Name Entity Recognition (NER).

B. The present study

Our present study is at the crossroads of Chinese morphology and semantics. It deals with sense tagging and sense discrimination of the meaning of Chinese polysemous free morphemes embedded in compound words occurring in a corpus. The following example illustrates the nature of our study in a speaking way. The character 公 is a Chinese polysemous morpheme with more than sixteen dimensions of meaning according to an etymological dictionary[1]. 公 is also a free morpheme: it can both appear as a standalone word or as a constitutent of a compound word. The Chinese words 公鹿 (male deer), 外公 (maternal grandfather) and 公平(fair) are compound words that all embed 公 as a common morpheme. However, for a Chinese native speaker, the meaning of 公 in each of the three words is different:

1) In 公鹿 (male deer), the meaning of 公 is male.
2) In 外公 (maternal grandfather), the meaning of 公 is grandfather.
3) In 公平 (fair), the meaning of 公 is fair.

In the three above examples, the meaning of 公 is defined each time by only one word: "male", "grandfather" and "fair". These one-word definitions exist as the etymological dimensions of meaning of 公 given in the aforementioned etymology dictionary of Chinese. Each etymological dimensions of meaning can be represented by a Chinese word - which we call a meaning dimension word -. Thereafter, we retain the definition of the meaning of 公 in a compound word as being one meaning dimension word.

C. Sense tagging and Sense discrimination

Sense tagging of 公 in a 公 compound word is the supervised task consisting in identifying a meaning dimension word as being the meaning of 公 in the compound word. Sense discrimination of 公 in a list of 公

[1] source: www.chineseetymology.org/ Some dimensions are overlapping.

compound words is the unsupervised task of discriminating clusters of 公 compound words, with the most informative clusters being the ones including an instance of the monomorphemic 公 word [2].

D. Potential applications of the study

Sense tagging of Chinese free morphemes in compound words will leverage the emergence of new features for Chinese WordNets: listing the listing of all the Chinese words embedding a same polysemous morpheme with a fixed identified meaning. Sense discriminating will be useful to linguists in preparing experimental materials for the cross-study of Chinese morphology and semantics.

III. Sense tagging of Polysemous Free Morphemes

The proposed computational approach put forward by [4] aims at identifying the meaning of 公 in 187 compound 公 words using knowledge of the etymology for 公. Shortly, the approach consists in detecting in a data structure - a minimum spanning tree built over the dissimilarity matrix of the reduced Singular Value Decomposition of a co-occurence matrix of 公 words - some semantic patterns which are instances of predefined abstract semantic patterns. These patterns can be interpreted through predefined meaning inference rules to identify which meaning dimension word can serve as the meaning of 公 in a 公 word captured in a semantic pattern. [4] approach can be viewed as a semisupervised learning approach or a light knowledge based approach to learning. The adjective "light" emphasizes that the amount of background knowledge for achieving the learning task is kept as low as possible: a dozen of meaning dimensions words and a few meaning inference rules. [4] work is generalizable to all Chinese polysemous monomorphemic words.

IV. Sense discrimination of Polysemous Free Morphemes

From a cognitive scientist standpoint, etymological knowledge - a kind of expert knowledge - could be considered as an unrealistic model of human background knowledge. Our current work aims at putting forward completely unsupervised - knowledge-free - approaches for sense discrimination of Chinese polysemous free morphemes embedded in compound words. As a preliminary step, we tried Principal Component Analysis (PCA), a popular linear dimensionality reduction methods to uncover clusters of 公 words. Computing a PCA on the correlation matrix of the co-occurence matrix and projecting the data set on the first three principal components (PCs) did not provide neither a satisfactory clustering - the majority of words are still drown in a cloud -, neither a reliable low dimensional representation of the data - the first three PCs account for only 4 percents of the total variance of the data -. Computing a PCA with the covariance matrix is not so informative because of the pecularities of the distribution of the term frequencies of the 公 words. The high frequency words load as expected the first components of the PCA. Latent Semantic Analysis (LSA) was also tried and provided unsatisfactory results too. Both PCA and LSA are linear methods of dimensionality reduction: they assume that there exists a linear manifold upon which lie the data. If this assumption of a linear manifold is wrong, non-linear manifolds learning techniques have to be considered [5]. Such methods - e.g ISOMAP - have been found to outperform PCA in semantic analysis [5]. Presently, we are considering to apply these methods on our data sets and later to assess the relevance of Discrete Component Analysis approaches [6] to our study.

[2] In the corpus, 公 has 5 Part-Of-Speech tags.

V. Conclusion

Sense tagging and sense discrimination of the meaning of free morphemes as constituents of Chinese compound words will not only pave the way for new features in Chinese WordNets that will be empowering for linguists, but will also advance work on Chinese morphology and semantics by offering new theoritical models and experimental predictions to test. We presented our research results for sense tagging and our current research effort for sense discrimination.

References

[1] S. Neuvel and S. Fulop, ``Unsupervised learning of morphology without morphemes,'' Association for Computational Linguistics, pp. 31--40, 2002.

[2] J. Goldsmith, ``Unsupervised learning of the morphology of a natural language,'' *Computational linguistics*, vol. 27, no. 2, pp. 153--198, 2001.

[3] R. Sproat and C. Shih, ``Corpus-based methods in Chinese morphology and phonology,'' *COOLING 2002*, 2002.

[4] B. Galmar and J. Chen., ``Identifying different meanings of a chinese morpheme through semantic pattern matching in augmented minimum spanning trees,'' *The Prague Bulletin of Mathematical Linguistics*, no. 94, pp. 15--34, 2010.

[5] V. Raykar and V. SHET, ``Unsupervised learning of semantic concepts,'' *Language*, vol. 3, no. 2, p. 4, 2004.

[6] W. Buntine and A. Jakulin, ``Discrete component analysis,'' *Subspace, Latent Structure and Feature Selection*, pp. 1--33, 2006.