Fighting Arbitrariness in WordNet-like Lexical Databases – A Natural Language Motivated Remedy

Shun Ha Sylvia Wong

Computer Science, Aston University, Aston Triangle, Birmingham B4 7ET, U.K.
Email: s.h.s.wong@aston.ac.uk

Abstract. Motivated by doubts on how faithfully and accurately a lexical database models the complicated relations that exist naturally between real-world concepts, we have studied concept organisation in WordNet 1.5 and EuroWordNet 2. Based on the arbitrariness in concept classification observed in these wordnets, we argue that concept formation in natural languages is a plausible means to improve concept relatedness in lexical databases. We also illustrate that word formation in Chinese exhibits natural semantic relatedness amongst Chinese concepts which can be exploited to aid word sense disambiguation.

1 Introduction

Research has shown that lexical databases are good sources of lexical knowledge for various Natural Language Processing (NLP) tasks. Over the years, several lexical databases have been developed, e.g. HowNet [1], WordNet [2], EuroWordNet [3] and CCD [4]. These knowledge bases differ in their detailed organisation of real-world concepts and how the knowledge base is structured. However, they all share one common feature – they all aim to specify a hierarchy of language-independent concepts which, in the developers' view, characterises important semantic distinctions between the concepts. These concepts are interrelated through a set of relations. Wong & Fung [5] observed that many of these concepts and relations are in common.

While such formalised knowledge bases are known to be well-defined hierarchical systems, there remains doubt as to how faithfully and accurately such artificial constructs model the complicated relations that exist naturally between real-world concepts. Based on the observation done on WordNet 1.5 and EuroWordNet 2, we discuss some common weaknesses in WordNet-like lexical databases. Motivated by Wong & Pala's studies [6,7], we propose a means to alleviate these weaknesses. We have carried out an experiment on the potential applicability of the proposed means of alleviation has been carried out. This paper gives a brief account of the results.

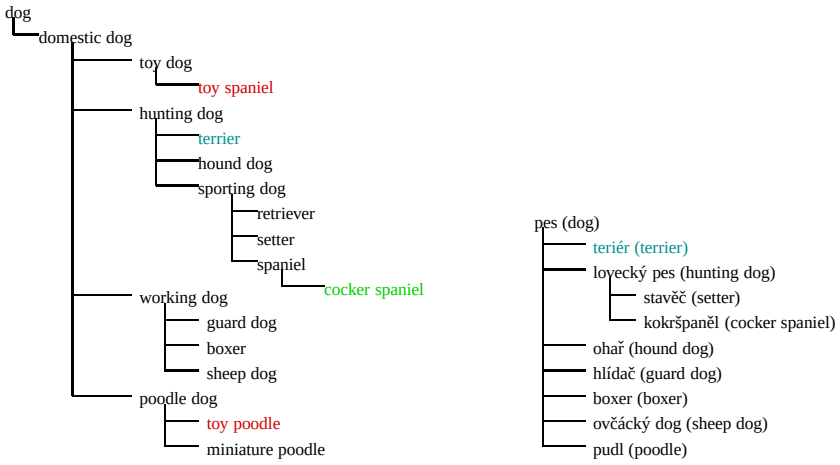
2 Some Common Weaknesses of WordNet-like Lexical Databases

In existing lexical databases, the classification of concepts is often based on hand-crafted guidelines and an individual's interpretation of the guidelines. Though exploiting existing electronic dictionary resources reduces the time involved in the manual classification process

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 234–241.

dramatically [3], by and large, given a set of relations and a set of concepts, to associate them with each other remains a subjective process.

Let us consider the concepts `toy poodle` and `toy spaniel` in Princeton WordNet 1.5 [2], i.e. "*the smallest poodle*" and "*a very small spaniel*", respectively. Both concepts are characterised by their smallness (in size) and they are also associated with the same set of top concepts in the 1stOrderEntity of the EuroWordNet 2 top ontology: `Animal`, `Form`, `Living`, `Natural`, `Object`, `Origin`. However, they are grouped under different hyperonyms (cf. Figure 1). While `toy dog` refers to "*any of several breeds of very small*



(a) WordNet 1.5 (b) Czech WordNet

Fig. 1. Extracts of `dog` concept hierarchies

dogs kept purely as pets" and poodles are also kept purely as pets, it is rather surprising that `toy poodle` is not classified as a kind of `toy dog` and that `poodle dog` is not a hyponym of `domestic dog`. Furthermore, WordNet 1.5 specifies that `toy spaniel` is "*a very small spaniel*" and `cocker spaniel` has hyperonym `spaniel`. While both `toy spaniel` and `cocker spaniel` are a kind of `spaniel`, this relation is not captured in WordNet 1.5. Imagine using such a concept hierarchy to aid a search on articles about various kinds of spaniels. Articles on toy spaniels would likely be ignored. As each system of concepts is defined according to the developers' view of the real-world, it is inevitable that the resulting ontology of concepts is fragmented and incoherent.

EuroWordNet was inspired by, and structured along the same line as, WordNet 1.5. WordNet 1.5 also serves as an interlingua within EuroWordNet. Real-world concepts and events exist regardless of the existence of natural languages. One would expect a concept to bear the same properties irrespective of its physical expression in different languages. However, while `terrier` in English is a `hunting dog`, its Czech counterpart `teriér`

is not[1] (Cf. Figure 1). The difference between the concept hierarchies in Figure 1 further exemplifies the existence of arbitrariness in concept classification.

Large-scale lexical databases are also prone to human errors. In EuroWordNet, the same synset in various European language wordnets are linked by Inter-Lingual-Index (ILI), which is in fact a list of meanings taken from WordNet 1.5 [3]. However, rather than relating `hunting dog` in English to `lovecký pes` (literally: *hunting dog*) in Czech, ILI incorrectly relates `lovecký pes` to `sporting dog`. This also explains why `teriér` and `ohař` (*hound dog*) are not considered as hunting dogs in Czech WordNet. If the association were to be done automatically based on the underlying component concepts, i.e. `lovecký` (*hunting*) and `pes` (*dog*), instead of relying on human classification, the mistake could have been avoided.

To attain a cohesive level of concept representation which is error-free from an human perception of the real-world is not an easy task. As a lexical knowledge base serves as the core foundation of various NLP tasks, a fragmented and incoherent knowledge base would, no doubt, hinder its effectiveness significantly.

3 A Natural Language Motivated Remedy

The aim of natural languages is to facilitate a concise communication of real-world concepts by means of sequences of symbols. This leads one to think whether the system for knowledge representation employed in a natural language could aid the development of a lexical database. Such a system is likely to be less subjective because, typically, it is a system developed, tested and agreed upon by millions of people over centuries. However, this system, though it exists, is hidden in most natural languages, especially those with phonetically-driven orthography.

Unlike most natural languages, the Chinese language displays a considerable amount of semantic information even at the character level. This distinctive feature suggests that the system of Chinese characters might contain a rich but concise system of inter-related concepts.

3.1 Chinese Characters

Chinese script has originated from picture-writing. Though over thousands of years of development, modern Chinese script is no longer dominated by pictographs [8,9], most Chinese characters continue to display some semantic information of the concept that it represents. Each Chinese character plays the role of a morpheme in the Chinese language. They all represent concepts that exist in the real-world.

According to Xu Shen's etymological dictionary, over 99% of the included Chinese characters display relevant semantic information to the concept that they represent [8,9]. The unique derivation of Chinese characters enables semantically related concepts to be grouped together naturally through their meaning component parts. For instance, concepts of psychological aspects like 怒 (*anger*), 恥 (*shame*), 想 (*think*) and 愛 (*love*) all possess the meaning component 心 (*heart / mind / feelings*) and concepts of trees like 橡 (*rubber tree*), 松 (*pine*), 杏 (*apricot*) and 樺 (*birch*) all share the component 木 (*tree /*

[1] Note that the words 'terrier' and 'teriér' are a pair of English-Czech cognates.

wood). Following this grouping, clusters of concepts displaying various semantic relations can be formed. While lexical databases often rely on subjective and even ad hoc judgement on concept classification, the semantic relatedness displayed by such clusters of Chinese characters provides a means to concept classification which is more objective, more explicit and, hence, easier to capture.

3.2 Chinese Concept Formation

There are over 50,000 characters in the Chinese script, but an average educated Chinese knows roughly about 6,000 characters [8]. Surprisingly, this rather limited knowledge of the Chinese script does not prohibit a Chinese from effective communication.

In English, the combination of letters to form words has little direct correlation with the meaning of words. With most Indo-European languages, it is possible to retrieve the composite meaning of a word by analysing its morphemic structure automatically [10] or semi-automatically [11]. However, with the presence of allomorphs and irregular morphology in words, to achieve reliable automatic analytical results is not an easy task.

Unlike Indo-European languages, Chinese words are typically composed of two Chinese characters. Each component character contributes part of the underlying meaning of a word, e.g. 噴射 (*jet*) = 噴 (*spurt*) + 射 (*shoot*). This characteristic holds even for words that are composed of more Chinese characters, e.g. 噴射式戰鬥機 (*fighter jet*) = 噴 (*spurt*) + 射 (*shoot*) + 式 (*model / style*) + 戰 (*battle / war*) + 機 (*machine / chance*). Thus, the knowledge of a few thousands characters allows a Chinese to deduce the meaning of words, even words which were previously unseen. Likewise, new words can also be formed by meaningful concatenation of characters.

Derivational morphology in Chinese is displayed naturally in Chinese word formation. Each Chinese character within a word corresponds to one morpheme. A study on the composite meaning of over 3,400 randomly selected Chinese words has been performed. This study revealed that the underlying meaning of over 99% of them correlates with the meaning of their component characters. Klimova & Pala [11] observed that morphemic structures of Czech words show sufficient regularity to shed light on improving the relatedness of concepts (which are organised as synsets) within EuroWordNet by means of Internal Language Relations (ILRs). This leads us to investigate the potential for Chinese word formation in enriching sense relations in existing lexical database.

With our collection of Chinese words, we grouped them according to their component characters. We found that each cluster of Chinese words displays a high level of sense relatedness. For instance, 假髮 (*wig*), 長假髮 (*peruke*), 長髮 (*long hair*), 短髮 (*short hair*), 直髮 (*straight hair*) and 曲髮 (*curly hair*) all end with 髮 (*hair*[2]) and they all describe various appearances of a person's hair. The Chinese words 牙齒 (*tooth*[3]), 牙膏 (*toothpaste*), 牙刷 (*toothbrush*), 牙線 (*dental floss*) and 牙醫 (*dentist*) begin with the component character 牙 (*a canine tooth*) which reveals that these Chinese words are all related to teeth.

Although word formation based on concatenation of morphemes exists in many natural languages, e.g. teach and teacher in English, učit (*teach*) and učitel (*teacher*) in Czech,

[2] 髮 often refers to hair on a person's head because its component part 髟 means (*long hair*).

[3] 牙齒 is composed of 牙 (*a canine tooth*) and 齒 (*a tooth, the upper incisors*).

lehren (*teach*) and Lehrer (*teacher*) in German, due to evolution of natural languages, the morphemic structure of a word might not be traceable without considering other influential natural languages. Furthermore, the set of morpheme involved in a general use of any natural language is larger than that in the Chinese language. Thus, the set of relations observed in these languages is likely not to be sufficiently representative for improving knowledge representation in a large scale lexical database.

4 Exploiting Concept Relatedness in Chinese

Concept relatedness naturally displayed among Chinese words enables clusters of semantically related Chinese words to be formed. One might argue that typical concept relations like *hyponymy/hyperonymy* also enable concept clustering. At a glance, the Chinese data shown in Section 3.2 simply correspond to a typical case of hyperonyms in WordNet, EuroWordNet and HowNet, and attributes in HowNet and CCD. In our view, the Chinese data also display the nature of multiple inheritance in concept formation. For instance, the Chinese concept 戰車 (*chariot*) is composed of 戰 (*battle / war*) and 車 (*vehicle*). These two component concepts contribute equally to the well-formedness of meaning for 戰車 (*chariot*). Hence, rather than simply considering the concept 戰車 (*chariot*) as a hyponym of vehicle with the attribute 戰 (*battle / war*), we also view 戰 (*battle / war*) and 車 (*vehicle*) as two distinct contexts in which the concept 戰車 (*chariot*) are likely to appear. This concept, when used in a text in conjunction with other concepts, shapes the overall context of the text. This characteristic also has a potential to assist in topic detection [12].

Concept relatedness in Chinese provides a ready means to exploit conceptual density in word sense disambiguation. Consider the polysemous English word fight in Figure 2. Each sense forms a cluster with their semantically related concepts. For example, the fight sense "*to hit, punch and beat (a person)*" (打) has a proximity to beat to death (打死); whereas the sense "*contending for, by or as if by combat*" (戰鬥) relates to the concept `battle`. The senses "*to engage in a quarrel*" (爭執) and "*to strive vigorously and resolutely*" (爭取) are semantically closer to each other than the fight sense "*to hit, punch and beat (a person)*" (打) because they both comprise the `argue` (爭) component.

We have implemented a Java program to perform word sense disambiguation on English texts based on the Chinese representation of each English sense expressed by an English word. Our disambiguation process is based solely on the relatedness of concepts that are expressed in each sample text. It does not take into account any part-of-speech information of the source word forms. The disambiguation process comprises three tasks: sample text preprocessing, dictionary lookup and word sense selection. The test preprocessing and dictionary lookup processes seek to locate all available Chinese interpretations of an English lexical unit in our dictionary of 2566 English-Chinese word pairs. Typically, an identifiable lexical unit in our sample texts is associated with 4–5 Chinese concepts. In word sense selection, the dominating context of each sample text is determined by counting the occurrence of each Chinese character which exists in the Chinese interpretations of each English lexical unit. The interpretation(s) which fall(s) in the determined dominating context is selected to be the intended sense of a lexical unit. A paper reporting on the implementation of the word sense disambiguation method is in preparation.

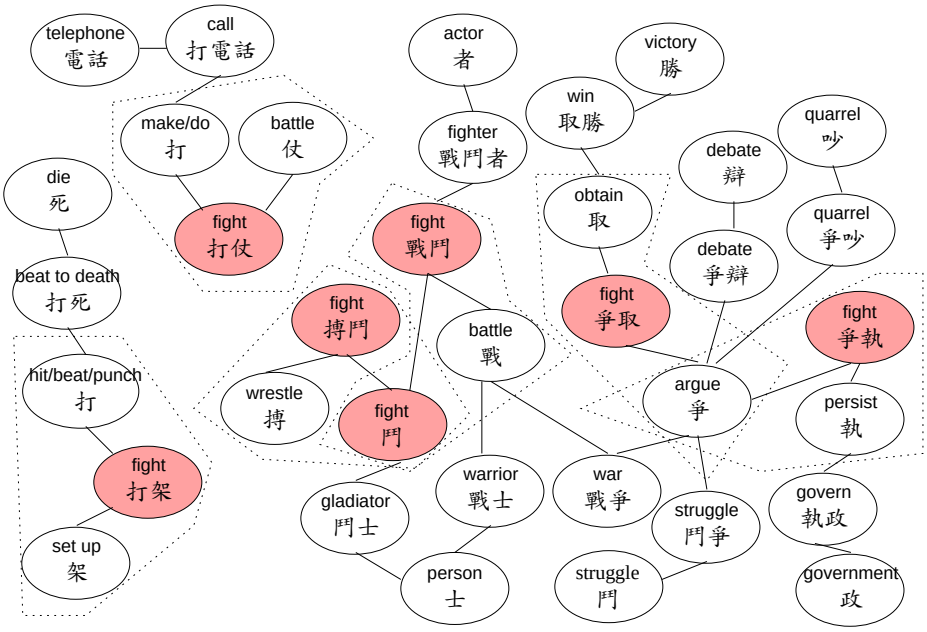


Fig. 2. Various senses of the English word 'fight' and their Chinese counterparts

5 Experimental Results

An experiment was carried out to test the effectiveness of the above approach to word sense disambiguation. Seven pieces of sample texts were taken from the NIV Bible [13], amounting a total of 62 sentences or 1313 words.

With regard to the target disambiguation lexical unit fight, the occurrences of fight in four pieces of the sample texts were referring to "*attempt to harm or gain power over an adversary by blows or with weapons*"[4]. The fight sense used in another two pieces of the sample texts refers to the sense of "*contending for, by or as if by combat*" or "*waging or carrying on (a battle)*.". The sense of fight used in the remaining piece of the sample texts can be interpreted as either of the above two main senses of fight. The disambiguation process correctly ruled out most of the inappropriate senses of fight in each case. In addition, the appropriate sense of lexical units like beat, blow, chariot, hit, loss, march, officer, plunder, strike are also correctly identified.

The disambiguation algorithm chooses the Chinese words 打, 打仗 and 打架 for intrepretting the sense "*to attempt to harm or gain power over an adversary by blows or with weapons*" of fight in Exodus 2:13, Joshua 10:2–7 and Exodus 21:20–22. The most frequently occurring character in the corresponding pieces of sample texts is 打 (*to hit*). In our approach, this character represents the prominent primary concept which dictates the context of the sample texts. Both 打 and 打仗 intrepret the sense of fight appropriately, but the interpretation 打架 is not desirable. This is chosen because this concept is also made

[4] Note that this sense of fight does not entail an organised military campaign.

up of the primary concept 打 (*to hit*) and, at present, the challenge poses by polysemous characters (e.g. 打) has been ignored.

In summary, taking note of the 45 lexical units whose interpretations were affected by the disambiguation process, 37 of them were appropriately interpreted within the context of our sample texts. Only 3 of them did not contain the best available interpretations.

Before the disambiguation, a total of 189 concepts were associated to the 45 lexical units; during the disambiguation, 125 of these concepts were ruled out. This means that, on average, our method reduced an ambiguous lexical unit of 4.2 interpretations to 1.4 interpretations even without considering part-of-speech information. Amongst the 64 remaining concepts, 56 of them appropriately interpreted the lexical units within the context of our sample texts[5]. Only 8 of them can be considered as inappropriate interpretations. Thus, by considering context information (as displayed by concept relatedness in Chinese) alone, our approach achieves 87.5% correctness in word sense disambiguation.

6 Conclusion

Based on the arbitrariness in concept classification observed in WordNet 1.5 and EuroWordNet 2, we have argued that concept formation in natural languages is a plausible means to improve concept relatedness in lexical databases. We have illustrated that word formation in Chinese exhibits natural semantic relatedness amongst Chinese concepts.

Lexical databases are good sources of lexical knowledge for domain-independent word sense disambiguation. To achieve good results, it is therefore vital for a lexical database to be as complete and coherent as possible. We have demonstrated that a method which simply exploits sense relatedness displayed naturally amongst Chinese words can aid word sense disambiguation. We believe enriching concept relations within existing lexical databases using relations inspired by sense relatedness in Chinese is worth pursuing. We propose that such a sense relatedness should be included in enhancing WordNet-like lexical databases.

7 Acknowledgments

The initial ideas of this paper spring from discussions with Doc. Karel Pala in summer, 2002. The author would like to thank him for his invaluable advice and encouragement on this research work.

References

1. Dong, Z., Dong, Q.: HowNet. [Online] Available at: `http://www.keenage.com/zhiwang/e_zhiwang.html` [7 June, 2001] (1999).
2. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998).
3. Vossen, P., et al.: Final report on EuroWordNet 2. Technical report, University of Amsterdam, Amsterdam (1999) [CD ROM].

[5] Note that 46 of the 56 appropriate interpretations came from the 37 appropriately interpreted lexical units.

4. Yu, J., Liu, Y., Yu, S.: The specification of the Chinese Concept Dictionary. Journal of Chinese Language and Computing **13** (2003) 177–194 [In Chinese].
5. Wong, P. W., Fung, P.: Nouns in WordNet and HowNet: An analysis and comparison of semantic relations. In Singh, U.N., ed.: Proceedings of the First Global WordNet Conference 2002, Mysore, 21–25 January 2002, Mysore, Central Institute of Indian Languages, Central Institute of Indian Languages (2002) 319–322.
6. Wong, S. H. S., Pala, K.: Chinese Radicals and Top Ontology in WordNet. In: Text, Speech and Dialogue—Proceedings of the Fourth International Workshop, TSD 2001, Pilsen, 10–13 September 2001. Lecture Notes in Artificial Intelligence, Subseries of Lecture Notes in Computer Sciences, Berlin, Faculty of Applied Sciences, University of West Bohemia, Springer (2001).
7. Wong, S. H. S., Pala, K.: Chinese Characters and Top Ontology in EuroWordNet. In Singh, U.N., ed.: Proceedings of the First Global WordNet Conference 2002, Mysore, 21–25 January 2002, Mysore, Central Institute of Indian Languages, Central Institute of Indian Languages (2002) 224–233.
8. Harbaugh, R.: Zhongwen.com – Chinese Characters and Culture. [Online]. Available at: `http://www.zhongwen.com/x/faq6.htm` [21 August, 2003] (1996).
9. Lu, A. Y. C.: Phonetic Motivation – A Study of the Relationship between Form and Meaning. PhD thesis, Department of Philology, Ruhr University, Bochum (1998).
10. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics **27** (2001).
11. Klímova, J., Pala, K.: Application of WordNet ILR in Czech word-formation. In M., G., et al., eds.: Proceedings of the Second International Conference on Language Resources & Evaluation (LREC 2000), Athens, 31 May – 2 JUNE 2002, ELRA – European Language Resources Association (2000) 987–992.
12. Alan, J., ed.: Topic Detection and Tracking – Event-based Information Organization. The Kluwer Internation Series on Information Retrieval. Kluwer Academic, Norwell, MA (2002).
13. International Bible Society, ed.: The NIV (New International Version) Bible. 2nd edn. Zondervan, Grand Rapids, MI (1983) [Online]. Available at: `http://bible.gospelcom.net/` [13 August, 2003].