Lang Resources & Evaluation
DOI 10.1007/s10579-016-9366-3

PROJECT NOTES

PQAC-WN: constructing a wordnet for Pre-Qin ancient Chinese

Yingjie Zhang[1] · **Bin Li**[2] · **Xinyu Dai**[1] · **Shujian Huang**[1] · **Jiajun Chen**[1]

Abstract The Princeton WordNet® (PWN) is a widely used lexical knowledge database for semantic information processing. There are now many wordnets under creation for languages worldwide. In this paper, we endeavor to construct a wordnet for Pre-Qin ancient Chinese (PQAC), called PQAC WordNet (PQAC-WN), to process the semantic information of PQAC. In previous work, most recently constructed wordnets have been established either manually by experts or automatically using resources from which translation pairs between English and the target language can be extracted. The former method, however, is time-consuming, and the latter method, owing to a lack of language resources, cannot be performed on PQAC. As a result, a method based on word definitions in a monolingual dictionary is proposed. Specifically, for each sense, kernel words are first extracted from its definition, and the senses of each kernel word are then determined by graph-based Word Sense Disambiguation. Finally, one optimal sense is chosen from the kernel word senses to guide the mapping between the word sense and PWN synset. In this

✉ Bin Li
libin.njnu@gmail.com

Yingjie Zhang
zhangyj@nlp.nju.edu.cn

Xinyu Dai
daixy@nlp.nju.edu.cn

Shujian Huang
huangsj@nlp.nju.edu.cn

Jiajun Chen
chenjj@nlp.nju.edu.cn

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

[2] School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097, China

Published online: 08 September 2016

research, we obtain 66 % PQAC senses that can be shared with English and another 14 % language-specific senses that were added to PQAC-WN as new synsets. Overall, the automatic mapping achieves a precision of over 85 %.

Keywords Definition-to-synset mapping · Pre-Qin ancient Chinese · Graph-based WSD · Global wordnet

1 Introduction

The Princeton WordNet® (PWN) (Fellbaum 1998) is a large lexical database of English that is widely exploited in many fields, such as natural language processing (NLP) (Ponzetto and Strube 2006; Budanitsky and Hirst 2006; Esuli and Sebastiani 2007) and information retrieval (IR) (Varelas et al. 2005; Hsu et al. 2008; Fernández et al. 2011). In PWN, words are grouped into sets of synonyms (synsets). Each synset expresses a distinct concept and is associated with other synsets by lexical relations.

Because PWN is useful for English information processing, wordnets of some other languages have been constructed as well. Two requirements for wordnets in other languages have been specified by the Global WordNet Association (GWA). First, any wordnet must include at least synsets and hyponymys, and second, it should be linked to PWN. For wordnets in other languages, the linkage to PWN provides a common semantic framework between this language and English. Simultaneously, the synsets and hyponymys represent a unique language-internal system of lexicalizations that maintains language-specific properties.

As a result of wordnet being extended to multilingual databases for other languages, methods and applications based on PWN can be applied to tasks in these languages. So far, GWA covers more than 50 languages in the world, including some ancient languages, e.g., Sanskrit (Kulkarn 2010), Hebrew (Ordan and Wintner 2007), and ancient Greek (Bizzoni et al. 2014). Given such a state of the art, we constructed a wordnet for Pre-Qin ancient Chinese (PQAC), called the *PQAC WordNet (PQAC-WN)*.

The Chinese language can be separated into modern Chinese and ancient Chinese. Modern Chinese is the Chinese used by the Han people nowadays, and Mandarin can be seen as modern Chinese in a narrow sense. Similarly, ancient Chinese is the Chinese used by the ancient Han people, which can be separated further into classical literary and classical vernacular Chinese. Classical literary Chinese is a written language formed based on the Pre-Qin spoken language in ancient times (around 1000–221 BC), while classical vernacular Chinese is based on the Northern dialect after the Six Dynasties (around AD 222 and later) and is closer to the spoken language.

In the past millennia, in contrast to the great changes in spoken language, classical literary Chinese has maintained a similar form. From the Pre-Qin period to the 1900s, books and records were almost all written in it. Thus, classical literary Chinese is the main focus of research in ancient Chinese, and PQAC, the most orthodox classical literary Chinese, plays a very important role in this study.

Although PQAC is the source of modern Chinese, there are many differences between PQAC and modern Chinese, such as graphemic, lexical, and syntactic differences. In this work, we focus on lexical differences, which are as follows:

1. Most of the words in PQAC are monosyllabic, while those in modern Chinese are polysyllabic;
2. A unambiguous word in PQAC that has the same word form as another unambiguous word in modern Chinese may convey totally different meanings;
3. Words in PQAC may be much more refined and condensed than the ones in modern Chinese;
4. Although Chinese is ideographic, in contrast to modern Chinese, there are many *Tongjiazi*, i.e. phonetic loan characters, in PQAC whose meanings should be understood by their pronunciation rather than their grapheme;
5. Part-of-speeches of words in PQAC are usually changeable but their senses are not reflected in most cases, which is an uncommon phenomenon in modern Chinese.

However, because there are no entirely structured lexical semantic resources in PQAC, Chinese linguists usually do their lexical research using enumeration. As a result, it is difficult to exhaustively cover the whole lexical system. Furthermore, owing to the huge differences between ancient and modern Chinese, the existing lexical resources in modern Chinese, e.g., the *Chinese Concept Dictionary* (CCD) (Yu and Yu 2002), cannot be used for the study of PQAC directly.

Therefore, given the above considerations, we decided to construct a lexical resource for PQAC, namely PQAC-WN, which is structured within the framework of PWN. In PQAC-WN, words are grouped into sets of synonyms (synsets) with each synset representing a distinct concept and associated with others by lexical relations. Once PQAC-WN has been constructed, the systematic and further lexical study in PQAC, such as the study of the distribution of lexical semantics, regular patterns of lexical structure, and so on, will be more accessible to researchers.

In previous work, a newly constructed wordnet could be established manually by experts, which was time-consuming. It could also be built automatically on resources from which word translation pairs between English and the target language could be extracted. However, because of the scarcity of language resources, these methods cannot be applied to PQAC. Specifically, there are no bilingual dictionaries or corpora between PQAC and English; what is worse, there are no resources from which word-to-word translation pairs can be extracted directly between ancient and modern Chinese. The materials in PQAC that we can utilize consist only of explanatory dictionaries, in which the senses of Pre-Qin words are explained by phrases or sentences in modern Chinese. Therefore, we would like to find a general method to map the Pre-Qin senses onto PWN synsets using their explanations or definitions.

In this paper, we present a strategy to construct PQAC-WN automatically based on the definitions in a monolingual dictionary. We employ the *Great Chinese Dictionary* (GCD) (Commercial-Press 2005), an explanatory diachronic dictionary, as our original resource to extract the Pre-Qin senses and their definitions. We then use these definitions to map Pre-Qin senses onto PWN synsets. Because the

definitions in the GCD are expressed in modern Chinese, the CCD, a modern Chinese translation of PWN, is also utilized as an intermediary to link PQAC and English.

We finally obtain the core part of PQAC-WN as our rudimentary result, which contains the major common concepts in both PQAC and English, together with some particular concepts in PQAC per se. The semantic relations in PQAC-WN follow those in PWN.

Because the GCD is a diachronic dictionary that contains almost all senses used from Pre-Qin to the present, the automatic method suggested in this paper can be simply extended to other periods of Chinese, even those later than Pre-Qin. Moreover, because all the resources used in this paper are explanatory dictionaries in the target language and an intermediary linking of the target language and English, this method could be applied to other languages that have such resources.

In the rest of this paper, Sect. 2 describes the work of mapping word senses in other languages onto PWN synsets. In Sect. 3, the employed resources are presented. Section 4 deals with the mapping procedures, Sect. 5 presents the mapping results, and Sect. 6 presents the conclusion.

2 Related work

In general, two main models are used to construct a new wordnet mapping with respect to PWN. One is the expand model, and the other is the merge model (Vossen 1998). In the expand model, the synsets of other languages are translated equivalently from those of PWN and the concepts or relations imposed on non-English are verified later. In the merge model, the new wordnets of other languages are first built separately from PWN, and the equivalent concepts and relations between the new wordnets and PWN are then generated. Previous work has shown that the mapping is more general and easier to achieve through the expand model than the merge model; however, its results will be biased to PWN and maintain fewer language-specific properties.

In recent years, there have been other wordnets constructed in ancient languages, such as Hebrew (Ordan and Wintner 2007), Latin (Monozzi 2009), Sanskrit (Kulkarn 2010) and ancient Greek (Bizzoni et al. 2014). These wordnets were almost all constructed using an expand model.

There are two approaches to implementing the expand model, one is via a translation procedure, i.e., from English to the target language, the other is via a mapping procedure, i.e., from the target language to English.

In the translation procedure, for each PWN synset, every word in it should be translated into the target language as a candidate, and the corresponding new synset for the target language consists of a subset of these candidates. The candidates can be obtained by manual translation as in the Finn WordNet (FiWN) (Lindén and Carlson 2010), or by using bilingual resources such as bilingual dictionaries (Pianta et al. 2002), bilingual parallel corpora (Fiser and Sagot 2008), and even existing bilingual wordnets (Bond et al. 2008). To obtain a confident subset of the translation candidates, Saveski and Trajkovski (2010) provided a strategy to obtain a subset via

the Google distance between the translation candidates and the machine translation results of the source PWN synset gloss.

In the mapping procedure, each target sense is expressed by one or more English translations, and the mapping results is chosen from the PWN synsets containing at least one of the many translations. Because the translations might be polysemous, Pianta et al. (2002) provides four heuristics to determine the exact mapping one, namely, generic probability, back translation, gloss matching, and synset intersection in the construction of MultiWordNet (MWN). When building up the wordnet for Korean, Lee et al. (2004) followed the first three heuristics of Pianta's work and changed the intersection heuristic into maximum likelihood. In addition, they proposed a co-occurrence heuristic and the relations among all these heuristics were learnt via a supervised method. A Romanian wordnet (Barbu et al. 2007) was built up using the synset intersection heuristic in the expand model, while two extra heuristics, namely, domain and the IS-A definition, were added as well.

In this paper, we also seek to use the expand model to build our PQAC-WN via a mapping procedure from PQAC to English. Because of the lack of bilingual resources between PQAC and English, the CCD, a modern Chinese translation version of PWN, was used as an intermediary and all Pre-Qin senses were extracted from a monolingual dictionary in modern Chinese.

3 Dictionary resources

To map PQAC word senses onto English, a bilingual resource, be it a parallel corpus or bilingual dictionary, is needed. However, because there are no such resources available, the PQAC senses cannot be mapped onto PWN synsets directly. In this case, we have to use the CCD as an intermediary for our mapping task. The mapping from PQAC to English is then converted into modern Chinese. In addition, a thesaurus in modern Chinese, the *HIT IR-Lab Tongyici Cilin (Extended)* (*Cilin-Extend*) (Mei et al. 1983; HIT 2012), is also used to extend the CCD vocabulary size.

Because the books in PQAC cover many domains such as history or poems with large differences in terms of their vocabularies, together with the fact that there are no common dictionaries for all Pre-Qin books, the GCD, a diachronic dictionary in modern Chinese, was utilized to extract the PQAC words and their senses.

3.1 CCD

The CCD is the modern Chinese version of PWN 1.6, which was translated manually from PWN and maintains all the synsets and semantic relations in PWN without any changes. Every CCD synset corresponds with a PWN synset, and the relations in CCD are exactly the same as those in PWN.

Although the CCD may be incomplete as a Chinese dictionary, it has good consistency with PWN and is a suitable intermediary between the GCD and PWN. The scale of the CCD is shown in Table 1. Table 2 presents an example of a CCD synset.

Table 1 Scale of CCD

	Noun	Verb	Adj.	Adv.
Synset number	66,025	12,127	17,915	3575
Average number of lexemes per synset	2.29	3.21	2.00	2.18
Word number	104,170	17,539	17,872	3903
Average number of synsets per word	1.45	2.22	2.01	1.99

Table 2 Content for noun synset 02876152 in CCD

Offset	02876152
Part-of-speech	n
Category	06
Synset	Jewelry, jewellery
CSynset	珍宝 珠宝 无价之宝 珍稀之宝
Definition	An adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems)
CDefinition	一件装饰品(作为一只手镯或者戒指或者项链)由贵金属并且镶珍宝制成(或者仿珍宝)
Hyperonym	02166920
Hyponym	02270135, 02287537, 02328059, 02450840, 02532019, 02620228, 02875598, 03034095, 03126308, 03241409, 03503247
Holonym	p10536941

From Table 2, we can see that there are two English synonyms in the noun PWN synset 02876152 that are translated into four Chinese synonyms. Note that the glosses of this synset that have been translated into Chinese but mechanically follow the original English version and do not result in an authentic Chinese expression are not used in this work.

Given the results of the CCD, we can utilize a monolingual PQAC dictionary as our source of senses and convert the PQAC-to-English mapping into a PQAC-to-modern-Chinese mapping.

3.2 GCD

As one of the most complete diachronic Chinese dictionaries, the GCD is the main resource we used to obtain the PQAC-WN, i.e., all the Pre-Qin senses mapped onto PWN are extracted from there.

Each sense of a Chinese word in GCD is explained in modern Chinese followed by examples from the books, especially the earliest occurrences, which can be helpful for extracting the senses of PQAC. The age of the books is used in our work to extract the senses that existed in the Pre-Qin Dynasty. In this work, we obtained 16,911 Chinese content words (54,073 Pre-Qin senses in total) from a GCD

Table 3 Example Chinese word 珠玉 (*zhuyu*) **and its senses in GCD**

No.	Definition	Example source	Example sentence
1	珍珠和玉。泛指珠宝。	庄子•让王	事之以珠玉而不受。
	Pearls and jades. Jewelry in general	*Zhuangzi - Rangwang*	(Foreigners) give him *zhuyu* but are rejected
2	小粒圆形的玉。	周礼•天官•玉府	共王之服玉、佩玉、珠玉。
	A tiny round jade	*The Rites of Zhou - Tianguan - Yufu*	(*Yufu*) provides the king *fuyu*, *peiyu*, and *zhuyu*
3	比喻妙语或美好的诗文。	晋书•夏侯湛传	咳唾成珠玉, 揮袂出風雲。
	A metaphor for witty remarks or beautiful poems.	*Jin Shu - The Story of Xiahou Zhan*	The saliva he spits becomes *zhuyu*. The sleeves he waves makes wind and clouds flow
4	比喻丰姿俊秀的人。	晋书•卫玠传	珠玉在側, 覺我形穢。
	A metaphor for an elegant and handsome person	*Jin Shu - The Story of Wei Jie*	Standing beside the *zhuyu* makes me feel that I am ugly
5	喻俊杰, 英才。	世说新语•容止	今日之行, 觸目見琳琅珠玉。
	A metaphor for a talented and smart person	*A New Account of the Tales of the World - Rong Zhi*	Today I saw a *linlang* and *zhuyu*

Table 4 Definition patterns in GCD

Pattern type	Relation	Pattern examples
EQUAL	Same synset	指 (*zhi*; "it refers to"); 比喻 (*biyu*; "a metaphor for"); ……的简称 (*jiancheng*; "an abbreviation of")
IS-A	@	是……的一种 (*yizhong*; "as a kind of")
IS-A-INS	@i	……名 (*ming*; "as a name of")
ANTONYM	!	……的对称 (*duicheng*; "the opposite of")
IS-PART-OF	%	……的一部分 (*yibufen*; "a part of")

Table 5 Three types of word sets in *Cilin-Extend*

Tags	Word list
Bm16D01@	琥珀(*hupo*; "amber")
Bm16B01=	玉石(*yushi*; "jade") 玉璧(*yubi*; "jade") 玉佩(*yupei*; "jade") 佩玉(*peiyu*; "jade pendant")
Bp33A11#	珠宝(*zhubao*; "jewellery") 猫眼(*maoyan*; "opal") 软玉(*ruanyu*; "nephrite") 珊瑚(*shanhu*; "coral")

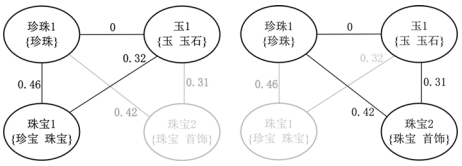
covering 84.67 % of the content words in 25 of the most representative Pre-Qin books.

Table 3 shows a word with its senses expressed in GCD. From Table 3, it is obvious that the Chinese word 珠玉 (*zhuyu*) has five senses, and each sense is explained by a definition and an example. Although a sense in GCD may have more than one example, only the first one is chosen, as shown in Table 3, including the example source column and the example sentence column. Because GCD is a diachronic dictionary that should show the origin of word senses, those examples, which are ordered by the age of their sources, always include the first occurrence of the sense in books. From the example source column, we find that the first and second senses of *zhuyu* are first used in 庄子 (*Zhuangzi*) and 周礼 (*The Rites of Zhou*) respectively, both of which are Pre-Qin books (Li 2004; Liu et al. 2014). Therefore, two senses that have been extracted in this work are Pre-Qin senses. The example sources of the three senses left are Post–Pre-Qin; these senses are out of the scope of this work.

3.2.1 Definition Patterns in GCD

We now focus on the definition column in Table 3. In the definition of the first sense (see No. 1 in Table 3), there are two sentences, each of which can represent it. The first sentence is expressed by a coordinate noun phrase. Both sides of the phrase are similar to this sense. As for the second definition sentence, there is a word "泛指" (*fanzhi*; "refers to something in general") which may be treated as a component of a sentence trunk when parsing the sentence. However, such a word has little to do with the extracted meaning of this definition.

Table 6 Mapping procedure of *zhuyu*

Step	Procedure	Result
	Original definition	珍珠和玉。泛指珠宝。 (*zhenzhu he yu. fanzhi zhubao*; "Pearls and jades. Jewelry in general.")
1	Sentence segmentation	珍珠和玉 (*zhenzhu he yu*; "Pearls and jades") 泛指珠宝 (*fanzhi zhubao*; "Jewelry in general")
	Definition patterns removed	EQUAL 珍珠 和 玉 (*zhenzhu he yu*; "pearls and jades") EQUAL 珠宝 (*zhubao*; "jewellery")
2	Parsing	(NP (NN 珍珠) (CC 和) (NN 玉)) (NP (NN 珠宝))
	Kernel words	SYN 珍珠; SYN 玉 SYN 珠宝
3	Sense candidates	珍珠1 {珍珠(*zhenzhu*)} ({pearl}) 玉1 {玉(*yu*) 玉石(*yushi*)} ({jade jadestone}) 珠宝1 {珍宝(*zhenbao*) 珠宝(*zhubao*)}({jewellery jewelry}); 珠宝2 {珠宝(*zhubao*) 首饰(*shoushi*)} ({bijou})
	Possible sense combinations	
	WSD results (score)	SYN 珍珠 1 (0.46); SYN 玉1 (0.32); SYN 珠宝 1(0.78)
4	Most possible kernel word	珠宝 (珠宝 1 (0.78))
	Is SYN?	Yes
	Definition type	EQUAL
	Mapped CCD Synset	珠宝1{珍宝 珠宝} ({jewellery jewelry})
	Mapped PWN Synset	Noun 02876152

We observe the definitions in GCD and find that there are many such words like "泛指" that are usually in the front or end of a sentence. These types of words, in the present work, are called definition patterns (Bond et al. 2004; Oliveira and Gomes 2013), and should be omitted to obtain pure definition sentences before extracting the kernel words. By counting the frequency of the front and end of definitions in GCD, we can determine the definition patterns, as shown in Table 4. The "Pattern type" column shows the names of the definition patterns defined in this work and the "Relation" column shows the semantic relations between the original and pure

Table 7 **a** Sense count with relations from syntax, **b** Count of relations from definition patterns for mapped SYNs

(a)

Relation type (count)	Result	Count
SYN (38,359)	Mapped	35,765
	OOV	2594
HYPER (15,714)	Added	7490
	Other	8224

(b)

Relation type	Count
EQUAL	35,476
IS-A	186
IS-A-INSTANCE	92
IS-PART-OF	8
ANTONYM	2

Table 8 Samples with their sense numbers in GCD and frequencies in 25 Pre-Qin books

Word	負 (*fu*)	逆 (*ni*)	繩 (*sheng*)	君 (*jun*)	命 (*ming*)	區 (*qu*)	攘 (*rang*)
Sense number	16	14	13	11	11	11	10
Frequency	210	516	113	6138	2626	24	46
Word	告 (*gao*)	珍 (*zhen*)	拊 (*fu*)	財 (*cai*)	缶 (*fou*)	對 (*dui*)	改 (*gai*)
Sense number	8	6	5	5	5	5	5
Frequency	1429	40	30	606	32	1885	224
Word	災 (*zai*)	戰 (*zhan*)	竹 (*zhu*)	席 (*xi*)	戕 (*qiang*)	叛 (*pan*)	
Sense number	4	3	3	3	3	2	
Frequency	265	1103	52	442	20	189	

definitions. The "Pattern examples" column provides some examples of the definition patterns in GCD.

3.3 *Cilin-Extend*

Cilin-Extend is a Chinese thesaurus, each line of which shows a set of synonyms. It contains 77,457 Chinese words, of which 54.65 % (42,334 words) are not contained in the CCD. In this work, *Cilin-Extend* is used to extend the vocabulary size of the CCD. Table 5 shows some word sets in *Cilin-Extend*.

Table 9 Mapping results from GCD into PWN for the samples

	This work	MWN
Total sense count	143	143
Kernel words with SYN-EQUAL	114	118
OOV of both GCD and *Cilin-Extend*	13	16
Mapped senses	101	102
Correct mappings	86	49
Mapping precision	85.15 %	48.04 %

From Table 5, we can see that there are three kinds of lines in Cilin-Extend, tagged by the marks "=," "#," and "@," respectively. Because a line marked by "@" contains only one word inside, there are no other words available to extend the CCD synset containing the single word. Furthermore, lines with the "#" mark are composed of equivalents rather than synonyms. As a result, it will introduce additional noise if these lines are used to extend CCD synsets. Accordingly, only the lines with "=" marks are utilized in this work.

4 Mapping PQAC words into PWN

In order to map a Pre-Qin word sense in the GCD onto a PWN synset in English, we should first map this sense onto a CCD synset in modern Chinese. In the GCD, the definition of a word sense is explained by sentences or phrases in modern Chinese, and it is difficult to obtain its relationship to the CCD synsets directly from the whole definition. However, definitions in a dictionary usually contain some kernel words that represent their general meanings (Amsler 1981; Eckard et al. 2012). Therefore, we would like to find these kernel words, via which we can obtain the relation between a GCD definition and CCD synset. To find the kernel words, we should first remove the definition patterns (Sect. 3.2.1) from the original definitions. After the kernel words are found, we compare them with the synonyms in the CCD synsets to obtain the mapping result. However, kernel words of a definition may be polysemy, so before we obtain the mapping result, it is necessary to perform a WSD procedure on them.

To sum up, we map a Pre-Qin sense in the GCD onto a PWN synset using the following four steps:

Step 1 Pre-processing: sentence and word segmentation and definition pattern removal;
Step 2 Parsing to obtain the kernel words of definition sentences;
Step 3 Performing a graph-based WSD method (Navigli and Lapata 2010; Hirst and St-Onge 1998; Sinha and Mihalcea 2007) on the CCD to determine the senses of all the kernel words;
Step 4 Obtaining the mapping result onto the CCD.

Table 10 Samples with their cosine distance between mapped pairs

Word	負	逆	繩	君	命	區	攘
Correct (min)	0.427	0.423	0.542	0.495	0.484	0.390	0.340
Wrong (max)	0.323	0.302	0.192	0.281	0.034	0.083	0.350
Word	告	珍	拊	財	缶	對	改
Correct (min)	0.424	0.570	0.254	0.490	0.821	0.380	0.623
Wrong (max)	–	0.330	0.174	0.374	–	0.312	0.198
Word	災	戰	竹	席	戕	叛	
Correct (min)	0.638	0.578	0.493	0.603	0.398	0.721	
Wrong (max)	0.102	–	–	–	–	–	

Because CCD synsets are in one-to-one correspondence with PWN ones, when we get the mappings between PQAC and CCD, we also determine the mapping results between PQAC and PWN.

4.1 Pre-processing

A definition in the GCD may consist of several sentences, each of which expresses the same meaning. In the pre-processing procedure, we would like to extract a list of pure definition sentences for each sense from its definition. This procedure is divided into two sub-steps: sentence segmentation and the removal of all definition patterns from every sentence.

In this work, we perform a simple sentence segmentation based on semicolons or periods. After the sentences have been segmented, the words matching the definition patterns in a definition sentence are removed so that all definition sentences are converted into pure definitions. Because some patterns may have only one character that may be a part of a word, the removal procedure should be done after word segmentation and the definition patterns should be matched between two-word boundaries. Details of the definition patterns are presented in Sect. 3.2.1.

4.2 Finding the kernel words

After obtaining the pure definitions of a sense, we extract the kernel words from them. The kernel words of a sentence are always in the trunk of the sentence, so we parse the pure definitions and find their trunks to extract the kernel words. Because the senses we would like to map are content words, the kernel words of these senses will also be content words. As for the trunk of a pure definition, it may be expressed as a coordinate noun (verb, adjective, or adverb) phrase or a non-coordinate verb phrase. The kernel words, however, are all the coordinate words with the same part-of-speech in a coordinate phrase, or those verbs in verb phrases.

Note that we divide two kinds of relations between the kernel words and pure definition in this work. One is called SYN, and the other is HYPER. The kernel word tagged by a SYN relation means that it has the same meaning as the pure definition, while a HYPER relation means this kernel word can only represent

general meanings of the pure definition. More formally, the kernel word with a SYN relation is always extracted from a coordinate phrase without any modifier. Furthermore, the kernel word with a HYPER relation may be extracted from a trunk of either a coordinate phrase with a modifier or verb phrase with an object.

4.3 Graph-based WSD on kernel words

To determine the position of the target Pre-Qin sense in CCD, it is necessary to determine the CCD sense of these kernel words first. Because each kernel word may be contained by more than one CCD synset, we should perform a WSD procedure to determine their sense.

Considering that most of the kernel words have to do with the target sense, there must be some close semantic relations among them. Therefore, in this work, we choose a graph-based WSD method (Navigli and Lapata 2010; Sinha and Mihalcea 2007) to determine their senses.

In a graph used to perform WSD, each node is a possible CCD synset containing a kernel word, and an edge between two nodes is based on the semantic relation in CCD, including synonym, antonym, similar_to, hypernym and hyponym, holonym, and meronym relations. All these relation types are chosen based on the types of definition patterns.

The weight of an edge represents the strength between two nodes and is calculated based on the shortest path between two CCD synsets. This weight is a revised version of the medium-strong function proposed by Hirst and St-Onge (1998). Considering that the path may consist of different types of semantic relations, we give different relations different weights and use the weighted sum of all relations in a path instead of the path's length. In addition, as we suppose the relations among kernel words must be close, the length of the shortest path is limited to five relations or less.

$$weight(Syn_w, Syn_{w'}) = \begin{cases} \frac{1}{d}\sum_{i=1}^{d} \omega_i - k \times (d-1), & d \in [1,5] \\ 0, & otherwise \end{cases} \quad (1)$$

where w and w' are two kernel words of a target sense, Syn_w ($Syn_{w'}$) is a CCD synset containing w (w'), d is the length of the shortest path in the CCD between Syn_w and $Syn_{w'}$, w_i is the weight of the ith semantic relation on the path, k is a penalty factor following the hypothesis that the longer the path between two synsets, the smaller the strength between them.

After the graph is built, we use a global approach (Navigli and Lapata 2010) to evaluate its connectivity, i.e., we calculate the edge densities of all different interpretations (sense assignments) of the set of kernel words. The interpretation with the largest density is chosen as our final WSD result.

Note that in this step when the kernel words are all out of the CCD vocabulary, *Cilin-Extended* is utilized to find their synonyms. Because a kernel word may be contained in several lines, we first perform a WSD procedure to determine the sense of the kernel word in *Cilin-Extended*. After all kernel words have found their

synonyms, we use these synonyms to replace the kernel words and put them back into CCD to perform the graph-based WSD procedure above.

4.4 Eliciting the mapping result

After the WSD method is complete, each kernel word of the target sense is mapped onto a CCD synset with a score and tagged by its relation with the original definition. All the mapped CCD synsets are mapping result candidates for the target sense. In this step, we would like to choose a CCD synset from among the candidates to be the basic synset and obtain our mapping result for the target sense based on the basic synset and its tagged relation.

Mapping the target sense onto CCD consists of three tasks.

First, we determine the synset with the largest score of all the candidates as the basic synset candidate. A threshold can also be used to ensure more confidence in the basic candidate. The basic synset candidate is tagged by a relation, e.g., EQUAL-HYPER, via Step 1 and 2. This relation has two parts. The former is obtained from the definition pattern shown in Table 4, and the latter is from the syntax structure described in Sect. 4.2.

Second, we check the latter relation tag of the basic synset candidate. If the latter relation tag is SYN, the basic synset candidate is the basic synset we need. Otherwise, there are two cases. If the basic synset candidate is a leaf synset in CCD, we add a new synset for PQAC-WN. The new synset should be related to the candidate with an IS-A relation and deemed to be the basic synset. In the other case, the target sense is not added in PQAC-WN in this work.

Third, after the basic synset is determined, we check the former part of the relation. If it is EQUAL, we determine that the basic synset is the exact mapping result for the target Pre-Qin sense. Otherwise, we add a new synset for the target sense, which is related to the basic synset by the former part of the relation tag.

4.5 Flow-lined example of a mapping procedure

In this section, we present a flow-chart of the mapping procedure for the first sense of 珠玉 (*zhuyu*; "jewelry") in Table 6. The original definition of this sense is "珍珠和玉。泛指珠宝。 (*zhenzhu he yu. fanzhi zhubao*; 'Pearls and jades. Jewelry in general.')". After the four mapping procedure steps are complete, we map it onto the noun synset 02876152 in PWN.

From Table 6, we can see that there are two sentences in the original definition of *zhuyu*, and every sentence can be seen as an explanation of this sense alone. There are no definition patterns in the first sentence, so it has already been purified and will be marked with an EQUAL relation. The second sentence contains a definition pattern "泛指 (*fanzhi*; 'refers to something in general')" in the EQUAL class. By removing this pattern, we obtain a pure definition sentence "珠宝 (*zhubao*; 'jewellery')" with an EQUAL relation. As a result, after Step 1, the first sense of *zhuyu* is expressed in two pure definition sentences, "珍珠 和 玉 (*zhenzhu he yu*; 'pearls and jades')" and "珠宝 (*zhubao*; 'jewellery')", both with EQUAL relations.

In Step 2, we parse the two pure definition sentences and find that the first one, *zhenzhu he yu*, is a noun phrase (NP) with two coordinate nouns (NNs) connected by a conjunctive (CC) "和 (*he*; 'and')", while the second one, *zhubao*, is an NP with only one NN. Hence, we obtain two kernel words, "珍珠 (*zhenzhu*; 'pearl')" and "玉 (*yu*; 'jade')", from the first definition and one other kernel word, "珠宝", from the second. Because there are no modifiers for either of these kernel words, they all have a SYN relation with their original pure definitions.

In Step 3, we check the kernel words in CCD and find that *zhenzhu* and *yu* are both monosemic, while *zhubao* is contained by two synsets. Therefore, there are two possible interpretations for this set of kernel words. Because the first interpretation has the largest score, its sense assignment is considered to be our WSD result. As a result, we obtain a list of kernel words, each of which has a determined sense that is weighted.

Step 4 of Table 6 shows the procedure of obtaining the mapping result for the first sense of *zhuyu*. From the WSD result in Step 3, we can see that the kernel word *zhubao* with its sense "珠宝 1" obtains the largest score and is chosen to be our basic synset. Because *zhubao* has a SYN relation with the pure definition and the definition pattern is EQUAL, this synset is considered to be our mapping result. Converting the CCD synset into a PWN one, we can map the first sense of *zhuyu* into a noun synset in PWN with the offset 02876152.

5 Experimental results

In this study, we performed our method on 16,911 ancient Chinese words with 54,073 Pre-Qin senses extracted from the GCD. They cover 84.67 % of the content words derived from 25 typical Pre-Qin books[1] (Li et al. 2012). Among the 16,911 ancient Chinese words, 8139 of them are not found in the CCD. As to the other words, only fewer than 100 of them have not changed their meanings. Words in the lexicographer files of act, artifact, attribute, person, contact and motion change much more than words in files of body, motive, phenomenon, possession, quantity, consumption and weather.

We observe the remaining 15.33 % content words of the Pre-Qin books that are not in the GCD vocabulary and find that they are all polysyllabic words. Except for the ones that are names of persons or places, most of them are made up of several monosyllabic words, which is common in PQAC. Because these monosyllabic words are collected by GCD, we can use further word segmentation to increase the

[1] The 25 books are *Zuo Zhuan* (*左传*), *Guanzi* (*管子*), *Hanfeizi* (*韩非子*), *Lv Shi Chun Qiu* (*吕氏春秋*), *Liji* (*礼记*), *Mohism* (*墨子*), *Xunzi* (*荀子*), *Guo Yu* (*国语*), *Yili* (*仪礼*), *Zhuangzi* (*庄子*), *The Rites of Zhou* (*周礼*), *Gongyang Zhuan* (*公羊传*), *Guliang Zhuan* (*谷梁传*), *YanziChun Qiu* (*晏子春秋*), *Mengzi* (*孟子*), *Book of Poetry* (*诗经*), *Shang Shu* (*尚书*), *Book of Changes* (*周易*), *Shang Jun Shu* (*商君书*), *The Analects* (*论语*), *Chu Ci* (*楚辞*), *The Art of War* (*孙子兵法*), *Taoism* (*道德经*), *Wuzi* (*吴子*) *and Xiao Jing* (*孝经*).

coverage of content words in the Pre-Qin books. The preliminary PQAC-WN result is available on the web and can be queried online.[2]

5.1 Coverage of the mapping results

As shown in Table 7a, we extracted 38,359 senses whose pure definitions can be represented exactly by kernel words along with 2594 senses omitted whose kernel words are out-of-vocabulary (OOV). The 35,765 Pre-Qin senses are mapped onto PWN. This map covers 66.14 % of the Pre-Qin senses. For the other 15,714 senses, their kernel words are considered to express general meanings and are mapped onto PWN synsets as well. We observe these mapping results and find that 7490 of them, i.e., 13.85 % of Pre-Qin senses, are leaf nodes in PWN that can be added directly into PQAC-WN as new synsets. These new synsets are language-specific to PQAC. In this work, there are 8224 senses left whose positions in PQAC-WN cannot be determined. To sum up, our PQAC-WN covers 79.99 % of the Pre-Qin senses, along with 4.80 % senses whose kernel words are OOV and 15.21 % senses for which their hypernyms but not exact positions can be found in the new resource.

Table 7b shows the statistical results of the relations tagged by definition patterns for the 35,765 senses with a SYN relation. We find that most of the definitions (35,476) have EQUAL relations with its pure form. These senses are considered to be concepts that are common to both PQAC and English. The 289 senses left are another part that is language-specific to PQAC and further, 279 of these are hyponyms of some PWN synsets.

5.2 Precision of the Mapping Results

We used two methods to evaluate the precision of our mapping results, sampling and automatic evaluation.

5.2.1 Evaluating by sampling

We sampled 20 words randomly according to the distribution of the sense counts and word frequencies. The sampling result is shown in Table 8. To evaluate our mapping results and compare it with the assign-procedure in MWN (Pianta et al. 2002), we used precision, defined as follows.

$$\text{precision} = \frac{\sum_{w}|correct\ mapping|}{\sum_{w}|mapped\ sense|} \tag{2}$$

Because the assign-procedure in MWN is based on translation equivalences (TEs), we focus on the kernel words with relation EQUAL-SYN, which can be considered to be the TEs from ancient Chinese to modern Chinese to some extent. The procedure depends on four main groups of rules, i.e., generic probability, back translation, gloss matching, and synset intersection.

[2] URL: http://langsphere.com/. License: CC BY-NC-SA 4.0.

The generic probability depends on the degree of ambiguity of TEs of the input word sense. The back translation rule increases the scores of TE sense candidates that contain more words that can be translated back into the input word. However, we cannot obtain TEs from modern to ancient Chinese, hence, the back translation rule is not considered in our comparison method. The gloss matching rule uses the gloss information, such as semantic field specification, synonym, hypernym, and context of use. In order to obtain such information, we relax the restriction of the EQUAL-SYN relation. In the comparison method, the kernel words with an EQUAL-HYPER relation can be accepted, when the kernel words and the modifiers of the kernel words have the same part-of-speech. The synset intersection rule takes different sets of candidates that are accessible through the different TEs and intersects them. The synsets that are in the intersection obtain an additional score.

The mapping result of the samples is shown in Table 9.

From Table 9, we can see that there are 143 Pre-Qin ancient senses for the 20 sample words. In this work, we obtain 114 senses whose kernel words have the relations of EQUAL-SYN with their original definitions. However, 13 of these senses represented by kernel words are not in the CCD vocabulary and cannot be extended by *Cilin-Extend* either. As a result, 101 senses are left to be mapped onto PWN, and only 86 obtain the correct mapping. This gives a precision of over 85 %, which is better than the results obtained by the MWN construction method.

5.2.2 Automatic evaluation

Because the sample evaluation only covers a small scale and evaluating all results manually takes a large amount of time, in this section, we evaluate our whole mapping result in an automatic way.

Word embedding (Mikolov et al. 2013a) is a kind of method that represents words from the vocabulary as vectors, and its representation results perform well on many NLP semantic tasks (Mikolov et al. 2013b; Socher et al. 2013). Therefore, in this evaluation, we represent our kernel words and synonyms of the CCD synsets by the result of word embedding and use them to evaluate our mapping results. All vectors are pre-trained on Chinese Gigaword (Parker et al. 2011) by the word2vec tool (Mikolov et al. 2013c). Though the vector of one word is extracted without considering its certain sense, the average of all vectors of words that share the same sense (i.e., words in the same synset) can decrease the noise efficiently. We use the cosine distance between two vectors of the mapped synsets to measure their similarity.

To check the ability of this automatic evaluation method, we first observed the results on the samples (see Table 10). We find that although the minimum cosine distances of the correct mapping pairs varies, the distances of the wrong mapping pairs are all below 0.4. Hence, we suppose that eliminating the mapping pairs that have a distance <0.4 can make our result more confident. As a result, we find that 81.02 % of the 43,255 (SYN-Mapped and HYPER-Added in Table 7a) Pre-Qin senses obtain a confident mapping.

5.3 Error analysis

Observing the false results of samples, we find that the errors can mostly be divided into two types: (1) partially correct (incomplete) or imprecise mappings and (2) completely wrong mappings. Detailed illustrations of each are as follows.

The incomplete mapping results are caused by the fact that the algorithm proposed in this paper always elicits only one mapping result. For instance, because one sense of "區(*qu*)" is explained to be the same as "驅(*qu*)", we use "驅" as its kernel word. Furthermore, "驅" has several meanings, e.g., "drive" and "walk" in CCD, so that all of them should be mapped. However, in our work, only "walk" is chosen as the result. The same is true of the word "叛(*pan*)", meaning "背叛者(*beipanzhe*)," which should be mapped to both "renegade" and "betrayer, traitor."

The imprecise mapping results are also caused by the tendency of our method when the meanings of kernel words in a gloss are similar but not exactly alike. Usually, as we choose the most common sense, we cannot distinguish between a common or special sense for a gloss, which should be decided by additional information, such as an example sentence of specific meaning.

There are two causes for the completely wrong mapping in the samples: One arises out of the language-specific parts of PQAC, for instance, "君(*jun*)" with the meaning "大夫(*dafu*)" is mapped onto the synset of "doctor." However, in the GCD, "大夫" means a kind of senior official in feudal China, a sense that is not included in English and not even used in modern Chinese. The other is due to a missing sense in CCD that is originally intact in PWN. For instance, "財(*cai*)" with a sense "裁(*cai*), 裁制(*caizhi*), 裁断(*caiduan*)" in the GCD should have two senses, i.e., "restrain" and "nip off" in the CCD, only to be informed that one sense, "nip off," has been kept without the coverage of the other sense, "restrain," leading to the failure of correct sense to be covered in the candidates and mapping errors in turn. However, *caizhi* in Chinese is not a general expression for "restrain." This kind of oversight might not be avoidable even we use a dictionary that is larger than the CCD.

To sum up, the incomplete or imprecise mappings errors are caused by properties of the method *per se*, which could be addressed by improving the algorithm. However, concerning the lack of resources in Chinese, the completely wrong mapping errors are caused by the incompleteness of knowledge in the *CCD* and *Cilin-Extend*, the dictionaries used in this paper, which is an issue that is difficult to address.

5.4 Future work

A more complete PQAC-WN would be preferable in future work. On one hand, in order to improve the coverage of the mapping result, kernel words tagged with HYPER relations should be distributed with certain positions in PQAC-WN; on the other hand, the problem of imprecise and incomplete mappings should be rectified for the sake of enhancing the precision of the mapping results as well.

6 Conclusion

In this paper, we presented a method to construct a lexical knowledge base of PQAC called *PQAC-WN*, which shares a structure that is identical to PWN. To achieve such an aim, the explanatory dictionary GCD was exploited as our original resource, from which we extracted the Pre-Qin senses to be mapped onto the PWN synsets. In this case, a graph-based WSD method was applied to our mapping procedure as follows. First, kernel words were extracted from the PQAC word definitions. Second, senses of those kernel words were disambiguated. Finally, the mapping results between PQAC words and PWN synsets were obtained based on disambiguation of the kernel words. The rudimentary results comprise 35,476 common concepts between PQAC and English, together with 7779 specific concepts of PQAC. The mapping results cover 79.99 % of the PQAC senses in the GCD and reach a precision of over 85 %.

Acknowledgments We are grateful for the comments of the reviewers. This work is the staged achievement of the projects supported by National Social Science Foundation of China (10&ZD117, 12&ZD177) and Ministry of Education of China (16YJC740034).

References

Amsler, R. A. (1981). A taxonomy for English nouns and verbs. In: *Proceedings of the 19th annual meeting on association for computational linguistics* (pp. 133–138), Stanford, California.

Barbu, E., Mititelu, V. B., & Molini, S. D. (2007). Automatic building of wordnets. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Proceedings of recent advances in natural language processing IV* (pp. 217–226), Bulgaria: John Benjamins Publishing Company Borovets.

Bizzoni, Y., Boschetti, F., Diako, H., Gratta, R. D., Monachini, M., & Crane, G. (2014). The making of ancient Greek wordnet. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 1140–1147), Reykjavik, Iceland.

Bond, F., Isahara, H., Kanzaki, K., & Uchimoto, K. (2008). Boot-strapping a wordnet using multiple existing wordnets. In *Proceedings of the sixth international conference on language resources and evaluation* (pp. 1619–1624), Marrakech, Morocco.

Bond, F., Nichols, E., Fujita, S., & Tanaka, T. (2004). Acquiring an ontology for a fundamental vocabulary. In *Proceedings of coling 2004* (pp. 1319–1325), Geneva, Switzerland.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics, 18*, 13–47.

Commercial-Press (Ed.). (2005). *The great Chinese dictionary 2.0*. Hong Kong: The Commercial Press.

Eckard, E., Barque, L., Nasr, A., & Sagot, B. (2012). Dictionary-ontology cross-enrichment using TLFi and WOLF to enrich one another. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon* (pp 81–94). Mumbai, India: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/W12-5107

Esuli, A., & Sebastiani, F. (2007). Pageranking wordnet synsets: An application to opinion mining. In: *Proceedings of the 45th annual meeting of the association for computational linguistics*, Prague, Czech.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach. *Journal of Web Semantics, 9*, 434–452.

Fiser, D., & Sagot, B. (2008). Combining multiple resources to build reliable wordnets. In *Proceeding of text, speech and dialogue*, Brno, Czech.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305–332). Cambridge: MIT Press

HIT (ed). (2012). HIT IR-Lab Tongyici Cilin (extended). Harbin Institute of Technology. http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=viewpage&PAGE_id=162.

Hsu, M., Tsai, M., & Chen, H. (2008). Combining wordnet and conceptnet for automatic query expansion: A learning approach. In *Proceedings of the fourth Asia information retrieval societies conference* (pp 213–224), Harbin, China.

Kulkarn, M. (2010). Introducing Sanskrit wordnet. In *Proceedings of the fifth global WordNet meeting GWC*, Mumbai, India.

Lee, C., Lee, G. G., & Seo, J. (2004). Multiple heuristics and their combination for automatic wordnet mapping. *Computers and the Humanities, 38*, 437–455.

Li, L. (2004). *Bamboo and silk books and academic origin*. Beijing: SDX Joint Publishing Company.

Li, B., Xi, N., Feng, M., & Chen, X. (2012). Corpus-based statistics of pre-qin Chinese. In *Proceedings of CLSW 2012* (pp 145–153), Wuhan, China.

Lindén, K., & Carlson, L. (2010). Finnwordnet — WordNet på finska via översättning. *LexicoNordica — Nordic Journal of Lexicography* (K. Lindén, Trans.), *17*, 119–140.

Liu, X. Y., Li, B., Zhang, Y. J., & Liu, L. (2014). Quantitative research on the origins of contemporary Chinese vocabulary based on the Great Chinese Dictionary. In *Proceeding of 15th Workshop CLSW 2014* (pp. 112–123), Macao, China.

Mei, J. et al. (1983). In J. Mei, Y. Zhu, Y. Gao & H. Yin (Eds.), *Tongyici cilin* (1st ed.). Shangai Lexicographical Publishing House.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of workshop at ICLR*.

Mikolov, T., et al (2013c). Word2vec. URL https://code.google.com/p/word2vec/.

Monozzi, S. (2009). The Latin wordnet project. *Proceedings of the 15th International Colloquium on Latin Linguistics* (pp. 707–716). Austria: Innsbruck.

Navigli, R., & Lapata, M. (2010). An experimental study on graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Anaylsis and Machine Intelligence, 32*(4), 678–692.

Oliveira, H. G., & Gomes, P. (2013). On the automatic enrichment of a Portuguese wordnet with dictionary definitions. In *Advances in artificial intelligence, local proceedings of the 16th Portuguese conference on artificial intelligence* (pp. 486–497), Azores, Portugal.

Ordan, N., & Wintner, S. (2007). Hebrew wordnet: A test case of aligning lexical databases across languages. *International Journal of Translation, 19*(1), 39–58.

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2011). *Chinese Gigaword fifth edition LDC2011T13*. Philadelphia: Linguistic Data Consortium.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*, Mysore, India.

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics*, New York, USA.

Saveski, M., & Trajkovski, I. (2010). Automatic construction of wordnets by using machine translation and language modeling. In *Proceedings of seventh language technologies conference, 13th international multiconference information society* (pp 707–716). Ljubljana, Slovenia.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE international conference on semantic computing (ICSC2007)* (pp. 363–369). Irvine, CA.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria* (Vol. 1: Long Papers, pp. 455-465). Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-1045.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., & Milios, E. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on web information and data management*, Bremen, Germany.

Vossen, P. (Ed.). (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic.

Yu, J., & Yu, S. (2002). The structure of Chinese concept dictionary. *Journal of Chinese Information Processing, 16*(4), 12–20.