A Maximum Entropy Approach to HowNet-Based Chinese Word Sense Disambiguation

WONG Ping Wai
Intendi Inc.,
Clear Water Bay, Hong Kong.
wongpw@intendi.com

YANG Yongsheng
Department of Computer Science,
HKUST, Clear Water Bay, Hong Kong.
ysyang@cs.ust.hk

Abstract

This paper presents a maximum entropy method for the disambiguation of word senses as defined in HowNet. With the release of this bilingual (Chinese and English) knowledge base in 1999, a corpus of 30,000 words was sense tagged and released in January 2002. Concepts meanings in HowNet are constructed by a closed set of sememes, the smallest meaning units, which can be treated as semantic tags. The maximum entropy model treats semantic tags like parts-of-speech tags and achieves an overall accuracy of 89.39%, outperforming a baseline system, which picks the most frequent sense.

1. Introduction

A word usually has more than one meaning or sense, which are listed in the dictionary. The task of Word Sense Disambiguation (WSD) is to make the choice between the senses for a particular usage of the word in context. There are, however, several difficulties to WSD (Yang et al, 2000): (i) The evaluation of word sense disambiguation system is not yet standardized. (ii) The potential for WSD varies by task. (iii) Sense-tagged corpora are crucial resources for WSD but they are difficult to obtain. Efforts in building large Chinese corpora started in the 90s, for example, the Sinica corpus (CKIP, 1995) and the Chinese Penn Tree Bank (Xia et al., 2000). However, these two corpora concentrate on the tagging of parts-of-speech and syntactic structures, while little work has been done on semantic annotation. Of the few efforts that were carried out, Lua[1] annotated 340,000 words with semantic classes defined in a thesaurus (Mei, 1983). This resource, however, was not publicly accessible. With the release of HowNet (Dong, 1999; Dong, 2000) in 1999, Gan and Tham (1999) manually annotated a Chinese corpus of 30,000 words with the senses from HowNet. The corpus is a subset of the Sinica balanced corpus, and consists of 103 narratives on news stories, in which the words have already been segmented and tagged with parts-of-speech. Gan and Tham (1999) added sense tagging and subsequently Gan and Wong (2000) annotated the corpus with semantic dependency relations as defined in HowNet. The corpus was released to the public in January 2002[2], providing essential resources for Chinese word sense disambiguation.

This paper is organized as follows: Section 2 gives an introduction of HowNet. Section 3 describes the WSD task and the experiment results. Section 4 describes the previous work, followed by a conclusion in Section 5.

2. An Introduction to HowNet

HowNet is a bilingual general knowledge base that encodes inter-concept semantic relations and the inter-attribute semantic relations. In contrast to WordNet (Miller, 1990), HowNet adopts a constructive approach of meaning representation (Miller, 1993). Basic meaning units called sememes, which cannot be decomposed further, combine to construct concepts in HowNet. So far, there are 65,000 Chinese concepts and 75,000 English equivalents defined with a set of 1503 sememes.

NO.=the record number of the lexical entries
W_X=concept of the language X
E_X=example of W_X
G_X=Part-of-speech of the W_X
DEF=Definition, which is constructed by sememes and pointers

Figure 1: A sample lexical entry in HowNet.

Figure 1 gives an idea of how word concepts are organized in HowNet. "X" represents some

[1] http://www.cslp.com.nus.edu/sg/cslp/

[2] http://godel.iis.sinica.edu.tw/CKIP/hk/index.html

language and each language has three specific items: W_X, E_X and G_X. The current version of HowNet has entries in two languages (Chinese and English) with the possibility of extending it to other languages. Therefore, W_C, E_C and G_C would be entries for the words, the examples and the parts-of-speech respectively in Chinese, whereas W_E, E_E and G_E are the corresponding entries for English.

NO.=040263
W_C=記者
G_C=N
E_C=
W_E=journalist
G_E=N
E_E=
DEF=human|人,#occupation|職位,*gather|采集,
*compile|編輯,#news|新聞

Figure 2: An example entry in HowNet.

Figure 2 shows an example word, "journalist", as entered in HowNet. As mentioned in Miller (1993), the definition of a common noun typically consists of (i) its immediate superordinate term and (ii) some distinguishing features. HowNet represents this with pointers[3] and the order of the sememes in concept definitions. In the example above, the sememe appearing in the first position 'human|人' is called the **categorical attribute**. It names the hypernym or the superordinate term, which gives a general classification of the concept. The sememes appearing in other positions: 'occupation|職位', 'gather|采集', 'compile|編輯', 'news|新聞' are **additional attributes**, which provide more specific, distinguishing features. Two types of pointers are used in this concept. The pointer "#" means "related" and thus '#occupation|職位' shows that there is a relation between the word "journalist" and occupations. The pointer "*" means 'agent', and thus, '*gather|采集' and '*compile|編輯' tell us that "journalist" is the agent of 'gather|采集' and 'compile|編輯'. The sememe '#news|新聞' that follows tells us that the function of "journalist" is to compile and gather news.

[3] The function of pointers is to describe various inter-concept and inter-attribute relations. Please refer to HowNet's homepage (http://www.keenage.com) or Gan and Wong (2000) for details.

2.1. Classification of content words

Concepts of content words in HowNet are classified into six categories: Entity, Event, Attribute, Quantity, Attribute Value and Quantity value. The sememes in each category are organized hierarchically in an ontology tree. The six categories can be grouped into four main types: (i) Entity, (ii) Event, (iii) Attribute and Quantity, (iv) Attribute Value and Quantity Value. Most nominal concepts, such as "journalist", belong to the Entity category and some of them belong to the Attribute category. Verbal concepts always belong to the Event category whereas adjectives are Attribute Values.

2.1.1. Convention of meaning representation of content words

The first sememe in concept definitions indicates which of the four categories the concept belongs to, and it is therefore called the **categorical attribute**. For Attribute, Quantity, Attribute Value and Quantity Value, the first sememe clearly names the categories, as illustrated in (iii) and (iv) of Table 1. Table 2 shows an example entry: the category of "亮度" (brightness) is indicated by the first sememe 'Attribute|屬性'. The second sememe is a node in the hierarchy of Attribute or Quantity that names the subcategory. For example, 'brightness|明暗' is a node under the ontological hierarchy of 'Attribute|屬性'[4], and can be viewed as a subcategory of Attribute.

Table 1: An overview of the order of sememes in concept definitions of HowNet

Category		Sememes in concept definitions	
		Categorical Attribute	*Additional Attribute*
		1st position	2nd and thereafter position (optional)
(i)	Entity	node in Entity	"secondary feature" OR "node in (iv)" [5] OR "pointer""node in (i), (ii), (iii) or (iv) "

[4] Sememes are organized hierarchically so that brightness|明暗 is the hyponym of Attribute|屬性, in other words, brightness|明暗 is a kind of Attribute|屬性.

[5] (i) stands for Entity, (ii) Event, (iii) Attribute and Quantity, (iv) Attribute Value and Quantity Value. Secondary features include the sememes that cannot be categorized into types (i) – (iv).

(ii)	Event	node in Event	"secondary feature" OR *"event role"*="node in (i), (ii), (iii) or (iv)"	
Category		Sememes in concept definitions		
		Categorical Attribute		*Additional Attribute*
		1st position	2nd position	3rd position
(iii)	Attribute Quantity	attribute|屬性 quantity|數量	node in Attribute node in Quantity	&*"Host"* &*"Host"*
(iv)	aValue[6] qValue	aValue|屬性值 qValue|數量值	node in Attribute node in Quantity	*"Value"* *"Value"*

Table 2: Examples of concepts of the categories of Attribute and Attribute value

Concepts		Sememes in concept definitions		
		1st position	2nd position	3rd position
(iii)	亮度 (brightness)	attribute|屬性,	brightness|明暗,	&physical|物質[7]
(iv)	光亮 (bright)	aValue|屬性值,	brightness|明暗,	bright|明[8]

For the categories of Entity and Event, it is not necessary to name the main categories, because this information is conveyed by their subcategories. Table 3 shows two examples. The first sememe of "信箋" (letter paper) is 'paper|紙張', a node in the Entity hierarchy and its function is to indicate the subcategory of Entity. '留存|SetAside', as the first sememe of the concept "存款" (deposit money), names the subcategory of Event.

Table 3: Examples of concepts of the categories of Event and Entity

Concepts		Sememes in concept definitions		
		1st position	2nd position	3rd position
(i)	信箋 (letter paper)	paper|紙張,	@write|寫	
(ii)	存款 (deposit money)	SetAside|留存,	patient=money|貨幣,	commercial|商

2.1.2. Categorical Attribute

The categories of Attribute and Attribute Value share parallel subcategories. As an example, Table 2 shows one of them: the subcategory 'brightness|明暗'. Therefore, it is not adequate to identify only the subcategory when dealing with Attributes or Attribute Values. That is why these two categories (along with Quantity and Quantity Value) use the first two sememes for the subcategorization of concepts, whereas Entity and Event can achieve this by using the first sememe only. We call such types of sememes "**categorical attributes**".

2.2. Function Words

Unlike WordNet, HowNet has a sense inventory for function words, and thus our WSD system includes both content words and function words. For function words such as prepositions, pronouns and conjunctions, the sememes in the definitions are marked by curly brackets in order to distinguish senses of function words from those of content words. For example, the pronoun "他" (he) is defined as {ThirdPerson|他,male|男}.

3. Task Description

3.1. Preprocessing of the corpus

The HowNet corpus is written in XML format, and contains the part-of-speech, sense and semantic dependency relation information for each word. There are 30,976 word tokens and 3,178 sentences [9] in the HowNet corpus, which is divided into two sets in the experiment: 2,400 sentences (23,191 word tokens) are reserved for training, and 778 sentences (7,785 word tokens) for testing. Since off-the-shelf software systems usually have a default cut-off value that may not be appropriate for such a small corpus, we create a larger corpus by concatenating 3 copies of the training data. As a result, the final training corpus consists of 7,200 sentences (69,573 words).

3.2. Experiments

3.2.1. Maximum Entropy Tagger

The goal of this work is to investigate the possibility of applying standard POS taggers to identify word sense tags. For this work, an off-the-shelf maximum entropy tagger [10] (Ratnaparkhi, 1996) was used. Each word is therefore tagged with a sememe (categorical attribute), which is treated equivalently to a POS tag by the tagger, whose goal it is to generate a

[6] "aValue" stands for attribute value whereas "qValue" stands for quantity value.

[7] "," separates one sememe from the other in the definitions, and is not part of the sememe. "&" represents attriubte-host relation.

[8] "光亮"(bight) is a **value** of the attribute 'brightness|亮度'. "Value" is the terminal node of Attribute Value. It is optional in some cases.

[9] Sentences are delimited by the following punctuations ，。:；!？

[10] ftp://ftp.cis.upenn.edu/pub/adwait/jmx/jmx.tar.gz

sense tag dictionary from the training data. In the following subsections, we will first explain the semantic tags used in the current research, its limitations and suggestion for resolving the problem, and then illustrate how to build the tag dictionary for the MaxEnt sense tagger.

3.2.2. Using categorical attributes as semantic tags

As illustrated in section 2, there are about 65,000 concepts in HowNet dictionary, defined by 17216 sense definitions. The number of definitions will still increase in future, but the closed set of 1503 sememes is not likely to expand. Definitions are represented by a sequence of sememes in HowNet. It is possible to use the whole sequences of sememes as semantic tags, but the complexity can be greatly reduced by using the 1503 sememes as semantic tags.

As illustrated earlier, in HowNet, the category for a particular word concept is determined by the first sememe (for Entities and Events) or the first two sememes (for Attributes, Quantities or Attribute Values). These sememes are thus referred to as categorical attributes. On observation, it became apparent that just picking the categorical attribute would be enough to differentiate one sense from the other. For example, none of the 27 senses for the polysemous word "打" (hit) in Chinese share the same first sememe.

Using sememes as semantic tags has an advantage over using a simple sense id. Assigning a sense id such as 打$_1$, 打$_2$….打$_{27}$ to each sense of the word "打" can distinguish different senses but will not give us any idea of the meanings of the ambiguous words. Sememes convey meanings while helping to differentiate senses. For example, the first sense is 'associate|交往', which indicates an association with friends or partners. The second sense is 'build|建造', which is self-explanatory.

3.2.3. Limitation of the semantic tags

There is a limitation to this strategy. It is found that this strategy can discriminate the senses for about 90% of the words in the corpus. The remaining 10% of the words are still ambiguous (Table 4).

Table 4: Word tokens still have ambiguity after the tagging of categorical attribute

	Training	Testing
Total word tokens	69573	7785
Word tokens still have ambiguity after the tagging of categorical attribute	7461	878
Percentage	10.72%	11.28%

Table 5 shows the senses for the word "一" (one). Since all the senses are Quantities (qValue|數量值) and Attribute Value (aValue|屬性值) types, the categorical attribute is defined as the first two sememes. However, there is still ambiguity to be resolved for two of the senses.

Table 5: Senses for the word "一" (one)

Categorical Attribute	Sense
qValue|數量值, amount|多少	qValue|數量值,amount|多少, cardinal|基
qValue|數量值, amount|多少	qValue|數量值,amount|多少, single|單
aValue|屬性值, range|幅度	aValue|屬性值,range|幅度,all|全
aValue|屬性值, frequency|頻率	aValue|屬性值,frequency|頻率
aValue|屬性值, sequence|次序	aValue|屬性值,sequence|次序, ordinal|第

3.2.4. Mapping categorical attribute to sense definition

In this work, the ambiguity problem is solved by building a mapping table which maps the (word ; categorical attribute) pairs to sense definitions. First a frequency table is built, which accounts for the frequency of occurrence that a (word ; categorical attribute) pair should be mapped to a sense in the training corpus. Table 5 shows the categorical attributes for the word "一" (one). The 'qValue|數量值,amount|多少,cardinal|基' sense appears 145 times, while the 'qValue|數量值,amount|多少,single|單' sense appears only 16 times. In this work, we simply disregard the second sense for this situation, and assume that when the word "一" (one) is tagged with the categorical attribute 'qValue|數量值,amount|多少', it corresponds to the 'qValue|數量值,amount|多少,cardinal|基' sense in all contexts. There is a one-to-one direct mapping of the categorical

attributes to the 3rd, 4th and the 5th senses, so frequency information is not needed for them.

3.2.5. Sense Tag dictionary for MaxEnt Tagger

Section 3.2.4 illustrates the mapping of a sense tag to a sense definition, and this section will briefly describe the building of the tag dictionary. There are two sources for the sense tag dictionary. One comes from the training corpus and one from the HowNet dictionary. The MaxEnt tagger automatically creates a tag dictionary from the training corpus. By default, this dictionary only includes words that appear more than four times in the training corpus (total 753 word types). [11] Another source is the HowNet dictionary, which has the information of semantic tags for 51275 word types. The two sources of information are combined in the sense tag dictionary for the maximum entropy tagger.

3.3. Testing results

The input of the testing component is the testing corpus, which is already segmented. The output is the most likely senses of words given by the WSD systems.

3.3.1. Baseline system

As a baseline system, the most frequent sense (MFS) of a word is chosen as the correct sense. The frequency of word senses is calculated from the occurrences of the word senses in the training corpus, with ties broken randomly. For all instances of unknown words, the baseline system just tags them with the most frequent sense for the rare words (that is, 'human|人,ProperName|專' as shown in Table 7).

3.3.2. Maximum entropy

The model first checks if the word in context can be found in HowNet dictionary. In case the word has only one sense in the dictionary, there is no need to perform disambiguation for this word and the system returns this sense as the answer.

For words with more than one sense, the maximum entropy model chooses one (categorical attribute) from the closed set of sememes. The categorical attribute is mapped to the corresponding sense according to the mapping table.

Table 6 shows the results for both the baseline and the maximum entropy system. It can be seen that the MaxEnt tagger achieves an accuracy of 88.94%, which outperforms that of the baseline system. An upper bound can also be calculated by imagining that we could employ an oracle system that would indicate, for each ambiguous semantic tag (described in Section 3.2.4), the correct sense of the word. In that case, the performance of the maximum entropy tagger would improve to 89.73%.

Table 6: The accuracy rate of MFS and MaxEnt for overall, polysemous and unknown word

		Accuracy	
		MFS	MaxEnt
Performance	Overall	84.63%	88.94%
	Unknown	45.83%	72.50%
	Polysemous	69.65%	77.33%
Semantic tag (categorical attribute only, effective upper bound)	Overall	86.48%	89.73%
	Unknown	46.39%	75.00%
	Polysemous	71.72%	77.42%

Table 7: Sense distribution and tagging accuracy of unknown words

Sense	Freq.	Accuracy	
		MFS	MaxEnt
人,專	165	100%	95.15%
數量值,次序,第	84	0%	96.43%
地方,專,(臺灣)	28	0%	75.00%
數量值,多少,基,眾	31	0%	93.55%
文字,姓名,專	20	0%	40.00%
Other senses	30	0%	3.33%
Total	360	45.83%	82.50%

Even though it does not look like the maximum entropy tagger outperforms the baseline system by much, it should be noted that the nature of the corpus makes the task simple for the baseline system. Since the corpus is composed of a collection of news stories, certain senses of polysemous words will tend to appear more often in the corpus --- indeed, it was observed that more than half of the word tokens appearing in the training and testing corpus have only one sense. The average sense per word token is 1.14 and 1.09 in the training and the testing sets, respectively. However, it should be noted that the MaxEnt

[11] Words occurring less than 5 times in the training corpus are treated as rare words. The tagging of rare words are illustrated in section 3.3.

model performs much better on polysemous words and unknown words, which bodes well for using the MaxEnt model with more diverse corpora.

Table 8: Average senses per word in the training data and the testing data

	Training	Testing
word tokens	69,573	7,785
word tokens with one sense only	4,2990 (61.78%)	4,905 (63.01%)
average sense per word token	1.14	1.09

One of the strengths of maximum entropy lies in its ability to use contextual information to disambiguate polysemous words and predict the senses of unknown words. The following shows an unknown word "廿八" with the context information:

Table 9: Example of an unknown word: "廿八"[12]

	previous	current	next
Word	元月	廿八	日
Tag	time|時間	Unknown	

Table 10: Features and possible tags of the unknown word "廿八"

Features	Possible tags of current word "廿八"
prefix is "廿"(twenty)	qValue|數量值,sequence|次序
suffix is "八"(eight)	qValue|數量值,sequence|次序
next word is "日"(day)	qValue|數量值,sequence|次序 *OR*
	qValue|數量值,amount|多少
previous tag is 'time|時間'	qValue|數量值,sequence|次序 *OR*
	qValue|數量值,amount|多少 *OR*
	time|時間

The MaxEnt tagger defines a set of feature patterns including the previous word, the next word, the previous tag, the prefix and the suffix of current word. In this example, the features extracted from the context are shown above. Accordingly, the MaxEnt tagger predicts 'qValue|數量值,sequence|次序' as the most likely sense tag for the word "廿八". The tag 'qValue|數量值,sequence|次序' is then mapped to the sense definition 'qValue|數量值,sequence|次序,cardinal|第' according to the mapping table.

[12] The meaning of the phrase 元月廿八日 is "the twenty-eighth of January". The correct sesne of "廿八" is 'ordinal number', defined by 'qValue|數量值,sequence|次序,cardinal|第' in HowNet.

4. Previous Work

To our knowledge, there currently exist three previous studies of word sense disambiguation using HowNet. Yang et al (2000) pioneered this work by using sememe co-occurrence information in sentences from a large corpus to achieve an accuracy of 71%. Yang and Li (2002), collecting sememe co-occurrence information from a large corpus, transferred the information to restricted rules for sense disambiguation. They reported a precision rate of 92% and 82% for lexical disambiguation and structural disambiguation, respectively.

Wang (2002) pioneered the work of sense pruning using the hand-coded knowledge base of HowNet. Unlike sense disambiguation, sense pruning seeks to narrow down the possible senses of a word in a text. Using databases of features such as information structure and object-attribute relations which were compiled from HowNet, Wang reported a recall rate of 97.13% and a per sentence complexity reduction rate of 47.63%.

The current study and Wang (2002) used the sense tagged HowNet corpus with different approaches. There is one similarity between our work and Wang (2002), though. Wang applied a sense pruning method to reduce the complexity of word senses. The strategy of the current study reduces the complexity of sense tagging by using the categorical attributes (first or the first two sememes) as semantic tags. About 10% of the words are still ambiguous, but the ambiguity can be reduced in future studies which extend to the tagging of the sememes in the third and the thereafter position of concept definitions. It is also interesting to see if the ambiguity can be resolved by integrating a diverse set of various knowledge sources, such as HowNet knowledge bases, sememe cooccurrence database and the tagged corpus.

5. Conclusion

This paper has presented the method of maximum entropy to perform word sense disambiguation in Chinese with HowNet senses. The closed set of sememes is treated as semantic tags, similar to

parts-of-speech tagging in the model. Our system performs better than the baseline system that chooses the most frequent sense. Our strategy of sememe tagging reduces the complexity of semantic tagging in spite of some limitations. Some possible ways to resolve the limitations are also suggested in the paper. Unlike the work of Yang et al (2000) and Wang (2002) that applied unsupervised methods using sense definitions in HowNet, the paper is the first study to use a supervised learning method with the availability of the HowNet sense tagged corpus. Much research remains to be done on the corpus and the HowNet knowledge base to get further improvement on the WSD task.

6. Acknowledgement

Our thanks go to Dr. Grace Ngai for her helpful comments. This work was supported and funded by Intendi Inc.

7. References

CKIP (1995) *The Content and Illustration of Sinica Corpus of Academia Sinica, Technical Report no. 95-02* (*中央研究院平衡語料庫的內容與說明, 技術報告 95-02*), Institute of Information Science, Academia Sinica.

Dong, Zhendong (1999) *Bigger Context and Better Understanding – Expectation on Future MT Technology*. In Proceedings of International Conference on Machine Translation and Computer Language Information Processing, 26-28 June, 1999, Beijing, China, pp. 17-25.

Dong, Zhendong (2000) *HowNet:* http://www.keenage.com.

Gan, Kok-Wee and Wai-Mun Tham (1999) *General Knowledge Annotation Based on HowNet* (*基於知網的常識知識標註*). Computational Linguistics and Chinese Language Processing, vol. 4, 1999, pp. 39-86.

Gan, Kok-Wee and Ping-Wai Wong (2000) *Annotating Information Structures in Chinese Text using HowNet.* In Proceedings of the 2nd Chinese Language Processing Workshop, Association for Computational Linguistics 2000 Conference, October 2000, Hong Kong, pp. 85-92.

Mei, Jiaju, Yiming Lau, Yunqi Gao, Yongxiang Ying (1983) *A Dictionary of Synonyms* (*同義詞詞林*), Shanghai Cishu Chubanshe.

Miller, George A. (1990) *WordNet: An Online Lexical Database.* In Special Issue of International Journal of Lexicography, Vol 3, No. 4.

Miller, George A. (1993) *Nouns in WordNet: a lexical inheritance system.* Five Papers on WordNet, CSL Report 43, Cognitive Science Laboratory, Princeton University.

Ratnaparkhi, Adwait (1996) *A Maximum Entropy Model for Part-of-Speech Tagging*. In Proceedings of the First Empirical Methods in Natural Language Processing Conference, pp. 133-141, Philadelphia, USA.

Yang, Erhong, Guoqing Zhang and Yongkui Zhang (2000) *The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of HowNet*. In Proceedings of the second Chinese language processing workshop, ACL 2000 Conference, October 2000, pp.60-65.

Wang, Chi-Yung (2002) *Sense Pruning by HowNet – a knowledge-based Word Sense Disambiguation.* MPhil Thesis. Hong Kong University of Science and Technology.

Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch and Mitch Mrcus (2000) *Developing Guidelines and Ensuring Consistency for Chinese Text Annotation.* In Proceedings of the second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.

Yang, Xiaofeng, Tangqiu Li (2002) *A Study of Semantic Disambiguation Based on HowNet*, International Journal of Computational Linguistics and Chinese Language Processing, vol. 7, no.1 , 2002, pp.47-78.