Annotating information structures in Chinese texts using HowNet

GAN Kok Wee
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
gankw@cs.ust.hk

WONG Ping Wai
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
wongpw@cs.ust.hk

Abstract

This paper reported our work on annotating Chinese texts with information structures derived from HowNet. An information structure consists of two components: HowNet definitions and dependency relations. It is the unit of representation of the meaning of texts. This work is part of a multi-sentential approach to Chinese text understanding. An overview of HowNet and information structure are described in this paper.

1 Introduction

Corpora are essential resources to any research in language engineering. For Chinese, efforts in building large corpora started in the 90s. For instance, the PH corpus of 4 million Chinese characters with word boundary information was released in 1993 (Guo, 1993). The first version of the Sinica corpus of two millions words marked with word boundaries and parts-of-speech was released in 1995 (CKIP, 1995). In 1996, a small corpus of 5266 distinct words (inclusive of punctuation marks) with a total occurrence frequency of 51870 was released (Yu et al., 1996). This corpus was derived from the Singapore Primary School Chinese Language Textbooks. It contained information on word boundaries, parts-of-speech and also syntactic structures. In 2000, two additional bracketed corpora have just been announced. The first one, the Chinese Penn Treebank, includes 100-thousand words (Xia et al., 2000). The second one, the Sinica Treebank, which is derived from the Sinica corpus, contains 38,725 sentences with 1000 of them released to the public[1] (CKIP, 2000).

The historical development of Chinese corpus construction has shown a consensus in incorporating more powerful linguistic structures into corpora. As noted by Marcus (1997), the more powerful linguistic structures will help in improving the accuracy of parsing. This is especially true to isolating language such as Chinese. However, there is very little work on annotating corpora with semantic information. To the best of our knowledge, there is only one report of this kind. The work by Lua[2] annotated 340,000 words with semantic class information as defined in a thesaurus of synonyms (Mei, 1983). With the release of HowNet[3], a bilingual general knowledge base, Gan and Tham (1999) reported the first corpus of 30,000 words that was annotated with the general knowledge structure defined in HowNet. This paper reported an extension of the work in Gan and Tham (1999) on the annotation of information structures in Chinese texts. In Section 2, an overview of HowNet is provided. Information structure and an illustration will be given in Section 3.

2 An Overview of HowNet

HowNet is a bilingual general knowledge-base describing relations between concepts and relations between the attributes of concepts. The latest version covers over 65,000 concepts in

[1] **http://godel.iis.sinica.edu.tw/CKIP/trees1000.htm**

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

[3] **http://www.HowNet.com (Dong Zhendong, Dong Qiang; HowNet)**

Chinese and close to 75,000 English equivalents. The relations include hyponymy, synonymy, antonymy, meronymy, attribute-host, material-product, converse, dynamic role and concept cooccurrence. The philosophy behind the design of HowNet is its ontological view that all physical and non-physical matters undergo a continual process of motion and change in a specific space and time. The motion and change are usually reflected by a change in state that in turn, is manifested by a change in value of some attributes. The top-most level of classification in HowNet thus includes: entity, event, attribute and attribute value. It is important to point out that the classification is derived in a bottom-up manner. First, a set of sememes, the most basic set of semantic units that are non-decomposable, is extracted from about 6,000 Chinese characters. This is feasible because each Chinese character is monosyllabic and they are meaning-bearing. Similar sememes are grouped. The coverage of the set of sememes is tested against polysyllabic concepts to identify additional sememes. Eventually, a total of over 1,400 sememes are found and they are organized hierarchically. This is a closed set from which all concepts are defined. The bottom-up approach takes advantage of the fact that all concepts, either current or new, can be expressed using a combination of one or more existing Chinese characters. It is yet to find a new concept that has to resort to the creation of a new Chinese character. Therefore, by deriving the set of sememes in a bottom-up fashion, it is believed that the set of sememes is stable and robust enough to describe all kinds of concepts, whether current or new. The fact that HowNet has verified this thesis over 65,000 concepts is a good proof of its robustness.

2.1 Types of Relation

The definition of a concept in HowNet expresses one or more of the following relations.

2.1.1 Dynamic Role

There are a total of 71 dynamic roles defined in HowNet. Dynamic role resembles case role in case grammar (Fillmore, 1968). However, it differs from case role in that it is concerned with all probable actants of an event and the roles they play in the event. The issue of whether these actants can be realized grammatically is not its concern. For example,

> Concept(1): 吃齋 (be a vegetarian for religious reasons)
> DEF=eat|吃, patient=vegetable|蔬菜, religion|宗教

At the syntactic level, "吃齋" is an intransitive verb. According to case grammar, it has only one case role: ***agent***. However, for this word, the ***patient*** is self-contained in its constituent (i.e. "齋"). HowNet specifies this explicitly and indicates the category (*'vegetable'*[4]) of prototypical concepts which fills up this role.

Another distinguishing feature of dynamic role is its use in defining concepts of '*entity*' class.

> Concept(2): 毛筆 (writing brush)
> DEF=PenInk|筆墨, *write|寫

Through the use of the "*" pointer, the above definition states that the concept being defined (毛筆) is the instrument of the event type 寫 '*write*'.

HowNet also uses dynamic role to specify the attributes that a concept contains. For example,

> Concept(3): 突如其來 (arise suddenly)
> DEF=happen|發生, manner=sudden|驟

The definition of concept (3) specifies that the manner of the event is *'sudden'*.

2.1.2 Hyponymy Relation

The '*event*' and '*entity*' classes in HowNet are organized in a hierarchical manner. The parent class is a hypernym of its children classes. Details of the organization are available from the HowNet site and are therefore omitted here.

2.1.3 Meronymy Relation

Meronymy relation is expressed through the

[4] We use single-quote and italic to mark sememes in HowNet.

pointer “%”. For example,

Concept (4): 中央處理器 (CPU)
DEF=part|部件, %computer|電腦, heart|心

The class of the concept “中央處理器” is ‘part’. It is a part of the class ‘computer’. The function of the part “中央處理器” is the ‘heart’ of the whole ‘computer’.

2.1.4 Material-Product Relation

Material-product relation is expressed through the pointer “?”. For example,

Concept (5): 毛線 (knitting wool)
DEF=material|材料, ?clothing|衣物

“毛線” belongs to the class ‘material’. It is a material for the product ‘clothing’.

2.1.5 Attribute-Host Relation

Attribute-host relation is expressed by the pointer “&”. For example,

Concept (6): 面子 (face)
DEF=attribute|屬性, reputation|名聲, &human|人, &organization|組織

“面子”is an attribute; in particular, it is about the attribute ‘reputation’. The hosts could be ‘human’ as well as ‘organization’.

2.1.6 Concept Co-occurrence Relation

Some concept typically co-occurs with certain concept. For example,

Concept (7): 不法之徒 (lawless person)
DEF=human|人, fierce|暴, crime|罪, #police|警, undesired|莠

The typical context where the concept “不法之徒” is used involves the concept ‘police’. This type of relation is expressed by the pointer “#”.

3 Information Structures

Dong (2000) uses the example “毒品走私集團” (Narcotic drugs smuggling group) to illustrate what information structure is. Describing the structure of this phrase at the syntactic level, such as the analysis of Penn Treebank (Xue, 1999: 72-77), only reveals that it is a noun phrase with the head of “集團” modified by a relative clause “毒品走私” which involves operator movement. At the semantic level of description, we would indicate that “集團” (group) is the ***agent*** of the event “走私” (smuggle) and “毒品” (Narcotic drugs) is the ***patient*** of “走私” (smuggle). The informaton structure of this example consists of two parts, the dependency relations and the HowNet definitions. The descriptions are as follows:

Dependency relations: 毒品[patient]←走私←[agent]集團

Definitions:
毒品: medicine|葯物, ?addictive|嗜好物
走私: transport|運送, manner=secret|秘, crime|罪
集團: community|團體

In this example, the descriptions specify that a ‘*community*’ is an ***agent*** involved in a ‘*transport*’ event transporting the ***patient*** ‘*medicine*’. Furthermore, the ‘*transport*’ event is a ‘*crime*’ and the manner is ‘*secret*’. The ‘*medicine*’ is a material of ‘*addictive*’ products. The arrow between two concepts is a dependency connection with the concept pointed to by the arrow denoting the dependent and the concept at the other end as the governor. The name of the dependency relation is enclosed in a square bracket and it could appear at either the dependent or the governor side.

Currently, over 60 types of information structure have been defined. The pattern of information structure is specified in the following format: (sememe) [DRel] → [DRel] (sememe), where *DRel* means the name of a dependency relation. For the dependency relation to apply, the governor and the dependent must satisfy the requirement of the sememes. Table 1 shows a

subset of the information structures. Information structures are derived in a bottom-up fashion from analysing the mechanisms used in the composition of words. This approach is based on the insight that mechanisms used in word formation are also applicable to phrase and sentence construction in Chinese. For example, the type "(時間|time) [時間|time] ← (事件|event)" applies to the formation of the following units at various levels of linguistic structure:

word level:	"午←睡" (afternoon nap)
phrase level:	"暑期←補習" (summer study)
sentence level:	"長期←商品短缺" (long-time shortage of commodities) "1999年12月9日星期四←發生洩漏" (leaking occurs on Thursday, December 9, 1999)

In the process of annotating the corpus, the coverage of information structure types at the phrase and sentence levels was evaluated and missing types are added. The new types arise mainly due to function words. For example, the type "(modality|語氣) [modality|語氣] ← (事件|event)" is due to the use of function words such as "務必" (must) and "應該" (must). These are words expressing the attitude of the speaker of an utterance towards an event.

3.1 An example

We annotated a subset of the Sinica corpus (version 3.0) of 30,000 words with information structures. The corpus includes 103 newspaper texts covering the crime domain. The annotation has been completed and is currently under verification. We expect to release the corpus and the annotation guideline at the end of this year. An example of our annotation is shown below and its information structures are shown in Figure 1 at the end of this paper. The difference between this work and the work reported in Gan and Tham (1999) lies in the addition of the dependency relations into the annotation.

(1) 台南 縣 新化 警分局
Tainan county Xinhua police branch

刑事組 小 隊長
criminal group junior captain

林文政 昨天 下午
Lin Wenzheng yesterday afternoon

舉 槍 自戕 後 ，[5]
raise gun suicide after,

"After Lin Wenzheng, a junior captain with the Criminal Investigation Department of the Xinhua police branch of Tainan county, committed suicide by shooting himself yesterday afternoon,"

The hierarchical structure in Figure 1 is another way to represent the relation between governor and dependent, as illustrated in Figure 2. C1 immediately dominates C2, indicating that C1 is the governor and C2 the dependent. The relation between them is either R1 or R2. R1 is located at the same level as C1 and R2 is located at the same level as C2. These two possibilities could also be represented linearly as shown in (2).

Concept	Relation	Concept
C1	R1	
	R2	C2

Figure 2: Relations between concepts

(2) C1 [R1] → [R2] C2

R2 between the two concepts C1 and C2 should be read as "C2 is the R2 of C1". For example, "下午" (afternoon) is the ***time*** of "舉" (raise). R1 between C1 and C2 should be interpreted as "C1 is the R1 of C2". For example, the "time" between "後" (after) and "自戕" (suicide) should be interpreted as "後" is the ***time*** of "自戕".

The HowNet definitions of the concepts in (1) are provided in Table 2:

[5] **A string of Chinese characters ending with a punctuation mark is regarded as a unit for information structure annotation.**

Table 2: An example of HowNet definitions

Concept	Definition
台南	place|地方, city|市, ProperName|專, (China|中國)
縣	place|地方
新化	place|地方, ProperName|專, (Taiwan|台灣)
警分局	institution|機構, police|警, branch|支
刑事組	part|部件, %institution|機構, police|警
小	aValue|屬性值, importance|主次, secondary|次
隊長	human|人, #occupution|職位, official|官
林文政	human|人, ProperName|專
昨天	time|時間, past|昔, day|日
下午	time|時間, afternoon|午
舉	lift|提升
槍	weapon|武器, *firing|射擊
自戕	suicide|自殺
後	time|時間, future|將
，	{punc|標點}

The structures in Figure 1 and Table 2 reveal the following information:

(a) example (1) is about the time after a '*suicide*' event;
(b) preceding the 'suicide' event is the event '*raise*';
(c) the ***time*** of the '*raise*' event is "昨天下午", the ***agent*** is "林文政" and the ***patient*** is a '*weapon*';
(d) the '*occupation*' of "林文政" is "隊長" which is an '*official*' of '*secondary*' importance and "林文政" belongs to the '*institution*' "警分局刑事組";
(e) the location of the '*institution*' "警分局刑事組" is at "台南縣新化".

This kind of representation enables a computer to analyse texts at a deeper level of understanding. As an English and Chinese bilingual common-sense knowledge system, HowNet can contribute much to better text understanding and machine translation (Dong 1999).

4 Conclusion

The work reported here constitutes part of our efforts to develop a new strategy for Chinese text understanding. The strategy was proposed by Dong (1999). It starts with tagging each concept with the most probable HowNet definitions. The second step is to determine the information structures as described in this paper. The last step is to recover the implicit information structures from the surface information structures based on two additional knowledge sources: (i) relations between the event types as defined in HowNet; and (ii) rules governing the interplay of dynamic roles between event types. For example, the '*suicide*' event in example (1) does not mention its agent directly. The means of committing '*suicide*' by '*firing*' is also implicit. The recovery of this information will be the main issues to be resolved at the final stage. In parallel with our annotation effort, we are also working on developing automatic algorithms for the disambiguation of HowNet definitions and the identification of information structures. The creation of the two additional knowledge bases will be our future plan.

Acknowledgements

This work was supported by the Hong Kong Research Grant Council. We would also like to thank Dong Zhendong. Without his advice, this work would not be possible.

References

CKIP (1995) *中央研究院平衡語料庫的內容與說明, 技術報告95-02* [Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica]. Institue of Information Science, Academia Sinica.

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

Dong Zhendong (2000) *中文信息結構模式* [The pattern of Chinese information structure], ms.

Fillmore C. J. (1968) *The case for case*. In "Universals in Linguistic Theory", E. Bach, R. Harms, eds., New York, Holt, Rinehart and Winston.

Gan, Kok-Wee and Wai-Mun Tham (1999) *基於知網的常識知識標註* [General Knowledge Annotation Based on HowNet], Computational Linguistics and Chinese Language Processing, vol. 4, 2, pp. 39-86.

Guo, Jin (1993) PH: A Chinese corpus. Communications of COLIPS, 3/1, pp. 45-48.

Marcus, Mitch (1997) Invited speech at the 5th Workshop on Very Large Corpora.

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

Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch and Mitch Marcus (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.

Xue, Nianwen, Fei Xia, Shizhe Huang, Anthony Kroch (1999) *The Bracketing Guidelines for the Penn Chinese Treebank (Draff II)*, http://www.upenn.edu/ctb/.

Yu, Shiwen, Qiang Zhou, Wei Zhang, Yunyun Zhang, Weidong Zhan, Baobao Chang, Zhifang Sui (1996) *Word segmented and POS tagged 12 volumes of Singapore Chinese primary school text*. Communications of COLIPS, 6(1), pp. 41.

Table 1: A subset of information structures

Information structures of Coordination

	Type	SYN_S	Dependent	R2		Head	Example
1	Entity	N←N	Entity	[coordinate]	←	Entity	花-草，花-卉，禽-獸，風-浪，姓-名，官-兵，師-生，師-徒，部-委, 飯-菜，夫-婦
		N←N	Part	[coordinate]	←	Part	皮-肉，骨-肉，手-腳，科-室，委-辦，章-節
2	Attribute Value	A←A	Attribute Value	[coordinate]	←	Attribute Value	鮮-嫩，白-淨，富-強，窮-困，清-新，老-弱，物美-價廉，天真-活潑，年富-力強
3	Event	V←V	Event	[coordinate]	←	Event	領-養，醫-護，醫-患，調-控，調-研，教-學，餐-飲，哭-喊，哭-訴，生-死，存-亡

Information structures of Patient

	Type	SYN_S	Head	R1		R2	Dependent	Example
1	Head initial	V→N	Event		→	[patient]	Thing	吃-飯，賣-國，理-髮，染-髮，害-人，害-己，殺-敵，輸-血，侵-權，侵-華
		V→N	Event, Act, Pass		→	[patient]	Time	度-日，過-年，過-冬，過-夜，越-冬，享-年，共度-難關，度過-難關
		V→N	Event		→	[patient]	Event	深化-改革，改善-供應，加強-管理，加強-領導，加強-團結，加強-學習
		V→N	Event, Act		→	[patient]	Attribute Value	刪-繁，就-簡，去-舊，增-新，吐-故，納-新，獎-勤，罰-懶，反-腐，倡-廉

	Type	SYN_S	Dependent	R2		R1	Head	Example
2	Head final	V←N	Event, Act		←	[patient]	Thing	僱-員，選-手，囚-犯，囚-徒，愛-人，棄-嬰，棄-婦，使-女，遺-言，遺-願
		V←N	Event		←	[patient]	Thing	獵-物，產-物，造-物，建築-物，產-品，制-品，廢-品，批-語，製成-品，半成-品
		N←V	Thing/ Part	[patient]	←		Event, Act	蔬-食，貨-運，客-運，水-污染, 文物-走私，毒品-走私，貨物-運輸，旅客-運輸

Figure 1: A Graphical View of Information Structures