Building a Conceptual GraphBank for Chinese Language

Ji Donghong, Tang Li, Yang Lingpeng

Institute for Infocomm Research
Singapore, 119613
{dhji, tangli, lpyang}@i2r.a-star.edu.sg

Abstract

In this paper, we introduce a new semantic resource, i.e., conceptual graph bank, for Chinese language. The resource is to provide conceptual information for each sample sentence, which includes concepts, conceptual relationship, head, and conceptual hierarchies. Compared with treebanks, the resource focuses on static semantic information. We also introduce recursive graph to denote the conceptual information.

1. Introduction

Treebank has long been a research focus in NLP since its inception (Marcus et al., 1993), and has been applied in many NLP areas, ranging from language modeling, grammar inference, and statistical parsing to information extraction, etc. In particular, the design and construction of treebanks themselves still remains an active topic up to now, and many treebanks for various languages have been built (e.g., Hajic, 1998; Stegmann, 1998), including three versions for Chinese language (Chen et al., 1999; Xia et al., 2000; Okurowski, et al., 1998). In this paper, however, we suggest building a more semantic alternative to Chinese treebank, i.e., conceptual graphbank, which is based on specific features of Chinese language and more powerful formal representation tools, i.e., recursive graphs.

From the point of view of semantic analysis, the rationale behind the construction of treebank lies in syntactic structures and their relation with semantic interpretations. According to Chomsky, associated with any linguistic form is a tree-like syntactic structure, with each node inside being labeled by some syntactic category, e.g., *S*, *NP*, *VP*, *N*, *V*, etc., and furthermore, with some interpretation rules, one can derive the semantic interpretation of the linguistic. However, both syntactic structure determination and its contribution to semantic interpretation meet challenges in the analysis of Chinese language.

One main application of treebanks is grammar learning or parser training due to its annotated syntactic information and the generalization over part-of-speeches or sub-part-of-speeches. The learned grammars or parsers can be applied to other phrases or sentences (Black, et al., 1996; Bikel et al., 2000; Charniak, 1996). Now, with large-scale Chinese thesaurus available (e.g., Mei et al. 1983), a natural extension of this idea is that if we can annotate example phrases or sentences with some semantic information somehow, we can use similar learning schemes to carry out semantic analysis directly based on the generalization over synsets. If so, we can avoid the difficulties of syntactic analysis to do some semantic analysis.

An immediate problem is to determine what kind of semantic information we should annotate. An unambiguous linguistic form in semantic level is always associated with unique static information but possibly with different dynamic information according to different semantic composition rules. In order to ensure the uniqueness of our semantic annotation for one linguistic form, we focus on the static semantic information, while discarding the dynamic combining procedures.

Another problem is to determine what kind of formal tools are appropriate for annotating the static semantic information. Normally, dependent tree is a traditional choice to describe semantic information by depicting binary relationship between concepts (Mel’čuk, 1988 Samuelsson, 2000). However, for Chinese language, its expression capability is in doubt, so in this paper, we propose a kind of more powerful formal tools to describe the static semantic information (See section 4).

The remainder of this paper is organized as the following. In section 2, we present a conceptual view of Chinese language. In section 3, we provide a formal framework to encode the conceptual information. In section 4, we give our conclusion.

2. Conceptual View of Chinese Language

2.1 Concept and Conceptual Relatedness

In general, a Chinese phrase or sentence describes some concepts and their relationships. Roughly, a *concept* is what a linguistic form, e.g., a word, a phrase or a sub-clause, denotes. In particular, a *lexical concept*, i.e., the concept denoted by a word can be represented by the synset that contains the word in a thesaurus due to the assumption that synonyms in a synset refer to the same concept (Dong, 2000; Fellbaum et al., 1998). On the other hand, a concept denoted by a phrase or a sentence is a *compound concept*. As an example, in 1), the three words 积极(actively, /jiji/), 走私(smuggle, /zousi/), and 汽车(car, /qiche/) denote lexical concepts respectively, while the phrases 积极走私(to actively smuggle, /jiji zousi/), 走私汽车(to smuggle cars, /zousi qiche/), and the whole phrase all denote compound concepts.

1) 积极走私汽车(actively smuggling cars, /jiji zousi qiche/)

In addition, the concept denoted by a sub-phrase or a sub-sentence is a *sub-concept* of that denoted by the phrase or the sentence. For example, 积极[1] (actively, /jiji/) and 走私(smuggle, /zousi/) are sub-concepts of 积极走私(to actively

[1] We use a word or phrase to denote its concept if no confusion occurs.

smuggle, /jiji zousi/), while 积极走私(to actively smuggle, /jiji zousi/) and 走私汽车(to smuggle cars, /zousi qiche/) are both sub-concepts of the whole phrase.

The conceptual relationship between concepts is normally defined manually, for example, 2) lists the conceptual relationship between the three lexical concepts in 1), which means that 积极 (actively, /jiji/) is the Manner of 走私(to smuggle, /zousi/), while 汽车(car, /qiche/) is the Patient of 走私(to smuggle, /zousi/).

2) Manner (积极, 走私); Patient (汽车, 走私)

So far, many kinds of semantic relationships have been proposed, e.g., Agent, Patient, Time, Locations, etc. (Dong et al., 1998). One common ground is that they all somewhat reflect the *relatedness* between two concepts, and different semantic relationship depicts *different relatedness*.

One may argue that the words in a context such as a phrase or a sentence may have more or less semantic relationship with each other, as is more evident in 3).

3) 外貌漂亮的姑娘(the girl with beautiful looking, /waimao piaoliang de guniang/)

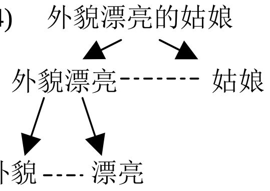
The three content words 外貌(appearance, /waimao/), 漂亮(beautiful, /piaoliang/) and 姑娘(girl, /guniang/) in 1) are mutually semantically related. Normally, an adjective acts as a *value* of some *feature* of a noun (Fellbaum et al., 1998). Thus in 3), 漂亮(beautiful, /piaoliang/) is a value of the feature 外貌(appearance, /waimao/) of the noun 姑娘(girl, /guniang/). So, the relatedness between 漂亮(beautiful, /piaoliang/) and 姑娘(girl, /guniang/) is via the relatedness between 漂亮(beautiful, /piaoliang/) and 外貌(appearance, /waimao/), and the relatedness between noun 姑娘(girl, /guniang/) and 外貌(appearance, /waimao/). In this sense, both 漂亮(beautiful, /piaoliang/) and 姑娘(girl, /guniang/) have a *direct relationship* with 外貌(appearance, /waimao/), while they themselves are *indirectly related*. However, in 14) due to the absence of the feature word, 漂亮(beautiful, /piaoliang/) and 姑娘(girl, /guniang/) are *directly related*.

Direct relatedness also applies between compound concepts or between lexical concepts and compound concepts. For example, in 1), the lexical concept 积极(actively, /jiji/) directly relates with the compound concept 走私汽车(to smuggle cars, /zousi qiche/), while the lexical concept 汽车(car, /qiche/) directly relates with the compound concept 积极走私(to actively smuggle, /jiji zousi/). Similarly in 13), the lexical concept 姑娘(girl, /guniang/) directly relates with the compound concept 外貌漂亮(beautiful appearance, /waimao piaoliang/).

2.2 Conceptual structure

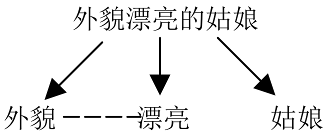
Seeing a phrase or a sentence as a compound concept, we can make a *partition* of it into some *sub-concepts* with direct relatedness holding between them, and a sub-concept, if still being a compound concept, can be further partitioned, and so on, until no concepts can be partitioned. Then, we get a concept hierarchy which we call *a conceptual structure* of the phrase or sentence. For example, 4) is a conceptual structure of 3), in which arrow line stands for constituency by partition and dot line stands for direct relatedness.

4)



Since a compound concept may have multi-partitions, one phrase or sentence may have several conceptual structures. For example, 5) is another conceptual structure of 3), in which the compound concept is directly partitioned into three directly related lexical concepts.

5)



Although one phrase or sentence may have several conceptual structures, there is some common information behind them. To see this, consider again the relatedness between the lexical concept 姑娘(girl, /guniang/) and the compound concept 外貌漂亮(beautiful appearance, /waimao piaoliang/) in 3). It can be *reduce to* that between 姑娘(girl, /guniang/) and 外貌(appearance, /waimao/), a sub-concept of 外貌漂亮(beautiful appearance, /waimao piaoliang/), in the sense that the former relatedness holds if and only if the latter does. We denote the above reductions in 6).

6) [姑娘(girl,/guniang/), 外貌漂亮(beautiful appearance,/waimao piaoliang/)]

⇩

[姑娘(girl, /guniang/), 外貌(appearance, /waimao/)]

For any two concepts in a phrase or sentence, if their relatedness cannot be reduced anyway, it is a *basic relatedness* in the phrase or sentence, and the concepts are *basic concepts*. For example, 7) lists the basic relatedness in 3) respectively.

7) [外貌(appearance, /waimao/), 漂亮(beautiful, /piaoliang/)]
[外貌(appearance, /waimao/), 姑娘(girl, /guniang/)]

All basic concepts together with their basic relatedness of a phrase or sentence form the *basic conceptual structure* of the phrase or sentence. For example, 5) is the basic conceptual structure of 3). So, although a phrase or sentence may have several conceptual structures, they can be generally reduced to one basic conceptual structure.

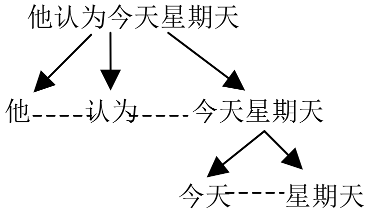
The reduction of relatedness also distinguishes between whether or not some concepts join together to relate with other concepts in a phrase or sentence. As an example, consider 8).

8) 他认为今天星期天(he thinks that it is Sunday today)

There exists direct relatedness between the lexical concept 认为(think, /renwei/) and the compound concept 今天星期天(today is Sunday, /jintian xingqitian/), however, the relatedness cannot be reduced anyway, because that there is no direct relatedness between 认为(think, /renwei/) and 今天(today, /jintian) or 星期天(Sunday, /xingqitian/). This means that the two lexical concepts, i.e., 今天(today, /jintian) or 星期天(Sunday, /xingqitian/) are combined together to relate with another lexical concept 认为(think, /renwei/). 9) and 10) list its basic relatedness and basic conceptual structure.

9) [他(he, /ta/), 认为(think, /renwei/)]
[认为(think, /renwei/), 今天星期天(today is Sunday)]
[今天(today, /jintian/), 星期天(Sunday, /xingqitian/)]

10)



2.3 Head

For any related concepts in a phrase or sentence, sometimes there is a *focus* concept, which we call *head* of the relatedness or the other concept. For example, consider 11) and 12).

11) 漂亮的姑娘 (beautiful girl, /piaoliang de guniang/)
12) 漂亮的外貌 (beautiful looking, /piaoliang de waimao/)
11’¯)姑娘的漂亮 (the girl’s beauty, /guniang de piaoliang/)
12’)外貌的漂亮 (the beauty of the looking, /waimao de piaoliang/)
11’’¡¯姑娘漂亮 (the girl is beautiful, /guniang piaoliang/)
12’’)外貌漂亮 (the looking is beautiful, /waimao piaoliang/)

姑娘(girl, /guniang/) is the head in 11), while 外貌(appearance, /waimao/) is the head in 12). Conversely, in both 11’¯) an 12’¯),漂亮(beautiful, /piaoliang/) is the head. But in 11’’¡¯) a 12’’¡), we assume that there ar*no heads*, otherwise we could not distinguish between 11’’¡¯-’’¡¯¡¯) an’ 11¡’)-12 11)-12).

The phenomena of none heads is more evident in coordinate phrases or sentences. As an example, in 13), the two parts 师(teacher, /shi/) and 生(student, /sheng/) joins together without any bias.

13) 师生(teacher and student, /shisheng/)

Due to the fact that a head is with regard to any two related concepts in a phrase or sentence, one concept may act as the head of one kind of relatedness, while not being a head of the other relatedness. Consider 1) again, 走私(smuggle, /zousi/) is the head of 积极(actively, /jiji/), while there is no head of the relatedness between 走私(smuggle, /zousi/) and 汽车(car, /qiche/).

In general, from the conceptual point of view, a phrase or sentence depicts some conceptual structures, which can be reduced to one basic conceptual structure, and a basic conceptual structure consists of basic concepts and conceptual relatedness. Furthermore, a basic concept can be further partitioned into several other basic concepts with some basic conceptual relationship holding between them. In addition, the basic conceptual relatedness can be directed or undirected, which reflects whether there is a focus between two related concepts.

3 Recursive Graph

An appropriate framework for formal representation of basic conceptual structures needs to encode concepts and conceptual relatedness, distinguish directed relatedness and undirected relatedness, and capture the conceptual hierarchies in basic conceptual structures.

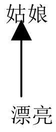
Dependent tree is a traditional apparatus to describe semantic information of phrases or sentences (Mel’¯cuk 1988; Lombardo et al., 1988; Samuelsson, 2000). However, it always enforces a head in the description of the relationship between concepts. This feature is inconsistent with the conceptual view of Chinese phrases ad sentences.

First, dependent trees don’¯t have enoug differentiating ability for Chinese phrases or sentences. As an example, a subject-predicate structure can be transformed into two definiteness-center structures in Chinese language. Across these structures, the conceptual relationship remains unchanged, and the difference only lies in the distribution of the heads. Consider 14)-16), which are all formed by the two concepts, 姑娘(girl, /guniang/) and 漂亮(beautiful, /piaoliang/).

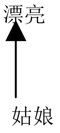
14) 姑娘漂亮 (the girl is beautiful, /guniang piaoliang/)
15) 漂亮姑娘 (beautiful girl, /piaoliang guniang/)
16) 姑娘的漂亮(the beauty of the girl, /guniang de piaoliang/)

14) is a subject-predicate structure, and 15) and 16) are two definiteness-center structures. The conceptual relationship holding between the two concepts, 姑娘(girl, /guniang/) and 漂亮(beautiful, /piaoliang/), remains the same across the three structures. However, the two concepts can only form two dependent trees, listed in 17) and 18).

17)



18)



In 17), 姑娘(girl, /guniang/) is the head, while in 18), 漂亮(beautiful, /piaoliang/) is the head. Obviously, these two dependent trees cannot encode the three semantically different phrases 14) to 16).

Second, in dependent theories, the head of a sentence is generally the verb or adjective in a sentence (Mel'cuk, 1988). However, in some Chinese sentences, there are no verbs at all. Intuitively, the sentence is just to reflect some relationship between some concepts, without any focus on any specific concept. For example, 19) only states the fact that I am 30 years old by relating one number, 30, with one of my feature, age, without any focus on the number or on myself.

19) 我30岁 (I am 30n years old, /wo sanshi sui/)

In such cases, it doesn’t make sense to assume one head for this sentence.

Trees are commonly used formal tools for encoding syntactic structures. One distinguished feature of syntactic trees is that each node inside has only one father node except the root, which makes them convenient to describe constituent hierarchies. So, the link between the nodes in the trees only reflects that one node is a constituent of the other, and there is no link between sister nodes, which means that they cannot encode the conceptual relationship between concepts except the constituent relationship.

Another feature of syntactic trees is that there is an ordering across the nodes inside a syntactic tree, which is in fact enforced by syntactic rules. However, in conceptual view, the focus is on whether there exists conceptual relationship between two concepts, and whether there is a head between them, so the ordering of concepts doesn’t make sense

Another choice is graph. We can expect the nodes in a conceptual graph to encode concepts and the edges between nodes to encode the relatedness between concepts with directed edges for specifying heads and undirected edges for specifying none heads. However, the conceptual hierarchies cannot be accommodated in conceptual graphs. To overcome this problem, we propose recursive graphs to replace graphs. The recursiveness feature of recursive graphs makes it possible to encode the conceptual hierarchies.

Formally, a recursive graph can be defined iteratively as the following: suppose P is a set of points, then:

i) $<P_1, E_1>$ is a 1-level recursive graph, where $P_1 \subseteq P$, $E_1 \subseteq P_1 \times P_1 \times Q$, and $Q=\{0, 1\}$

Let $G_1=\{<P_1, E_1>: P_1 \subseteq P, E_1 \subseteq P_1 \times P_1 \times Q,\}$. Intuitively, 0 in Q stands for directed edge, while 1 in Q stands for undirected edge, and G_1 is the set of all 1-level recursive graphs produced by P.

ii) $< P_k, E_k>$ is a k-level recursive graph, where $P_k \subseteq P \cup G_1 \cup \ldots \cup G_{k-1}$, $P_k \cap G_{k-1} \neq$NIL, and $E_k \subseteq P_k \times P_k \times Q$.

Let $G_k=\{<P_k, E_k>: P_k \subseteq P \cup G_1 \cup \ldots \cup G_{k-1}, P_k \cap G_{k-1} \neq \text{NIL}, E_k \subseteq P_k \times P_k \times Q,\}$. Intuitively, G_k is the set of all 1-level recursive graphs produced by P, and at least one point in P_k is a k-1 level recursive graph.

Compared with syntactic trees, recursive graphs can describe both the relationships between concepts and the constituency of concepts through edges and embed graphs respectively, while the syntactic trees only provide the constituency of syntactic categories. On the other hand, syntactic trees need syntactic category labels, while recursive graphs doesn’t need semantic category labels

In contrast with dependent trees, recursive graphs encode the concept hierarchies by embed graphs, while dependent trees encode the hierarchies through the relatedness between heads. On the other hand, recursive graphs extend the control relationship in dependent trees in several ways. First, the head is with regard to the two concepts with direct relatedness, rather than a whole phrase or sentence; second, for two directly related concepts, there may be no heads; third, the direct relatedness may exist between lexical concepts or non-lexical concepts.

4. Conclusion and Future Work

In this paper, we discuss about the conceptual graph bank for Chinese phrases and sentences. Now the bank has included 5,000 phrases or sentences selected from a textbook about Chinese sentence patterns (Li, 1987), and will include 5,000 more sentences from LDC Chinese corpus. For each selected phrase or sentence, we annotate the static part of its semantic information. To do so, we give a conceptual view of Chinese phrases and sentences, and for each element in this view, we try to find its syntactic markers in order to ensure the consistency of the annotation. Furthermore, we also introduce recursive graph, a kind of formal representation tools, to encode the semantic information.

Future work includes some linguistic issues as well as NLP applications. One future work is concerned with the syntactic markers for irreducible relatedness, or the markers for the linguistic forms that relate with others as a whole. For example, clauses are such linguistic units, and they are generally introduced by ‘that’ in English, but for Chinese language, there are no such markers, so we can only turn to other aspects, e.g., transformation, etc. to specify the markers. Another future work is about the capability of the conceptual graphs in describing the semantic information, and we need to further explore whether this encoding can differentiate various meaningfully different phrases or sentences. Finally, with large-scale treebanks available and static information of syntactic annotation derived from the treebanks, and with the help of large-scale thesaurus and machine learning techniques, we can explore the correspondence between linguistic forms, syntactic information and semantic information in the three different levels.

References

Bikel, D. M. and David Chiang. 2000. Two statistical parsing models applied the the Chinese treebank. In M. Palmer et al., Second Chinese Language Processing Workshop, Hong Kong, October. ACL.

Black, E. et al. 1996. Beyond skeleton parsing: producing a comprehensive large-scale general English treebank with full grammatical analysis, In Proceedings of COLING, pages 107--112, Copenhagen, Denmark.

Keh-Jiann Chen et al., The CKIP Chinese Treebank. In Journ ees ATALA sur les Corpus annot es pour la syntaxe. Talana, Paris VII, 1999.

Charniak, E. 1996. Treebank Grammars. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, Portland, OR. AAAI Press / MIT Press. 1031--1036.

Chomsky, N., 1968, Language and Mind, Harcourt Brace Jovanovich.

Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. The MIT Press.

Dong, Z. D., et al., 2000, Hownet, http://www.keenage.com

Gazdar, G, Klein, E., Pullum, G., & Sag, I, 1985, Generalized Phrase Structure Grammar, Harvard, Cambridge, Mass.

Jan Hajic: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning, pp. 106-132 Karolinum, Charles University Press, Prague, 1998.

Li L. D., Modern Chinese Patterns, Yuwen Press, 1987.

Mei J. J. et al., Tongyici Cilin (A Chinese Thesaurus), 1983, Shanghai Cishu Press, Shanghai.

I. Mel'cuk, 1988 : Dependency Syntax : Theory and Practice. Albany. State Univ. of New York Press.

M. Marcus, B. Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19:313--330.

Okurowski, M. E., Dolan, R. Kovarik, J. The Chinese Treebank Project, 1998.

Samuelsson, C. (2000a). A statistical theory of dependency syntax. In Proceedings of COLING-2000. ICCL.

Fei Xia et al., Developing guidelines and Ensuring Consistency for Chinese Text Annotation, Proceedings of the 2nd LREC, Greece, June 1-4, 2000.

Zhu D. X., Yufa Dawen, (in Chinese), 1983