IOP Conf. Series: Materials Science and Engineering **242** (2017) 012109 doi:10.1088/1757-899X/242/1/012109

Semantic computing and language knowledge bases

Lei Wang[1,2] Houfeng Wang[1] Shiwen Yu[1]

[1]Key Lab of Computational Linguistics of Ministry of Education, Peking University, Beijing, China

[2]School of Foreign Languages, Peking University, Beijing, China

{wangleics, wanghf, yusw}@pku.edu.cn

Abstract. As the proposition of the next-generation Web – semantic Web, semantic computing has been drawing more and more attention within the circle and the industries. A lot of research has been conducted on the theory and methodology of the subject, and potential applications have also been investigated and proposed in many fields. The progress of semantic computing made so far cannot be detached from its supporting pivot – language resources, for instance, language knowledge bases. This paper proposes three perspectives of semantic computing from a macro view and describes the current status of affairs about the construction of language knowledge bases and the related research and applications that have been carried out on the basis of these resources via a case study in the Institute of Computational Linguistics at Peking University.

1 Introduction

Semantic computing is a technology to compose information content (including software) based on meaning and vocabulary shared by people and computers. Its goal is to bridge the semantic gap through this common ground, to let people and computers cooperate more closely as in Fig. 1. The task of semantic computing is to explain the meaning of various constituents of sentences (words or phrases) in a natural language. We believe that semantic computing is a field that addresses two core problems: First, to map the semantics of user with that of content for the purpose of content retrieval, management, creation, etc.; second, to understand the meanings (semantics) of computational content of various sorts, including texts, videos or audios and expressing them in a form that can be processed by machine.

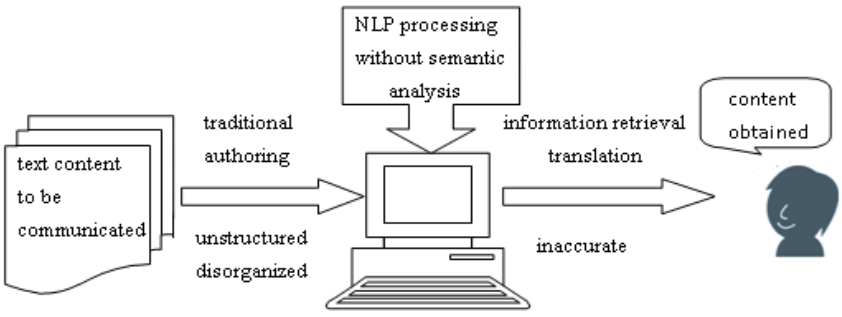

Fig. 1. Human-computer interaction is handicapped without semantic computing.

But the way to the success of semantic computing is not even and it has taken a quite long time for researchers to make some progress in this field. The difficulties of semantic computing involve many aspects: ambiguity, polysemy, domain of quantifier, metaphor, etc. Different individuals will have different understanding of the same word or the same sentence. Research on the theory and methodology of semantic computing still has a long way to go.

2 Related Work on Semantic Computing

Semantics is an interesting but controversial topic. Many a theory has been proposed in attempt to describe what meaning really means. But up until now there has not been a theory that can describe the meaning of various language units (words, phrases and sentences) so perfectly that was accepted universally, even though Fillmore's proposition of Framework semantics [1] is successful enough. Since Gildea [2] initiated the research on automatic semantic role labeling, many evaluations have been conducted internationally, such as Senseval-3 and SemEval 2007, as well as CoNLL SRL Shared Task 2004, 2005 and 2008. And word sense disambiguation remains a popular research topic. [3][4]

In recent years, semantics-based analysis such as data and web mining, analysis of social networks and semantic system design and synthesis have begun to draw more attention from researchers. Applications using semantics such as search engines and question answering [5], content-based multimedia retrieval and editing, natural language interfaces [6] based on semantics have also been attracting attentions.

3 The Theory and Methodology of Semantic Computing

Why semantic computing (or NLU) has posed so great a challenge? We may attribute this to two major reasons: First, it is based on the knowledge of human language mechanism. If fully-developed complicated brains are often seen as a crowning achievement of biological evolution, the interpersonal communication is no simpler than human biological mechanism. Language has to be a crucial part of the evolutionary process, which has not been fully understood by scientific research. Second, in NLP research the language is both the target and the tool. Current NLP research focuses on either speech or written texts only. However, in the real world scenario, reading and interaction between humans are multi-dimensional (through different forms of information such as text, speech, or images and utilizing our different senses such as vision, hearing).

There are complex many-to-many relations between the form and the meaning of a language. Semantic computing is not only the way but also the ultimate goal of natural language understanding. Although it is hard, we should not give up. Here we propose that the main contents of semantic computing include the following three aspects: ontological, cognitive and pragmatic. As for ontologies, much progress has been made worldwide. The remarkable achievements in English include: WordNet by Princeton University, PropBank by University of Pennsylvania [7], etc. Also there are quite a number of efforts made on building ontologies in Chinese, which will be elaborated in Section 4.

As to WSD tasks on the word level, some problems can be solved when ontology is applied. But ambiguity can also appear on the syntactic level. For this, it is usually difficult for ontologies to do much, so we may seek help from language knowledge bases. The following examples of syntactic semantic analysis will illustrate how different syntactic structures will change the meaning of sentences:

Example 1

这样的电影不是垃圾是什么?	--该电影是垃圾。
zhè yàng de diàn yǐng bú shì lā jī shì shén me?	-- gāi diàn yǐng shì lā jī
If a movie as such is not rubbish, what is it?	-- It is rubbish.
这样的电影怎么能说是垃圾呢?	-- 该电影不是垃圾。
zhè yàng de diàn yǐng zěn me néng shuō shì lā jī ne?	-- gāi diàn yǐng bú shì lā jī
How can a movie as such be rubbish?	-- It is not rubbish.

With respect to semantic computing on cognitive level, we will use metaphor as an example. For a long time, NLP research has focused on ambiguity resolution. Can NLU be realized after ambiguity resolution? Metaphor, insinuation, pun, hyperbole (exaggeration), humor, personification, as well as

intended word usage or sentence composing, pose a great challenge to NLU research. If the computer can deal with metaphors, it will greatly improve the ability of natural language understanding.

The linguistic function of metaphor is also important. Metaphor is the base of new word creation and polysemy production (sense evolution), for example, 垃圾箱 lā jī xiāng (recycle) and 病毒 bìng dú (virus) are used in a computer setting and words like 高峰 gāo fēng (peak), 瓶颈 píng jǐng (bottleneck) and 线索 xiàn suǒ (clue) are endowed with new meanings which have not been included in traditional Chinese dictionaries. As for the NLP tasks of metaphor computing, we can conclude that there are three tasks to be accomplished: First, metaphor recognition. For instance, how can we distinguish 知识的海洋 from 海洋资源考察 hǎi yáng zī yuán kǎo chá (investigation of ocean resources); Second, metaphor understanding and translation. For instance, 知识的海洋 actually means 知识像海洋一样丰富。zhī shí xiàng hǎi yáng yí yàng fēng fù (Knowledge is as rich as the ocean.). Third, metaphor generation. For instance, how phrases such as 信息的海洋 xìn xī de hǎi yáng (ocean of information) and 鲜花的海洋 xiān huā de hǎi yáng (ocean of flowers) can be generated successfully by computer?

Second, empirical (statistical) method i.e., providing machine with a large number of samples and training a model. From a statistical point of view, metaphor recognition can be seen as a problem to compute the conditional probability $p(m|c)$ to decide whether 海洋 is a metaphor in context c. The reversed order of two variants m and c will not change the value of unified probability of $p(m|c)$ and $p(c|m)$, while the relation between unified probability and conditional probability can be written as:

$$p(c)p(m \mid c) = p(m)p(c \mid m) \tag{1}$$

Then,

$$p(m \mid c) = p(m)p(c \mid m) / p(c) \tag{2}$$

Given c，$p(c)$ is a constant. Then,

$$p(m \mid c) \propto p(m)p(c \mid m) \tag{3}$$

Given a threshold δ, if $p(m)p(c \mid m) > \delta$, then we can deem this 海洋 is a metaphor.

Then the problem becomes how to compute $p(m)p(c \mid m)$. We can compute it based on large-scale annotated corpus and get

$$p(m) = N_m / N \tag{4}$$

N_m — the times of 海洋 as a metaphor in the corpus;

N — the total times of 海洋 in the corpus.

Then we simplify 海洋 and its context c into: W_{-k} … W_{-1} 海洋 W_1 … W_i, where W_{-k}, …, W_{-1}, W_1,…, W_i represent the n-gram of 海洋 and its syntactic and semantic attributes respectively.

$$p(c \mid m) = p(W_{-k} \mid m) \cdots p(W_{-1} \mid m)p(W_1 \mid m) \cdots p(W_i \mid m) \tag{5}$$

$$p(W_s \mid m) = N(W_s) / N_w, \quad (s = -k, \cdots, -1, 1, \cdots, i) \tag{6}$$

$N(W_s)$ stands for the times of co-occurrence of 海洋 as a metaphor and word W with designated attributes at position. Here an important hypothesis of independence is: words at different position s is not correlated with the word 海洋 as is in Fig. 2.

Last, we will discuss semantic computing on the pragmatic perspective, which is more or less unique of Chinese language. First, the change of construction in Chinese will affect the meaning of a sentence even though the words themselves are not changed. The emphasized meaning of the construction is not equal to the combination of the underlying meaning from each element in the construction. The meaning reflects the distribution of quantity of entities and the relative locations among entities. Although the underlying syntactic relationship among the main verb, the agent and the object(s) still exists, such syntactic relationship is only secondary.

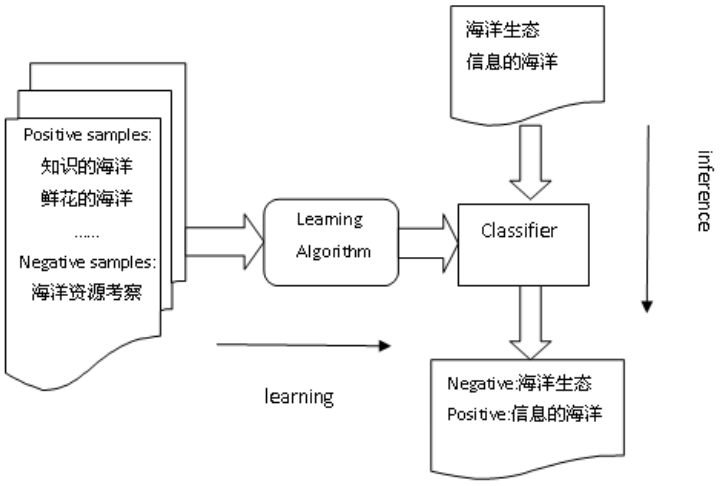

Fig. 2. Empirical (statistical) method of metaphor processing

4 Language Knowledge Bases as the Foundation of Semantic Computing

Language knowledge base is an indispensable component for NLP system, and its quality and scale determines the failure or success of the system to a great extent. For the past two decades, a number of important language knowledge bases have been built through the effort of people in Institute of Computational Linguistics (ICL) at Peking University. Among them, the Grammatical Knowledge Base of Contemporary Chinese (GKB) [8] is the most influential. Based on GKB, various research projects have been initiated. For instance, a project on the quantitative analysis of "numeral-noun" construction of Chinese was conducted by Wang [9] to further analyze the attributes of Chinese words. A project aiming at the emotion prediction of entries in CIKB was completed by Wang [10] to further understand how the compositional elements of a fossilized construct like an idiom function from the token level.

Following GKB, language knowledge bases of large scale, high quality and various type (words and texts, syntactic and semantics, multi-lingual) have been built, such as the Chinese Semantic Dictionary (CSD) for Chinese-English machine translation, the Chinese Concept Dictionary (CCD) for cross-language text processing, the multi-level Annotated Corpus of Contemporary Chinese, etc. When we want to do cross-lingual information retrieval, the two senses need to be distinguished. Hence, CCD can serve as a useful tool to complete the task for it organizes semantic knowledge from a different angle. Concepts in CCD are represented by Synsets, i.e. sets of synonyms as in Table 2. For instance, the concept 教师 is in a Synset {教师 教员 老师 先生 导师 老板 孩子王 臭老九 …} and all the concepts form a network to associate the various semantic relations between or among the concepts: hypernym-hyponym, part-whole, antonym, cause and entailment, by which we can retrieve information in either an extensive or a contractive way so as to improve the precision or recall of a search engine. It can also provide support for WSD tasks.

Table 1. The Synset of the word 教师 jiào shī and its related Synsets

Offset	**Synset**	**Csynset**	**Hypernym**	**Hyponym**	**Definition**	**Cdefinition**
07632177	teacher instructor	教师 教员 老师 先生 导师 孩子王 臭老九	07235322	07086332 07162304 07209465 07243767	a person whose occupation is teaching	以教学为职业的人

		…		07279659 07297622 07341176 07401098 …		

In 2009, the various knowledge bases built by ICL were integrated into the CLKB. The integration of heterogeneous knowledge bases is realized by a resolution of "a pivot of word sense". Three basic and important knowledge bases, GKB, CSD and CCD have been integrated into a unified system which includes language processing module, knowledge retrieval module and knowledge exploration module. Although there are some fundamental resources on semantic computing, it needs further improvement, updating, integration and specification to form a collective platform to perform more complicated NLP tasks. To further improve the result of semantic computing, innovative projects for new tasks should also be launched, for instance:

- metaphor knowledge base
- ultra-ontology dynamic knowledge base (generalized valence mode)
- the integration of information based on multi-lingual translation

5 Conclusion

Why semantics is so useful in the first place? Linguists and psychologists are interested in the study of word senses to shed light on important aspects of human communication, such as concept formation and language use. Lexicographers need computational aids to analyze in a more compact and extensive way word definitions in dictionaries. Computer scientists need semantics for the purpose of natural language processing and understanding. Therefore, the significance of semantic computing in NLP is obvious and more research needs to be done with this respect.

Acknowledgement

Our work is supported by National High Technology Research and Development Program of China (863 Program) (No. 2015AA015402).

References

[1] Fillmore, C. J.. Frame Semantics and the Nature of Language. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, pp 20-32. New York (1976).
[2] Gildea, Denial and Denial Jurafsky. Automatic Labeling of Semantic Roles. Computational Linguistics, 28(3), 245-288(2002).
[3] Ide, Nancy and Jean Véronis. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics, 24(1), 2-40(1998).
[4] Schutze, Hinrich. Automatic Word Sense Discrimination. Computational Linguistics, 24(1), 97-124(1998).
[5] Li, Sujian, Zhang Jian, Huang Xiong and Bai Shuo. Semantic Computation in Chinese Question-Answering System, Journal of Computer Science and Technology, 17(6), 993-999(2002).
[6] Neo, Ee Sian, Takeshi Sakaguchi and Kazuhito Yokoi. A Humanoid Robot that Listens, Speaks, Sees and Manipulates in Human Environments. Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems Seoul, Korea, pp 419-425. Seoul (2008).
[7] Xue, Nianwen and Martha Palmer. Automatic Semantic Role Labeling for Chinese Verbs. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp 1160-1165. Edinburgh (2005).
[8] Yu, Shiwen et al.. Introduction to Grammatical Knowledge Base of Contemporary Chinese (Second

Edition), Tsinghua University Press, Beijing, China (2003).
[9] Wang, Meng et al. Quantitative Research on Grammatical Characteristics of Noun in Contemporary Chinese. Journal of Chinese Information Processing, 22(5), 22-29(2009).
[10] Wang, L., Yu, S., Zhu, X., Li, Y.. Chinese Idiom Knowledge Base for Chinese Information Processing. In Proceedings of the Chinese Language Semantics Workshop (2012), Wuhan, pp 85-90, China (2012).