Semantic Computing and Language Knowledge Bases[1]

Lei Wang
Key Laboratory of Computational Linguistics
of Ministry of Education
Department of English, Peking University
wangleics@pku.edu.cn

Shiwen Yu
Key Laboratory of Computational
Linguistics of Ministry of Education,
Peking University
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 and thereby to design and operate information systems (i.e., artificial computing systems). Its goal is to plug the semantic gap through this common ground, to let people and computers cooperate more closely, to ground information systems on people's life world, and thereby to enrich the meaning and value of the entire life world. (Hasida, 2007) The task of semantic computing is to explain the meaning of various constituents of sentences (words or phrases) or sentences themselves 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, but is not limited to, text, video, audio, network, software, and expressing them in a form that can be processed by machine.

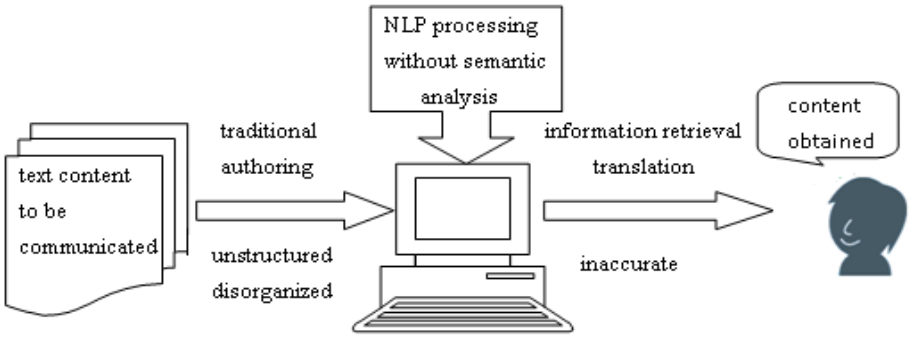

Figure 1. Human-computer interaction is handicapped without semantic computing.

[1] This work is supported by the National Natural Science Foundation of China (No. 60970083) and Chiang Ching-kuo Foundation for International Scholarly Exchange (2009).

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.

Now we provide an example in a search engine to show how difficult for the computer to understand the meaning of a word. We input two sentences into Google.com Translate and the following results were returned:

Example 1

I bought a table with three dollars.（*20091016 Google: 本人买了3 美元一表*）
I bought a table with three legs. （*20091016 Google: 本人买了3 条腿的表*）

We know that the word "table" has two common meanings in English (a wooden object and a structured data report). But in Chinese they correspond to two different words (表 biǎo and 桌子 zhuō zi[2]). From Example 1, we can see that the search engine cannot distinguish the two senses and translate them both as 表. Thus, without semantic analysis queries in a search engine may result in very poor performance. The first principle of a search engine is based on shallow Natural Language Processing (NLP) techniques, for instance, string matching, while future direction of search engines should aim at content index and the understanding of user's intention. Semantic computing becomes applicable only with the development of deep NLP techniques. Machine Translation (MT) is the first application of digital computers in the non-digital world and semantic information is indispensable in MT research and applications. However, there has been no breakthrough to the extent of Natural Language Understanding (NLU) and semantic computing may serve as the key to some success in this field.

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 (1976) is successful enough. Since Gildea et al. (2002) 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. Word Sense Disambiguation (WSD) is also a very important research subject and a lot of work has been done in this regard, such as Lesk (1986), Gale et al. (1998), Jin et al. (2007) and Qu et al. (2007) as the Chinese counterpart. As to the research on computing word sense relatedness, Dagan et al (1993) did some pilot work and Lee (1997) and Resnik (1999) contributed to the research on semantic similarity.

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 (Li et al., 2002), content-based multimedia retrieval and editing, natural language interfaces (Yokoi et al., 2005) based on semantics have also been attracting attentions. Even semantic computing has been applied to areas like music description, medicine and biology and GIS systems and architecture. The whole idea is how to realize human-centered computing.

[2] Pinyin is currently the most commonly used Romanization system for standard Mandarin. The system is now used in mainland China, Hong Kong, Macau, parts of Taiwan, Malaysia and Singapore to teach Mandarin Chinese and internationally to teach Mandarin as a second language. It is also often used to spell Chinese names in foreign publications and can be used to enter Chinese characters on computers and cell phones.

3 The Theory and Methodology of Semantic Computing

3.1 Important Questions That Need to Be Asked about Semantic Computing

In the past few years there has been a growing interest in the field of semantics and semantic computing. But there are questions that have been always lingering on researchers' minds. What on earth semantics is? What is the best way to describe the meaning of a language unit? How can natural languages be processed so that we are able to benefit from human-computer interaction, or even interpersonal communication? It seems that no one can give satisfactory answers to these questions. But it is now commonly agreed that the study of semantic computing or knowledge representation is a central issue in computational linguistics. The major contributions on this topic are collected in *Computational Linguistics* (1987-2010) and *International Journal of Semantic Computing* (2007-2010). Research in computing semantics is, however, rather heterogeneous in scope, methods, and results. The traditional "wh" and "how" questions need to be asked again to understand the consequences of conceptual and linguistic decisions in semantic computing:

What? What should be computed in terms of semantics? Each word is a world and its meaning can be interpreted differently. Despite the interest that semantics has received from the scholars of different disciplines since the early history of humanity, a unifying theory of meaning does not exist, no matter whether we view a language from a lexical or a syntactic perspective. In practice, the quality and type of the expressed concepts again depend upon the one who uses it: any language speaker or writer, a linguist, a psychologist, a lexicographer, or a computer. In psycholinguistics and computational linguistics, semantic knowledge is modeled with very deep and formal expressions. Often semantic models focus on some very specific aspect of language communication, according to the scientific interest of a researcher. In natural language processing, lexical entries or semantic attributes typically express linguistic knowledge as commonsensically understood and used by humans. The entries or attributes are entirely formatted in some knowledge representation and can be manipulated by a computer.

Where? What are the sources of semantic knowledge? Traditionally, individual introspection is often a source of obtaining word senses. However, individual introspection brings about both theoretical and implementation problems. Theoretically, it is because "different researchers with different theories would observe different things about their internal thoughts..." (Anderson 1989). With regard to implementation, it is because consistency becomes a major problem when the size of the lexicon or the syntactic tree bank exceeds a few thousands entries or annotation tags. Despite the scientific interest of such experiments, they cannot be extensively repeated for the purpose of acquiring mass word sense definitions. On-line corpora and dictionaries are widely available today and provide experimental evidence of word uses and word definitions. The major advantage of on-line resources is that in principle they provide the basis for very large experiments, even though at present the methods of analysis and application are not fully developed and need further research to get satisfactory results.

How? Semantic computing can be realized at various levels. The hard work is to implement a system in a real domain, or the more conceptual task of defining an effective mathematical framework to manipulate the objects defined within a linguistic model. Quite obviously the "hows" in the literature about semantic computing are much more important than the "whats" and "wheres". The methodology that really works in semantic computing is deeply related to the ultimate objective of NLP research, which still cannot be defined adequately so far.

3.2 The Perspectives of Semantic Computing from a Macro View

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). It is necessary to rely on the advancements of brain science, cognitive science and other related fields and work in collaboration to produce better results. Linguistics, especially computational linguistics, has made its own contribution, and semantic computing will play an important role in NLP.

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:

- semantic computing on the ontological perspective
- semantic computing on the cognitive perspective
- semantic computing on the pragmatic perspective

As for ontologies, much progress has been made worldwide. The remarkable achievements in English include: WordNet by Princeton University, PropBank by University of Pennsylvania, etc. Also there are quite a number of efforts made on building ontologies in Chinese, which will be elaborated in Section 5.

In the last few years, the main direction of semantic computing is to disambiguate language units and constructions. In the following Example 2, the word 仪表 yí biǎo has two meanings in different contexts. In Chinese, word segmentation is also a problem that needs to be addressed. In Example 3, segmenting the word 白天鹅 bái tiān é as 白/天鹅 or 白天/鹅 can result in different understanding of the sentences.

Example 2

她的仪表很端庄。 tā de yí biǎo hěn duān zhuāng (*She has a graceful appearance*.)
她的仪表很精确。 tā de yí biǎo hěn jīng què (*Her meters are very accurate*.)

Example 3

白天鹅飞过来了。bái tiān é fēi guò lái le (*A white swan flies toward us.)*
白天鹅可以看家。bái tiān é kě yǐ kān jiā (*A goose can guard our house at daytime.)*

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 (See Section 5). The following examples of syntactic semantic analysis will illustrate how different syntactic structures will change the meaning of sentences:

Example 4

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.

Example 5

蚂蚱是蚂蚱，蛐蛐是蛐蛐。 -- 蚂蚱不是蛐蛐。

mà zhà shì mà zhà, qū qū shì qū qū -- mà zhà bú shì qū qū

A grasshopper is a grasshopper, while a cricket is a cricket. -- A grasshopper is not a cricket.

Rule：A is A, while B is B. ——〉A is not B.

丁是丁，卯是卯。dīng shì dīng, mǎo shì mǎo

Ding is ding, while mao is mao. — being conscientious

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.

First, let's discuss the rhetorical function of a metaphor. Metaphor is extensively and skillfully used in the Chinese classic "Book of Songs" to boost expressiveness.

Example 6

Simile: 自伯之东，首如飞蓬[3]；岂无膏沐？谁适为容。 --（卫风 伯兮）

zì bó zhī dōng ，shǒu rú fēi péng ；qǐ wú gào mù ? shu ísh ìw éi r óng。-- （wèi fēng bó xī）

(Your hair is like disordered grass.)

Metaphor：它山之石，可以攻玉。 --（小雅 鹤鸣）

tā shān zhī shí ，kě yǐ gōng yù。 --（xiǎo yǎ ·h è m íng）

(Rocks from another mountain can be used to carve jade. Metaphorically this phrase means a change of method may solve the current problem.)

Also, many Chinese idioms are metaphorical expressions: 同舟共济 tóng zhōu gòng j ì(Literally, to cross the river in the same boat; metaphorically, to work together with one heart while in difficulty), 铜墙铁壁 t óng qi áng tiě bì (Literally, walls of brass and iron; metaphorically, impregnable). The Chinese language makes use of lots of idioms or idiomatic expressions that are derived from ancient Chinese stories and fables. These idioms and idiomatic expressions are often used metaphorically and reflect historical and cultural background of the language. They are the most precious relics to the Chinese language and culture. Therefore the Chinese Idiom Knowledge Base (CIKB) was also built in 2009. CIKB consists of 38,117 entries and describes many attributes of Chinese idioms. Among the attributes, "literal translation", "free translation" and "English equivalent" are very valuable.

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 b ìng d ú(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. Besides, metaphor creates new meanings in sentence level, for instance, in 地球是人类的母亲。dì qiú shì rén lèi de mǔ qīn (The earth is the mother of humanity.), the word 母亲 (mother) has a different meaning. So, metaphor understanding is beyond the scope of ambiguity resolution. Metaphor, linguistics, and human cognitive mechanisms are inextricably interlinked. So metaphor becomes a fort that must be conquered in NLU research.

From an NLP perspective, metaphors can be summarized into the following categories as in Table 1. As for the NLP tasks of metaphor computing, we can conclude that there are three tasks to be accomplished: First, metaphor

[3] For the purpose of conciseness, only the underlined parts that contain metaphors are translated.

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?

Perspective of grammatical properties		**Perspective of language unites of metaphorical expressions**	
Nominal	祖国的花朵 zǔ guó de huā duǒ (flower of the country), 生命的旅程 shēng mìng de lǚ chéng (life journey)	Word-formation level	卵石 luǎn shí(egg-like stone), 杏仁眼 xìng rén yǎn (apricot-like eyes)
Verb	心潮澎湃 xīn cháo péng pài (heart wave), 放飞理想 fàng fēi lǐ xiǎng (let f dream fly)	Word level	潮流 cháo liú (tide), 朝阳 zhāo yáng (morning sun)
Adjective	这篇文章写得干巴。zhè piān wén zhāng xiě de gān bā(This article is written drily), 这篇文章清汤寡水。zhè piān wén zhāng qīng tāng guǎ shuǐ (This article is like plain soup and water.)	Phrase level	知识的海洋 zhī shí de hǎi yáng (ocean of knowledge), 播种幸福的种子 bō zhǒng xìng fú de zhǒng zi (to sow the seeds of happiness)
Adverb	纯粹胡说 chún cuì hú shuō(absolute nonsense)	Sentence level	汽车喝汽油。qì chē hē qì yóu (Cars drink gasoline.), 女人是水 nǚ rén shì shuǐ (A woman is water.)
		Discourse level	打起黄莺儿，莫叫枝上啼。啼时惊妾梦，不得到辽西。dǎ qǐ huáng yīng ér, mò jiào zhī shàng tí. tí shí jīng qiè mèng, bù dé dào liáo xī 。(To scare away the nightingales for their noise has my dream in which I went to the west to meet my dear husband.)

Table 1. Categories of metaphors from NLP perspective.

Currently we focus on recognition and understanding of metaphors on phrase and sentence level. The automatic processing methods of metaphors can be summarized as two: First, rule (or logic)-based method, i.e., finding the conflicts between the target and the source, and search their common properties.

Example 7

这个人是一头狮子。zhè gè rén shì yī tóu shī zi (This man is a lion)

— only the target and the source

那个人是老狐狸。nà gè rén shì lǎo hú li (That man is an old fox.)

— only the target and the source

森林里既有勇猛的狮子，也有狡猾的狐狸。sēn lín lǐ jì yǒu yǒng měng de shī zi, yě yǒu jiǎo huá de hú li (In the forest, there are both brave lions and sly foxes.)

--- find out properties of the sources

这个人是勇猛的，那个人是狡猾的。zhè gè rén shì yǒng měng de, nà gè rén shì jiǎo huá de (This man is brave, while that man is sly.)

The utterance 河北有个老太太吃土块。hé běi yǒu gè lǎo tài tài chī tǔ kuài (An old lady in Hebei eats clay.) is not in conformity with common sense, but it is not a metaphor; whereas 男人都是动物。nán rén dōu shì dòng wù (All men are animals.) is logical but it may be a metaphor in certain context and may not be in another context.

Second, empirical (statistical) method i.e., providing machine with a large number of samples and training a model. Yu Shiwen presided over the national 973 project "Database for text content understanding" (2004-2009), which includes a subtask named "Analysis of Metaphorical Expressions and Their Pointed Contents in Chinese Texts". In this project, various machine learning methods have been applied to do semantic analyses from the token level. Among them, Wang Zhimin completed her doctoral thesis "Chinese Noun Phrase Metaphor Recognition" in 2006. Jia Yuxiang studied verb metaphor recognition and "X is Y" type metaphor understanding and generation. Qu Weiguang presided over the National Natural Science Fund Project "Research on Key Technologies in Chinese Metaphor Understanding" (2008-2010).

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:

$$p(c)\,p(m \mid c) = p(m)\,p(c \mid m) \qquad (1)$$

Then,

$$p(m \mid c) = p(m)\,p(c \mid m) / p(c) \qquad (2)$$

Given c，$p(c)$ is a constant. Then,

$$p(m \mid c) \propto p(m)\,p(c \mid m) \qquad (3)$$

Given a thresholdδ , 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

$$p(m) = N_m / N \qquad (4)$$

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: $W_{-k} \ldots W_{-1}$ 海洋 $W_1 \ldots W_i$, where W_{-k}, …, W_{-1}, W_1,…, W_i represent the n-gram of 海洋 and its syntactic and semantic attributes respectively.

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

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

$N(W_s)$ 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.

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. As in the sentence 这张床可以睡三个人。zhè zhāng chuáng kě yǐ shuì sān gè rén (This bed can sleep three people.) is different in meaning from the sentence 三个人可以睡这张床。(Three people can sleep on this bed.). Second, the

semantic direction of the complement in verb-complement constructions and the adverbial phrase in verb-adverbial constructions also change the semantic roles of each constituent. For instance, （文章）写完了。（ wén zhāng ） xiě wán le ((The article) is completed.) or （老师）写累了。（ lǎo shī ） xiě lèi le ((The teacher) is tired for writing.) or 香喷喷地炸了一盘花生米。xiāng pēn pēn dì zhà le yī pán huā shēng mǐ(aromatically fried a plate of peanuts). Here the ontology cannot provide enough information to reflect the process and result of change in semantic roles. Thus the Generalized Valence Mode (GVM) is proposed to describe not only participants of the action, but also the change of participants' states. Third, our ultimate goal will be to achieve "semantic harmony". For instance, in both English and Chinese we can say 拔出来 bá chū lái (pull out) or 插进去 chā jìn qù (thrust into), but we never say 插出来 (thrust out) or 拔进去 (pull into). It is alright to say 那个大苹果他都吃了。nà gè dà pín guǒ tā dōu chī le (That big apple he eats it all.) , but it is awkward to say 那颗小核桃他都吃了。nà kē xiǎo hé táo tā dōu chī le (That small chestnut he eats it all.). In fact we can say 那颗小核桃松鼠都吃了。nà kē xiǎo hé táo sōng shǔ dōu chī le (That small chestnut the squirrel eats it all.).

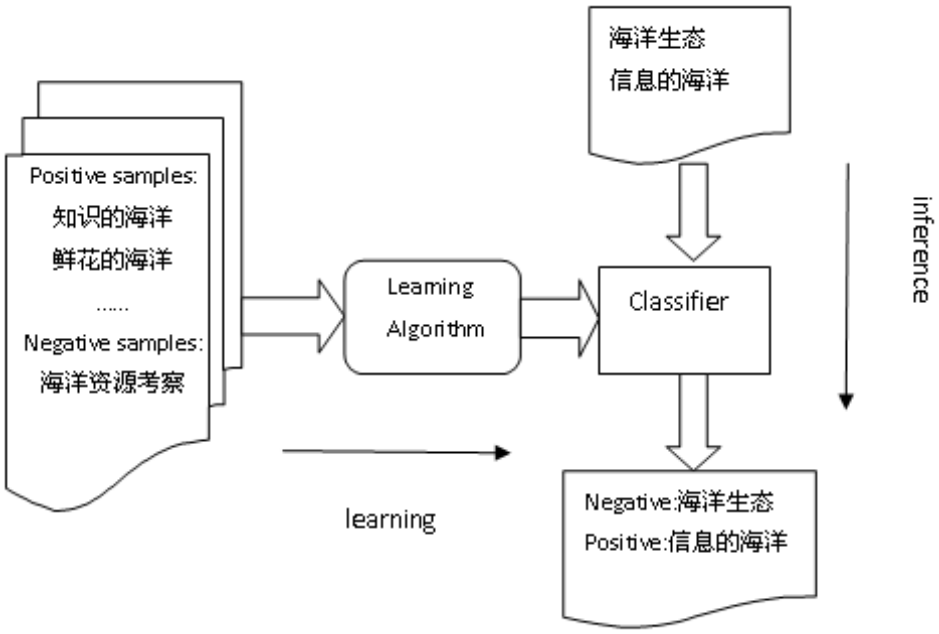

Figure 2. Empirical (statistical) method of metaphor processing.

Professor Lu Jianming (2010) remarked on the realization of semantic harmony. The principle of semantic constraint of words essentially requires that the words in sentences should be harmonic in terms of meaning. Analysis of ill-formed sentences and automatic language generation will benefit from the research in semantic harmony. Semantic computing on the pragmatic level has unique characteristics with respect to Chinese language. The solution of these problems poses a great challenge and will make great contribution to the understanding of the essence and universality of languages.

4 Potential Applications of Semantic Computing – a Case Study on Automatic Metaphor Processing in Search Engines

Nowadays, search engines are developing very rapidly and some of them have won great economic success. In terms of semantic computing, Baidu.com takes the lead and has unveiled the search concept "Box computing" which introduces semantic analysis. The precision and recall of a search engine are

always the essential issue that a user is concerned. Therefore we will find the value of semantic computing first in a search engine.

Certainly, if metaphor can be understood properly by a computer, the precision of search engines will be improved. Let's take the phrase 起飞 qǐ fēi(take off) as an example. Literally 起飞 means an aircraft takes off such as in 航班起飞时间 háng bān qǐ fēi shí jiān (the time for the airplane to take off). Sometimes we also use it in phrases like 经济起飞 jīng jì qǐ fēi (economic take-off) or 东方美女歌坛起飞 dōng fāng měi nǚ gē tán qǐ fēi (Oriental beauties take off in the music arena.) to mean metaphorically. If the literal sense and its metaphorical sense can be distinguished successfully, we will find the exact information that we need. Meanwhile, we hope that through this the recall of search engine will also be improved. For example, in Chinese we often use the phrase 祖国的花朵 zǔ guó de huā duǒ (flowers of the country) metaphorically to refer to 儿童 ér tong (children). So web pages describing 祖国的花朵 should also be related to the query word 儿童.

We also observe that the phrases 金融风暴 jīn róng fēng bào (financial storm) and 金融海啸 jīn róng hǎi xiào(financial tsunami) metaphorically refer to 金融危机 jīn róng wēi jī (financial crisis). But when we input the query 金融危机 into a search engine, the results were only web pages with 金融危机 or 金融//危机. But when we use the query 金融风暴 or 金融海啸, there were no web pages with the results 金融危机. We know that the phrase 炒鱿鱼 chǎo yóu yú has literal usage (to fry squids) and metaphorical usage (to fire sb. from his/her job). When we input the phrase into the search engine, we find the result with metaphorical usage takes up 65% while other usage only accounts for 35% (Wang, 2006). Therefore we may conclude that whether metaphor is understood will seriously affect precision and recall.

Another important application lies in machine translation and cross-lingual search. Correct metaphor recognition and understanding is the precondition of correct translation. Machine translation can be a framework to evaluate the performance of metaphor recognition and understanding, and also is a tool to realize cross-lingual search. For instance, a well-known Chinese female volleyball player got a nickname as 铁榔头 tiě láng tou. Shall we translate it literally as "iron hammer" or more metaphorically as "iron fist" in order to let a user of search engine have a better sense of what it actually means? Translation is culture-bound. When we see the sentence 该电影是鸡肋。gāi diàn yǐng shì jī lèi, how should we translate the word 鸡肋 (a chicken's rib) here? And how shall we distinguish its literal meaning with its metaphorical meaning (食之无味弃之可惜。shí zhī wú wèi qì zhī kě xī, tasteless to eat but a waste to cast away) in order to understand better the sentence "The movie is a chicken's rib"?

Therefore when we investigate the feasibility analysis of applications of automatic metaphor recognition, we propose there are still three solutions to the above-mentioned problems:

- To overcome the limitedness of source domain words
- To recognize metaphors in web pages and build metaphor indexes. Offline processing often makes good use of the advantages of a search engine.
- Before realizing query understanding, let users choose metaphorical or literal meaning of the query through human-computer interaction.

5 Language Knowledge Bases as the Foundation of Semantic Computing

As the foundation of semantic computing, language knowledge bases are in great demand. The achievements on language knowledge bases for Chinese-centered multilingual information processing include: Chinese LDC, Comprehensive Language Knowledge Base (CLKB) by ICL at Peking University, HowNet by Zhendong Dong, Chinese Dependency Tree Bank by Harbin Institute of Technology, etc.

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) (Yu et al., 2000) 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 (2009) to further analyze the attributes of Chinese words. A project aiming at the emotion prediction of entries in CIKB was completed by Wang (2010) to further understand how the compositional elements of a fossilized construct like an idiom function from the token level.

Offset	Synset	Csyncet	Hypernym	Hyponym	Definition	Cdefinition
07632177	teacher instructor	教师 教员 老师 先生 导师 老板 孩子王 臭老九 …	07235322	07086332 07162304 07209465 07243767 07279659 07297622 07341176 07401098 …	a person whose occupation is teaching	以教学为职业的人

Offset	Synset	Csyncet	Hypernym	Hyponym	Definition	Cdefinition
07331418	husband hubby married_man	丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷 …	07391044	071094820 719596807 255726073 28008	a married man; a woman's partner in marriage	已婚男子； 婚姻中女性一方的伴侣
Offset	**Synset**	**Csyncet**	**Hypernym**	**Hyponym**	**Definition**	**Cdefinition**
07414666	Mister Mr.	先生 师傅 同志 大哥 老兄 老弟	07391044		a form of address for a man	对男子的一种称呼

Table 2. The Synset of the word 教师 jiào shī and its related Synsets.

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. The projects as a whole won the Science and Technology Progress Award issued by Ministry of Education of China in 2007.

As mentioned in Section 3, the word 病毒 (virus) has two senses in both English and Chinese: one is in biology and the other is in computer science. 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.

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

6 Concluding Remarks

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.

All in all, we may conclude that the methods of semantic computing can be summarized as the following:

- The research of applicable language model
- The research of effective algorithms
- To build language knowledge bases as its foundation

Semantic computing is a long-term research subject. We hope more progress can be made if a clearer view can be provided for the direction of its development and the pavement for future research can be constructed more solidly with more work done.

Acknowledgements

Our work is based on the long-term accumulation of the language resources that have been built by the colleagues of ICL and it is their contributions that make our achievement possible today. Parts of the content in this paper were presented by Shiwen Yu on the conferences in Hangzhou (International Workshop on Connected Multimedia 2009) and Suzhou (the 11th Chinese Lexical Semantics Workshop 2010), and many thanks should be given to those who offered valuable thoughts and advice. The authors also want to extend their gratitude toward CIPS-Sighan for this valuable opportunity to demonstrate our viewpoints and work.

References

Anderson, J. R. 1989. A Theory of the Origins of Human Knowledge. *Artificial Intelligence*. 40(1-3): 313-351.

Carreras, X. and Marques L. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. *Proceedings of the CoNLL 2004:* 89-97.

Dagan, I. et al. 1993. Contextual Word Similarity and Estimation from Sparse Data. In *Proceedings of the 31st Annual Meeting on the Association for Computational Linguistics (ACL):*164-171

Fillmore, C. J.. 1976. 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*:20-32

Gale, William A., Kenneth W. Church, and David Yarowsky. 1993. A Method for Disambiguation Word Senses in a Large Corpus. *Computers and the Humanities*. 26(5-6): 415-439

Gildea, Denial and Denial Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics,* 28(3): 245-288.

Hasida, K. 2007. Semantic Authoring and Semantic Computing. Sakurai, A. et al. (Eds.): JSAI 2003/2004, LNAI 3609, 137–149.

Ide, Nancy and Jean Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational Linguistics*, 24(1) : 2-40.

Jin, Peng, Wu Yunfang, Yu Shiwen. SemEval-2007 Task 05: Multilingual Chinese-English Lexical Sample. In *Proceedings of SemEval-2007*: 19-23.

Johansson, Richard and Pierre Nugues. 2008. Dependency-based Syntactic-semantic Analysis with PropBank and NomBank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning:* 183-187.

Lee, Lillian. Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis. Harvard University.

Lesk, Michal. 1986. Automatic Sense Disambiguation: How to Tell a Pine from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation:* 24-26.

Li, Sujian, Zhang Jian, Huang Xiong and Bai Shuo. 2002. Semantic Computation in Chinese Question-Answering System, *Journal of Computer Science and Technology*, 17(6) : 993-999.

Lu, Jianming. 2010. Foundations of Rhetoric -- The Law of Semantic Harmony. *Rhetoric Learning*, 2010(1): 13-20.

Qu, Weiguang, Sui Zhifang, et al. 2007. A Collocation-based WSD Model: RFR-SUM. In *Proceedings of the 20th International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems*:23-32.

Schutze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-124.

Resnik, Philip. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Journal of Artificial Intelligence Research 11*: 95-130.

Wang, Lei and Yu Shiwen. Forthcoming 2010. Construction of Chinese Idiom Knowledge Base and Its Applications. In *Proceedings of Coling 2010 Multi-word Expressions Workshop*.

Wang, Meng et al. 2009. Quantitative Research on Grammatical Characteristics of Noun in Contemporary Chinese. *Journal of Chinese Information Processing*, 22(5): 22-29.

Wang, Zhiming. 2006. Recent Developments in Computational Approach to Metaphor Research. *Journal of Chinese Information Processing*, 20(4): 16-24.

Xue, Nianwen and Martha Palmer. 2005. Automatic Semantic Role Labeling for Chinese Verbs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence:*1160-1165

Yu, Shiwen et al.. 2003. *Introduction to Grammatical Knowledge Base of Contemporary Chinese* (Second Edition) (in Chinese), Tsinghua University Press, Beijing, China.