Investigating Semantic Formalization from the Computational Perspective

Huangfu Wei
School of Foreign Languages
North China Electric Power University
Beijing, China
mailtohf@163.com

***Abstract*—Senses are the extensions of language and interrelations between word meanings within language and the basis of flexible language use. Computers are basically made to process symbols. Computational linguistics is to simulate part of, if not all of, the natural language competence. From the perspective of computational linguistics, this paper sheds some light on semantic formalization, and proposes tagging of senses along with tagging weight values when processing word meanings. Also, this paper makes attempts to apply sense weight values to sentence processing in computational linguistics.**

Keywords-semantics, sense, weight, formalization, computational linguistics

I. SEMANTIC KNOWLEDGE IN LANGUAGE PROCESSING

According to prototype theory, semantic knowledge is part of human's cognitive and belief systems. Semantics is also the critical research area since its initiation in 1893 by Michael Breal, and then by Trier and Lakoff who contributes to semantic research from both the descriptive and generative perspective. In linguistic semantics, rules and principles explain how literal meaning is coded in language and how to understand the "senses" of an expression or sense-relations that hold between one word and other expressions within the same language (Ceusters et al, 1999).

Language is the carrier of both thoughts and forms. Thoughts are to induce, deduce, infer and judge, while senses are the contents and basis of thoughts. Senses are not the repeated addition of forms, rather a psychological process. Semantics is the mapping of cognitive structure from language expressions. Based on concepts, which are the abstract and general categorization of objects, people can apply semantic principles, situational contexts and encyclopedic knowledge to construct cognitive schema of the outside world. With frequent interactions and cognitive strategies, limited forms can be used to express unlimited connotation of thoughts (Tesser and Leone, 1977). For example, same meanings can be expressed with different semantic forms.

a) 他这个人真有**意思** (funny);

b) 于是人们以为他们有了**意思** (wish);

c) 你向他**意思意思** (express)如何;

d) .我根本没有那个**意思** (thought);

e) 他常说人生真**没有意思** (nonsense);

f) 你们这么说是什么**意思** (intention)

These sentences are easy to comprehend. People not only use the grammatical structures but also the sense weights to process the words. The sense weight schema are stored in memory, and activated to represent and distill other sense weights. In this chain activation, sense weights are restructured and created to reflect prototypes. For example, "你好？", "你好么？", "好么你？". Different structures are used in the above examples, but their essential meaning remains the same. So, semantic knowledge is of utmost importance in language comprehension.

II. THE STATUS QUO OF COMPUTATIONAL LINGUISTICS

Three domains and three corresponding levels of study are involved in building natural language processing systems, i.e. theoretical linguistics and grammar level, computational linguistics and formalization level, computer technology and realization level. Among them, grammar level concerns the innate characteristics of language and paves the way for the latter modules. Realization level focuses on how to control computer operation and provides the efficient development tools and environment. Formalization level is the interface of grammar level and realization level, with the task to readjust the general grammar modules to form grammar modules for convenient computer processing. Computational linguists' mission is to establish formal grammars (Kaufmann and Pfister, 2012).

Computational linguistics is also to establish formalized mathematical modules to analyze and process natural language, and develop computer programs to make it possible that machines can simulate part of, if not all of, human behavior and competence. Computers can only be used to process symbols and all those processed by computers are symbolized with no exception of languages (Yushi, 2003). Meanwhile, there have been breakthroughs in processing Chinese characters and phrases, so sentence level processing has become critical. Basic researches in syntactic, semantic and pragmatic knowledge are the frontier studies in Chinese processing. And how to acquire syntactic and semantic knowledge is the key, while the latter constitutes the major difficulty in processing the word level senses.

As aforementioned, semantic formalization can take on many impossible missions in artificial intelligence and computational linguistics, and face challenges at the same time. Thus, the researcher will use semantic field to investigate formalization issues. Semantic fields are the sememe systems, which include the semantemes to represent similarities and differences, and are classified by the subordinates, i.e. degree and number of abstraction of

Published by Atlantis Press, Paris, France.

semantemes. Discourse meanings are made up of the sentence meaning set, while the latter are expressed by the semantemes' set. The compositionality of the same set of sememes or different compositionality of sememes leads to different sentence meanings. Thus, the matching relations among semantemes will affect the paradigmatic relations among meanings. Look at the structural formula bellow:

问题：需要研究讨论并且加以解决的矛盾 、疑难

[(矛盾) V (疑难)] {X (需要) K [((研究) V (讨论)) (解决)] }

Note:[] and {} represent levels, and contents inside () are semaments or semament variants, and V represents "or", and X represents the existence of the actions or changes, and K represents the objects of the actions or changes.

From this example, semanteme structural formula marks both the semantemes and the logic connecting them, distinctively indicating word meaning at the formal level. With the semanteme structural formula, individual and series of cross-correlational semantic fields can be obtained.

g) 他把散落的硬币包在手绢里面

h) 他把散落的硬币包在铅笔里面 (X)

i) 他把小王包在手绢里面 (X)

Sentence g, h, i all have the same structures, but sentence g is right, while sentence h and i are wrong. Computers cannot make correct judgment of their correctness only from the formal level, but it will be possible when using the cross-matching modules.

包 ：用布,纸或类似的薄片把东西裹起来

X [(使裹在) [方] ([工] 2 里面)] [施] (人) [工] 1[(手) V(机器)] [工] 2{(纸) V (布) V [(其他) (片状) (东西)] [受] (东西) (1)

铅笔：条形状

小王：人

After matching, it is found that semantemic features of "包" are not in congruence with the semantic features of "铅笔" and "小王", thus sentence h and i are illogical.

III. Formalizations of Senes

Natural language is used in human communication. Large amount of information is stored in natural languages with individual characteristics. Due to the imperfection of natural languages, perfect description is impossible.

To express vague or ambiguous meanings, people can always use cognitive abilities to classify categories, such as using schema to know the world (Sharifi and Mayamei, 2012). So, grammars, with only simple form description, such as words and phrases, are not applicable in natural language processing.

Formalized language is different from natural language because the former requires absolute clarity and conforms to strict formal specifications. It is human capacity to use metaphor, metonymy and iconicity, etc. in coding and decoding natural languages, and to know how to understand emotive implicatures and performatives.

There are two major problems in rendering natural language formalized, i.e. formal and epistemology. The limitations of formalization are first discussed in Tarski' true value theory and Montague's grammar, which stress metalanguage. But any deduction regulation set must be recursive and any semantic form of metalanguage is incomplete(Guangwei, 2011). Similarly, there are limitations to epistemology, i.e. two speakers have their different epistemology in languages. Therefore, only by overcoming the obstacles of forms and epistemology can total formalization be achieved.

Senses cannot be easily symbolized, because the meaning extension is unlimited, but language symbols are limited. Until now, it is unrealistic to symbolize senses because semanteme and sememes are complex. Without senses, it is not enough to use word segmentation, taxonomy, POS tags, feature description, so that it is necessary to formalize senses in natural language. Formal linguistics use formal languages and formal grammars to formalize language and specify senses.

IV. Annotation and sentence processing

A. Sense Annotation

Generally speaking, the steps can be as follows: First, distinguish separate entries for senses of words to be tagged; and then, specify a confirmed entry of senses for polysemants in actual contexts to technically realize word sense disambiguation. Some additional measures can be taken to tag sense entries for polysemants in a given context:

- Compare with those tagged entries of senses for polysemants;
- Apply the preset principles of word disambiguation;
- Use knowledge base, such as dictionaries and cyclopedias.

B. Syntactic Parsing

Syntactic parsing will use computer algorithm to obtain syntactic structures in natural languages (Aarne, 2003). So, a parser has to be compiled to judge the legibility of an input sentence, i.e. outputting its syntactic structure if it is syntactically legible. Currently, syntactic parsing can use the following methods: standard LR analysis algorithm, general LR analysis algorithm, chart-based analysis, context-free grammar probability-based methods and link grammar oriented analysis, etc.

At present, Chinese language syntactic parsing is the focal point, among which the syntactic and semantic knowledge is the key, and primary word sense is the basis. In sense tagging, general practice is to use comprehensive corpora as the basis. This will lead to two facets of problems: firstly, there are low processing efficiency and difficulties in set operations due to the excessive overload of sense entries; and moreover, there are difficulties in establishing syntactic rules due to the multiple part of speeches and sense entries, such as those in using LR analysis algorithm. It is generalization of languages that contributes to multiple senses of words and contexts that result in different sense uses. What can be done is to categorize pragmatic contexts and base sense tagging in corpora and assign weight values to different senses.

C. The Methods to Calculate and Annotate

Generally speaking, it is to convert frequency of occurrences of different senses in contexts into weight values and store this information in corpora. Enlightened by Bayes' methods of word sense disambiguation, our study adjusts his methods to processing both senses and contexts, and makes optimal use of the maximum probability, i.e. calculating probability for every sense and converting the data into percentage to get the weight vale and storing the value in a corpus. The steps are as follows:

1) Establish Corpora with Scientific Classification

As Bayes' methods of word sense disambiguation depend on sentences, similarly our methods rely on different context-oriented corpora, namely C1, C2, C3, and the relation of Si with C1, C2, C3 is clearly specified.

2) Use Bayes' Mthods to Calculate Weight Values

Weights (W1,W2,W3...) in those corpora(C1, C2, C3…) and maximum probability of multiple senses(Si) can be obtained by using the following formula with a tested result of accuracy level of more than 90% in this study:

$$\text{Formula 1: } P(Si \mid C)\frac{P(C \mid Si)P(Si)}{P(C)}$$

$$\text{Formula 2: } P(C \mid Si) = \prod_{w \in c} P(w \mid Si)$$

$$\text{Formula 3: } P(w \mid Si) = \frac{N(w, Si)}{N(Si)},\ \ P(Si) = \frac{N(Si)}{N(A)}$$

$$\text{Formula 4: } P(Si \mid C) = \left[\prod_{w \in c} \frac{N(w, Si)}{N(Si)}\right]\frac{N(Si)}{N(A)}$$

The obtained probability values will then be converted to weight values in terms of percentage, so P(C) can also be omitted.

3) Convert P(Si|C) of a Word to Weight Value in Percentage and Store the Data in the Corpora

Sense weight values can be obtained by these procedures, in which the maximum probability and weight value for every sense are all calculated. This process can be illustrated step by step with the following example of "吃":

Step one: Pragmatic contexts can be classified into three categories, i.e. military, study and daily life. This Chinese word "吃" has eight senses: chew and swallow, also suck and drink:吃饭 |吃奶 |大吃大喝 ; edible food:小吃 |有吃有穿; absorb liquid:这种纸不吃墨 |这种菜很吃油; live by:吃老本|靠山吃山, 靠水吃水; eliminate(in chess playing and military affairs):吃掉敌人的一个团 |我的车(jū)把他的炮吃了; endure and take on:吃苦 |吃官司 |吃惊; use or consume: 吃力 |吃劲; grasp or comprehend:吃透文件精神|他的话我吃不准 (http://www.ourdict.cn /) .

Step two: Compile three corpora with texts classified as military, study and daily life.

Step three: Tag the word senses in the three corpora built according to pragmatic contexts. Attach tags to the eight senses if or if not found in one of the corpus.

Step four: Choose one corpus based on a certain pragmatic contexts and assign weight values to senses in it. Hidden Markov model in POS tagging can be used: W represents the observed frequency of occurrences in the pragmatic context, such as military; λ represents model parameter, and T represents the frequency of occurrences of different sense entries. Then, calculate the T'=argmaxP(T|W, λ) and retain the data of P(T|W, λ). So, probability values can be converted into weight values regarding to this sense in the corpus of military context (other tagging methods can also be used, but are not of concern in this paper). Similarly, senses in other two corpora can be tagged in this way. Obviously, weight values will vary with different senses across the corpora.

After the tagging, set calculation for eigenvalue can be improved to better use LR aligoriam in sentence processing. For example, it will be unrealisitic to use LR algorithm and can be slightly better to use general LR algorithm to process a large amount of sentences such as those with "吃" in "被我们吃掉了" before using our aforementioned methods. After applying this methods and when using LR algorithm in processing, the maximum weight value will be passed to the parsing table. Consequently, in military context, weight value of the sense "eliminate (in chess playing and military affairs)" is 60% and thus this sense entry was passed into the parsing table. In general LR algorithm, weight values are also taken as the references to better control the parser and improve efficiency, but not discussed in this paper.

V. REMAINING PROBLEMS TO BE SOLVED

As to the word senses which co-occur in the same context, it is still hard to deal with them by following our methods. Conversely, Bayes' methods can be applied as in dealing with the following example of "锁": gadgets used on doors, closets, cases, boxes and drawers and opened with keys or passwords:锁具|铁锁|开锁|上锁; fastened with locks: 锁门|把车锁好; closed to or kept away from the outside:封锁|闭关锁国; chains:锁链|枷锁|拉锁; using needles to sew clothes with over-lock stitches:锁边|锁扣眼.

All of these five senses have close relevance to the same context, so it is of limitations to reach this problem just from the perspective of correlation between senses and contexts. A comprehensive approach has to be used to better solve the problem, such as combining Bayes' POS tagging methods with scientific categorization of contexts and abstraction of textual features.

To sum up, we have to base our future study on the scientific categorization of various pragmatic contexts, and incorporate grammatical and phonetic knowledge into sentence processing in a comprehensive way. Facing a vast sea of linguistic knowledge and multifaceted natural language processing, we are still attempting by standing on the shoulders of giants and proposing these humble suggestions which concern merely a small portion of a much bigger issue.

References

[1] Ceusters W, Rogers J, Consorti F and Rossi-Mori A. “Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN,” Artificial Intelligence in Medicine, vol 15, Jan. 1999, pp. 5-23, http://dx.doi.org/10.1016/S0933-3657(98)00043-8.

[2] A.Tesser and C. Leone, “Cognitive schemas and thought as determinants of attitude change,”Journal of Experimental Social Psychology, vol.13, July 1977, pp 340–356, http://dx.doi.org/10.1016/0022-1031(77)90004-X.

[3] T.Kaufmann and B.Pfister, “Syntactic language modeling with formal grammars,”Speech Communication,vol.54, Issue 6, July 2012, pp.715–731, http://dx.doi.org/10.1016/j.specom.2012.01.001.

[4] Yushi. Wen,An Introduction to Computational Linguistics, Beijing: The Commerical Press, 2003.

[5] S.Sharifi and N. Mayamei, “Cognitive study of schema in two poems by Sohrab Sepehri, ”Procedia - Social and Behavioral Sciences,Vol. 32, 2012, pp. 329–333,http://dx.doi.org/10.1016/j.sbspro.2012.01.048.

[6] Guangwei Hu, “Metalinguistic knowledge, metalanguage, and their relationship in L2 learners, ”System,vol.39, Issue 1, Mar. 2011, pp. 63-77,http://dx.doi.org/10.1016/j.system.2011.01.011.

[7] Aarne Ranta, “Computational semantics in type theory”, Department of Computing Science, Chalmers University of Technology and the University of Gothenburg, 2003:55.

Published by Atlantis Press, Paris, France.