A Word Selection Model Based On Lexical Semantic Knowledge In English Generation[1]

CHEN Yi-Dong, LI Tang-Qiu, ZHENG Xu-Ling
Department of Computer Science, Xiamen University
Xiamen, Fujian, China, 361004
{ydchen, tqli, xlzheng}@xmu.edu.cn

Abstract

Word selection is an vital factor to improve the quality of machine translation. This paper introduces a new model for word selection based on lexical semantic knowledge, which could deal with the problem significantly better. Meanwhile, the construction of the English lexical semantic knowledge base required for the model in our Chinese-English machine translation system is also discussed in detail.

1 Word selection Methods Based on Lexical Semantic Knowledge in Generation

The task of *Vocabularies Handling* in machine translation is to map source language words or phrases to their corresponding ones in target language. The task should be performed in almost every stage of machine translation, since words are basic elements of a sentence. A word in a source language can be translated into many different ones in the corresponding target language, since there exist 1 to N mapping between words in different languages due to the homophony and synonyms. But only one of them should be chosen according to the context. Such work is called *Word selection*. It is common practice that if one target word is selected improperly during the word selection, the sentence of the translation becomes quite unreadable, or even its meaning is much different from the source sentence. Word selection is regarded as one of the most important and difficult problem in machine translation. (Liu Xiaohu et al., 1998).

With the development of machine translation, researchers realized that it is more important to consider its semantic constraints in dealing with the problem of word selection than syntax constraints of each word candidates, and are now paying more and more attention to applying of semantic knowledge in machine translation. The following (in 1.1 and 1.2) are two frequently used methods of this kind.

1.1 Semantic Pattern Based Method

In this method, a semantic pattern consists of a headword and its one or more slots of semantic constraints. The semantic pattern base with a great number of such patterns should be constructed first. In word selection, the probability of each candidate can be calculated by comparing the semantic slot constraints of the pattern with the actual semantic environment of a concept, the interlingua structure. The interlingua structure is structurally similar to the pattern but contains the concept to be expressed with proper target word. Finally, one pattern with the highest probability will be chosen as the base of the word selection.

This method is usually referred to as *Rationalist Method* and was first used in DOGENES (Nirenburg et al., 1998) developed in Carnegie Mellon University.

There are a few main weak points of this method. First, the pattern base is usually constructed manually, and it is hard to construct a good one without losses. Also, subjective factors will be introduced while constructing such a pattern base. Secondly, the semantic slot constraints in patterns manually made are usually high level concepts, so the variety and particularity in the natural language

[1] This paper was supported by the Chinese 863 High Tech Research Fund (2001AA114110), and the Fund of Key Research Project of Fujian Province (2001H023)

could not be reflected easily. Therefore, there will often be more than one result chosen after this stage, because many candidates have the same probability. Third, the semantic slot constraints are qualitative constraints and the quantitative differences of language phenomena could not be embodied.

1.2 Example-Based Method

Example based method is an *Empiricist Method*. It was proposed as a new model of machine translation at first. In performing a sentence translation in example based method, the most similar example to the input sentence, together with its corresponding translation, will be found out from a large scale bilingual corpus. Then the corresponding translation will be used as the result, or with some necessary adjustments.

This same idea could also be used in many stages of machine translation, especially those involved with disambiguation. For example, this idea was used to deal with word sense disambiguation in a Chinese-English machine translation system (Yang Xiaofeng et al., 2001). Similarly, the idea could also be used to deal with the word selection problem. The key advantage of using this idea is that the variety and particularity in the natural language could be taken account of in the course of word selection.

There are also some main problems with the method. First, the examples in the example base should be selected carefully, and should be somewhat representative. Otherwise the result of the word selection using this example base would become unreliable. The problem could be resolved by constructing example base as large as possible with examples extracted from a real corpus randomly. Meanwhile, the "Combination Explosion" problem will likely to occur in the process of word selection, if the scare of the example base becomes very large.

1.3 The Hybrid Method

Although the two methods of word selection mentioned above have their merits and demerits respectively, they complement each other well: The *Semantic-Pattern-Based Method* is somehow simpler and has a smaller computation complexity compared to *Example-Based Method*. On the other hand, the *Example-Based Method* can take into account the variety and particularity of languages while the *Semantic-Pattern-Based Method* does it relatively deficient in this aspect.

In order to make use of the complement-ness of these two methods, we put forward a Hybrid method. In this method, word selection is divided into two stages. In first stage, the *Semantic-Pattern-Based Method* is used, and many improper candidates will be got rid of in this stage, hence the sides of the candidate set will become smaller. Then, in second stage, the *Example-Based Method* is used, and the quantitative language knowledge will be utilized to select the best result (Chen Yidong et al., 2001). Figure 1 shows the process briefly.

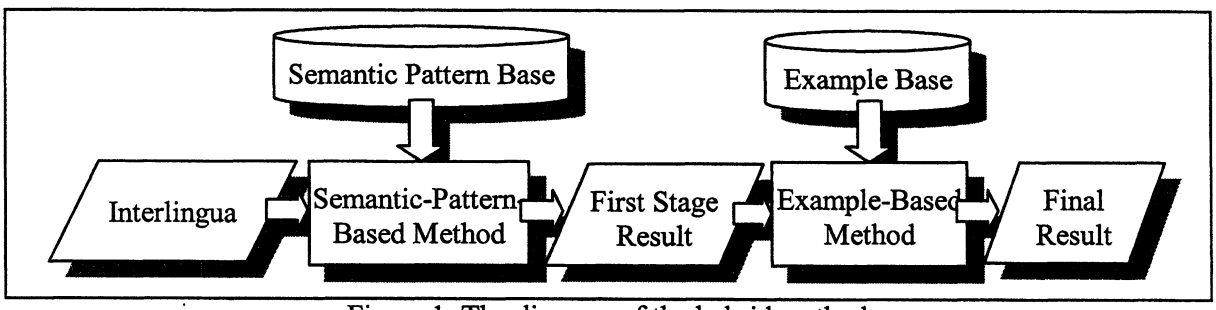

Figure 1. The diagram of the hybrid method

It is clear that the basic component of the *Hybrid Method* is a well-organized *Lexical Semantic Knowledge Base*, which consists of a semantic pattern base and an example base. The details of constructing such a knowledge base may be different from system to system, due to the different language pairs and different semantic representation adopted etc. But the main idea should be common. The design of the structure and organization of the base is very important to realize our goal. The rest of this paper will introduce in detail the construction of the lexical semantic knowledge base

used in our interlingua-based Chinese-English machine translation system. Here, the interlingua is a frame-like representation which utilizes *HowNet* (Dong Zhendong) as its semantic basics.

2 Construction of the Example Base

2.1 The Knowledge Source

In order to fulfil the need of natural language translation, the corpus should be large enough and come from a real corpus. The examples used in our system are extracted from a real corpus named *SEMCOR* (Miller et al., 1993), which is a text corpus and is normally released with *WordNet* (Chen Qunxiu, 1998). The text of *SEMCOR* stems from *Brown Corpus* (Francis et al., 1982) and is semantically tagged according to the *WordNet*. *SEMCOR* is tagged mainly by hand with the help of some tagging tools.

In SEMCOR released with WordNet 1.6, there are about 352 tagged SMGL-like files, among which about 183 ones are entirely tagged and others only have their verbs tagged. In these 183 entirely-tagged files there are altogether 359,732 words or so, with around 193,373 ones semantically-tagged. The proper name in SEMCOR such as names of person, group and location and etc. are tagged with additional taggers according to their types. For instance, names of person are tagged with "person", names of group are tagged with "group" and names of location are tagged with "location" and etc.

2.2 Knowledge Organization

To build an example base, we must acquire required knowledge from the knowledge source and organize them into a proper form. To do so, two inconsistent problems should be resolved in our system:

First, the tag information of knowledge source is inconsistent with the semantic tag information required in the semantic disambiguation: as mentioned above, *SEMCOR* is tagged with tags used in *WordNet*, while the structure of our interlingua is constructed according to *HowNet*. The difference should be resolved one way or the other so that the knowledge acquired from the corpus can be utilized in reasoning. In our system, we adopt the tag system used in *HowNet* and all the examples extracted from the source must be converted into a proper form accordingly.

Secondly, there are another inconsistence, the inconsistence of relation names and structures. SEMCOR is neither syntactically nor semantically analyzed, only linear collocations could be extracted from the corpus. This kind of structure is not easily utilized in the disambiguation process. To facilitate the reasoning process, the examples should be constructed as frame-like structure, which is similar to the structures adopted in our interlingua. In other word, we must build frame-like examples from linear collocation relations abstracted from the corpus. It means that there is much information to be resolved.

2.3 Approach to Construct the Example Base

To ensure that the examples can be easily be utilized for disambiguation, its structure should be similar to the structure of our interlingua. Following is its formal definition as shown in Figure 2:

The steps of the construction of the example base can be described informally as follows:

First, to change the corpus tagged in WordNet formalism into one tagged in HowNet formalism. To accomplish the step, we designed an algorithm that can map WordNet senses in the corpus to HowNet concepts effectively.

The main idea of the algorithm is that: Each sense in the WordNet is represented as a set of synonym, called Senset. Given a WordNet Senet, each word in the Senset has a series of possible corresponding words, each of which has series corresponding concepts in the HowNet. All the probable combinations will be enumerated and the common concept occurred with in each word or the one occurred most often will be chosen as the meaning representation of the sense.

Using the algorithm, a mapping list from WordNet senses to HowNet concepts was constructed. Further, this list is used to transform *SEMCOR*, the corpus tagged with tags used in *WordNet*, to the corpus, tagged with tags used in the *HowNet*.

```
<Example>:  :  =(<Concept>  (<  Sem_Slot  >  +  ))
<Concept>:  :  =(<Word>  <Cat>  <Def>)
<Sem_Slot>::=(<SS_Name>  (<SS_Val>+))
<SS_Val>::=(<Concept>  <Prob>)
<Word>::=eat  |  see  |  look  |  .....
<Cat>::=N  |  V  |  ADJ  |  .....
<Def>::=  a  valid  HowNet  concept
<SS_Name>::=AGENT  |  THEME  |  .....
<Prob>::= a number value between 0 and 1
```

Figure 2. The structure definition of the examples

Secondly, linear collocation relations would be extracted from the corpus already tagged in HowNet formalism, and their semantic relationships among the words in each collocation be inferred and then transformed into frame-like forms. To do this, we designed an example transforming procedure which can transform the collocations automatically into frame-like forms. After a process of automatic transformation, a manual adjustment process is performed to correct some errors in the results.

Thirdly, to reorganize the examples and make the example base optimized, two procedures are performed: The first is to delete the redundant entries and count the frequency. The second is to merge the examples that has identical headword. Doing so, the redundancy can be reduced and the base be packed into a smaller size.

2.4 Results and Some Instances

Following the approach described in 2.3, an example base with 4362 examples was constructed. To list a few, some examples are shown below:

```
((keep  (SENSE  keep|保持)  (CAT  V))
 ((THEME
  (((lead  (SENSE  surpass|强过)  (CAT  N))  0.1)
     ((fashion  (SENSE  attribute|属性,SocialMode|风气,&entity|实体)  (CAT  N))  0.1)
     ((faith  (SENSE  experience|感受,believe|相信)  (CAT  N))  0.1)
  .....))
   .....))
((keep  (SENSE  SetAside|留存)  (CAT  V))
 ((THEME
  (((moisture  (SENSE  attribute|属性,dampness|湿度,&physical|物质)  (CAT  N))  0.1)
     ((package  (SENSE  physical|物质)  (CAT  N))  0.1)
     ((letter  (SENSE  letter|信件)  (CAT  N))  0.1)
  .....))
   .....))
((reserve  (SENSE  SetAside|留存)  (CAT  V))
 ((THEME
  (((complaint  (SENSE  thought|念头,different|异,#oppose|反对)  (CAT  N))  0.1)
     ((power  (SENSE  attribute|属性,ability|能力,&physical|物质)  (CAT  N))  0.1)
     ((right  (SENSE  rights|权利)  (CAT  N))  0.1)
  .....))
   .....))
((conserve  (SENSE  SetAside|留存)  (CAT  V))
```

((THEME
(((*energy*
(SENSE attribute|属性,strength|力量,$function|活动,&AnimalHuman|动物)
(CAT N)) 0.1)
((*resources* (SENSE material|材料,generic|统称) (CAT N)) 0.1)
.....))
.....))

3 Construction of the Semantic-Pattern Base

3.1 Improvement of Selection Method Based on Semantic Pattern

As mentioned in the first part of the paper, the pattern based method plays an important part in the first stage of the hybrid method. But the traditional semantic pattern based method has some demerits and it may influence the performance of the whole system. Two important improvements are proposed to overcome its difficulties and to improve its performance. One is in the way to build semantic pattern base. The other is the improvement of the structure of the pattern representation itself.

3.1.1 Automatic Approach to construct the Pattern Base

With the traditional semantic pattern based method, the most difficulty is how to build the pattern base. Manual approach to build vast number of semantic patterns not only results in a big workload but also leads to the introduction of subjectivity of the author who builds the patterns. So the way to construct the pattern base has to be changed.

In our system, the semantic pattern base will be constructed automatically from the example base mentioned in 2.4. Since the examples in the base have been semantically tagged and the relations among words in them have been well determined, it is not difficult to utilize the example base as the training knowledge source to extract the semantic patterns. Obviously the automatic approach to construct the pattern base well overcome the short-comes of great workload and subjectivity.

3.1.2 Fuzzy Semantic Pattern

The semantic patterns used in the traditional method are all yes or no rules. If the corresponding semantic pattern to a valid collocation is not included in the pattern base, the relevant candidate will be rejected incorrectly. It lacks the flexibility characterized by a quantitative matching.

To solve the problem, the structure and content of the semantic pattern should be improved. In our system, an additional field, *Probability*, is introduced into a semantic slot constraint, and the amended semantic patterns are so called *Fuzzy Semantic Patterns* (Chen Yidong et al., 2002). By introducing this additional field, the semantic pattern will be able to support inexact match. Obviously it makes the method become more flexible. Since our pattern is extracted from the real example base, it is possible for the probability field to be calculated from the corpus statistically without difficulty.

3.2 Organization of the Semantic Pattern Base

As described in 3.1.2, fuzzy semantic patterns consist of semantic slot constraints that indicate the collocation relation of a headword. The structure of fuzzy semantic patterns is shown below.

It can be seen that, similar to the example base, the semantic slot constraints with the same headword will be merged into the same pattern, and similarly, the values of the semantic slot constraints with the same name in a pattern will be merged into the same value list.

<Sem_Pattern>::=(<Concept> (<SS_Constraint>$^{+}$))
<Concept>::=(<Word> <Cat> <Def>)
<SS_Constraint>::=(<SS_Name> (<SSC_Val>$^{+}$))
<SSC_Val>::=(<Atom> <Prob>)

<Word>::=eat | see | look | ...
<Cat>::=N | V | ADJ | ...
<Def>::=a valid HowNet concept
<SS_Name>::=AGENT | THEME | ...
<Atom>::=EAT|吃 | HAPPY|福 | ...
<Prob>::=a number value

Figure 3. The structure definition of fuzzy semantic patterns

3.3 Train Algorithm of Fuzzy Semantic Patterns

As we can see in Figure 2 and Figure 3, the structure of the semantic patterns and examples in our system are very similar to each other. So it's easy to train fuzzy semantic patterns from the example base. The algorithm to train fuzzy semantic patterns could be described as follows (Chen Yidong et al., 2002):

Given an example (CH (SS_1 SS_2 ... SS_n)), where CH is the concept of the headword and each SS in the list stands for a semantic slot of this headword respectively (the detail definition is shown in Figure 2), a fuzzy semantic pattern with a structure that meets with the definition shown in Figure 3 will be trained using the following steps:

1° For each semantic slot SS, which is of the form (SSN (SSV_1 SSV_2 ... SSV_m)), with SSN as its name and the SSVs list as the list of collocation instances of CH in it, construct a corresponding semantic slot constraint, SSC, whose form is (SSN ($SSCV_1$ $SSCV_2$ ... $SSCV_m$)). In this step, two sub-steps are to be performed:

1.1° Get each $SSCV$ from the corresponding SSV and form the list of $SSCV$s. Meanwhile, the probability field of each $SSCV$ will be calculated respectively. (See the more detail description in Chen Yidong et al., 2002)

1.2° Use the value list, ($SSCV_1$ $SSCV_2$ ... $SSCV_m$), and the semantic slot name, SSN, to construct a semantic slot constraint.

2° Construct the final fuzzy semantic pattern using CH and SS that is formed in 1°.

Figure 4. The algorithm to train fuzzy semantic patterns

3.4 Results and Some Instances

Using the algorithm above, 4362 semantic patterns were trained automatically from the example base described in 2.4. As is mentioned above, when constructing the fuzzy semantic pattern base, all the collocation information related to the same headword will be merged into the same pattern, and so is the construction of the example base, hence, although the number of entries in the pattern base and in the example base is identical, the actual scale of them are not the same. Some instances of the semantic patterns are shown below.

((*keep* (SENSE SetAside|留存) (CAT V))
((THEME
((physical|物质 1.2) (attribute|属性 1.1) (letter|信件 1.1) (thing|万物0.17)
(readings|读物 0.14) (entity|实体 0.034) (artifact|人工物 0.013) (inanimate|无生物 0.0042)
.....))
.....))
((*reserve* (SENSE SetAside|留存) (CAT V))
((THEME
((thought|念头 1.1) (attribute|属性 1.1) (rights|权利 1.1) (thinking|思想 0.14)
(mental|精神 0.12) (thing|万物0.021) (entity|实体 0.0041)
.....))
.....))

((*conserve* (SENSE SetAside|留存) (CAT V))
((THEME
((attribute|属性 1.1) (material|材料 1.1) (artifact|人工物 0.092)
(inanimate|无生物 0.031) (physical|物质 0.0076) (thing|万物0.0013) (entity|实体 0.0003)
.....))
.....))

4 Conclusion

In order to improve the quality of machine translation, semantic knowledge should be utilized, especially in word selection, an important process in machine translation. Based on the investigation and analysis of several commonly used methods, a hybrid method is proposed in this paper. The method combines the advantages of the two traditional methods and shows its extra flexibility. In implementing such a method, how to obtain lexical semantic knowledge becomes the key. Therefore, the main part of this paper focuses on the construction of the example base and semantic pattern base used. Using the algorithms presented in this paper, a lexical semantic knowledge base with considerable scale was constructed successfully in our system.

References

Chen Qunxiu. 1998. An On-line Thesaurus: WordNet. *Application of Language and Literary*, (2):93-99

Chen Yidong, Li Tangqiu, Hong Qingyang and Zheng Xuling. 2001. Designing of a Model of Word selection in English Generation. *Journal of Chinese Information Processing*, 15(6):19-26.

Chen Yidong, Li Tangqiu and Zheng Xuling. 2002. Fuzzy Semantic Pattern and Its Application in the Word selection for English Generation. *Journal of Chinese Information Processing*, 16(5):15-22.

Dong Zhendong and Dong Qiang. HowNet. http://www.keenage.com

Francis W. N., and Kucera H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin Company, Boston.

Liu Xiaohu and Li Sheng. 1998. Target Word Selection in Machine Translation Based on Machine Learning. *Computer Research & Development*, 35(10):946-950

Miller G.A., Leacock C., Tengi R. and Bunker R.T. 1993. A Semantic Concordance. *In: Proceedings of the ARPA Workshop on Human Language Technology*.

Nirenburg, Sergei and Nirenburg, Irene. 1998. A framework for word selection in NLG. In: *Proceedings of the 12th International Conference on Computational Linguistics*.

Yang Xiaofeng, Li Tangqiu and Hong Qingyang. 2001. Word Sense Disambiguation Method in Chinese-English Machine Translation System. *Journal of Chinese Information Processing*, 15(3):22-28.