Automatic Information Transfer Between English And Chinese

Jianmin Yao, Hao Yu, Tiejun Zhao
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China, 150001
james@mtlab.hit.edu.cn

Xiaohong Li
Department of Foreign Studies
Harbin Institute of Technology
Harbin, China, 150001
goodtreeyale@yahoo.com.cn

Abstract

The translation choice and transfer modules in an English Chinese machine translation system are introduced. The translation choice is realized on basis of a grammar tree and takes the context as a word bag, with the lexicon and POS tag information as context features. The Bayes minimal error probability is taken as the evaluation function of the candidate translation. The rule-based transfer and generation module takes the parsing tree as the input and operates on the information of POS tag, semantics or even the lexicon.

Introduction

Machine translation is urgently needed to get away with the language barrier between different nations. The task of machine translation is to realize mapping from one language to another. At present there are three main methods for machine translation systems [Zhao 2000]: 1) pattern/rule based systems: production rules compose the main body of the knowledge base. The rules or patterns are often manually written or automatically acquired from training corpus; 2) example based method. The knowledge base is a bilingual corpus of source slices S' and their translations T' Given a source slice of input S, match S with the source slices and choose the most similar as the translation or get the translation from it. 3) Statistics based method: it is a method based on monolingual language model and bilingual language model. The probabilities are acquired from large-scale (bilingual) corpora.

Machine translation is more than a manipulation of one natural language (e.g. Chinese). Not only the grammatical and semantic characteristics of the source language must be considered, but also those of the target language. To sum up, the characteristics of bilingual translation is the essence of a machine translation system.

A machine translation system usually includes 3 sub-systems [Zhao 1999]：(1) Analysis: to analyse the source language sentence and generate a syntactic tree with syntactic functional tags; (2) Transfer: map a source parsing tree into a target language parsing tree; (3) Generation: generate the target language sentence according to the target language syntactic tree.

The MTS2000 system developed in Harbin Institute of Technology is a bi-directional machine translation system based on a combination of stochastic and rule-based methods. Figure 1 shows the flow of the system.

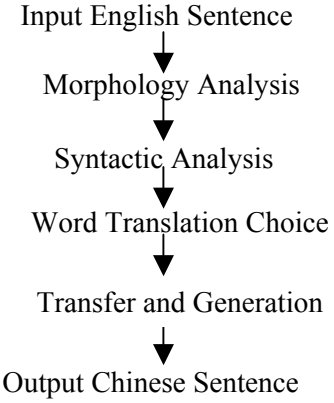

Figure 1 Flowchart of MTS2000 System

Analysis and transfer are separated in the architecture of the MTS2000 system. This modularisation is helpful to the integration of stochastic method and the rule based method. New techniques are easier to be integrated into the modularised system. Two modules implement the transfer step and the generation step after analysis of the source sentence. The specific task of transfer and generation is to

produce a target language sentence given the source language syntactic tree. In details, given an English syntactic tree (e.g. S[PP[In/IN BNP[our/PRP$ workshop/NN]] BNP[there/EX] VP[is/VBZ NP[no/DT NP[NN[machine/NN tool/NN] SBAR[but/CC VP[is/VBZ made/VBN PP[in/IN BNP[China/NNP]]]]]]]]), using knowledge sources such as grammatical features, simple semantic features, construct a Chinese syntactic tree, whose terminal nodes compromise in sequence the Chinese translation. The input sentence are analysed using the morphology analyser, part-of-speech tagger, and syntactic analyser. After these steps, a syntactic parsing tree is obtained which has multiple levels with functional tags [Meng 2000]. Followed is the parser flow:

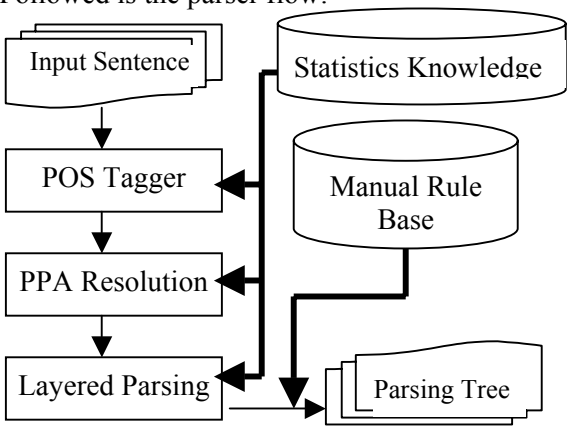

Figure 2. Parser based on Hybrid Methods

At present, our English parser is able to generate syntactic tree in comparative usable way. The English parsing tree, with the basic information about relationship among the nodes in the source sentence, also with semantic information of the nodes, is input to the module of transfer and generation. The information of the nodes is the starting point of transfer and generation. After syntactic parsing, the task of transfer and generation includes word translation choice of ambiguous words, word order adjustment and insertion/deletion of some functional words. Transfer and generation are implemented using two modules: one is for word translation choice, the other for structure transfer and translation modification.

1 Parsing Based Translation Choice

First we will give a formal description for translation choice in machine translation [Manning 1999]: Suppose the source sentence to be translated to be ES. In the sentence the ambiguous word EW has M target translations CW1, CW2, ... CWM. And the translations occurs in a specific context C with probabilities P(CW1 | C)、P(CW2 | C), ... P(CWM|C)。From the Bayes minimum error probability formula, we get:

$$CW = \arg\max[P(CWk|C)]$$
$$= \arg\max[\log P(CWk) + \log P(C|CWk)] \quad (1)$$

Generally when the condition fulfills $P(CW1|C)>P(CW2|C)>...>P(CWM|C)$, we may choose CW1 as the translation for EW. From the Naïve Bayes formula：

$$P(C|CWk) = P(\{vj \mid vj \text{ in } C\}|CWk)$$
$$= \prod_{Vj \text{ in } C} P(vj|sk) \quad (2)$$

So formula (1) can be rewritten as:

$$CW = \arg\max[P(CWk|C)]$$
$$= \arg\max[\log P(CWk)+\sum_{Vj \text{ in } C}\log P(vj|CWk)] \quad (3)$$

Where P(CWk) denotes the probability that CWk occurs in the corpus; P (vj| CWk) denotes the probability that the context feature vj co-occurs with translation CWk。

A general algorithm of supervised word sense disambiguation is as follows:

```
1.  comment: Training
2.  for all senses sk of w do
3.      for all words vj in the vocabulary do
4.          P(vj|sk) = C(vj, sk)/C(vj)
5.      end
6.  end
7.  for all senses sk of w do
8.      P(sk) = C(sk)/C(w)
9.  end
10. comment: Disambiguation
11. for all sense sk of w do
12.     Score(sk) = logP(sk)
13.     for all words vj in the context window c do
14.         score(sk) = score(sk) + logP(vj|sk)
15.     end
16. end
17. choose s' = argmaxskscore(sk)
```

Figure 5. Bayesian disambiguation

From the above formal description we can see that the key to the stochastic word translation is to select proper context and context features Vj. Present methods often define a word window of some size, i.e. to suppose only words within the window contributes to the translation choice of the ambiguous word. For example, [Huang 1997] uses a word window of length 6 words for word sense disambiguation; [Xun

1998] define a moveable window of length 4 words; [Ng 1997] uses a word window with offset ±2. But two problems exist for this method: (1) some words that are informative to sense disambiguation may not be covered by the window; (2) some words that are covered by the word window really contribute nothing to the sense choice, but only bring noise information.

After a broad investigation for large-scale ambiguous words, we choose the context according to the correlation of the context words with the ambiguous word, but not only the distance from the word.

From the above analysis, we choose the translation choice method based on syntactic analysis. Place the module of translation choice between the parser and the generator; acquire a context set for the ambiguous word. When choosing the translation, we may take the context set as a word bag, i.e. the grammatical context as word bag. No single word is considered but only that lexical and part-of-speech information are taken as context features. Bayes minimum error probability is taken as evaluation function for word translation choice.

In this paper, grammatical context is considered for word translation choice. The structure related features of the ambiguous words are taken into account for fully use of the parsing result. It has the characteristics below: (1) The window size is not defined by human but on basis of the grammatical structure of the sentence, so we can acquire more efficiently the useful context features; (2) The unrelated context features in sentence structure are filtered out for translation choice; (3) The features are based on the structure relationship, but not 100% right parsing result. From the above characteristics, we can see the method is really practical.

2 Rule Based Transfer & Generation

For MTS2000, structural transfer is to start from the syntactic parsing tree and construct the Chinese syntactic tree. While the generation of Chinese is to generate a word link from the Chinese tree and build the translation sentence [Yao 2001]. This module has adopted the rule-based knowledge representation method. The design of the rule system is highly related to the performance of the machine translation system.

The rule description language of the machine translation system is in the form of production rules, i.e. a rule composed of a conditional part and an operational part. The conditional part is a scan window of variable length, which uses the context constraint conditions such as phrases or some linguistic features. The operational part generates the corresponding translation or some corresponding generation features in the operational part. If the conditions are met, the operations will be performed. The representation of the rule system has shown a characteristic of the system, that is the integration of transfer and generation. The rule description language is similar to natural language and consistent with human habits. Multiple description methods are implemented.

The conditional part of the rules is composed of node numbers and "+" symbols that is used to link the nodes. The operation part consists of corresponding conditional parts and translations and also, if necessary, some action functions.

For example, the rule to combine an adjective and a noun to generate a noun phrase is as follows:

0:Cate=A + 1:Cate=N

->0:* + 1:* + _NodeUf(N，0，1)

in which, "*" stands for corresponding translation of the nodes, _NodeUf() is a function that combines the nodes to generate a new node. The new translation is generated at the same time with the combination of nodes.

In general, the English Chinese machine translation system has the following features in the transfer and generation phase:

1) The grammatical and semantic features are described by a string composed of frame name and values linked with "=";
2) The conditions may be operated by "and", "or" and "not";
3) Nodes in the same level of the sentence may be scanned and tested arbitrarily;
4) The action functions and test functions can generate corresponding features for feature transmission and test.

The rules are organized into various levels. All the rules are put in the knowledge base with

part-of-speech as the entry feature. The rules have different priorities, which decide their sequence in rule matching. In general, the more specific the rule, the higher is its priority. The more general the rule, the lower is its priority. The levels of the rules help resolve rule collision.

Conclusion

The system prototype has been implemented and large-scale development and refinement are under progress. From our knowledge of the system, knowledge acquisition and rule base organization is the bottleneck for MTS2000 system and similar natural language processing systems. The knowledge acquisition for word translation choice needs large-scale word aligned bilingual corpus. We are making research on new word translation methods on basis of our 60,000-sentence aligned bilingual corpus. The transfer and generation knowledge base are facing much knowledge collision and redundancy problem. The organization technique of knowledge base is also an important issue in the project.

References

Tie-Jun Zhao, En-Dong Xun, Bin Chen, Xiao-Hu Liu,Sheng Li, Research on Word Sense Disambiguation based on Target Language Statistics, Applied Fundamental and Engineering Journal, 1999，7（1）：101-110

Meng Yao, Zhao Tiejun, Yu Hao, Li Sheng, A Decision Tree Based Corpus Approach to English Base Noun Phrase Identification, Proceedings International conference on East-Asian Language Processing and Internet Information Technology, Shenyang, 2000: 5-10

Christopher D. Manning, Hinrich Sch ü tze, Foundation of Statistical Natural Language Processing. The MIT Press. pp229-262. 1999.

Chang-Ning Huang, Juan-Zi Li, A language model for word sense disambiguation, 10th anniversary for Chinese Linguistic Society, October, 1997, Fuzhou

En-Dong Xun, Sheng Li, Tie-Jun Zhao, Bi-gram co-occurrence based stochastic method for word sense disambiguation, High Technologies, 1998, 10(8): 21-25

Hwee Tou Ng. Exemplar-Based Word Sense Disambiguation: Some Recent Improvements. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2), August 1997

Tie-Jun Zhao etc, Principle of Machine Translation, Press of Harbin Institute of Technology, 2000.

Jian-Min Yao, Jing Zhang, Hao Yu, Tie-Jun Zhao,Sheng Li, Transfer from an English parsing tree to a Chinese syntactic tree, Joint Conference of the Society of Computational Linguistics, 2001, Taiyuan.-138.