Pre-reordering Model of Chinese Special Sentences for Patent Machine Translation

Renfen Hu, Zhiying Liu, Lijiao Yang, Yaohong Jin
Institute of Chinese Information Processing,
Beijing Normal University,
Beijing 100875, China
bnuhurenfen@126.com
{liuzhy,yanglijiao,jinyaohong}@bnu.edu.cn

Abstract

Chinese prepositions play an important role in sentence reordering, especially in patent texts. In this paper, a rule-based model is proposed to deal with the long distance reordering of sentences with special prepositions. We firstly identify the prepositions and their syntax levels. After that, sentences are parsed and transformed to be much closer to English word order with reordering rules. After integrating our method into a patent MT system, the reordering and translation results of source language are effectively improved.

1 Introduction

As typical technical documents, Patents have proven to be suitable for automatic translation for its strict format and united writing pattern (Jin and Liu, 2011), and patent machine translation (MT) is one of the major application fields of MT. However, sentences in patent are known for their complicated structures with multiple verbs and prepositions. Some Chinese prepositions are used to change the original S-V-O order of sentences, such as 把(BA), which make it more difficult for reordering in Chinese-English machine translation. In ancient Chinese, these prepositions are mostly verbs or other notional words, and in modern Chinese they became grammatical markers after diachronic grammaticalization. Huang(1998) and Miao(2005) discussed the reordering function of these prepositions, and defined them as Logic-0 (L0) words.

A linguistic study by Zhang(2001) shows that more than 20% Chinese sentences are reordered by the prepositions, including 把(BA), 将(JIANG), 向(XIANG), 与(YU), 对(DUI), 给(GEI), 被(BEI), 由(YOU) and 为(WEI). After analyzing sentences of 500 Chinese patent documents, we find that L0 words appear more frequently in patent texts. Sentences with 1 L0 word occupy 30.75%, sentences with 2 L0 word occupy 9.05%, and sentences with ≥3 L0 words occupy 2.10%. Therefore, Chinese special sentences with L0 words are concerned in this paper, and we will present a pre-reordering model of these special sentences for patent translation.

Figure 1 and Figure 2 show an example illustrating some of the differences in word order between Chinese and English.

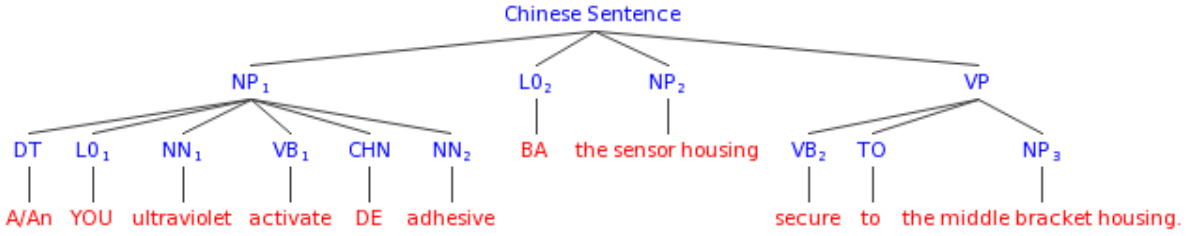

Figure 1. Chinese syntax tree of the example sentence

Proceedings of SADAATL 2014, pages 40–47,
Dublin, Ireland, August 24, 2014.

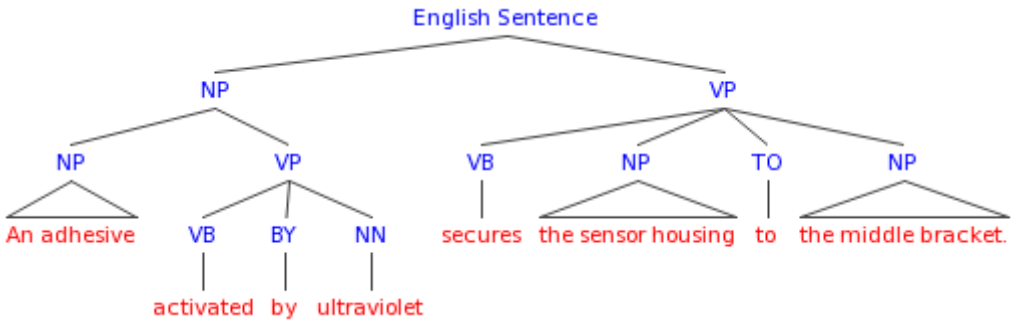

Figure 2. English syntax tree of the example sentence

The example shows a Chinese sentence whose literal translation in English is:

A/an **YOU(L0)** ultraviolet activate DE adhesive **BA(L0)** the sensor housing secure to the middle bracket housing. *(一种由紫外线激活的粘合剂把传感器壳体固定在中支架上。)*

And where a natural translation in English would be

An adhesive activated by ultraviolet secures the sensor housing to the middle bracket.

As exemplified by this sentence, differences of word order between Chinese and English are determined by BA and YOU, and they are in two different levels in the syntax tree.

In order to produce a good English translation, we firstly identify L0 words in two levels, and parse the sentence into chunks with core predicate and L0 words. Based on the sentence parsing, chunks are reordered according to related rules, transforming Chinese special sentence into a word order that is closer to that of English. After integrating into a patent MT system running in SIPO (State Intellectual Property Office of People's Republic of China)[1], our model performs better than the baseline system and Google Translate in an open test, and it greatly improves the performance of patent translation.

After a discussion of related work in section 2 and an introduction to semantic features in section 3, we will discuss the reordering model in section 4. Section 5 presents the processing steps, and section 6 gives the experiment and evaluation. Finally we draw some conclusions in section 7.

2 Related Works

Nowadays statistical machine translation (SMT) is the most widely used method in MT field, and reordering approaches are proved necessary in SMT performance(Xia and McCord, 2004; Collins et al, 2005). Most SMT systems employ some mechanism that allows reordering of the source language during translation(Wang et al, 2007), and researchers find that reordering based on syntactic analysis are effective for handling systematic differences in word order between source and target languages (Xia and McCord, 2004; Collins et al, 2005).

Although sentence structure of source language has been taken into consideration, most SMT systems make use of syntax information in decoding stage (Lin, 2004; Ding and Palmer, 2005; Quirk et al, 2005; Liu et al, 2006, Huang et al, 2006). Wang et al.(2007) firstly incorporate a Chinese syntactic reordering method into preprocessing stage of a statistical MT system, and achieve a significant improvement in reordering accuracy. Zhang et al.(2007) propose a chunk-level method with reordering rules automatically learned from source-side chunks, it shows improvement of BLEU score and better computational efficiency than reordering during decoding in Chinese-English task. Genzel(2010) applies this approach to 8 different language pairs in phrase-based machine translation, and demonstrates that many important order transformations (SVO to SOV or VSO, head modifier, verb movement) can be captured by this approach. An automatic reordering model in preprocessing also works effectively in Japanese-English patent machine translation(Katz-Brown and Collins, 2008).

However, existing methods face difficulties in Chinese-English patent translation. A Chinese patent sentence often contains multiple nested phrases with a number of verbs, prepositions and correlations. In addition to that, ambiguity of L0 words turns it more difficult for language parsers to make syntax analysis. Moreover, reordering rules can hardly be automatically learned from patent sentences with complicate structures. To deal with the long-distance reordering of special sentences in patent texts,

[1] http://c2e.cnpat.com.cn/sesame.aspx

we must fully consider semantic features of L0 words, including their positions, correlations, functions, ambiguities and levels. With the identification of L0 words and their levels, we can parse and reorder a sentence more explicitly.

3 Semantic Features

A linguistic survey shows that S-V-O account for more than 75% of the world's languages, suggesting it may be somehow more initially "obvious" to human psychology(Crystal, 1997). Both modern Chinese and English are S-V-O languages, however, word order in Chinese sentence is often changed by L0 words to emphasize a part of the sentence, or to make nuance of the meaning. Our work aims at reordering Chinese special sentences to organize phrases or words in English order without L0 words. We have defined 9 Chinese prepositions as L0 words in Section 1. To deal with the reordering of Chinese special sentences with these words, we use semantic features from the Hierarchical Network of Concepts theory (HNC theory). In the opinion of HNC researchers, L0 words and verbs are important clues of syntactic and semantic analysis(Jin, 2010). Therefore, we will introduce the features of L0 words and verbs in the following part.

3.1 Types of L0 Words

According to HNC theory, L0 words can be divided into 2 types, L01 and L02(Huang, 1998; Miao, 2005).

Sentence 1 Tom eats a banana.

*Sentence 2 Tom **BA(把)** a banana eat.*

*Sentence 3 A banana **BEI(被)** Tom eat.*

In sentence 1, we know that *Tom* is an agent, and *a banana* is a patient. Sentence 2 and 3 are expressing the same meaning in Chinese with L0 words. It can be seen that the two L0 words reorder the sentence in different ways, BA(把) just exchanges the location of the predicate and the object, while BEI(被) changes the sentence from active to passive voice. BA(把) is a L02 word, and BEI(被) belongs to L01. The types of L0 words are presented in Table 1.

Types	Members	Semantic format	Example sentence
L01	被 (BEI), 由(YOU) , 为(WEI)	Patient+L01+Agent+Predicate	Tom BA(把) a banana eat.
L02	把(BA), 将(JIANG),向(XIANG), 对(DUI), 给(GEI), 与(YU)	Agent+L02+Patient/Recipient+ Predicate	A banana BEI(被) Tom eat.

Table 1. Two types of L0 words

3.2 Levels of L0 Words

By comparing the syntax trees in Figure 1 and Figure 2, we can note that *YOU(由)* appears in a NP(noun phrase), while *BA(把)* appears independently in the sentence. To distinguish the two kinds of L0 words, we define a LEVEL value for L0 words according to their node locations in the syntax tree. L0 word as LEVEL[1] is a child node of S(sentence), while L0 as LEVEL[2] is a child node of NP.

Accordingly, our reordering model includes two modules, the reordering of Sentence with L0 as LEVEL[1] and the reordering of NP with L0 as LEVEL[2]. Therefore, we need firstly identify the level of each L0 word in sentences.

3.3 Collocation with Predicates

More than one L0 word may appear in a sentence, but each L0 is in combination with a certain predicate. As exemplified by the sentence in Figure 1, *YOU(由)* goes with the verb *activate*, while *BA(把)* goes with the verb *secure*. For this reason, L0 and its level can also help to determine the core predicate when a sentence has more than one verb.

Predicates are also classified into 2 types according to their levels in the syntax tree. We give 2 simple English sentences to explain the 2 types, P1 and P2.

Sentence 4 Bob tells(P1) me a secret.

Sentence 5 A secret told(P2) by Bob is spreading(P1).

As labelled in the 2 examples, P1 refers to the core predicate, and P2 is the predicate in a noun phrase. L0 word in level 1 is in combination with P1, while L0 word in level 2 is in combination with P2. The identification of the two types of predicates is introduced in detail by Zhu(2012). Table 2 shows the levels of L0 words and their collocations with predicates.

L0 Level	Parent Node	Collocated Predicates
LEVEL[1]	Sentence	P1
LEVEL[2]	NP	P2

Table 2. The collocations of L0 words and predicates

4 Reordering Model of Chinese Special Sentences

Our model aims at the reordering of Chinese special sentences with L0 words for patent machine translation. With the identification of predicates, L0 words, and their levels (Hu et al, 2013), we can parse the sentence and get a Chinese syntax tree. In this section, we will firstly introduce the transformations and rules in the reordering, and then discuss how to transform the syntax tree to make the word order closer to English sentence. Semantic features of L0 and verbs that we discussed in section 3 will be applied into the model.

4.1 Transformations in the Reordering

There are 5 types of transformations in the processing of our reordering model.

Deletion: L0 words are Chinese prepositions, so we need to delete or substitute them at first.

Addition: Some L0 words have no real meanings, such as 把(BA) and 将(JIANG), we can make other transformations after deleting them. However, some L0 words have preposition meanings that cannot be neglected, so we need to add English prepositions to the new tree. This operation can also be interpreted as "substitution".

Copying: In long distance reordering, some chunks do not need any transformation, so we just copy it to the new syntax tree.

Rearrangement: We need to rearrange the chunks to make the word order closer to English.

Voice Transition: A research by Liu(2011) shows that 95.6% English patent sentences use passive voice. Considering the voice difference between Chinese and English, we transform some active sentences to passive sentences.

4.2 Rule Description

The above 5 transformations are integrated in our reordering rules. We will describe how rules work with Rule 1 as an example.

Rule 1:

(b){(-1)CHK[NP]}+(0)CHN[由]&CHK[L0] +(1)CHK[NP]+(2)CHK[P1]&VV[2]=>(-1)+ COPY[-1,0]+ DEL_NODE(0)+(2){VOI=P}+ ADD_NODE(ENG=[by])+(1)

Each reordering rule includes a left part and a right part, with arrow "=>" as the boundary. *CHK* is short for *chunk*, and *VV* is short for *verb valency*, which is a feature for verbs in our semantic knowledge base. The left part describes chunks in the Chinese syntax tree, and each chunk is marked with a node number. The right part describes the reordering result. In Rule 1, L0 由 is deleted, English preposition *by* is added, *P1* is transformed to passive voice, contents between node 0 and node 1 are copied to the new syntax tree, and the chunks orders are rearranged from *(-1)+(0)+(1)+(2)* to *(-1)+COPY[-1,0]+(2)+by+(1)*.

4.3 Reordering Analysis

After an introduction to our rules and their functions, we will present two examples to illustrate the reordering work.

4.3.1 Reordering of Sentences

After analyzing sentences from 500 patent texts, 51 rules are made to deal with the sentences with L0 as LEVEL[1]. We will discuss this reordering work with Sentence 6 as an example.

*Sentence 6 BA data to be transmitted divide into plural blocks. (**把**待发送的数据分为多个数据块。)*

After identification of L0 words and predicates, we can get a syntax tree as shown in Figure 3.

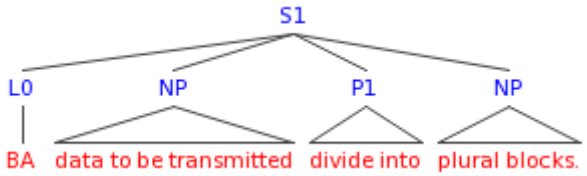

Figure 3. Syntax tree of Sentence 6 before reordering

In reordering stage, the sentence will be transformed by matching the following rule.

Rule 2:

(b){!CHK[NP]}+(0)CHN[把]&CHK[L0]+(1)CHK[NP]+(2)CHK[P1]&VV[3]+(3)CHK[NP] => DEL_NODE(0)+(1)+(2){VOI=P}+(3)

After the transformation, we get a new syntax tree as shown in Figure 4.

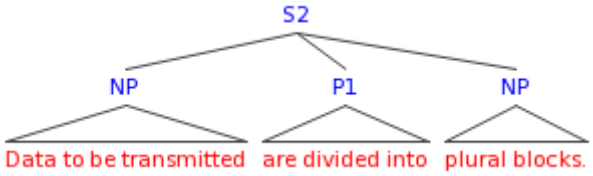

Figure 4. Syntax tree of sentence 6 after reordering

4.3.2 Reordering of Noun Phrases

In patent texts, noun phrases are often long and complicated, as well as sentences. We have made 31 rules to deal with the reordering of NPs with L0 words. Taking Sentence 7 as an example, we will present the reordering of NPs.

*Sentence 7 The fastener has YU mounting hole formed on the blade fit DE projection. (紧固件具有**与**锯条上安装孔相配合的凸台。)*

YU mounting hole formed on the blade fit DE projection is a NP with L0. In this NP, *YU* is identified as L0 in LEVEL[2], and *fit* is P2 in combination with L0. By matching Rule 3, we can transform the syntax tree in Figure 5 to a new syntax tree in Figure 6.

Rule 3:

(-3){CHK[L0]&CHN[与]}+(-2)CHK[NP]+(-1)CHK[P2]+(0)CHN[的] +(1)CHK[NP] => DEL_NODE(-3)+(1)+ADD_NODE(ENG=[which])+(-1){VOI=P}+ADD_NODE(ENG=[with])+(-2)+ DEL_NODE(0)

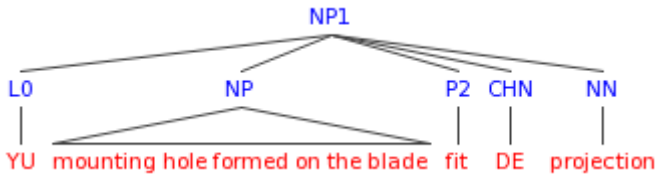

Figure 5. Syntax tree of the NP before reordering

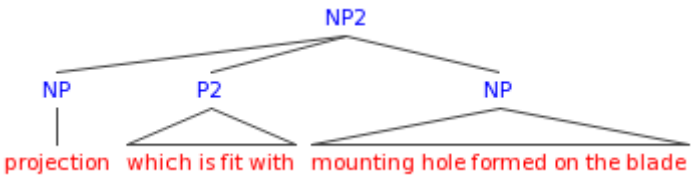

Figure 6. Syntax tree of the NP after reordering

Obviously, the word order in Figure 6 is much closer to English sentence than in Figure 5.

5 Processing Steps

The reordering is processed in steps as follows.

Step 1: To preprocess the Chinese sentence, including word segmentation and word-sense-disambiguation.

Step 2: To identify predicates, L0 words and their levels.

Step 3: To segment the sentence into chunks with L0 words(Level[1]) and predicates(P1) as boundaries.

Step 4: To reorder the sentences with L0 words(LEVEL[1]) based on transformation rules.

Step 5: To reorder the NPs with L0 words(LEVEL[2]) based on transformation rules.

Step 6: To generate a new syntax tree closer to English language order.

6 Experiment and Discussion

The experiment takes 500 authentic patent texts provided by SIPO (State Intellectual Property Office of China) as the training set. The evaluation will use the development data for the NTCIR-9 Patent Machine Translation Pilot Task[2], containing 2,000 bilingual Chinese-English sentence pairs.

After integrating into a rule-based patent machine translation system(Zhu et al, 2012), we will take a closed test on the training set, and an open test on the evaluation set. To evaluate the effects of the reordering rules, precision and recall are calculated by manual evaluation for both two tests. In the open test, NIST (Doddington et al, 2002) and BLEU score (Papineni et al, 2002) are also employed to evaluate the translation performance. Table 3 shows the result of the closed test.

Types	Precision (%)	Recall (%)	F-score (%)
Sentences with L0	97.14	88.20	92.45
NPs with L0	91.80	73.91	81.89

Table 3. Experiment Result on the Training Set

It can be seen from table 3 that the reordering rules have higher accuracy and reliability than coverage, and the module of sentence reordering performs better than NP reordering.

In the open test, comparison is made as shown in table 4. RB-MT is the baseline system. RB-MT+PRM is the system integrated with our reordering model. GOOGLE is an online statistical MT system, the reordering result of which is inferred from its translation result. Table 4 shows the comparison in reordering of the three systems.

Systems	Precision (%)	Recall (%)	F-score (%)
RB-MT	71.23	62.02	66.31
RB-MT+PRM	88.11	75.90	81.55
GOOGLE	60.71	51.20	55.56

Table 4. Compared Result of the Open Test

The result of the open test shows that our model has effectively improved the reordering result of Chinese special sentences, and Google performs poorly in this test. It is mainly because statistical methods face difficulties in long distance reordering, and technical texts (including patent texts) account for a fairly low proportion in the training bilingual corpus. Thus, our method is advantageous in processing technical texts with long and complicated sentences.

After calculating the precision and recall, we give NIST and BLEU scores of the three systems. In order to learn the impact of the pre-reordering model in statistical machine translation, we also put the

[2] http://research.nii.ac.jp/ntcir/ntcir-9/

pre-reordered Chinese sentences into GOOGLE Translate and get the English translation result as a comparison. The reordered sentences are obtained from intermediate outputs of RB-MT+PRM system.

Systems	NIST	BLEU(%)
RB-MT	4.85	19.97
RB-MT+PRM	5.36	22.33
GOOGLE	7.84	35.24
GOOGLE+PRM	7.90	36.07

Table 5. NIST and BLEU-4 Scores

From table 5, we can see that after integrating the reordering model, NIST score of RB-MT system has increased by 10.52%, and BLEU score has increased by 11.82%. Google also has an improvement when input texts are replaced by reordered sentences. Since statistical machine translation has already worked efficiently in short-distance reordering, its improvement is slighter than rule-based systems.

Besides, Google Translate performs better in this evaluation. It is mainly because the corpus domain is not limited, unknown terms or entities may result in a bad translation performance for rule-based systems. In addition, the module of word selection in RB-MT needs to be improved urgently. From the experiment, we also find that the pre-reordering model is strongly dependent on the completeness of rules and the accuracy of the knowledge base, which still need to be improved in the future work.

7 Conclusion

To deal with the reordering of Chinese special sentences, we use a source-language parser to distinguish the levels of L0 words and make transformations in the syntax tree.

Our model improves the performance of patent machine translation. In the future, the rule set and knowledge base need to be improved, and our reordering method can be extended to machine translation of technical texts in other fields.

ACKNOWLEDGMENT

The authors are grateful to National High Technology Research and Development Program of China (No. 2012AA011104) for financial support.

Reference

Collins M., Koehn P., and Ivona K. 2005. *Clause restructuring for statistical machine translation*. In Proceedings of ACL: 531–540.

Crystal D. 1997. *The Cambridge Encyclopedia of Language*, 2nd ed. Cambridge University Press, Cambridge, UK.

Ding Y. and Palmer M. 2005. *Machine translation using probabilistic synchronous dependency insertion grammars*. In Proceedings of ACL: 541–548.

Doddington G. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*, In Proceedings of Human Language Technology Research: 138-145.

Genzel D. 2010. *Automatically learning source-side reordering rules for large scale machine translation*. In Proceedings of COLING: 376-384.

Hu R.F., Zhu Y., and Jin Y.H. 2013. *Semantic Analysis of Chinese Prepositional Phrases for Patent Machine Translation*. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer Berlin Heidelberg.

Huang Z.Y. 1998. *HNC (Hierarchical network of concepts) Theory (in Chinese)*. Tsinghua University Press, Beijing, China.

Huang L. Knight K. and Joshi A. 2006. *Statistical syntax-directed translation with extended domain of locality*. In Proceedings of AMTA:223-226.

Jin Y.H. 2010. *A hybrid-strategy method combining semantic analysis with rule-based MT for patent machine translation*. In Proceedings of NLP-KE: 1-4.

Jin Y.H. and Liu Z. Y. 2011. *Improving Chinese-English patent machine translation using sentence segmentation*, In Proceedings of NLP-KE: 620-625.

Katz-Brown J. and Collins M. 2008. *Syntactic reordering in preprocessing for Japanese→English translation: MIT system description for NTCIR-7 patent translation task*. In Proceedings of NTCIR-7 Workshop Meeting: 409-414.

Lin D. 2004. *A path-based transfer model for machine translation*. In Proceedings of COLING, Geneva, Switzerland: 625–630.

Liu Y., Liu Q. and Lin S. 2006. *Tree-to-string alignment template for statistical machine translation*. In Proceedings of ACL: 609–616.

Liu Z.Y. 2011. *The research of passive voice in Chinese-English patent machine translation*. In Proceedings of NLP-KE: 300-303.

Miao C.J. 2005. *HNC (Hierarchical network of concepts) theory introduction (in Chinese)*. Tsinghua University Press, Beijing, China.

Papineni K., Roukos S., Ward T, et al. 2002. *BLEU: a method for automatic evaluation of machine translation*. In Proceedings of ACL: 311-318.

Quirk C. Menezes A. and Cherry C. *Dependency tree let translation: Syntactically informed phrasal SMT*. In Proceedings of ACL: 271–279.

Wang C. , Collins M. and Koehn P. 2007. *Chinese Syntactic Reordering for Statistical Machine Translation*. In Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: 737–745.

Xia F. and McCord M. 2004. *Improving a statistical MT system with automatically learned rewrite patterns*. In Proceedings of ACL: 508.

Zhang Y., Zens R. and Ney H. 2007. *Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation*. In Proceedings of the NAACL-HLT: 1-8.

Zhang Y.H. 2001. *Format transformation in English-Chinese translation (in Chinese)*. HNC theory and language research. Wuhan University of Technology Press, Wuhan, China.

Zhu Y. and Jin Y.H. 2012. *A Chinese-English patent machine translation system based on the theory of hierarchical network of concepts*. The Journal of China Universities of Posts and Telecommunications, 19(2): 140-146.