A Hybrid System for Chinese-English Patent Machine Translation

Hongzheng Li lihongzheng@mail.bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China
Kai Zhao 765136570@qq.com
Institute of Chinese Information Processing, Beijing Normal University, 19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China
Renfen Hu irishere@mail.bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China
Yun Zhu zhuyun@bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China
Yaohong Jin jinyaohong@bnu.edu.cn
Institute of Chinese Information Processing, Beijing Normal University, 19, Xinjiekou Wai St., Haidian District, Beijing, 100875, China

Abstract

This paper presents a novel hybrid system, which combines rule-based machine translation (RBMT) with phrase-based statistical machine translation (SMT), to translate Chinese patent texts into English. The hybrid architecture is basically guided by the RBMT engine which processes source language parsing and transformation, generating proper syntactic trees for the target language. In the generation stage, the SMT subsystem then provides lexical translation according to the defined structures and generates final translation. According to our empirical evaluation, the hybrid approach outperforms each individual system across a varied set of automatic translation evaluation metrics, verifying the effectiveness of the proposed method.

1. Introduction

As one of important applications in Natural Language Processing (NLP), Machine translation (MT) has developed several paradigms in past decades, basically including rule-based MT (RBMT) and statistic-based MT (SMT). Both the two approaches have strengths and weaknesses.

RBMT systems tend to produce better translations and deal with long distance dependencies, agreement and constituent reordering in a more principled way (Gorka et al., 2014), since they perform the analysis, transfer and generation steps based on syntactic principles. However, they usually have problems in word translation selection preferences, which usually have negative impacts on the translation quality. Also, in cases in which the input sentence has an unexpected syntactic structure, the parser may fail and the quality of the translation will decrease dramatically.

Contrary to RBMT, SMT models are more robust and usually better in fluent lexical selection since they exploit explicit probabilistic language models trained on very large corpora (Xuan et al., 2012). On the downside, SMT has difficulties in dealing with requirements of linguistic knowledge, such as syntactic functions and long distance word reordering, especially in the translation between distant language pairs such as Japanese and English (Isozaki et al., 2010), which may generate translations with improper even worse structures. While SMT has been recognized as the main stream approach of translation, RBMT has tended to be more effective for limited subject domains than SMT (List, 2012). As a result, hybrid MT (HMT) models have become increasingly popular in recent years, aiming to improve final translation effects and qualities. An typical example that can reflect the the increasing interest in hybrid approaches to MT is the workshop on hybrid approaches to translation (Hytra), which was first held at the EACL2012 conference, since then, it was continuously held in 2013 and 2014, in this year, it took part in conjunction with the ACL2015 conference held in Beijing.

It is well known that MT can be applied to various domains. With continued growth in the number of patent applications and the need of exchanging related information, patent domain MT has become one new application of MT, and attracted worldwide attentions of researchers and governments. In this article, we present a novel hybrid translation combination architecture thay takes advantage of RBMT and phrase-based SMT to translate Chinese patent texts into English. As juridical and official documents, Chinese patent documents are usually featured by formal fixed expressions, and much longer sentences with more complex syntactic structures, compared with SMT, rule-based method is more suitable to describe the structures more precisely. Thus, our HMT system is constructed based on the RBMT system (Zhu and Jin, 2012). The main idea is that the RBMT guides main steps in performing source language parsing and transfer, generating proper transferred and reordered syntactic trees for the output, and the SMT system “Moses” (Koehn et al. 2007) then helps the lexical selection by providing more alternative translations according to the trees for target language generation. The final decoding also accounts for fluency by using language models. Since the structures of the translation are already decided by the RBMT subsystem, decoding of SMT will be more fast and efficient in turn.

We performed some experiments on the HMT system with several automatic evaluation metrics to test its performance. After comparing the HMT with individual RBMT and SMT systems, as well as Google online translation, the HMT outperformed all individual MT systems and gained much improvements in evaluation metrics, indicating the hybrid approach is indeed beneficial and effective for translation qualities.

The rest of the article is organized as follows. Section 2 overviews the related literatures on MT system combination and hybridization. Section 3 presents the individual systems and the architecture of system combination in detail. Section 4 describes the experimental work carried out with the hybrid architecture and discusses the obtained results. Finally, Secetion 6 concludes the work.

2. Related Work

This section mainly includes two parts: the first part overviews syntactic reordering in MT and the second part will discuss some previous work on MT system combination.

2.1. Syntactic Reordering

In SMT, reordering positons of chunks in source languages to generate proper and acceptable translation has been a hot issue, and syntactic reordering is effcetive in improving the performance of MT. Xia and McCord (2004) proposed an approach for French-English translation by automatically extracting rewrite patterns after parsing the source and target sides of the training corpus. Collins et al., (2005) described a method for reordering German clauses in German-English translation. Some lexicalized reordering models(Tillman, 2004; Galley and Manning, 2008; Cherry et al., 2012) were employed to predict reordering by taking advantage of lexical information. Different with lexicalized models, a hierarchical phrase-based translation model (Chiang, 2007; Nguyen and Vogel, 2013) based on synchronous grammar was also used in reordering the chunks.

For Chinese-English MT, Wang et al., (2007) described a set of syntactic reordering rules that exploited systematic differences between Chinese and English word order and introduced a reordering approach. Zhang et al., (2007) described a sourceside reordering method based on syntactic chunks for phrase-based statistical machine translation. The source language sentences were first shallow parsed. Then, reordering rules were automatically learned from source-side chunks and word alignments. During translation, the rules were used to generate a reordering lattice for each sentence. Cao et al., (2014) proposed a novel lexicalized reordering model which is built directly on synchronous rules. For each target phrase contained in a rule, they calculated its orientation probability conditioned on the rule. Based on a set of dependency-based preordering rules, Cai et al., (2014) presented a dependency-based preordering approach for C-E MT, improved the BLEU score by 1.61 on the NIST 2006 evaluation data.

2.2. MT System Combination

System combination has been shown to improve classification performance in various tasks in the field of NLP (Rosti et al., 2007). Frederking and Nirenburg (1994) first applied system combination to MT. They integrated outputs of three different translation system (knowledge-based MT, example-based MT and a lexical transfer system) with Chart Walk Algorithm, then performed post-editing processing on the integrated outputs to generate final translation results. Bangalor et al. (2001) introduced recognizer output voting error(ROVER) (Fiscus, 1997) into MT, using a multiple string alignment (MSA) approach to align the hypotheses together, their experiments proved that integrated output was better than single system translation. Since then, system combination has aroused more attention around the world.

Confusion networks is one of the common methods used in combination strategies, which try to combine fragments from a number of different systems and use consensus network decoding to search for the best output from a list of n-best translations (Bangalore et al., 2001; Matusov et al., 2006; Chen et al., 2008; Ayan et al., 2008).

In most hybrid systems, the statistical components are usually selected as basic skeletons and in charge of the translation, correspondingly, the companion system provides complementary information. On the other hand, hybrid architectures where the RBMT system leads the translation and the SMT system provides complementary information to adjust the output from the RBMT, has been less explored. Such systems are applied to relative small domains (Simard et al., 2007), the output tends to be grammatical, and the main effect of the combination is an increase in lexical selection quality (Dugast et al., 2007). Following are some typical works led by RBMT in patent domain.

Jin (2010) proposed a hybrid approach which combined semantic analysis with rule-based method to translate Chinese patent to English. Alexandru et al., (2011) conducted some experiments on English–French patent domain adaptation of the MT systems used in the PLuTO project, both manual and automatic evaluations showed a slight preference for the hybrid system over the two individual baseline engines. Enache et al., (2012) also presented a system for English-French patent translation on the basis of large scale corpora with statistic method. Sheremetyeva (2013) discussed a Russian-English patent MT system which integrated hybrid and rule-based components for several complementary levels of output. There also exists some hybrid systems participating the patent evaluation workshop of the NTCIR conference held in Japan (Isao et al., 2013).

Unlike many previous works using SMT system as the basic skeleton and adapting confusion networks, our hybrid system, oriented for patent domain, is constructed based on the RBMT system and does not involve a confusion network. In the RBMT, we build a considerable knowledge base and manually write rules to help the system analyse and reorder the source sentences according to the grammatical expressions in target language, and the SMT is responsible for the target words selection. As a result, the hybrid system can guarantee both proper syntactic structures and lexical selection qualities that are consistent with target language. The hybrid system will be clarified in detail in following sections.

3. System Architecture

Before presenting the system combination architecture, we need to first introduce the individual RMBT and SMT.

3.1. RBMT

The RBMT engine is based on the traditional translation model which is mainly divided in three steps: (i) analysis of the source language into syntactic-semantic tree structures, (ii) transfer and transformation from source language to target language, and (iii) generation of the target language. It is well known that rule-based approach is featured by the knowledge base and rules which can describe linguistic information. In the system, we have built a considerable knowledge base with more than 50,000 words which cover most patent texts. In the knowledge base, the words are annoated with various syntactic and semantic information. We also manually wrote numerous formal targeted rules to help the engine process the sentences in each step. These rules provide a hierarchical parsing and reordering access to deal with various structures and chunks in source sentences. By using the information both in the knowledge base and the rules, the MT system will finish the processing of three steps. We will describe each of the stages in the following.

Analysis of Source Language

As mentioned, sentences in patent texts are usually much longer. A sentence (S) ended with a full stop may include several sub-sentences (marked as SS) and chunks separated by punctuations (marked as SST, most are commas, colons and semicolons also included). That is, *S = SS1, SS2……SSn*. Considering the expression features of patent documents, a parser is specially developed for the patent texts and integrated into the translation engine, aiming to regards a whole long sentence as the basic processing unit.

The analysis is conducted in three syntactic levels: first, the sentence is separated into several SS according to the punctuations; then, parsing each SS into chunks served as direct componts of SS, including identifying the subject, predicate verb, object and adverbial etc.; last, further anlyse the chunks into terminals (leaf nodes on syntactic trees). Thus, S is the root of sentence, and it has several SS nodes and separators SST, then each SS node is composed of several chunks such as NP, VP, and adverbial phrases (ADVP) etc, further, chunks are composed of terminal nodes.

In the first level, the main purpose is to divide the complex long sentence into several SS mainly by commas. But not all commas can seperate the sentence, because some of them may follow by phrases. In our research, we determine that, for commas following NP and ADVPs, they cannot separate the sentence, and they will be marked as DBT.

In the subsentence level, the system parses the subsentences to get the syntactic components. Parsing rules basically include rules for indentifying predicate verb, ADVP, special NP with long modifiers, and prepositional phrases (PPs) introduced by unique prepositions (such as “把BA” “被BEI” “将JIANG”,etc.)in Chinese. We want to mainly discuss about identification of predicate verbs, which play more important role in parsing.

As Chinese lacks necessary lexical changes, when severel verbs appear in the same sentence, it is usually more difficult to identify the proper core verb. During the beginning stage of the identification process, we first write some rules according to the context information to exclude some verbs that cannot be selected as core verb, next, we then design various high and low weights for remaing possible verbs, the weights represent the possibilities that verbs serve as predicate. When matching kinds of rules, the verbs will be added with corresponding weights, as a result, after comparing the weights, the verbs with the highest weight will be selected as final predicate verb.

The final stage is chunk-level parsing. In this level, the system will continue to analyse the non-terminal chunks into leaf nodes. Since some chunks may contain complex and nested structures, the system needs to exploit the rules in a circular manner and perform the syntactic analysis hierarchically until each node is parsed.

After finishing parsing, the system will generate a syntactic tree for each source sentence, and each node on the tree possesses several marks and symbols representing various syntactic and semantic information.

Here is an example sentence in patent texts, including two subsentences, which followed by the syntactic tree.

E.g.1: Source sentence: 在上述结构中，单电池由突起部支撑，因此可以提高耐振动性。

Target sentence: In above structure, the single cell is supported by the protrusions, therefore the vibration resistance can be improved.)

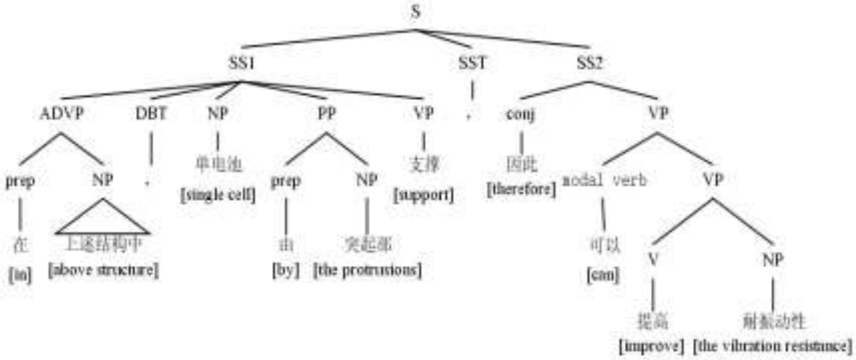
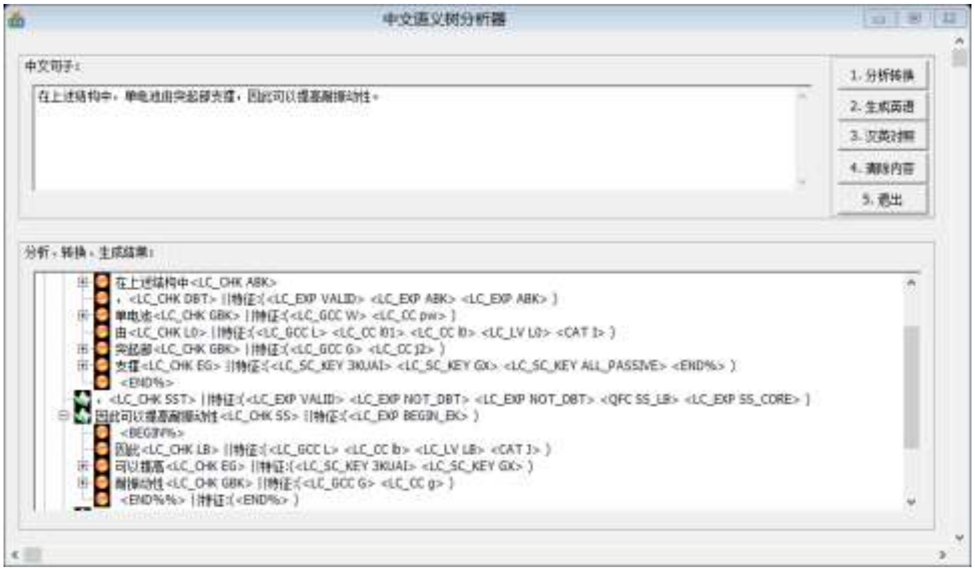


Figure 1. Syntactic Tree of the Example Sentence

Fig.2 shows the form of the tree of the example sentence in our MT system. It can be seen from the vertical tree, there exists some orange circles under each subsentences, and each of them represents the direct component chunk of the subsentence, symbols in the angle brackets “< >” indicate syntactic information of the chunks and punctuations, and when click the “+”buttern in front of some circles, the chunks will extend and show the detailed information of each terminal under the chunk.



中文句子：

在上述结构中，单电池由突起部支撑，因此可以提高耐振动性。

分析、转换、生成结果：

```
在上述结构中<LC_CHK ABK>
，<LC_CHK DBT> ||特征:{<LC_EXP VALID> <LC_EXP ABK> <LC_EXP ABK> }
单电池<LC_CHK GBK> ||特征:{<LC_GCC W> <LC_CC pw> }
由<LC_CHK L0> ||特征:{<LC_GCC L> <LC_CC l01> <LC_CC l0> <LC_LV L0> <CAT 1> }
突起部<LC_CHK GBK> ||特征:{<LC_GCC G> <LC_CC j2> }
支撑<LC_CHK EG> ||特征:{<LC_SC_KEY 3KUAI> <LC_SC_KEY GX> <LC_SC_KEY ALL_PASSIVE> <END%> }
<END%>
，<LC_CHK SST> ||特征:{<LC_EXP VALID> <LC_EXP NOT_DBT> <LC_EXP NOT_DBT> <QFC SS_LB> <LC_EXP SS_CORE> }
因此可以提高耐振动性<LC_CHK SS> ||特征:{<LC_EXP BEGIN_BK> }
<BEGIN%>
因此<LC_CHK LB> ||特征:{<LC_GCC L> <LC_CC lb> <LC_LV LB> <CAT 1> }
可以提高<LC_CHK EG> ||特征:{<LC_SC_KEY 3KUAI> <LC_SC_KEY GX> }
耐振动性<LC_CHK GBK> ||特征:{<LC_GCC G> <LC_CC g> }
<END%%> ||特征:{<END%> }
```

Figure 2. Tree Structure of the Example Sentence in Our MT System

Transfer from Source Language to Target Language

When transforming Chinese sentences into English, it is necessary to reorder and transfer chunks and words to guarantee fluent and proper expressions. Corresponding to the analysis phases above, the transformation process also includes three levels in a top-down order: ① transformation of relationships between subsentences, ②structural reordering of chunks in subsentences and ③reordering inside the chunks. Basic operations in the reordering include add or delete nodes, chunks position adjustment etc. In the following, we will introduce the three stages with some rules and examples.

Transformation between subsentences mainly refers to transfering the source sentences into expressions commonly used in target language according to the semantic relationships between subsentences.

Rule1: *{SS1&CHN[在于, 包括]&END%}+ CHN[，]+ SS2→SS1+ that + SS2*

The rule means that, if the Chinese characters(CHN) such as "在于(lie in), 包括(include)" appear in the end of the first SS1, and followed by the comma and another SS2, then the comma will be replaced by the word "that" when transferred into English.

E.g.2: Source sentence:本发明的特征在于，它可以调节输出装置的参数。

Target sentence: The feature of this invention lies in *that* it can adjust the parameter of the output device.

In the example, the sourece sentence includes two subsentences, in which the second one is actually the object of the first one. Considering that, when transformed into English, it is better to transfer the two subsentences into a single sentence by using object clause and replacing the comma with the connecting word "that".

Generally, Chunk-level reordering and transformation inside the chunks play more important roles in generating grammatical target language. Which mainly includes following types:

Changing form, tense and voice etc. of core verbs. As for form, for example, if some modal verbs appear before the core verbs, the verbs should be transformed in the form of prototype. As for tense, simple present is considered as default tense in most cases, but if some words such as “已, 已经(have already)” appear before the core verbs, then they need to be changed into the perfect tense. As for voice, the default voice is active voice, but verbs should be changed into passive voice if they are followed by words representing passive voice such as the typical preposition “被(BEI)”. On the other hand, for some sentences without subjects, the predicates may also be changed into passive voice, and the objects will serve as subject in English (*VP+NP → NP+VP (passive voice)*).

Transforming the ADVPs introduced by prepositions. Such transformation includes two aspects: (1) reordering positions of adverbials. For those located between subject and predicate verb in Chinese, they need to be reordered to the end of the sentence in English. If some parallel adverbials appeare in the same sentence, it is better to reorder them in reverse order. (2) Transformation of long-distance fixed structures. In some adverbial chunks, “当……时(when……)”and “在……中(in……)”, for example, the left and right boundary words are usually collacations and appear together, which can be directly replaced with corresponding words in English. We have wrote rules to cover the fixed collacations as much as possible.

Rule2: *NP+ADVP+VP→NP+VP+ADVP*

Rule3: *NP+ADVP1+ ADVP2+VP→NP+VP+ADVP2+ADVP1*

Rule4: *(0)CHN[当]+(f){(1)CHN(时)} → DELETE (0)+DELETE(1)+ADD[when]*

E.g.3: Source sentence: [NP本发明][ADVP1在实验中][ADVP2通过一种有效的方法][VP提高产品的性能]。

Target sentence: [NP This invention] [VP improves the performance of the products] [ADVP2 by an effective method] [ADVP1 in the experiment].

Reordering special prepositional phrases (PPs) in Chinese. Some prepositions, such as “把(BA)” “将(JIANG)” and “被(BEI)” etc., are unique in Chinese and lack corresponding translation in English. PPs composed of such prepositions and NPs always appear in front of VP, when transfer them into English, these prepositions should be deleted, and NPs behind the prepositions must be reordered to proper positions, usually after the VP.

Rule5: *NP1 + prep. + NP2 + VP → NP1 +VP + NP2*

E.g.4: Source sentence:这些计数器**对**这些数据输入/输出装置的数量进行计数。

Target sentence: These counters count the number of these data input/output devices.

Before syntactic reordering	After syntactic reordering
NP1 这些计数器(These counters)	NP1 这些计数器(These counters)
PP prep. 对(DUI)	VP 进行计数 (count)
NP2 这些数据输入/输出装置的数量(the number of these data input/output devices)	NP2 这些数据输入/输出装置的数量(the number of these data input/output devices)
VP 进行计数 (count)	

Figure 3. Original and Reordered Trees of Example 4

Rule6: *NP1 + prep. + NP2 + VP +NP3 → NP1 + VP +NP2 +NP3*

E.g.5: Source sentence:第二通信模块**将**第二表示数据发送到计算机系统。

Target sentence: The second communication module sends the second indicating data to the computer system.

Before syntactic reordering	After syntactic reordering
NP1 第二通信模块(The second communication module)	NP1 第二通信模块(The second communication module)
PP prep. 将(JIANG)	VP 发送(sends)
NP2 第二表示数据(the second indicating data)	NP2 第二表示数据(the second indicating data)
VP 发送到(sends to)	Prep. 到(to)
NP3 计算机系统(the computer system)	NP3 计算机系统(the computer system)

Figure 4. Original and Reordered Trees of Example 5

Structural reordering inside chunks, especially in NPs. The most common structure of NPs in Chinese is “modifiers + 的(DE) + head NP”. In which the modifiers can include NP, VP, quantifier phrase (QP), determiner phrase (DP), adjective phrase (ADJP) or even relative clauses. The placement of QP, DP, and ADJP modifiers is somewhat similar to English that these phrases typically occur before the nouns they modify, and they need not reordering.

For *NP1*+DE+*NP2*, althouth it is analogous to the English possessive structure of “NP1’s NP2” and does not require reordering, the Chinese possessive structure “*NP1* DE *NP2*” can express more sophisticated relationships, additionally, the “*NP2* of *NP1*” expression is more general and can replace “*NP1*’s *NP2*” in many cases, except for the case that the NP1 is a pronoun. Thus, the reordering rules will state that and map the following rule.

Rule7: *NP1+DE+NP2 →NP2+DE+NP1*

NPs modified by relative clauses (CP), with long distance structures, are quite different with those in English. For such NPs, we apply the rules to reposition the child CP after its sibling head NP under a parent NP.

Rule8: *CP+DE + NP →NP+that+CP*

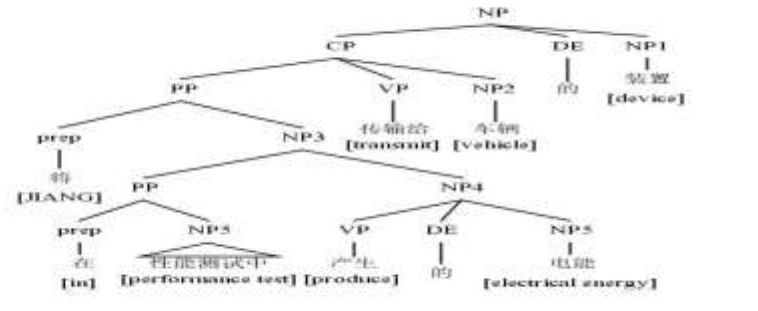
E.g.6: Source sentence: 将在性能测试中产生的电能传输给车辆的装置。

Target sentence: The device *that* transmits the electrical energy produced in the performance test to the vehicle.

From the syntactic trees, it can be seen that the CP modifier is reordered to the position just after its sibling head NP, and the whole NP is transformed into a NP modified by an attributive clause. In the transformation, it is necessary to add an additional word “that” between the antecedent and the clause.

The example is also a nested chunk with multi-level structures, it clearly outlines the various types needed to be transformed, including ADVP, PP, NP and VP mentioned above. Reordering rules sequentially process the elements in a top-down order, and the rules will be exploited circularly.

Before syntactic reordering



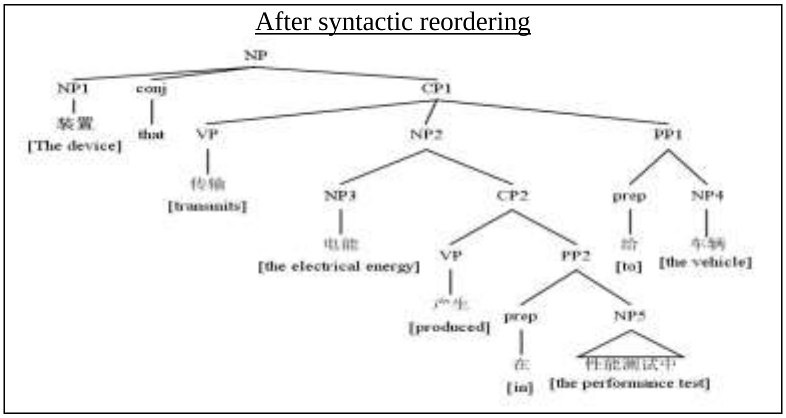

After syntactic reordering

Figure 5. Original and Reordered Parising Tree of Example 6

Generation of Target Language

Generation can be decomposed into two steps. First, word selection. According to the reordered syntactic transformation tree, the MT engine selects target words for each node from the Chinese-English parallel translation dictionary. Second, morphological generation, which consists of generating the target surface forms from their associated morphological information.

3.2. SMT System

The phrase-based SMT baseline system Moses is built on the basis of freely available state-of-the-art tools: the GIZA++ toolkit (Och 2003) to estimate word alignments, the IRST Language Modelling toolkit (IRSTLM) (Federico, et al., 2008) with modified Kneser-Ney smoothing (Chen and Goodman 1999) to guarantee more fluent target language outputs. And in the paper, we use the IRSTLM toolkit to train a 5-gram language model with the patent texts corpus. Last, as decoding is the central stage of SMT, the Moses decoder (Koehn et al. 2007) is employed to find the highest scoring sentence in the target language corresponding to given source sentence.

3.3. Hybrid System Architecture

Many previous works use SMT as basic skeleton of the hybrid system. In our work, considering the pros and cons of RBMT and SMT, as well as the special features of patent texts, we try to build a hybrid patent MT system guided by the RBMT. Just as mentioned before, RBMT usually performs better in dealing with long distance structure and reordering. In the system combination, the RBMT is responsible for parsing source language and generate grammatical syntactic reordering lattice for the target language by applying the knowledge base and the rules, the main task of SMT is to generate translation for each node according to the reordered tree determined by the RBMT. While the RBMT guarantees basic proper structures of target languange, the SMT provide more lexical selection, as a result, the hybrid system is supposed to generate more fluent and acceptable translation.

In the hybrid system, after word segmentation, the RBMT first analyses Chinese sentences and transform positions of chunks according to the corresponding expressions in English by matching kinds of rules. Next, instead of generating all the translation for the words, the RBMT just genernates partial translation for special words (most are functional words) in the sentences. On the other hand, the RBMT also adds some connecting words such as “that”to make the final translation more fluent. Let’s take a sentence for example.

E.g.7: Source: 权利要求1的滑门机构，其特征在于，所述轴部配置于所述卡的插入方向上的所述滑门构件的里端。

Transformed: 滑门/机构/ *of* 权利要求/1/ ，其/特征/在于/ *that* 所述/轴部/配置于/里端/*of* /所述/滑门/构件/*on*/插入/方向/ *of* /所述/卡/。

The segmented sentence sequences with partial translation are output results of the RBMT, which will be sent to the SMT as input files.

In the SMT stage, the standard statistical decoder Moses is used for monotone decoding to limit reordering, which can not only speed up the decoder, but also increase the translation performance. The system will search proper translation for remaining words in the source sentences, generating final target hypothesis of complete sentences.

Sum up, the RBMT is responsible for analyzing and transforming the source sentences, and give translation for partial words in advance, the SMT is mainly responsible for generating translation for most words and chunks. Following is the architecture of the hybrid system.

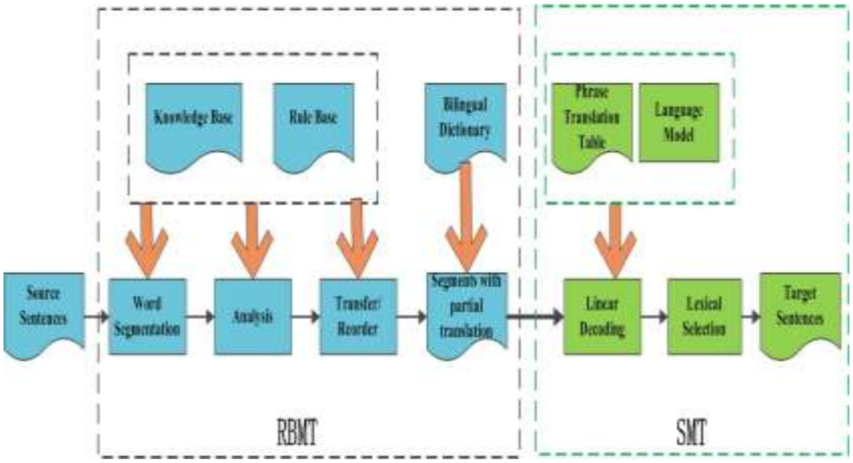


Figure 6. Architecture of the Hybrid System

4. Experiments

In this part, we conducted some evaluations on the hybrid system to test its performances, which were measured by several popular automatic evaluation metrics : WER (Nießen et al., 2000), PER (Tillmann et al., 1997), TER (Snover et al., 2006), BLEU (Papineni et al., 2002), NIST (Doddington, 2002), GTM (Melamed et al., 2003), METEOR(Banerjee and Lavie., 2005) and ROUGH (Lin and Och, 2004). All measures were calculated with the Asiya toolkit [1] for MT evaluation (Giménez and Màrquez, 2010). We also compared the scores of the hybrid system with those of RMBT, SMT and Google online translation.

4.1. Experimental Setting

We exploited the training set of NTCIR-9[2] (Sakai and Joho, 2011), including 1 million bilingual Chinese-English patent sentence pairs, to train the Moses decoder. In the develop set of NTCIR-

[1] http://nlp.lsi.upc.edu/asiya/

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

9 which included 2000 sentence pairs, 1000 sentence pairs were randomly extracted as development set, and remaining 1000 pairs as test set.

Following is the statistic data of the experiments.

	sentence pairs	size	
Traning Set	1 million	188MB(Chinese)	227MB(English)
Development Set	1000	188KB(Chinese)	212KB(English)
Test Set	1000	191KB(Chinese)	214KB(English)

Table 1. statistic data of the experiments

4.2. Experimental results

Table 2 gave the comparative evaluation scores of the four MT systems.

	WER	PER	TER	ROUGE	BLEU-4	NIST-5	METEOR	GTM
RBMT	93.08	71.45	87.55	40.19	11.44	4.44	19.03	13.90
SMT	61.33	32.62	52.73	53.44	29.30	7.57	34.47	22.58
HMT	**59.41**	**30.63**	**51.68**	**56.46**	**31.22**	**7.84**	**35.16**	**24.65**
Google	68.12	45.99	61.05	59.66	37.36	8.80	27.96	22.99

Table 2. Comparison of Several Systems

4.3. Analysis

As clearly shown in table2, the hybrid system outweighted other two individual systems in all the evaluation metrics. The experimental data have proved that the proposed method performed well in improving the translation results significantly.

After analysing the results of each system, we can come to some conclusions. Take the BLEU-4 scores for example, compared with SMT and HMT, the score of RBMT is quite lower. Several reasons are supposed to account for the result: (1) some errors occured in the word segmentation process and ambiguous structures resulted in improper or even wrong syntactic parsing, futher affected the transformation and generation of target language. (2) some bugs of the system also made the final translation worse. (3) last, each node in Chinese usually has only one corresponding English translation in the bilingual dictionary, and the lexical selection is quite limited, so the candidate traget words are more likely different with the reference translation.

The large difference in BLEU scores between RBMT(11.44) and SMT(29.30) has reflected a well known phenomenon of automatic lexical-matching evaluation metrics overestimating the quality of statistical systems on in-domain test sets. (Giménez and Màrquez, 2007). Despite the difference, the hybrid system is still able to take advantage of the combination and consistently improve results over the individual SMT system.

Note that, although the score is low, the syntactic structures of many target sentences generated by the RBMT are good and grammatical, especially some complex structures with long distance. which indicates the rule-based approach is more easily to describe and consider the linguistic information. Following is a comparison of a sentence transformed by the RBMT and Google.

Source sentence	在活塞缸52与基底构件48b之间设置第二弹簧58，而另外在第二弹簧58与基底构件58b之间设置一个或一组垫片60。

Reference	A second spring 58 is positioned between the piston cylinder 52 and the base member 48b , and a shim 60 or series of shims 60 is further positioned between the second spring 58 and the base member 58b.
RBMT	Second spring 58 *is arranged* between the piston jar 52 and the basement member 48b, the other one or one group gasket 60 *is arranged* between second spring 58 and the basement member 58b.
Google	*Disposed* between the base member 52 and the piston cylinder 48b of the second spring 58, while the other ***set*** of one or a group of the spacer 60, between the second spring 58 and the base member 58b.

Table 3. Comparison of an Example Translated by the RBMT and Google

Since both the two subsentences lack subjects, thus it is better to tranfer the predicate verbs into passive vocie, which will be more fluent and be consistent with English expression. The RBMT has transfered successfully and generate fluent translation. On the contrary, as for Google, the translation is quite bad and unacceptable. The orders of chunks are actually ungrammatical, and, what's worse, the core verb "设置(dispose)" in the second subsentence is even translated into a noun(set), resulting in the absence of predicate verb.

Compared with Google, the BLEU score of the HMT is lower less than 5 percent. As a typical representative of SMT approaches, Google has always possessed powerful and state-of-the-art algorithms on NLP and language technology, we guess it's natural for Google to gain a better score. However, scores of error rates(WER, PER and TER) and METEOR metrics were all better than Google. The main reason is that, the syntactic reordering structures of target language generated by the RBMT benefits the HMT to a large extent, on the contray, Google tends to have much difficulties in reordering complex chunks, especially for those with long distances and dependency.

5. Conclusion

This paper presents a hybrid MT system, which combines an individual RBMT and phrase-based SMT system, for Chinese-English patent translation. Since RBMT is generally able to produce grammatically better translations, different with many previous system combinations mainly guided by SMT, our HMT is built based on the RBMT, its analysis and transfer modules are exploited to generate the backbone of the translation and provide syntactic reordering structures of target language. In the generation stage, statistical decoding of SMT-based translation generate translation for the source sentences, which can improve lexical selection and fluency of the final translation.

We conducted experiments to evaluate the system on patent texts with several evaluation metrics, the hybrid system outperformed both the RBMT and SMT, indicating the proposed approach is a good choice and efficient in improving final translation performance. The experimental results also confirmed that syntactic reordering provided by the RBMT is essential.

The work still sleaves some issues that derserve further research. In the future, we need to expand the size of patent corpus to design more proper processing rules and improve the RBMT system to produce better transferred and reordering structures. On the other hand, we would like to train the SMT models with larger data set and optimize the related parameters.

Acknowledgement

This work was supported by the National Hi-Tech Research and Development Program of China (2012AA011104), and we also want to express our sincere thanks to the anonymous reviewers of this article.

References

Necip Fazil Ayan, Jing Zheng and Wen Wang. (2008). Improving Alignments for Better Confusion Networks for Combining Machine Translation Systems. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 33-40, Manchester, UK.

Satanjeev Banerjee and Alon Lavie. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Cor-Relation with Human Judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, Ann Arbor.

Srinivas Bangalore, German Bordel and Giuseppe Riccardi. (2001). Computing Consensus Translation from Multiple Machine Translation Systems. In *Proceedings of Automatic Speech Recognition and Understanding*, pages 351–354, Waikoloa, Hawaii, USA.

Jingsheng Cai, Masao Utiyama, Eiichiro Sumita and Yujie Zhang. (2014). Dependency-based Pre-ordering for Chinese-English Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 155–160, Baltimore, USA.

Hailong Cao, Dongdong Zhang, Mu Li, Ming Zhouand Tiejun Zhao. (2014). A Lexicalized Reordering Model for Hierarchical Phrase-based Translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1144–1153, Dublin, Ireland.

Alexandru Ceauşu, John Tinsley, Jian Zhang and Andy Way. (2011). Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project. In *Proceeding of the 15th Annual Conference of the European Association for Machine Translation*, pages 30-31, Leuven, Belgium.

Boxing Chen, Min Zhang, Aiti Aw and Haizhou Li. (2008). Regenerating Hypotheses for Statistical Machine Translation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 105–112, Manchester, UK.

Stanley F. Chen and Goodman J. (1999). An Empirical Study of Smoothing Techniques for Language Modeling. Comput Speech Lang, 4(13):359–393.

Michael Collins, Philipp Koehn, and Ivona Kucerova. (2005). Clause Restructuring for Statistical Machine Translation. In *Poceedings of the 43rd annual meeting of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, USA.

Loïc Dugast, Jean Senellart, and Philipp Koehn. (2007). Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220-223, Czech Republic.

R. Enache, España-Bonet C., Ranta A., and Màrquez L. (2012). A Hybrid System for Patent Translation. In *Proceedings of the 16th annual conference of the European Association for Machine Translation* (EAMT12), pages 269–276, Trento, Italy.

Jonathan. G. Fiscus. (1997). A Post-Processing System to Yield Reduce Word Error Rates: Recognizer Output Voting Error (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Reorganization and Understanding*, pages 347-354.

Marcello Federico, Nicola Bertoldi and Mauro Cettolo. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of 9th Annual Conference of the International Speech Communication Association*, pages 1618- 1621, Brisbane, Australia.

Markus Freitag, Matthias Huck and Hermann Ney. (2014). Jane: Open Source Machine Translation System Combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden.

Michel Galley and Christopher D. Manning. (2008). A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu.

Jesús Giménez and Lluís Márquez. (2007). Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the second workshop on statistical machine translation*, pages 256–264, Prague.

Jesús Giménez and Lluís Márquez. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *Prague Bull Math Linguist*, 94:77–86.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita and Benjamin K. Tsou. (2013). Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260-287, Tokyo, Japan.

Hideki Isozaki, Tsutomu Hirao and Kevin Duh. (2010). Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Massachusetts, USA.

Yaohong Jin. (2010). A Hybrid-Strategy Method Combing Semantic Analysis with Rule-Based MT for Patent Machine Translation. In *Proceedings of the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing.

Philip Koehn., Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Federico M., Bertoldi N., Cowan B., ShenW., Mo-ran C., Zens R., Dyer C., Bojar O., Constantin A. and Herbst E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume*, pages 177–180, Czech Republic.

Gorka Labaka, Cristina España-Bonet, Lluís Màrquez and Kepa Sarasola. (2014). A Hybrid Machine Translation Architecture Guided by Syntax. *Machine Translation*, 28:91–125.

Maoxi Li and Chengqing Zong. (2010). A Survey of System Combination for Machine Translation. *Journal of Chinese Information Processing*, 4:74-84.

CY Lin, and Franz Josef Och. (2004). Automatic Evaluation Of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, pages 605–612, Barcelona.

Jane List. (2012). Review of Machine Translation in Patents–Implications for Search. *World Patent Information*, 34:193–195.

Evgeny Matusov, Nicola Ueffing and Hermann Ney. (2006). Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy.

I.Dan Melamed, Ryan Green and Joseph P. Turian. (2003). Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 61–63, Edmonton.

Simard Michel, Ueffing Nicola, Isabelle Pierre and Kuhn Roland. (2007). Rule-based Translation With Statistical Phrase-based Post-editing. In *Proceedings of ACL 2007 Second Workshop on Statistical Machine Translation*, pages 203-206, Prague, Czech Republic.

Thuylinh Nguyen and Stephan Vogel. (2013). Integrating Phrase-based Reordering Features into Chart-based Decoder for Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1587–1596, Sofia, Bulgaria.

Franz Josef Och and Hermann Ney. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of 40th Annualmeeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, USA.

Franz Josef Och. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of 41st Annual meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo.

K. Papineni, S. Roukos, T. Ward and WJ. Zhu (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Antti-Veikko I. Rosti, Spyros Matsoukas and Richard Schwartz. (2007). Improved Word-Level System Combination for Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, Prague, Czech Republic.

Tetsuya Sakai and Hideo Joho. (2011). Overview of NTCIR-9. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 1-7, Tokyo, Japan.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Jone Makhoul. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Visions for the future of machine translation*, pages 223–231, Cambridge, MA.

Svetlana Sheremetyeva. (2013). On Integrating Hybrid and Rule-Based Components for Patent MT with Several Levels of Output. In *Proceedings of the 5th Workshop on Patent Translation*, pages 8-15, Nice.

Nie ßen Sonja, Franz Josef Och, G. Leusch and Hermann Ney. (2000). an Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd international conference on language resources and evaluation*, pages 39–45, Athens.

Christoph Tillmann, S.Vogel, H.Ney, A.Zubiaga, and H.Sawaf (1997). Accelerated DP Based Search for Statistical Translation. In *Proceedings of the fifth European conference on speech communication and technology*, pages 2667–2670, Rhodes.

Christoph Tillmann. (2004). A Unigram Orientation Model for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Chao Wang, Michael Collins and Philipp Koehn. (2007). Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 737–745, Prague.

Fei Xia and Michael McCord. (2004). Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, Switzerland.

Hongwei Xuan, WeiWei Li and Guangyi Tang. (2012). an Advanced Review of Hybrid Machine Translation (HMT). *Procedia Engineering*, 29:3017-3022.

Yuqi Zhang, Richard Zens and Hermann Ney. (2007). Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 1–8, Rochester, New York.

Yun Zhu and Yaohong Jin. (2012). A Chinese-English Patent Machine Translation System Based on the Theory of Hierarchical Network of Concepts. *Journal of China Universities of Posts and Telecommunications*, 19(Suppl. 2): 140–146.