NTCIR-7 Patent Translation Experiments at Hitachi

Hiroyuki Kumai, Hirohiko Sagawa, and Yasutsugu Morimoto
Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
{hiroyuki.kumai.gf, hirohiko.sagawa.cu, Yasutsugu.morimoto.zf}@hitachi.com

Abstract

Statistical Machine Translation (SMT) is a new paradigm in machine translation, which enables high-quality translation. However, many translation errors occur in the translation of complex and compound sentences because of the lack of grammatical knowledge about the global structure of a sentence. We adopt the pre-editing method, which divides sentences into clauses, and translate these clauses using the Moses SMT engine. The translation accuracy, BLEU, was 29.33%, so pre-editing has a small effect. Translation quality is degraded because the order of words is changed by not using information about other clauses. We also performed an experiment to confirm the optimum distortion-limit parameter of Moses. The Maximum BLEU was 29.45 for an English-Japanese patent translation when the distortion limit was 20 instead of -1.

Keywords: *Patent information, Statistical machine translation, Clause dividing, NTCIR*

1. Introduction

We studied rule-based machine translation for many years [1, 2]. However, we could not attain sufficient accuracy for practical business use. However, many people need to communicate with foreigners, especially in business. Statistical Machine Translation (SMT) is a new paradigm in machine translation, which enables high-quality translation. Therefore, we participated in the NTCIR-7 Patent Translation Task [3] to evaluate the possibility of SMT.

We used the Moses SMT engine to perform experiments. One of them is an intrinsic evaluation of the task, and the other is an evaluation of the effects of the distortion-limit parameter, which controls a range of translated-phrase rearrangements. In the intrinsic evaluation, we proposed a clause-separation method to improve accuracy.

2. Intrinsic Evaluation

First, language and translation models were constructed by using tools such as, ChaSen [4], GIZA++ [5], SRILM [6], and Moses [7]. These models were evaluated on the basis of test sentences. The evaluation was only performed on English-to-Japanese translation to understand the behavior of SMT and because of our resource limit. As a result, the Bilingual Evaluation Understudy (BLEU) was only 29.22% [8]. We used the NIST BLEU scoring tool for evaluation. The result is inferior than that of Koehn's evaluation (BLEU 50%) for translation from Japanese, Chinese, or Korean, to English [9]. The reason is that patent sentences are longer and more complex than travel expression sentences. In analyzing the translation error, we find that incorrect ordering of prepositional phrases and subject clauses, for example, occur frequently.

Examples of translation errors are shown in Table 1. In example #1, a conditional clause is exchanged for an introductory clause. In #2, a noun phrase is misplaced in the wrong clause. In #3, an adjective is misplaced in the wrong clause.

For two reasons, the translation model in Moses cannot represent the difference in the global structure of the sentence such as that between Japanese and English. The first reason is that the main targets of Moses are European languages, whose structures are similar to each other. Therefore, the order of chunks is not dealt with appropriately when translating between Japanese and English. The second reason is that punctuation symbols, e.g., period and comma, are not distinguished from letters nor adequately processed to perform the translation. This leads to errors in translating compound and complex sentences, but it brings the advantage that the model is independent of specific languages.

2.1. Proposed method

The proposed method is based on pre-editing, which divides sentences into clauses at the location of commas. The order of clauses is maintained because the translation is executed depending on clauses. However, the comma is not used only as a separator between clauses but also as a separator between listed nouns. In that case, pre-editing causes translation errors.

Detecting clause-separating commas is necessary. Detection of comma-separated clauses is achieved by using the part of speech of words around the comma.

Table 1. Examples of better translations achieved by proposed method. (BLEU)

#1 Conditional clause is exchanged for introductory clause.

Source	If the comparison shows coincidence among all the data, then **the process ends at normal processing step st88** .
Before (0.0)	通常の処理ステップＳＴ８８で，処理が終了すれば，全てのデータのうちの比較一致を示している 。
After (0.233)	全てのデータのうちの比較一致を示し，そして，通常処理ステップＳＴ８８で処理を終了する。
Reference	比較で全データが一致すれば，正常処理ステップＳＴ８８として処理を終了する。

#2 Noun phrase is misplaced in wrong clause.

Source	(5) in a state where the signal input of the lfsrmode terminal is lfsrmode & equals; 1 , **the same operation as that of fig. 6** is performed.
Before (0.356)	（５）lfsrmode＆＝１，図６と同様の動作 lfsrmode 端子は，信号入力の状態で行われる。
After (0.36)	（５）lfsrmode 端子は，信号入力の状態で lfsrmode＆＝１，図６と同様の動作が行われる。
Reference	（５）そして，ＬＦＳＲＭＯＤＥ端子の信号入力がＬＦＳＲＭＯＤＥ＝１の状態で，図６と同様の動作を行う。

#3 Adjective is misplaced in wrong clause.

Source	Accordingly, by arbitrarily setting the voltages of the reference voltage sources 30 and 31 as in the construction of the shown embodiment, the conversion range can be set arbitrarily, and even in a **small** conversion range, resolution can be maintained.
Before (0.438)	従って，本実施例の構成のように，基準電圧源３０及び３１の電圧を任意に設定することができるので，小さな変換範囲を任意に設定し，変換範囲においても，分解能を維持することができる 。
After (0.461)	従って，本実施の形態の構成のように，基準電圧源３０及び３１の電圧を任意に設定することにより，変換範囲を任意に設定することができ，小変換範囲においても，分解能を維持することができる 。
Reference	従って，本実施例で示す構成のように基準電圧源３０，３１の電圧を任意に設定出来るようにすることで，変換レンジを任意に設定することができ，且つ小さい変換レンジに於いても分解能を維持することができるようになる。

Table 2. Examples of tagging

# 1	if/IN the/DT comparison/NN shows/VBZ coincidence/NN among/IN all/PDT the/DT data/NNS ,/, then/RB the/DT process/NN ends/VBZ at/IN normal/JJ processing/NN step/NN st88/CD ./.
#2	however/RB ,/, rechargeable/JJ batteries/NNS 1/CD may/MD also/RB be/VB nickel/JJ hydrogen/NN batteries/NNS ,/, nickel/NN cadmium/NN batteries/NNS ,/, or/CC some/DT other/JJ type/NN of/IN rechargeable/JJ batteries/NNS ./.

Table 3. Examples of poor translations by proposed method. (BLEU)

#1 Order of phrases in each clause is changed and meaning is changed because of different clause separation.

Source	**a chain line** 21 indicates an angle of view of the image sensor 16, and a range surrounded by a chain line 22 indicates an image range at a position of the passenger-side door 9 .
Before (0.427)	一点鎖線で囲まれた範囲の画像範囲を示し，画像センサ１６と，この助手席ドア９の位置での画角２１は，一点鎖線２２である。
After (0.243)	チェーンライン２１の画角は，画像センサ１６，この助手席ドア９の位置に一 点鎖線で囲まれた範囲と画像範囲２２で表している。
Reference	一点鎖線２１はイメージセンサ１６の画角，一点鎖線２２で囲まれた範囲は，イメージセンサ１６の，助手席側ドア９の位置での撮像範囲を示す。

#2 In addition to changing the order of phrases, incorrect inflected forms of verb in separated clause degrade BLEU.

Source	an annular protruding portion 343 is formed on the end face of the cylindrical portion 34, and the seal member 35 is arranged on the radically inner side of the annular protruding portion 343.
Before (0.506)	環状突部３４３は，円筒部３４の端面に形成されており，シール部材３５の内周側には環状突部３４３が配置されている。
After (0.371)	環状突部３４３の端面には，筒状部３４が形成されている。とシール部材３５の内周側には環状突部３４３が配置されている 。
Reference	筒部３４の端面には環状の突条３４３が形成されており，シール部材３５は環状の突条３４３の内側に配置されている。

Table 3. Examples of poor translations by proposed method. (BLEU) (continued)

#3 Order of phrases in clause is incorrect; however, clauses are separated correctly.

Source	**with reference to memory chip 100** electrically connected to node na as an example , the memory chip will be described schematically .
Before (0.433)	ノードＮＡに電気的に接続され，一例として，メモリチップのメモリチップ１００を参照して概略的に説明する。
After (0.292)	メモリチップ１００を参照し，一例としてノードＮＡに電気的に接続されている。メモリチップ概略的に説明する。
Reference	ノードｎａと電気的に接続されるメモリチップ１００を一例として，メモリチップの概要を説明する 。

We adopt the rule that if both a noun and verb are included before or after a comma, the word sequence is a clause.

In our method, pre- and post-editing is executed before and after translation respectively, and Moses is used as is.

"POS tagger"[10] is used for part-of-speech tagging.

Examples of tagging are shown in Table 2.

The sentence is separated into two chunks in #1 because the word sequence after the comma includes both a noun and verb, but it is not separated in #2.

We implemented the proposed method and evaluated the method using test sentences. As a result, the BLEU was 29.33% for test sentences. We also evaluated each sentence by BLEU with scripts provided from the organizer.

Of the 1381 sentences, translation accuracy was improved in 173 cases, worsened in 159 cases, and unchanged in 1049 cases.

This is the official formal run result for the NTCIR-7 Patent Translation Task.

(GROUP-ID=HCRL, RUN=1, BLEU=29.97)

2.2. Discussion about proposed method

The clause-separating method has only a small effect.

Examples of sentences worsening BLEU are shown in Table 3.

In example #1, phrases belonging to a clause reordered due to clause separation. In #2, in addition to phrase reordering inflected verb forms in separated clauses undergo BLEU degradation. In #3, phrase order in a clause is incorrect, however clause separation is correct.

We constructed the translation model by training it against a parallel corpus, which consists of sentence pairs including complex sentences, but not "clause pairs". Our clause-separation method is improved by applying clause division to the corpus to make "clause pairs," and constructing a translation model with "clause pairs."

3. Optimization of distortion-limit parameter

The Moses decoder has a distortion-limit parameter, which restricts phrase position range within the produced sentence when phrase position in the foreign sentence differs from that of a corresponding phrase in the source sentence. The Moses decoder deals with the difference between phrases positions in source and foreign sentences as "distortion." The value of the distortion-limit parameter is the number of phrases between the phrases in the produced sentence that are adjacent in the source sentence. In translating between English and other European languages, the distortion-limit value is often set to 6. However, it is known that a good translation result can be obtained by setting the distortion-limit value to -1 meaning "unlimited" in translations between Japanese and English. In the baseline system, the distortion-limit value was also set to -1.

3.1. Experiment

We investigated how the BLEU value representing translation accuracy changed according to the distortion-limit value in English-to-Japanese translation for patent documents. We subsequently changed the distortion-limit value from 5 to 60.

The result is shown in Figure 1. The BLEU value reaches a maximum of 29.45% when the distortion-limit value is set to 20 and is improved by 0.23% compared with 29.22% of the baseline system.

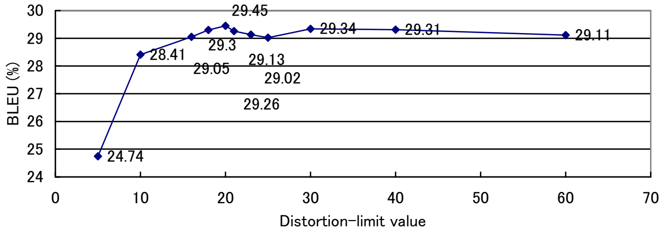

Figure 1. Relationship between distortion-limit value and BLEU

3.2. Discussion about the distortion-limit parameter

The optimum distortion-limit value is different for each foreign sentence. Therefore, we investigated the distortion-limit value that yields the maximum BLEU for each English sentence and the relationship between the distortion-limit value and the number of words in an English sentence. In this examination, the distortion-limit values are 5, 10, 20, 25, 30, 40, 60, and "unlimited," which is located at 100 for convenience.

The result is shown in Figure 2. There is a relationship between the distortion-limit value and the minimum number of words in the English sentence, which means the best translation accuracy cannot be obtained when using a bigger distortion limit than the number of words in the sentence. However, the correlation coefficient is 0.4 when "unlimited" is not used, and this means that improving the translation accuracy by using the relationship between the distortion-limit value and the number of words in the sentence is difficult. There are 489 sentences whose BLEU value became the maximum value when the distortion-limit value is "unlimited." In other cases, there are 739 sentences. From this result, the optimum distortion-limit value is not necessarily "unlimited" in the English-to-Japanese translation of patent documents.

The difference in the word order between European languages and English is small and the distortion-limit parameter in the Moses decoder is applied to these languages. On the other hand, the structure of Japanese is much different from that of English. In such a case, the definition of the distortion-limit value in the Moses decoder needs to be thoroughly revised.

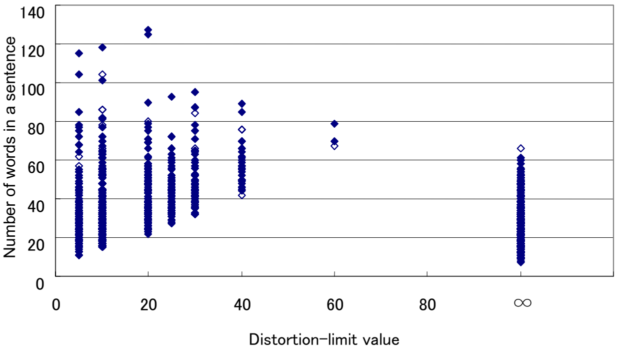

Figure 2. Relationship between number of words in sentence and distortion limit

5. Conclusion

We proposed and evaluated the clause-separation method. However, the proposed method did not improve the BLEU value, and a side effect that offset the improvement caused by the proposed method occurred. We also experimented to find the optimum value of the distortion-limit parameter. The optimal value of the distortion limit was traditionally considered to be unlimited in the translation between Japanese and English. However, as a result of the experiment, the BLEU value became maximum when the distortion-limit value was set to 20.

In the future, we will analyze the effect of the clause-separation method in detail and study a method to limit the side effect. We will also investigate how to control the distortion-limit parameter.

References

[1] Hiroyuki Kaji. An efficient execution method for rule-based machine translation. In *Proceedings of the 12th Conference on Computational Linguistics*, Association for Computational Linguistics Morristown, NJ, USA, 1988

[2] Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. Learning translation templates from bilingual text. In *Proceedings of the 14th Conference on Computational Linguistics,* Association for Computational Linguistics Morristown, NJ, USA, 1992

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2008.

[4] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano. Japanese Morphological Analysis System ChaSen version 2.0 Manual. *NAIST Technical Report*, *NAIST-IS-TR99009*, April 1999

[5] Franz Josef Och, Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume 29, number 1, pp. 19-51,March 2003

[6] A. Stolcke. SRILM - An extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, *vol. 2*, *pp. 901--904*, Denver, CO, September 2002.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic, June 2007

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU:a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, Philadelphia, 2002

[9] Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. *IWSLT2005*, 2005

[10] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. In *Proceedings of HLT/EMNLP*, 2005