Rule-based Reordering Constraints for Phrase-based SMT

Chooi-Ling Goh and **Takashi Onishi** and **Eiichiro Sumita**
Multilingual Translation Laboratory
National Institute of Information and Communications Technology
619-0289 Kyoto, Japan
{chooiling.goh,takashi.onishi,eiichiro.sumita}@nict.go.jp

Abstract

Translation results suffer when a standard phrase-based statistical machine translation system is used for translating long sentences. The translation output will not preserve the same word order as the source, especially between a language pair that has different syntactic structures. When a sentence is long, it should be partitioned into several clauses, and the word reordering during the translation should be done within clauses, not between clauses. In this paper, we propose splitting the long sentences using linguistic information, and translating the sentences with the split boundaries. In other words, we constrain the word reordering so that it cannot cross the boundaries. We propose two types of constraints: split condition and block condition. By doing so, word order can be preserved and translation quality improved. Our experiments on the patent translation between Japanese and English are able to achieve better translations measured by BLEU score, NIST score and word error rate (WER).

1 Introduction

Translating long and complex sentences has been a critical problem in machine translation. A standard phrase-based statistical machine translation (SMT) system cannot solve the problem of word reordering in the target when the source sentence has a complex structure. A syntax-based machine translation system could solve the problem by running a parser on the source sentence in order to get the syntactic structure, but when a sentence is long and complex, the parser may fail to give a correct parse tree. Klein and Manning (2003) have shown that the accuracy of parsing decreases as sentence length increases, and the parsing time increases exponentially. However, in this research, we found that even when a sentence is long and complex, it is possible to split a sentence into smaller units which can be translated separately with minor consideration of the context. The main problem here is locating the best locations for the split. We use linguistic information such part-of-speech (POS) tags and commas as clues to determine the split positions. After splitting a sentence into small clauses, the clauses are translated almost independently. This means that word reordering can only be done within a clause, not between clauses. This constraint can be specified using "wall" tag in Moses[1] (Koehn and Haddow, 2009). Furthermore, a long sentence may include some long and complex noun phrases. These noun phrases should also be translated individually regardless of the context. We try to locate the area of these noun phrases by looking at the sequences of nouns in the sentences. We then bracket these noun phrases using the "zone" constraint in Moses.

We use the NTCIR-8 (Fujii et al., 2010) Patent Translation shared task data between Japanese and English in our experiment. The sentences in the patent are always long and have complex structures, so even the humans have difficulties understanding if the texts are not segmented into the proper portions. In our experiment, the results show that splitting the long sentences into small independent clauses and bracketing the noun phrases helps to improve the translation quality. Automatic evaluation using BLEU scores, NIST scores and WER shows that our rule-based constraints can improve the translation in a phrase-based SMT system.

[1] http://www.statmt.org/moses/. The Moses toolkit is a statistical machine translation system that allows automatic training of translation models for any language pair.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 113–120
Leuven, Belgium, May 2011

2 Previous Work

Research has been done on splitting long sentences into smaller segments in order to improve the translation. Splitting can be done either at the translation model training phase or the translation testing phase.

Furuse et al. (1998) and Doi and Sumita (2003) tended to split speech output instead of text data. For speech output, the main issue is that one utterance may contain a few short sentences instead of one long sentence. Therefore, the main problem is splitting them into proper sentences for translation. However, since there is no punctuation in the speech data, it is difficult to locate the sentence boundaries, so, parsing results and word characteristics were used to determine the sentence boundaries.

Kim and Ehara (1994) proposed using a rule-based method to split long sentences into multiple sentences. Furthermore, after splitting the sentences, they tried to identify a subject and inserted it into the subsequence sentences wherever needed. When a sentence is split, the ending grammar of the former part is changed so that its conjugation (tense, aspect, modality) matches the ending of the original complete sentence. This process is especially necessary for Japanese, where the ending grammar in the middle of the sentence is usually not complete.

Xu et al. (2005) proposed to separate a sentence pair into two sub-pairs based on a modified IBM Model 1. This process continues recursively over all the sub-pairs until their lengths are smaller than a given threshold. Finally, these sub-pairs are used for training a statistical translation model. The difference with our proposed method is that they split the sentences used for training, whereas we split only the input test data: they split the sentences using a bilingual corpus, but we split the sentences monolingually.

Sudoh et al. (2010) proposed dividing the source sentence into small clauses using a syntactic parser. Then, a non-terminal symbol serves as a place-holder for the relative clause. However, they would also have to train a clause translation model which can translate the non-terminal symbols. They proposed a clause alignment method using a graph-based method to build the non-terminal corpus. The advantage of their method is that it can perform short and long distance reordering simultaneously.

Domain	Japanese			English		
	MIN	MAX	AVG	MIN	MAX	AVG
Travel	2	162	10.75	2	116	9.46
News	1	132	28.77	1	135	26.52
JST	1	250	30.98	1	114	25.36
Patent	3	636	39.95	1	474	33.92

Table 1: Minimum/Maximum/Average sentence lengths in various domains

Xiong et al. (2010) used Maximum Entropy Markov Models to learn the translation boundaries based on word alignments in hierarchical trees. The obtained beginning and ending translation boundaries are integrated into the decoder as soft constraints. A new feature is introduced to the decoder's log linear model: translation boundary violation counting. This feature prefers the hypotheses that are consistent with the translation boundaries.

Our research in this paper is different in the sense that we only want to split long sentences for translation where the context before and after the splitting points are independent, and we specify the translation zones as the boundary constraints. Our method is simple and does not require complicated processes like clause alignment, parsing, subject supplement or sentence ending completion.

3 Translation of Long Sentences in Patent

Patent translation is a difficult task, as the sentences are usually very long and consist of many complex noun phrases. Table 1 shows the minimum, maximum and average sentence lengths in various domains for Japanese and English, such as the travel domain[2] (Fordyce, 2007), JENAAD news articles (Utiyama and Isahara, 2003), JST[3] scientific paper abstracts and patents (Fujii et al., 2010). For the word count statistics, the Japanese texts are segmented using ChaSen[4] (Matsumoto et al., 2007) and the English texts are tokenized using a standard tokenizer provided by WMT workshop[5]. As we can see, the patent text is much longer than any other domain and the maximum length is 3-4 times longer as well. For the

[2]BTEC - http://iwslt07.fbk.eu/.
[3]http://www.jst.go.jp
[4]http://chasen-legacy.sourceforge.jp/
[5]http://www.statmt.org/wmt08/scripts.tgz

Source	また、図１中の第１のスイッチ素子１３は放電用のＮＭＯＳトランジスタ１７からなり、このＮＭＯＳトランジスタ１７のゲートは制御回路２３により制御される。
Reference	in addition , the first switch element 13 in fig . 1 comprises an nmos transistor 17 , and a gate electrode of the nmos transistor 17 is controlled by a control circuit 23 .
Baseline	further , the first switch element 13 is controlled by the control circuit 23 , and the gate of the nmos transistor 17 from the nmos transistor 17 shown in fig . 1 for discharge .
Split into multiple clauses	また、図１中の第１のスイッチ素子１３は放電用のＮＭＯＳトランジスタ１７からなり、 further , the first switch element 13 in fig . 1 , from the discharging nmos transistor 17 and このＮＭＯＳトランジスタ１７のゲートは制御回路２３により制御される。 the gate of the nmos transistor 17 is controlled by the control circuit 23 .
Source	accordingly , in the radiation image generating system 100 , it is possible to reduce electric power consumption under an image generation standby mode while an image generation is immediately performed , and it is possible to realize electric power saving and a long life duration thereof .
Reference	したがって、放射線画像撮影システム１００では、迅速に撮影を行いながらも撮影待機モードにおける消費電力の削減し、省電力化及び長寿命化を図ることが可能となる。
Baseline	従って、スタンバイモード時の消費電力を低減することができるので、放射線画像生成システム１００では、画像の生成が行われると、直ちにその寿命が長い省電力化を実現することができる画像生成されている。
Split into multiple clauses	accordingly , in the radiation image generating system 100 , 従って、放射線画像生成システム１００では、 it is possible to reduce electric power consumption under an image generation standby mode while an image generation is immediately performed , 消費電力を低減することができる画像生成スタンバイモードでは、画像の生成が即座に実行し、 and it is possible to realize electric power saving and a long life duration thereof . 省電力で長寿命を実現することができる。

Figure 1: Long sentence translation examples

travel domain, long sentences exist because multiple small sentences are joined into one utterance. However, in the patent text, one single sentence by itself can be very long.

A standard phrase-based statistical machine translation system does not work well for translating long sentences. This is because the longer the sentence, the larger the search space for reordering becomes. Therefore, the word order in the translation may not be arranged in the correct order as in the source. Figure 1 shows two examples of long sentence translation using a standard phrase-based SMT system. In both sentences, the word order of the translation does not follow the source sentence and the translation is not satisfactory. However, if we can split the sentences into small clauses such as those shown below the baseline translation, each clause can be translated in a better word order, and the overall translation improves. Furthermore, for a noun phrase like "図 1 中の第 1 のスイッチ素子 1 3", the baseline model has translated it into "the first switch element 13" and "in fig . 1", and reordered them separately into different locations. If we could put them in a translation block, then they might be translated correctly as a set as "the first switch element 13 in fig . 1" and not mixed with the outside material. Our research here is to find out where best to split the sentences into small clauses and the area for the translation blocks.

4 Proposed Method

4.1 Split Conditions

Many previous research showed that punctuation is very useful when parsing a text (Jones, 1994; Briscoe and Carroll, 1995; Collins, 1996; Jin et al., 2004). A comma is one such useful mark. Basically, a comma has two roles: as a delimiter to separate different syntactic types, or as a separator to separate the elements of the same category type (Nunberg, 1990). However, this information alone is not enough to distinguish whether the comma is suitable to be a split position for machine translation. A comma and the information around the comma could help to find a proper place for a split. Whether or not it is a proper place for a split depends upon if the information on the left and right sides of the comma are able to be translated independently.

Punctuation can be very useful in written texts for aiding in comprehension. According to Murata et al. (2010), there are more than 8 uses for commas in Japanese written text, and 36.32% of commas are used when the context before and after are independent of each other. This indicates to us that a Japanese comma can be used as a clue for a split positions. However, while a comma is usually used in Japanese to improve readability if a sentence is long and complicated, its use is not compulsory and there are no strict rules on usage, so research

POS tag	Description
Head Position	
副詞-助詞類接続	adverb-particle_conjunction
接続詞	conjunction
Tail Position	
名詞-副詞可能	noun-adverbial
名詞-非自立-副詞可能	noun-affix-adverbial
動詞-自立	verb-main
動詞-非自立	verb-auxiliary
動詞-接尾	verb-suffix
助動詞	auxiliary
助詞-格助詞-連語	particle-case-compound
助詞-接続助詞	particle-conjunctive
助詞-係助詞*	particle-dependency
助詞-副詞化	particle-adverbializer
助詞-格助詞-一般*	particle-case-misc

Table 2: POS tags used for split in Japanese

is being done on inserting missing punctuation into the text (Murata et al., 2010; Guo et al., 2010).

Similar to Kim and Ehara (1994), a rule-based approach is proposed to split a sentence into multiple clauses. First, the sentence is part-of-speech (POS) tagged by ChaSen using the IPAdic dictionary. In many cases, if there is a comma, the context before and after the comma may be independent and can be translated separately, making a comma a very important clue for locating splitting position candidates. However, not all commas are suitable to be used as split boundaries. We therefore combine the POS tags and commas as clues to determine the split position for long sentences. Table 2 shows some of the POS tags that have been used for splitting Japanese text. These POS tags were analyzed and found to be good markers for splitting position candidates, as the clauses before and after they occur may be independent of each other, and thus able to be translated independently. Two simple rules are used:

1. If a POS tag in the head position is found after a comma, then the head will be a split position.
2. If a POS tag in the tail position is found before a comma, then the word after the comma will be a split position.

Most of the POS tags indicate places that are very likely to be good for splitting. There are two exceptions: the dependency particle (助詞-係助詞) and the general case marker (助詞-格助詞-一般). If a comma is found after a dependency particle, it is hard to say if the context before the particle is independent of the context after the comma. However, by observing the corpus data, we found that when the sentence is long, it is better if the text before the dependency particle to be translated seperately if it happens to be a very long noun phrase. Therefore, we leave it here as one of the POS tags for splitting the sentences. The only condition is that the dependency particle can only be used if no other conditions are fulfilled. For the general case marker, a POS tag at the tail position must precede the case marker to make it valid for a split, such as "ため/名詞-非自立-副詞可能 に/助詞-格助詞-一般 、/記号-読点" (because of).

POS tag	Description
Head Position	
CC	coordinating conjunction
DT*	determiner
EX	existential *there*
IN	preposition or subordinating conjunction
PP	personal pronoun
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
WRB	Wh-adverb

Table 3: POS tags used for split in English

For English, we used Treetagger[6] to obtain the POS tags. Table 3 shows the POS tags we used for splitting. For English, we only have POS tags in the head position because we could not find any clues that could be used at the tail position. There is a special case for a split with a determiner (DT): a split can be done if and only if the head position of the preceding clause is on the list of POS tags for a split. It is because in this case, there is a high possibility that the determiner be the new subject for the next clause. For both Japanese and English, a split cannot be done in between a pair of brackets. We also specify a threshold for the minimum number of words a clause must contain after splitting. We do not split if a clause is too short because when a sentence/clause is short, word reordering can usually be done correctly without problems.

[6]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

POS tag	Description
Japanese	
名詞	all types of nouns
接頭詞	prefix
記号-アルファベット	symbol-alphabetic
記号-空白*	symbol-space
記号-一般*	symbol-misc
助詞-連体化*	particle-adnominalizer
未知語	unknown word
English	
N+	all types of nouns
CD	cardinal number
DT#	determiner
JJ	adjective
,*	comma (only when left-right contexts are the same type)

Table 4: POS tags used inside a noun phrase block

4.2 Block Conditions

Blocks are the areas where long noun phrases are found. While complex noun phrases often occur in patent, we are only concentrating on extracting simple long noun phrases. This is because complex noun phrases normally consist of simple noun phrases plus functional words or even verbs. If we can extract the simple noun phrase as a block, it will ease translating the whole complex noun phrase as well. We look for continuous sequence of words that fall into some prefedined POS tags. Table 4 shows the POS tags used for determining whether a word should be grouped in blocks for Japanese and English, respectively. The POS tags that are marked with an asterik (*) are not allowed to appear either at the begining or at the end of a block and the hash mark (#) is not allowed at the end of a block. Finally, we also insert blocks for open and close brackets, as they should not be mixed with the outside material during the translation process, similar to what was proposed by Koehn and Haddow (2009).

4.3 Reordering Constraint Marker

The Moses decoder (Koehn et al., 2007; Koehn and Haddow, 2009) provides a way to specify the word reordering constraints. Two types of constraints were introduced:

1. Words within zones have to be translated without reordering with outside material.
2. Walls form hard reordering constraints, over which words may not be reordered.

The wall and zone constraints are compatible with our split and block conditions. We apply the wall constraint as our proposed split condition in Section 4.1 and the zone constraint as our block condition in Section 4.2. By adding the <zone> and <wall> markers, the input source sentences from Figure 1 will look like this.

Japanese

また 、 <zone> 図 1 中 の 第 1 の スイッチ 素子 1 3 </zone> は <zone> 放電 用 の ＮＭＯＳ トランジスタ 1 7 </zone> から なり 、<wall /> この <zone> ＮＭＯＳ トランジスタ 1 7 の ゲート </zone> は <zone> 制御 回路 2 3 </zone> により 制御 さ れる 。

English

accordingly , in <zone> the radiation image generating system 100 </zone> , <wall /> it is possible to reduce <zone> electric power consumption </zone> under <zone> an image generation standby mode </zone> while <zone> an image generation </zone> is immediately performed , <wall /> and it is possible to realize <zone> electric power saving </zone> and <zone> a long life duration </zone> thereof .

5 Experiment Results

We used the patent corpus provided by the NTCIR-8 Translation Campaign[7] for Japanese and English translation. The training corpus contains about 3 million sentence pairs, the development set has 2,000 sentence pairs and the test set has 1,251 and 1,119 sentence pairs for J-E and E-J translation directions, respectively.

We used Moses as a baseline system, with the following settings:

- grow-diag-final-and heuristic
- 5-gram language model, interpolated Kneser-Ney discounting
- msd-bidirectional-fe lexicalized reordering
- distortion-limit = -1 (unlimited).

The distortion limit is set to unlimited, based on the findings in Kumai et al. (2008). Since Japanese

[7] http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html

	Japanese	English
# of zone	5604	5696
# of wall	914	434
# of bracket zone	280	347
# of clause	2165	1553

Table 5: Number of zones, walls, bracket zones and clauses in the test data

and English are fairly different in the word order, the distortion limit has to be more than 20 words and the difference with the unlimited value is less than 0.5% of the absolute BLEU score. The feature weights were optimized using Minimum Error Rate Training (MERT) with the development set. We used a minimum length of 10 words as a threshold to split the sentences and insert the "wall" constraint. In other words, a split segment must contain at least 10 words. Gerber and Hovy (1998) have fixed the minimum sentence length as 7 words for splitting, but they could not prove that this is the best length. However, based on our preliminery experiment, a threshold of 10 words could give optimum results between BLEU and WER (Goh and Sumita, 2011). For the block condition, a zone must contain at least two words.

Table 5 shows the statistics of the zones, walls, bracket zones and clauses in the test data for the word reordering constraints. The number of clauses shows the number of independent clauses after splitting. We can see that splitting Japanese text generates a lot more clauses than the English text. Our rules for English still not be able to cover most of the cases, and more attention should be given to this problem in the future.

From Table 6, we know that Japanese text splits most with the dependency particle, followed by the main verb and conjunctive particle. These three POS tags cover more than 57% of all split conditions. For English text, the most used is the coordinating conjunction, followed by the determiner, which covers more than 67%. For the zones (excluding bracket zones), the longest zone for Japanese is 25 words and the shortest is 2 words, with an average of 4.64 words. For English, the counts are 16 words, 2 words and 3.23 words, respectively.

Table 7 shows the translation evaluation results. The English reference is lowercased and tokenized, and the Japanese reference is segmented by ChaSen. We evaluated the results using BLEU

POS tag	# of split
Japanese	
particle-case-compound	72
particle-dependency	238
particle-conjunctive	125
particle-adverbializer	50
auxiliary	69
auxiliary+particle-case-misc	2
conjunction	28
verb-main	163
verb-suffix	67
verb-auxiliary	17
adverb-particle_conjunction	2
noun-affix-adverbial	38
noun-affix-adverbial+particle-case-misc	26
noun-adverbial	15
noun-adverbial+particle-case-misc	2
TOTAL	914
English	
coordinating conjunction	164
determiner	128
existential *there*	1
preposition or subordinating conjunction	56
personal pronoun	21
adverb	39
adverb, comparative	1
Wh-adverb	24
TOTAL	434

Table 6: Number of split condition by POS tags

(Papineni et al., 2002), NIST (Doddington, 2002) and WER (Nießen et al., 2000). The baseline shows the results without using any reordering constraints. "Wall" shows the results where only the wall constraint was used and "zone" shows the results where only the zone constraint was used. Finally, the "wall+zone" shows the results where both the wall and zone constraints were used. From the results, we can see that by adding either zone or wall constraints, we can improve the quality in all three metrics. However, WER improves the most with the wall constraint. This means that wall constraint contributes more to control long-distance word reordering. By adding both constraints together, the improvements piles up and the best translation quality is obtained. The ab-

Method	BLEU	NIST	WER
Japanese-English			
baseline	0.3034	7.5754	0.7960
wall	0.3102	7.6286	0.7533
zone	0.3115	7.7327	0.7817
wall+zone	**0.3179**	**7.7436**	**0.7408**
English-Japanese			
baseline	0.3780	8.0584	0.7096
wall	0.3876	8.1349	0.6741
zone	0.3878	8.1309	0.6977
wall+zone	**0.3932**	**8.1635**	**0.6692**

Table 7: Translation results

solute improvements in BLEU percentage scores are +1.45/+1.52, NIST scores are +0.17/+0.11 and WER percentage values are -5.52/-4.04, at a significance level of 95% confidence using bootstrap method[8]. Figure 2 and Figure 3 show some good and bad translations. For the bad translation, the split condition did not provide a suitable split point for translation. However, the block condition worked well for most of the cases.

6 Conclusion and Future Work

It is difficult to translate long sentences using a standard phrase-based statistical machine translation system due to source word order being badly preserved in the target. We proposed splitting the long sentence into multiple short clauses and several block areas that could be translated independently. POS tags and commas are used as clues to determine the splitting positions and the block areas. "Zone" and "wall" markers in Moses are used to specify these constraints in the source text. Our experiment results for the patent translation between Japanese and English showed some improvements in the translation quality measured by BLEU score, NIST score and WER. In the future, automatic sentence clause splitters or noun phrase chunking by statistical approach will be considered to replace the human-crafted rules.

References

Briscoe, Ted and John Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies*, pages 48–58.

Collins, Michael John. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the ACL*, pages 184–191.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the HLT*, pages 138–145.

Doi, Takao and Eiichiro Sumita. 2003. Input Sentence Splitting and Translating. In *Proceedings of the HLT/NAACL: Workshop on Building and Using Parallel Texts*.

Fordyce, Cameron S. 2007. Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proceedings of NTCIR-8 Workshop Meeting*, pages 371–376.

Furuse, Osamu, Setsuo Yamada, and Kazuhide Yamamoto. 1998. Splitting Long and Ill-formed input for Robust Spoken-language Translation. In *Proceedings of COLING-ACL*, pages 421–427.

Gerber, Laurie and Eduard Hovy. 1998. Improving Translation Quality by Manipulating Sentence Length. In *Proceedings of AMTA*, pages 448–460.

Goh, Chooi-Ling and Eiichiro Sumita. 2011. Splitting Long Input Sentences for Phrase-based Statistical Machine Translation. In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pages 802–805.

Guo, Yuqing, Haifeng Wang, and Josef van Genabith. 2010. A Linguistically Inspired Statistical Model for Chinese Punctuation Generation. *ACL Transactions on Asian Language Information Processing*, 9(2):6:1–6:27.

Jin, Meixun, Mi-Young Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2004. Segmentation of Chinese Long Sentences Using Commas. In *Proceedings of SIGHAN Workshop On Chinese Language Processing*, pages 1–8.

Jones, Bernard. 1994. Exploring the Role of Punctuation in Parsing Natural Text. In *Proceedings of the COLING*, pages 421–425.

Kim, Yeun-Bae and Terumasa Ehara. 1994. A Method for Partitioning of Long Japanese Sentences with Subject Resolution in J/E Machine Translation. In *Proceedings of International Conference on Computer Processing of Oriental Languages*, pages 467–473.

Klein, Dan and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS)*, pages 3–10.

Koehn, Philipp and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164.

[8] http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm

Source	従って 、 <zone> 上記 角度 表示 バー 1 8 </zone> を <zone> 上記 角度 表示 片 1 7 の 任意 の 位置 </zone> に 固定 する こと が できる と共に 、 <wall /> <zone> 上記 規制 片 1 6 </zone> を <zone> 接合 部材 3 2 </zone> を 介し て <zone> 上記 板 部 2 9 </zone> および <zone> 板 部 3 0 </zone> に対し スライド さ せる こと が できる 。
Reference	hence , the angle indicating bar 18 can be fixed to an arbitrary position of the angle indicating piece 17 , and the regulating piece 16 can be let slide relative to the plate section 29 and plate section 30 via the joining member 32 .
Baseline	therefore , the angle of the angle display bar 18 can be fixed at an arbitrary position of the restricting piece 16 is joined member 32 , and the plate portion 29 and the plate portion 30 can be slid via a display piece 17 .
Proposed	accordingly , the angle display bar 18 can be fixed in an arbitrary position of the angle display piece 17 and the restricting pieces 16 can slide relative to the plate portion 29 and the plate portion 30 through a joint member 32 .
Source	after <zone> the redundancy information </zone> is latched into <zone> the shift register 53 </zone> , <wall /> for example , <zone> the redundancy information </zone> is serially read out from <zone> the shift register 53 </zone> and serially input to <zone> an even-odd determination circuit 49 </zone> .
Reference	リダン ダン シ 情報 を シフトレジスタ 5 3 に ラッチ し た 後 、 リダン ダン シ 情報 は 、 この シフトレジスタ 5 3 から 、 例えば 、 シリアル に 読み出さ れ 、 偶 奇 判定 回路 4 9 に 、 シリアル に 入力 さ れる 。
Baseline	リダン ダン シ 情報 は 、 例えば 、 シリアル に 読み出さ れる と 、 シフトレジスタ 5 3 に シリアル に 入力 さ れ 、 シフトレジスタ 5 3 に ラッチ さ れ た 後 に 、 リダン ダン シ 情報 は 偶数 判定 回路 4 9 を 備え て い る 。
Proposed	リダン ダン シ 情報 は 、 シフトレジスタ 5 3 に ラッチ さ れ た 後 、 例えば 、 リダン ダン シ 情報 を シリアル に 読み 出し て シフトレジスタ 5 3 から 偶数 判定 回路 4 9 に シリアル に 入力 さ れる 。

Figure 2: Good translation examples

Source	この <zone> 酸素 含有 ガス </zone> は 、 例えば 、 空気 で あっ て も よい し 、 <wall /> <zone> 純 酸素 </zone> で あっ て も よい し 、 空気 又は <zone> 純 酸素 </zone> を 、 窒素 、 アルゴン 、 <zone> ヘリウム の よう </zone> な <zone> 不 活 性 ガス </zone> で 希釈 し た もの で あっ て も よい 。
Reference	the oxygen-containing gas may be , for example , air or pure oxygen , or may be diluted air or pure oxygen with an inert gas such as nitrogen , argon and helium .
Baseline	this oxygen containing gas diluted with inert gas such as nitrogen , argon , helium , and may be , for example , may be a pure oxygen may be in the air or pure oxygen of the air .
Proposed	this oxygen containing gas may be air , for example , the pure oxygen in the air or an inert gas such as nitrogen , argon or helium gas may be diluted with pure oxygen may be .
Source	therefore , <zone> a polarization speed </zone> is largely affected by , <wall /> for example , <zone> the size , moment </zone> , and shape of <zone> the molecule </zone> .
Reference	従って 、 分子 の 大き さ 、 モーメント 、 形状 等 によって 分極 速度 が 大きく 影響 を 受ける 。
Baseline	そこで 、 偏 光 、 例えば 、 分子 の 形状 、 大き さ 、 モーメント の 影響 を 大きく 受ける 速度 で ある 。
Proposed	そこで 、 偏 光 速度 に 大きく 影響 さ れる が 、 例えば 、 分子 の 形状 、 サイズ 、 モーメント で ある

Figure 3: Bad translation examples

Koehn, Philipp, 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL, demo and poster session*, pages 177–180.

Kumai, Hiroyuki, Hirohiko Sagawa, and Yasutsugu Morimoto. 2008. NTCIR-7 Patent Translation Experiments at Hitachi. In *Proceedings of NTCIR-7 Workshop Meeting*, pages 441–444.

Matsumoto, Yuji, Kazuma Takaoka, and Masayuki Asahara, 2007. *ChaSen Morphological Analyzer version 2.4.0 User's Manual*. Nara Institute of Science and Technology, March.

Murata, Masaki, Tomohiro Ohno, and Shigeki Matsubara. 2010. Automatic Comma Insertion for Japanese Text Generation. In *Proceedings of the EMNLP*, pages 892–901.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research. In *Proceedings of the 2nd LREC*, pages 39–45.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation*. CSLI lecture notes: no. 18.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL*, pages 311–318.

Sudoh, Katsuhito, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. 2010. Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation. In *Proceedings of the Joint 5th Workshop on SMT and MetricsMATR*, pages 418–427.

Utiyama, Masao and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the ACL*, pages 72–79.

Xiong, Deyi, Min Zhang, and Haizhou Li. 2010. Learning Translation Boundaries for Phrase-Based Decoding. In *Proceedings of NAACL*, pages 136–144.

Xu, Jia, Richard Zens, and Hermann Ney. 2005. Sentence Segmentation Using IBM Model Alignment Model 1. In *Proceedings of the EMNLP*, pages 280–287.