Patent Claim Translation based on Sublanguage-specific Sentence Structure

Masaru Fuji fuji.masaru@nict.go.jp
National Institute of Information and Communications Technology, Kyoto, Japan,
and Nara Institute of Science and Technology, Nara, Japan
Atsushi Fujita atsushi.fujita@nict.go.jp
Masao Utiyama mutiyama@nict.go.jp
Eiichiro Sumita eiichiro.sumita@nict.go.jp
National Institute of Information and Communications Technology, Kyoto, Japan
Yuji Matsumoto matsu@is.naist.jp
Nara Institute of Science and Technology, Nara, Japan

Abstract

Patent claim sentences, despite their legal importance in patent documents, still pose difficulties for state-of-the-art statistical machine translation (SMT) systems owing to their extreme lengths and their special sentence structure. This paper describes a method for improving the translation quality of claim sentences, by taking into account the features specific to the claim sublanguage. Our method overcomes the issue of special sentence structure, by transferring the sublanguage-specific sentence structure (SSSS) from the source language to the target language, using a set of synchronous context-free grammar rules. Our method also overcomes the issue of extreme lengths by taking the sentence components to be the processing unit for SMT. The results of an experiment demonstrate that our SSSS transfer method, used in conjunction with pre-ordering, significantly improves the translation quality in terms of BLEU scores by five points, in both English-to-Japanese and Japanese-to-English directions. The experiment also shows that the SSSS transfer method significantly improves structural appropriateness in the translated sentences in both translation directions, which is indicated by substantial gains over 30 points in RIBES scores.

1. Introduction

Advances in reordering techniques based on syntactic parsing (Isozaki et al., 2010b; de Gispert et al., 2015), with growing volumes of parallel patent corpora available, have brought significant improvements in the performance of statistical machine translation (SMT) for translating patent documents across distant language pairs (Goto et al., 2012; Goto et al., 2015). However, among various sentences within a patent document, patent claim sentences still pose difficulties for SMT resulting in low translation quality, despite their utmost legal importance.

A patent claim sentence is written in a kind of *sublanguage* (Buchmann et al., 1984; Luckhardt, 1991) in the sense that it has the following two characteristics: (i) comprising a patent claim by itself with an extreme length and (ii) having a typical sentence structure composed of a fixed set of components irrespective of language, such as those illustrated in Figures 1 and 2. The difficulties in patent claim translation lie in these two characteristics. Regarding the first characteristic, the extreme lengths cause syntactic parsers to fail with consequent low reordering accuracy. Regarding the second characteristic, the high regularity of the claim-specific sentence structure cannot be captured and transferred properly by the models trained only on the other parts of patent documents, such as the *abstract* and *background description*.

This paper presents a method for improving the SMT translation quality of patent claims. We have developed a system that is used as an add-on to state-of-the-art, off-the-shelf SMT systems to deal with the sentence structure specific to the patent claim sublanguage. Our method based on this *sublanguage-specific sentence structure* (henceforth, *SSSS*) has two major effects. (1) Pre-ordering and SMT are applied for each sentence component, rather than for the entire long sentence. This in effect shortens the input to pre-ordering and SMT, thus improves translation quality. (2) Claim sentences are translated according to the sentence structure, producing structurally natural translation outputs. We manually extracted a set of language independent claim components. Moreover, using these components, we constructed a set of synchronous rules for English and Japanese to transfer the SSSS in the source language to the target language.

The results of an experiment demonstrate these two major effects of our SSSS transfer method. Regarding the first effect, when used in conjunction with pre-ordering, our method improves translation quality by five points in BLEU score (Papineni et al., 2002), in both English-to-Japanese and Japanese-to-English translations. Regarding the second effect, gains in RIBES score (Isozaki et al., 2010a) of over 30 points are obtained, indicating that our SSSS transfer is effective in transferring an input sentence structure to the output sentence.

Components		Example strings
Preamble		An apparatus,
Transitional phrase		comprising:
Body	Element	a pencil;
	Element	an eraser attached to the pencil; and
	Element	a light attached to the pencil.

Figure 1. Example of an English patent claim (WIPO, 2014)

Components		Example strings
Body	Element	鉛筆と；
	Element	鉛筆に取り付けられた消しゴムと；
	Element	鉛筆に取り付けられたライトと
Transitional phrase		を備える
Preamble		装置

Figure 2. Japanese patent claim corresponding to Figure 1

2. Transferring Claim-Specific Sentence Structure

While patent claims share a common vocabulary and phrases with the rest of the patent document, they are written in a distinctive way that is different from the rest of the patent document, comprising a sublanguage of its own. This writing style of patent claims developed through the history of filing patent applications, and is now described in the literature. According to the WIPO Patent Drafting Manual (WIPO, 2014), the fundamental structure of an English claim is that it is a single sentence consisting of three components:

$$S \rightarrow \text{PREA TRAN BODY}$$

where S denotes the claim sentence, PREA the *preamble*, TRAN the *transitional phrase* and BODY the *body*. The *preamble* is an introductory phrase that identifies the category of the invention, the *body* is the main component of the claim that describes the elements or purposes of the invention, and the *transitional phrase* is the component that connects the *preamble* and the *body*.

Figure 1 shows one of the typical structures of English claim sentences, in which the *body* of the claim comprises claim elements. Each of the *elements* is a claim component comprising the invention. Figure 2 shows the structure of a Japanese claim sentence corresponding to the English claim sentence shown in Figure 1. Note that the sets of components comprising the claims in the two languages are identical, although the order of components is different in the two languages.

Our manual analysis revealed that a claim consists of a fixed set of components and the set is common to the two languages. We also found that there are strict generation rules in each language. For example, the English patent claim sentence in Figure 1 is represented by the set of rules in Figure 3, where ELEM denotes the *element* component shown in Figure 1. The symbol "+" denotes a non-null list of the preceding components. The corresponding Japanese sentence is represented by another set of rules comprising the same components, as shown in Figure 4.

Having observed a strong regularity in the structure of patent claim sentences across languages, we represent the structural transfer in the form of synchronous context-free grammar (SCFG). For example, we derive the SCFG rules in Figure 5 by connecting the corresponding rules in Figures 3 and 4, where the numeric indices indicate correspondences between non-terminals in both constituent trees. We handcrafted a set of SCFG rules for translating patent claim sentences. The details of the process are presented in Section 3.1

$$\begin{aligned} &\text{S} &&\rightarrow &&\text{PREA TRAN BODY} \\ &\text{TRAN} &&\rightarrow &&\text{“comprising:”} \\ &\text{BODY} &&\rightarrow &&\text{ELEM+} \end{aligned}$$

Figure 3. Example of generation rules for an English claim sentence

$$\begin{aligned} &\text{S} &&\rightarrow &&\text{BODY TRAN PREA} \\ &\text{TRAN} &&\rightarrow &&\text{“備える”} \\ &\text{BODY} &&\rightarrow &&\text{ELEM+} \end{aligned}$$

Figure 4. Example of generation rules for a Japanese claim sentence

$$\begin{aligned} \text{S} &\rightarrow \langle \text{PREA}_{\textcircled{1}}\ \text{TRAN}_{\textcircled{2}}\ \text{BODY}_{\textcircled{3},}\ \text{BODY}_{\textcircled{3}}\ \text{TRAN}_{\textcircled{2}}\ \text{PREA}_{\textcircled{1}} \rangle \\ \text{BODY} &\rightarrow \langle \text{ELEM+, ELEM+} \rangle \\ \text{TRAN} &\rightarrow \langle \text{“comprising:”, “備える”} \rangle \end{aligned}$$

Figure 5. SCFG rules derived from English rules in Figure 3 and Japanese rules in Figure 4

3. Pipeline for Patent Claim Translation

While patent claim sentences have a distinctive structure, their components, such as the *elements* and *purposes* of the claimed inventions, are described with the same vocabulary and phrases in the other parts of patent documents. We therefore implemented the SSSS transfer as an add-on to off-the-shelf SMT systems. More specifically, given a patent claim sentence in the source language, our method translates it through the following three-step pipeline (see also Figure **6**).

A button comprising: a plurality of first ribs integrally formed on the surface of the plate-like base portion, each rib radially extending from a center towards the circumference of the plate-like base portion; and an annular portion integrally formed on the surface of the plate-like base portion, to which each center ends of the plurality of first ribs are coupled.

(a) Input English sentence

[S [PREA A button] [TRAN comprising:] [BODY [ELEM a plurality of first ribs integrally formed on the surface of the plate-like base portion, each rib radially extending from a center towards the circumference of the plate-like base portion;] [ELEM and an annular portion integrally formed on the surface of the plate-like base portion, to which each center ends of the plurality of first ribs are coupled.]]]

(b) Synchronously obtained English SSSS

[S [BODY [ELEM a plurality of first ribs integrally formed on the surface of the plate-like base portion, each rib radially extending from a center towards the circumference of the plate-like base portion;] [ELEM and an annular portion integrally formed on the surface of the plate-like base portion, to which each center ends of the plurality of first ribs are coupled]] [TRAN を備える] [PREA A button]]

(c) Synchronously generated Japanese SSSS

[S [BODY [ELEM plate like base portion of circumference towards center from extending plate like base portion of surface on formed integrally first ribs of plurality , each rib radially;] [ELEM and plate like base portion of surface, plurality of first ribs of each center ends coupled are which to on formed integrally annular portion]] [TRAN を備える] [PREA A button]]

(d) Each SSSS component pre-ordered

[S [BODY [ELEM 前記板状ベース部の前記表面で一体に形成され、各々が前記板状ベース部の中心から外周に向かって放射状に延在する複数の第１リブと、] [ELEM 前記板状ベース部の前記表面で一体に形成され、 前記複数の第１リブ各々の中心端が連結された環状部と、]] [TRAN を備える] [PREA ボタン]]

(e) Each SSSS component translated by English-to-Japanese SMT

Figure 6. Overview of our translation pipeline

1. **Step 1. SSSS transfer** (Figure 6: (a) → (b), (c)): The given sentence is analyzed using a set of handcrafted SCFG rules. The goal of this step is not to obtain a fine-grained parse tree of the input sentence, but to identify its sublanguage-specific structure, and transfer it to the target language. By the use of the set of SCFG rules, the components in the given sentence are identified, and simultaneously the sentence structure in the target language is generated.
2. **Step 2. Pre-ordering** (Figure 6: (c) → (d)): The words of each component are reordered so that the order becomes close to that in the target language. This process is performed using a constituent parser. As a result of Step 1, shorter word sequences are the input to this process, resulting in higher parsing and reordering accuracy.
3. **Step 3. Translation by SMT** (Figure 6: (d) → (e)): Each component is translated by an SMT system, and the translated components joined up to form a sentence, with words conjugated and conjunctions added as necessary. Again, as a result of Step 1, shorter components are input that are easier to translate.

The rest of this section elaborates Steps 1 and 2 in turn.

3.1. SSSS Transfer

As described in Section 1, one of the major issues in patent claim translation is that, despite the high regularity, the claim-specific sentence structure cannot be captured and transferred properly by models trained only on the other parts of patent documents.

This step is introduced to identify the structure of the given patent claim sentence and to generate the structure in the target language simultaneously. This process is performed using a set of handcrafted SCFG rules. We created the rules in the following manner. First, we manually analyzed the English and Japanese claim sentences in our development set (described in Section 4.1) and found that each claim sentence is composed of a fixed set of components and that the set is common to the two languages. The set of components U we have identified is as follows:

$$U \in \{\text{PREA, TRAN, BODY, ELEM, PURP}\},$$

where the first four are explained in the previous section, i.e., *preamble*, *transitional phrase*, *body* and *element*. PURP denotes the *purpose* component, which is similar to the *element* component in the sense that they comprise the *body* component.

We then constructed a set of generation rules for English and Japanese claims using U as a set of non-terminal symbols, and obtained 8 and 16 generation rules respectively. We obtained a larger number of rules for Japanese, because the writing style of Japanese claim sentences is more flexible than that of English claim sentences. Finally, we handcrafted a total of 16 SCFG rules by combining the generation rules for the two languages that have the same set of symbols on both the left- and right-hand sides, respectively. Table 1 shows the entire SCFG rule set for English-to-Japanese translation. Our SCFG rules for Japanese-to-English translation are produced by reversing the above English-to-Japanese generation rules.

In the actual implementation of the SCFG rules, we designed each of the rules in the rule set to be deterministic, by using regular expressions for obtaining a unique match for a terminal symbol. For example, to analyze input sentences containing more than one occurrence of the string "*comprising:*" we prepared a regular expression to match the first occurrence. This heuristic rule correctly matches the claim string in most cases.

Table 1. SCFG rule set for English-to-Japanese translation

ID	SCFG rules		
R1	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$, PREA$_{①}$ BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R2	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$, BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R3	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$, PREA$_{①}$ BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R4	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$, BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R5	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$ TRAN$_{②}$ BODY$_{③}$, BODY$_{③}$ TRAN$_{②}$ BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R6	S	→	⟨PREA$_{①}$ TRAN$_{②}$ BODY$_{③}$ TRAN$_{②}$ BODY$_{③}$, PREA$_{①}$ BODY$_{③}$ TRAN$_{②}$ BODY$_{③}$ TRAN$_{②}$ PREA$_{①}$⟩
R7	BODY	→	⟨ELEM+, ELEM+⟩
R8	BODY	→	⟨PURP+, PURP+⟩
R9	TRAN	→	⟨“comprising:”, ”備えることを特徴とする”,⟩
R10	TRAN	→	⟨“comprising:”, ”備える”⟩
R11	TRAN	→	⟨“including:”, ”備えることを特徴とする”⟩
R12	TRAN	→	⟨“including:”, ”備える”⟩
R13	TRAN	→	⟨“having:”, ”備えることを特徴とする”⟩
R14	TRAN	→	⟨“having:”, ”備える”⟩
R15	TRAN	→	⟨“wherein:”, ”ことを特徴とする”⟩
R16	TRAN	→	⟨“wherein:”, ”する”⟩

3.2. Pre-ordering

Another major issue in patent claim translation is that the extreme lengths cause syntactic parsers to fail with consequent low reordering accuracy. To evaluate the effect of introducing our SSSS transfer on the translation quality, we also implemented a pre-ordering tool using state-of-the-art techniques (Isozaki et al., 2010b; Goto et al., 2012; Goto et al., 2015).

Our pre-ordering method is based on syntactic parsing. First, the input sentence is parsed into a binary tree structure by using the Berkeley Parser (Petrov et al., 2006). For example, when “He likes apples.” is inputted into our English-to-Japanese translation system, it is parsed as shown in Figure 7. Second, the nodes in the parse tree are reordered using a classifier. For example, according to the classifier's decision, the two children of the “VP” node, i.e., “VBZ” and “NP”, are swapped, whereas the order of the two children of the “S” node, i.e., “NP” and “VP”, is retained. Once such a decision is made for every node with two children (henceforth, *binary mode*), the word order of the entire sentence becomes very similar to that in Japanese, i.e., “He (*kare wa*) apples (*ringo ga*) likes (*suki da*) . (.)”

The pre-ordering model is trained on a given parallel corpus through the following procedure (Section 4.5 of Goto et al., 2015):

1. Parse the source sentences of the parallel corpus.[1]
2. Perform word alignment on the parallel corpus.
3. Reorder words in each source sentence by swapping some binary nodes so that Kendall's τ over the aligned source and target sentences is maximized. As a

[1] Note that we used the parse model trained from the source treebank, while Goto et al. (2015) used the parse model learned via cross-language syntactic projection.

result, every binary node is classified as either SWAP, i.e., the two children of the node are swapped, or STRAIGHT, i.e., they are not swapped.

4. With the above data, a neural network classifier is trained for predicting whether a given node is SWAP or STRAIGHT.[2]

The constituent parser is also domain-adapted. The initial parsing model for English was trained on the sentences in the Penn Treebank[3] as well as 3,000 patent sentences manually parsed by the authors. The initial model for Japanese was trained on the EDR Treebank[4] consisting of approximately 200,000 sentences. In contrast to what we did for English, we did not use patent sentences in Japanese because no annotator was available.

We first parsed 200,000 patent sentences using the initial parsing model. We then built a patent-adapted (not claim-adapted) parsing model by applying a self-learning procedure (Huang et al., 2009) to the above automatic parses.

```
(ROOT
   (S
      (NP (PRP He))
      (VP (VBZ likes)
         (NP (NNS apples)))
   (. .)))
```

Figure 7. Parsing result of "He likes apples."

4. Experiments

We evaluated to what extent our SSSS transfer and pre-ordering improved the translation quality. As mentioned in Section 3, these methods are implemented as an add-on to off-the-shelf SMT systems. In particular, we used phrase-based SMT (Koehn et al., 2003) as the base system. We also regard it and its hierarchical version (Chiang, 2005) as baseline SMT systems.

4.1. Data

The training data for SMT consists of two subcorpora. The first is the Japanese-English Patent Translation data comprising 3.2 million sentence pairs provided by the organizer of the Patent Machine Translation Task (PatentMT) at the NTCIR-9 Workshop (Goto et al., 2011). We randomly selected 3.0 million sentence pairs. Henceforth, we call this Corpus A. SMT systems trained on the corpus are reasonably good at lexical selection in translating claim sentences, because the vocabulary and phrases are commonly used in entire patent documents, and Corpus A is of a substantial size to cover a large portion of them. However, the claim-specific sentence structure would never be taken into account, as Corpus A does not contain any claim sentences.

To bring claim-specific characteristics into the SMT training, even for the baseline systems, we also used Corpus B comprising 1.0 million parallel sentences of patent claims. These were automatically extracted from pairs of English and Japanese patent documents published between 1999 and 2012 using a sentence alignment method (Utiyama and Isahara, 2007). The concatenation of Corpora A and B was used to train baseline SMT systems, as well as those for our extensions.

[2] Note that Goto et al. (2015) learned the SWAP/STRAIGHT classification problem jointly with the parsing source sentences.
[3] https://www.cis.upenn.edu/~treebank/
[4] https://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html

Development and test data were constructed separately from the training data in the following manner. First, we randomly extracted English patent documents from patents filed in the USA in 2014 and extracted up to the first five claims from each patent document. Then, we randomly selected 2,000 sentences from the results and asked professional translators specializing in patent translation to translate them into Japanese, without informing them that their translations would be used for tuning and testing SMT systems. Finally, the resulting set of 2,000 sentence pairs was randomly divided into development and test data respectively consisting of 1,000 English-Japanese claim sentence pairs.

4.2. Systems

In this experiment, we regard the implementation of phrase-based SMT in the Moses toolkit (Koehn et al., 2007) with distortion limit of six as the baseline. We examined each of our SSSS transfer, and pre-ordering modules and their combination over the baseline. For reference, we investigated the performance of phrase-based SMT with a larger distortion limit 20, as well as hierarchical phrase-based SMT.

Throughout the experiments, we used KenLM (Heafield et al., 2013) for training language models and SyMGIZA++ (Junczys-Dowmunt and Szał, 2010) for word alignment. We used the grow-diag-final method for obtaining phrase pairs. Weights of the models were tuned with n-best batch MIRA (Cherry and Foster, 2012) regarding BLEU (Papineni et al., 2002) as the objective. For each system, we performed weight tuning three times and selected for the test the setting that achieved the best BLEU on the development data.

4.3. Evaluation Metrics

Each system is evaluated using two metrics: BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010a). Although our primary concern in this experiment is the effect of long distance relationship, in general, n-gram based metrics such as BLEU alone do not fully illustrate it. RIBES is therefore used alongside BLEU.

RIBES is an automatic evaluation method based on rank correlation coefficients; RIBES compares the word order in the SMT translation output with those in the reference. Hence it readily depicts the effects of drastic rearrangement in sentence components that often occurs between distant languages. In fact, RIBES has shown high correlation with human evaluation in both English-to-Japanese and Japanese-to-English translation tasks including those in the PatentMT at the NTCIR-9 Workshop (Goto et al., 2011).

Table 2. BLEU and RIBES scores for all systems

ID	Settings			Results			
	SSSS transfer	Pre-ordering	SMT	English-to-Japanese		Japanese-to-English	
				BLEU	RIBES	BLEU	RIBES
P1			**PB d=6**	**23.9**	**43.9**	**21.4**	**40.2**
P1’			PB d=20	23.4 (-0.5)	49.1 (+5.2)	22.4 (+1.0)	46.3 (+6.1)
H1			HPB	24.3 (+0.4)	53.4 (+9.5)	23.2 (+1.8)	49.6 (+9.4)
P2	✓		PB d=6	24.7 (+0.8)	67.9 (+24.0)	20.8 (-0.6)	63.8 (+23.6)
P3		✓	PB d=6	23.7 (-0.2)	55.1 (+11.2)	22.3 (+0.9)	74.4 (+34.2)
P4	✓	✓	**PB d=6**	**28.8 (+4.9)**	**74.9 (+31.0)**	**27.5 (+6.1)**	**74.7 (+34.5)**

4.4. Results

Table 2 summarizes the BLEU and RIBES scores for all systems, where the numbers in the brackets show the improvement over P1, the vanilla PBSMT system. The letter “d” in the SMT column denotes the distortion limit of the SMT decoder. In both English-to-Japanese and Japanese-to-English directions, the combination of SSSS transfer and pre-ordering, i.e., P4, substantially improved the translation quality in terms of BLEU and RIBES scores. While both SSSS transfer alone (P2) and pre-ordering (P3) alone also led to drastic increases of RIBES scores, they achieved only marginal improvement of BLEU scores. Thus the substantial BLEU improvement derived by their combination suggests that SSSS transfer also contributes to improving the performance of pre-ordering.

4.5. Analysis

Experimental results confirm that translation quality can be improved significantly by using our SSSS transfer, irrespective of the existence of the pre-ordering process and translation directions. In this section, we first explain how our initial issues, i.e., extreme lengths and sub-language-specific structures in claim sentences, are resolved by SSSS transfer and pre-ordering. Subsequently, we provide and an in-depth analysis of the additional benefit of our SSSS transfer, i.e., making SMT inputs short. Finally, we discuss the different trends of the observed gains in the two translation directions.

Complementary contribution of SSSS transfer and pre-ordering: Figure 8 illustrates a typical sequence of example translations generated by the four configurations, P1 to P4, in our Japanese-to-English experiment. Throughout the figure, a labelled bracketing scheme is used to illustrate claim components. The contributions of SSSS transfer and pre-ordering are summarized as follows.

(1) **Contribution of SSSS transfer**: The order of components is not changed from the input Japanese sentence in P1. However, in P2, with the introduction of SSSS transfer, the components are well arranged in the order of English. The entire translation can be better understood by properly generating the transitional phrase “*comprising*”. Regarding the translation quality of each component, P1 and P2 do not seem significantly different. In contrast, we obtain a better translation for the second element in P4 than in P3. This is an evidence that SSSS transfer improves pre-ordering effectively.

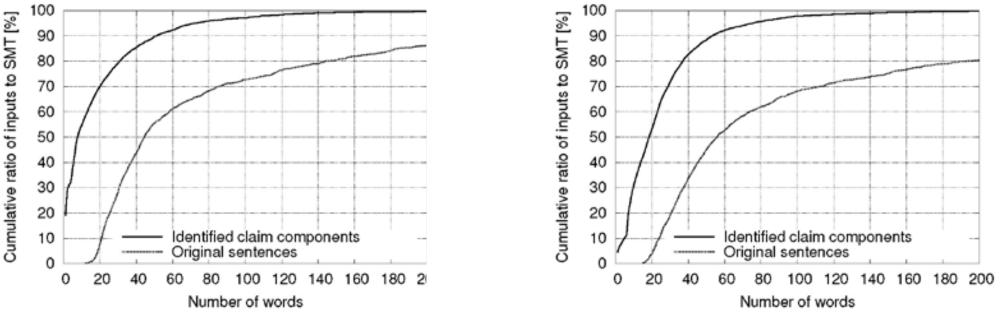
(2) **Contribution of pre-ordering**: As already demonstrated in the previous work, pre-ordering techniques are effective in generating translations with a reasonable word order in the target language. In fact, the words in P3 are better arranged than in P1: the word order is closer to that of the English reference. However, from the viewpoint of sentence structure, the components are not arranged well, and somehow the *preamble* is generated twice. Conversely, explicitly teaching the sentence-level structure through SSSS transfer, i.e., as in P4, suppresses such an undesirable error. Furthermore, dividing the input into shorter components, results in the words in each component being properly reordered.

In summary, SSSS transfer and pre-ordering complement each other in generating translations that are natural both structurally and component-wise.

Effects of shortening SMT inputs: As seen above, pre-ordering works better on components obtained through SSSS transfer rather than on the entire input sentence. To estimate the shortening effect of SSSS transfer, we compared the distributions of lengths of the processing unit of the succeeding steps, i.e., the entire sentence for P1 and automatically identified claim components in P2. Figure 9 shows the cumulative ratio of original sentences and identified claim components in English and Japanese, respectively. For example, a point (20,70) on the graph indicates that the sentences having lengths between 1 and 20 comprise 70% of all the sentences. As clearly illustrated, SSSS transfer considerably shortened the input to pre-ordering and SMT. This not only makes SMT easier, but also improves the accuracy of syntactic parsing and pre-ordering. Table 3 shows the sentence-wise accuracy of the English parser invoked by our pre-ordering module, calculated on the basis of 100 sentences sampled randomly from the test set. The parse tree of each sentence is manually checked for correctness. The shorter sentences show higher accuracy. This means that SSSS transfer does contribute to improving pre-ordering accuracy in addition to transferring sublanguage-specific sentence structure.

Input	[S [ELEM ベース管と、] [ELEM 前記ベース管を囲む少なくとも部分的に拡張されたスクリーンジャケットと] [TRAN を含む] [PREA 井戸スクリーンシステムであって、] [PURP 前記スクリーンジャケットが、前記ベース管上 に位置決めされる前に拡張され、前記スクリーンジャケットのフィルタ層が、前記スクリーンジャケットの拡張により、前記スクリーンジャケットの外側シュラウドに接触する。]]
Reference	[S [PREA A well screen system,] [TRAN comprising:] [ELEM a base pipe;] [ELEM and an at least partially expanded screen jacket surrounding the base pipe,] [PURP the screen jacket being expanded prior to being positioned on the base pipe, wherein a filter layer of the screen jacket contacts an outer shroud of the screen jacket due to expansion of the screen jacket.]]
P1 Baseline	[S [ELEM The base tube and] [ELEM the base tube is extended to at least partially surround a screen jacket] [PREA and a screen system well,] [PURP the screen jacket is positioned on the base tube before the expansion of the cooling jacket of the screen, the screen filter layer by an extension of the jacket of the cooling jacket of the screen outer shroud contact well screen.]]
P2 w/ SSSS transfer	[S [PREA Well screen system] [TRAN comprising:] [ELEM base tube;] [ELEM and the base tube is extended to at least partially surround a screen jacket,] [PURP wherein: the screen jacket is positioned on the base tube before extended to; and the screen jacket filter layer , said screen jacket by the expansion of the screen jacket outer shroud contact.]]
P3 w/ pre-ordering	[S [PREA Well screen system] [ELEM base tube,] [ELEM well screen system] [PREA including a screen jacket extended at least partially surrounding the base tube,] [PURP and the screen jacket, expansion before the positioning to the base tube, the screen jacket filter layer, the expansion of the screen jacket contacts the outer shroud of the screen jacket.]]
P4 Pipeline	[S [PREA Well screen system] [TRAN comprising:] [ELEM base tube;] [ELEM and at least partially extended screen jacket surrounding the base tube,] [PURP wherein: the screen jacket, expansion before the positioning to the base tube; and the screen jacket filter layer contacts the outer shroud of the screen jacket by the expansion of the screen jacket.]

Figure 8. Example Japanese-to-English translation: The bracket information in the input, reference, P1 and P3 are not determined automatically. We indicated them for explanation purpose only.



(a) English input (b) Japanese input

Figure 9. Cumulative ratio of inputs to SMT with respect to the number of words, with and without SSSS transfer

Table 3. Parsing accuracy of English parser used for English-to-Japanese pre-ordering

Number of words in sentence	Number of sampled sentences	Number of correctly parsed sentences	Sentence-wise accuracy
1-20	10	10	100%
21-40	35	32	91%
41-60	18	11	65%
61-80	5	2	40%
81-100	9	1	11%
101-120	5	0	0%
120-140	2	0	0%

Different trends for translation directions: In terms of RIBES score, pre-ordering improved Japanese-to-English translation substantially, while showing less improvement in the English-to-Japanese setting. We speculate that the difference lies in the difficulty of pre-ordering, and more specifically, in the difficulty of parsing sentences in the source language. As Japanese is a strictly head-final language, parsing sentences is easier than in English. Consequently, pre-ordering alone achieved almost the entire gain in the Japanese-to-English setting. Conversely, English sentences are much more difficult to parse than Japanese. As a result, the pre-ordering module can sometimes fail to bring the English word order close to that in Japanese. Nevertheless, as a result of SSSS transfer, which divides an input English sentence into shorter pieces, pre-ordering became more accurate, and the RIBES score was further improved.

5. Related Work

The quality of machine translation across distant languages has been improved as a result of the recent introduction of syntactic information into SMT (Collins et al., 2005; Quirk et al., 2005; Katz-Brown and Collins, 2008; Sudo et al., 2013; Hoshino et al., 2013; Cai et al., 2014; Goto et al., 2015). One of the promising avenues for further improvement appears to be the incorporation of sublanguage-specific information (Buchmann et al., 1984; Luckhardt, 1991). This is particularly important for translating formalized documents that tend to form sublanguage-specific document structures and sentence structures. In dealing with structures across close language pairs, an early study of sublanguage introduced the notion of *flat trees* which represents both source and target sentences using minimal depth structures for facilitating the transfer between the source and target structures (Buchmann et al., 1984). Much of the recent work relating to document and sentence structures between close languages focuses on structures centered on discourse connectives (Miltsakaki et al., 2005; Pitler and Nenkova, 2009; Meyer et al., 2011; Hajlaoui and Popescu-Belis, 2012; Meyer et al., 2012) and on resolving the ambiguity of discourse connectives connecting structural components.

Conversely, when dealing with structures across distant language pairs, a more comprehensive approach is more appropriate. A wide range of research has been conducted in this direction. A study by Marcu et al. (2000) proposed a method for improving Japanese-to-English translation by transforming the source structure generated by a rhetorical structure theory (RST) parser, to the corresponding target structure. Some work in this direction has been conducted in translations across distant languages, in which the source text is parsed using an RST parser, and translation rules are automatically extracted from the source and target pair (Kurohashi and Nagao, 1994; Wu and Fung, 2009; Joty et al., 2013; Tu et al., 2013). There are also approaches of simplifying long sentences by capturing the overall structure of a sentence, or a group of sentences. The skeleton-based approach (Mellebeek et al., 2006; Xiao, 2014) attempts to extract the key elements/structure (or *skeleton*) from the input sentence using a syntactic parser. The divide-and-translate approach (Shinhori et al., 2003; Sudo et al., 2010; Hung et al., 2012) also makes use of syntactically motivated features, such as phrases and clauses, for extracting subcomponents to be translated by SMT. There are also studies on pattern translation (Xia et al., 2004; Murakami et al., 2009; Murakami et al., 2013) and sentence segmentation (Xiong et al., 2009; Jin and Liu, 2010) for dealing with long input sentences with complex structures.

Our approach is similar to the above models in the sense that it incorporates structural information into SMT, but differs in that it uses sublanguage-specific sentence structures, rather than syntactically motivated structures. This results in significant improvement in translation quality for the claim sublanguage using only a handful of rules.

6. Conclusion

In this paper, we described a method for transferring sublanguage-specific sentence structure for English-to-Japanese and Japanese-to-English patent claim translations. The experimental results show that our proposed method, a combination of SSSS transfer and pre-ordering based on syntactic parsing, achieved five point gains in BLEU scores, in both English-to-Japanese and Japanese-to-English directions. In addition, a substantial gain of more than 30 points in RIBES scores was observed in both SMT settings, indicating a significant contribution of SSSS transfer. We achieved these results with only a handful of SCFG rules.

Our proposed method successfully improved the translation of patent claims with quality comparable to that of the other parts of patent documents. In our future work, we will concentrate on the translation of *independent claims* which are the longest and most complex of claim sentences.

References

Buchman, Beat, Susan Warwick and Patrick Shann. (1984). Design of a Machine Translation System for a Sublanguage. In *Proceedings of the 9th International Conference on Computational Linguistics*, pages 334-337.

Cai, Jingsheng, 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.

Cherry, Colin and George Foster. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427-436.

Chiang, David. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263-270.

Collins, Michael, Philipp Koehn and Ivona Kucerova. (2005). Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540.

de Gispert, Adrià , Gonzalo Iglesias and Bill Byrne. (2015). Fast and Accurate Preordering for SMT using Neural Networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 1012-1017.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou. (2011). Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of the 9th NII Test Collection for Information Resources (NTCIR) Conference*, pages 559-578.

Goto, Isao, Masao Utiyama and Eiichiro Sumita. (2012). Post-ordering by Parsing for Japanese-English Statistical Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 311-316.

Goto, Isao, Masao Utiyama, Eiichiro Sumita and Sadao Kurohashi. (2015). Preordering using a Target-Language Parser via Cross-Language Syntactic Projection for Statistical Machine Translation. In *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 14, No. 3, Article 13, pages 1-23.

Hajlaoui, Najeh and Andrei Popescu-Belis. (2012). Translating English Discourse Connectives into Arabic: A Corpus-based Analysis and an Evaluation Metric. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, pages 1-8.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark and Philipp Koehn. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 13-21.

Hoshino, Sho, Yusuke Miyao, Katsuhito Sudoh and Masaaki Nagata. (2013). Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 1062–1066.

Huang, Liang, Hao Zhang, Daniel Gildea and Kevin Knight. (2009). Binarization of Synchronous Context-Free Grammars. In *Journal of Computational Linguistics*, Volume 35 Issue 4, pages 559-595.

Hung, Bui Thanh, Nguyen Le Minh and Akira Shimazu. (2012). Divide and Translate Legal Text Sentence by Using its Logical Structure. In *Proceedings of 7th International Conference on Knowledge, Information and Creativity Support Systems*, pages 18-23.

Isozaki, Hideki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh and Hajime Tsukada. (2010a). 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.

Isozaki, Hideki, Katsuhito Sudoh, Hajime Tsukada and Kevin Duh. (2010b). Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251.

Joty, Shafiq, Giuseppe Carenini, Raymond Ng and Yashar Mehdad. (2013). Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 486–496.

Jin, Yaohong and Zhiying Liu. (2010). Improving Chinese-English Patent Machine Translation Using Sentence Segmentation. In *Proceedings 2010 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1-6.

Junczys-Dowmunt, Marcin and Arkadiusz Szał. (2010). SyMGiza++: A Tool for Parallel Computation of Symmetrized Word Alignment Models. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 397–401.

Katz-Brown, Jason and Michael Collins. (2008). Syntactic Reordering in Preprocessing for Japanese-English Translation: MIT System Description for NTCIR-7 Patent Translation Task. In *Proceedings of the NTCIR-7 Workshop Meeting*, pages 409-414.

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 Association for Computational Linguistics Demo and Poster Sessions*, pages 177–180.

Koehn, Philipp, Franz Josef Och and Daniel Marcu. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54.

Kurohashi, Sadao and Makoto Nagao. (1994). Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1123-1127.

Luckhardt, Heinz-Dirk. (1991). Sublanguages in Machine Translation. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, pages 306-308.

Marcu, Daniel, Lynn Carlson and Maki Watanabe. (2000). The Automatic Translation of Discourse Structures. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 9-17.

Mellebeek, Bart, Karolina Owczarzak, Declan Groves,Josef Van Genabith and Andy Way. (2006). A Syntactic Skeleton for Statistical Machine Translation. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, pages 195-202.

Meyer, Thomas, Andrei Popescu-Belis, Najeh Hajlaoui and Andrea Gesmundo. (2012). Machine Translation of Labeled Discourse Connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*.

Meyer, Thomas, Andrei Popescu-Belis, Sandrine Zufferey and Bruno Cartoni. (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, pages 194–203.

Miltsakaki, Eleni, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi and Bonnie Webber. (2005). Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*.

Murakami, Jin'ichi, Isamu Fujiwara and Masato Tokuhisa. (2013). Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT. In *Proceedings of the 10th NTCIR Conference*, pages 350-355.

Murakami, Jin'ichi, Masato Tokuhisa and Satoru Ikehara. (2009). Statistical Machine Translation adding Pattern-based Machine Translation in Chinese-English Translation. In *Proceedings of 6th International Workshop on Spoken Language Translation*, pages 107-112.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

Pitler, Emily and Ani Nenkova. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Short Papers, pages 13–16.

Petrov, Slav, Leon Barrett, Romain Thibaux and Dan Klein. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433-440.

Quirk, Chris, Arul Menezes and Colin Cherry. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279.

Shinhori, Akihiro, Manabu Okumura, Yuzo Marukawa and Makoto Iwayama. (2003). Patent Claim Processing for Readability - Structure Analysis and Term Explanation -. In *Proceedings of the Workshop on Patent Corpus Processing, Association for Computational Linguistics*, pages 56-65.

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 Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 418-427.

Sudoh, Katsuhito, Jun Suzuki, Hajime Tsukada, Masaaki Nagata, Sho Hoshino and Yusuke Miyao. (2013). NTT-NII Statistical Machine Translation for NTCIR-10 PatentMT. In *Proceedings of 9th NTCIR Conference 2013*, pages 294-300.

Tu, Mei, Yu Zhou and Chengqing Zong. (2013). A Novel Translation Framework Based on Rhetorical Structure Theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 370–374.

Utiyama, Masao and Hitoshi Isahara. (2007). A Japanese-English Patent Parallel Corpus. In *Proceedings of the Eleventh Machine Translation Summit*, pages 475-482.

The World Intellectual Property Organization (WIPO). (2014). WIPO Patent Drafting Manual. In *IP Assets Management Series*.

Wu, Dekai and Pascale Fung. (2009). Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Companion Volume: Short Papers, pages 13–16.

Xia, Fei and Michael McCord. (2004). Improving A Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514.

Xiao, Tong, Jingbo Zhu and Chunliang Zhang. (2014). A Hybrid Approach to Skeleton-based Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 563–568.

Xiong, Hao, Wenwen Xu, Haitao Mi, Yang Liu and Qun Liu. (2009). Sub-Sentence Division for Tree-based Machine Translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing Conference Short Papers*, pages 137–140.