NAIST-IS-DD1561018

Doctoral Dissertation

Improving Formal Document Translation Using Sublanguage-Specific Sentence Structure

Masaru Fuji

March 16, 2017

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Masaru Fuji

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)
Assistant Professor Hiroshi Noji	(Co-supervisor)

Improving Formal Document Translation Using Sublanguage-Specific Sentence Structure[*]

Masaru Fuji

Abstract

Advances in reordering techniques based on syntactic parsing, with growing volumes of parallel corpora available, have brought about significant improvements in the performance of statistical machine translation (SMT) for translating across distant language pairs. However, formal documents such as patent, law, and operations manual documents still pose difficulties for SMT owing to the extreme sentence lengths and characteristic sentence structures.

These formal documents are often regarded to form sublanguages because they have their own characteristic writing styles. As the sentences comprising these formal documents are often very long and complex, a characteristic writing style has been devised for each sublanguage in daily practice among writers so that readers can easily comprehend the documents. This thesis describes methods for incorporating features specific to each sublanguage into the translation mechanism to recognize the sentence structure correctly and improve translation quality. The correct recognition of sentence structure is particularly important for translating long sentences between distant language pairs, because not only the syntactic order but also the sentence structure is different between these language pairs.

Section 3 describes translation experiments for the patent claim sublanguage, where sentences are extremely long but have very high regularity in the writing style. As the writing style of this patent claim sublanguage is consistent, I chose to handcraft rules for detecting sentence segments and performed translation experiments using these detected sentence segments. The proposed method resulted

[*] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1561018, March 16, 2017.

in substantial improvement in translation quality between distant language pairs, such as English-to-Japanese and Chinese-to-Japanese translations.

Section 4 describes a method for capturing the sentence structure with moderate regularity of writing style and higher occurrence frequency compared with patent claim sentences. Because there exists some variation in the writing style of these documents, I chose to automatically recognize sentence structures. I proposed and constructed automatic reordering of segments in translating from the source to the target languages, which I call *global reordering*. Substantial improvement in translation quality was observed by incorporating global reordering along with conventional reordering especially for Japanese-to-English translation.

To summarize, my proposed method for translating formal documents between distant language pairs covers the range of sentences with very high regularity as well as sentences with moderate regularity.

Keywords:

sentence structure, sublanguage, statistical machine translation, formal documents

サブ言語に特有の文構造を用いたフォーマル文書の翻訳精度向上[*]

富士　秀

内容梗概

日英間や日中間のような、語順の大きく異なる言語間における機械翻訳の精度は、利用可能なコーパスデータ量の増加に加え、構文解析にもとづく単語並べ替え技術の進展によって大きく向上した。しかし、法令文、特許文、技術マニュアルなどのフォーマルな文書は、内容が複雑でかつ長文で記述される場合が多いため、他の対象文書と比べても依然として翻訳が困難である。

これらフォーマルな文書は、分野や文種によって特有の表現形態を持っており、このためそれぞれが独自の「サブ言語」を構成しているとみなすことができる。これら文書は内容が複雑でかつ長文である反面、読者の理解を助けるために、サブ言語毎に特有の表現が発達し、執筆者は慣習的にこの表現をを用いて文の構造をわかりやすく記述する場合が多い。このことから、サブ言語に特有の表現に着目した処理を機械翻訳に導入することで、文の構造を正確に把握して、翻訳精度を向上させることが考えられる。特に、語順の異なる言語間では、長文において、構文的な順番が異なるだけではなく文の構造も異なるため、正確な文構造の把握が重要となる。本研究では、サブ言語を構成するようなフォーマルな文書の翻訳において、サブ言語に特有の表現を用いて入力文の構造を把握し、その構造を単位とした処理を導入することで高精度な翻訳を実現する。

第3章では、様々なサブ言語の中でも特に1文が特異的に長く、表現の規則性が高いという特徴を持つ、特許請求項文のサブ言語を対象とした翻訳実験について述べる。特許請求項では、表現の規則性が極めて高いため、文の構造を認識するための規則を人手で記述して、これをもとに統計的機械翻訳を動作させた。そ

[*] 奈良先端科学技術大学院大学 情報科学研究科 博士論文, NAIST-IS-DD1561018, 2017年3月16日.

の結果、英日・日英・中日・日中という語順の大きく異なる言語間の翻訳において、極めて大きな翻訳精度の向上が認められた。

次の第 4 章では、表現の規則性はそれほど高くないが、適用範囲の幅広い特有表現を持つ文が存在するサブ言語をを対象とした実験を行った。ここでは、特有表現にある程度の揺れがあるため、人手による規則作成ではなく、大量のデータからの統計学習による方法を用いた。語順の異なる言語対を対象とし、原言語と目的言語間の構造部品の自動的な並べ替えを実現した。この構造部品の自動的な並べ替えを「グローバルな並べ替え」と呼んでいる。特に、従来の構文解析による並べ替えの精度が低い日英翻訳において、グローバルな並べ替えと従来の並べ替えの併用によって大きな翻訳精度の改善がみられた。

以上のようにして本研究では、特有表現の規則性の高さが異なるサブ言語を対象として、サブ言語に特有の文構造を導入した手法の有効性を確認することができた。

キーワード

文構造、サブ言語、統計的機械翻訳、フォーマル文書

Acknowledgements

本論文は、私が富士通研究所から情報通信研究機構に出向し、また同時に奈良先端科学技術大学院大学に在学している間に行った研究成果をまとめたものです。

情報科学研究科の松本裕治教授には、指導教官として本研究の実施の機会を与えていただき、その遂行に当たってご指導をいただいたことに深く感謝いたします。また、お忙しい中審査をお引き受けいただいた審査員の先生方に感謝いたします。松本研究室のスタッフおよび学生の皆様と研究について有意義な意見交換をさせていただきました。

情報通信研究機構では、先進的音声翻訳研究開発推進センターの隅田英一郎副センター長、内山将夫研究マネージャー、藤田篤主任研究員にお導きいただき、また機構内の多くの方々にお世話になりました。

さらには、奈良先端科学技術大学院大学への就学の機会および支援をいただいた富士通研究所に感謝するとともに、長年にわたって研究活動を支援していただいた富士通研究所および富士通の関係者のみなさまに感謝いたします。

最後に、日頃から私を応援してくれている私の家族と多くの方々、そしてすべてを受け入れてともに歩んでくれる妻に感謝します。

Contents

List of Figures

List of Tables

1. Introduction

Ever since the advent of computers, continuous research attempts have been made to develop machine translation for assisting or even replacing human translation. Recent advances in translation technology along with the accumulation of text data for training machine translation has resulted in remarkable improvement both in the speed and quality of machine translation. Although the speed and quality of machine translation have proved useful for browsing and understanding the gist of foreign sentences in certain situations, they are still not regarded as being useful in most translation processes involved in the translation services industry. The situation is much worse in the case of translation between a language pair having largely different word orders, such as between English and Japanese and between Chinese and Japanese.

Many of the documents to be translated in the translation services industry are said to form sublanguages in the sense that the vocabulary, sentence structure, and expressions used in each translation domain and applications are noticeably different from those of general documents. It is this sublanguage that makes machine translation difficult. On the other hand, the key to improving machine translation for the translation services industry is to devise methods for incorporating information specific to the sublanguages into the translation mechanism.

Past studies suggest the necessity of introducing mechanisms for dealing with sublanguage-specific features to improve machine translation for the translation industry (Buchmann, Warwick, and Shann 1984; Luckhardt 1991) [5, 48]. Among the various aspects of sublanguage-specific features, this research focuses on the sentence structure aspect of sublanguages, because this is the aspect that is currently not sufficiently dealt with for adapting general purpose machine translation to meet the translation quality required by the translation services industry.

1.1 Research target

In light of the above mentioned background, my research is focused on the development of machine translation with the aim of improving the translation quality in the following ways:

Translation of formal documents forming sublanguages I focus on the translation method for formal documents forming sublanguages. The documents forming sublanguages often have the characteristics described in Section 2.4.

Incorporation of sublanguage-specific sentence structures (SSSSs) Among the characteristics possessed by sublanguage documents, my focus is on the SSSSs. The term *sentence structure* signifies the semantic structure within each sentence, rather than the semantic structure extending over the series of sentences within a document.

Translation between distant languages The challenge of SSSS often poses difficulties for machine translation between distant languages possessing largely different word orders. The research focuses on SSSS in translating between distant languages.

1.2 Contributions

This dissertation addresses the issues stated in the research target and makes the following contributions:

1. The effectiveness of incorporating SSSSs into the mechanism of statistical machine translation (SMT) by conducting experiments is shown.

2. The effectiveness of incorporating handcrafted rules for recognizing SSSSs when high regularity in the writing style is present in the formal document in question is shown.

3. The effectiveness of incorporating an automatic detection method for recognizing SSSSs when moderate regularity in the writing style is present in the formal document in question is shown.

1.3 Thesis outline

This dissertation is organized as follows:

- Section 2 describes the background information leading to the motivation of this research. It describes aspects such as the translation services industry, formal documents, machine translation in the industry and SSSSs.

- Section 3 describes translation experiments for patent claim sentences which are extremely long but have exceptionally high regularity in the writing style. The proposed method resulted in substantial improvement in translation quality because the handcrafted rules perform accurately to detect sentence segments, the detected segments are reordered in the order appropriate to the target language, and the original input sentence is effectively shortened as a result of segment detection.

- Section 4 describes a method for capturing the sentence structure with moderate regularity of writing style and higher occurrence frequency compared with patent claim sentences. A substantial improvement in translation quality was observed by incorporating global reordering along with conventional reordering.

2. Preliminaries

This section describes the background information related to the motivation of my research, that is, the incorporation of features specific to the target sublanguage into the translation mechanism to recognize the sentence structure correctly and improve translation quality in formal documents forming a sublanguage.

2.1 Translation services industry

The translation services industry has a large market size, and machine translation is expected to make contribution. The outcome of my research is expected to be useful in this industry.

2.1.1 Definitions

The term "translation services industry" usually refers to an industry that includes companies which (i) translate written material and (ii) offer interpreting services from one language to another. My research mainly concerns the former industry segment, namely, translation of written material. Translation of written material is further categorized into translation of paper-based documents and translation of digital documentation, and the latter includes the "localization"[1] industry. Localization constitutes a large portion of the translation services industry and its definition is given by Wikipedia as follows:

> **Language localization** *is the process of adapting a product that has been previously translated into multiple languages to a specific country or region (from Latin* locus *(place) and the English term* locale*, "a place where something happens or is set"). It is the second phase of a larger process of product translation and cultural adaptation (for specific countries, regions or groups) to account for differences in distinct markets, a process known as internationalization and localization.*

[1] Language localization https://en.wikipedia.org/wiki/Language_localisation

2.1.2 Market size

According to studies by Common Sense Advisory (DePalma et al. 2016) [13], a translation industry think tank, the size of the translation industry is estimated to be $33.5 billion in 2012. The market is expected to continue growing to reach $37 billion in 2018, according to IBISWorld (IBISWorld 2016) [31]. The United States represents the largest single market, and Europe is a close second. Asia is the largest growth area. Currently, business is generated from both the government and private industries. According to the U.S. Bureau of Statistics, the translation industry is expected to grow by 42% between 2010 and 2020. The most important reason for this growth is globalization. The market size in Japan is estimated to be ￥25.8 million in 2015 according to studies by Japan Translation Federation.

2.1.3 Language pairs in the Japanese translation industry

Figure 1 shows the language pairs involving the Japanese language in the Japanese translation services industry. Among the language pairs involving Japanese in the Japanese translation services industry, 90% is in/from English, 4% in/from Chinese, 2% in/from German, 2% in/from other Asian languages and another 2% in/from other European languages. The category “other European languages” includes French, Spanish, Italian, Portuguese and Russian. The category “other Asian languages” includes Korean, Vietnamese, Thai and Indonesian.

Considering that the only major language that is linguistically close to Japanese is Korean, most of the languages in and from Japanese are distant languages in the Japanese translation services industry.

2.1.4 Future prospects of the industry

The industry of translation services has been exhibiting steady growth in spite of the worldwide economic instability since 2008. This is due to the proven correlation between the industry market size and the volume of Internet content and online publications, which has grown steadily irrespective of the economic situation. The increase in migration into Europe and United States in turn increases the need for multilingual translation. There is also an increasing trend of globalization of companies worldwide, where companies sell their products

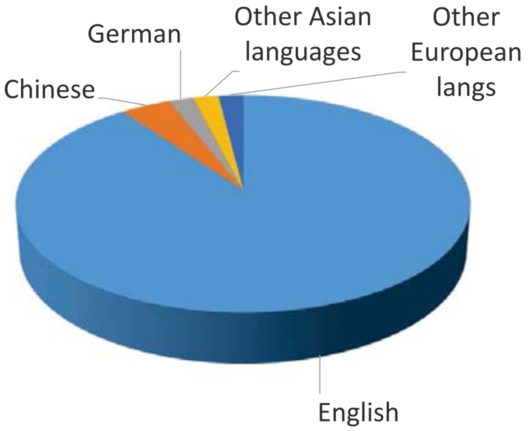


Figure 1. Language pairs in the Japanese translation industry where "English" represents both English-to-Japanese and Japanese-to-English translation pairs, etc.

overseas and run multilingual marketing. Despite the downward cost pressure in the industry, the increase in demand prevails over the downward trend of the price. The upward trend of the translation volume along with the downward trend of the translation cost calls for the deployment of machine translation.

2.2 Formal documents in translation services industry

The vast majority of documents handled in the translation services industry are formal documents. Figure 2 shows that in the Japanese translation services industry, 90% of the documents are formal documents, whereas only 7% are literature and novels, and 3% are visual media translation. The formal documents handled in the industry can be classified into two categories: domains and text types. Some of the text types are common to many of the domains, whereas others are found only in specific domains.

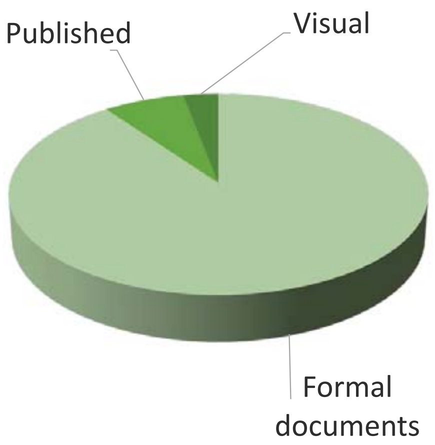


Figure 2. Documents in the translation industry

2.2.1 Domains

The main domains comprising the industry currently are (i) technical and IT, (ii) pharmaceutical and medical, and (iii) financial. Each of these domains has the following subcategories:

Technical and IT This domain includes information technology, such as informatics, communication and networks. There is a huge demand for translation in this domain especially, in “localization” of documents. Because words and expressions appear repeatedly and large volumes of documents have to be translated in short periods of time, computer-assisted tools are readily introduced in this domain. The following are the domains in which translation memory tools are most actively introduced:

- Communication technology
- Semiconductor devices
- Information science
- Games
- Automobiles
- Mechanical engineering
- IT and ICT

- Electronic and electrical engineering
- Aeronautical engineering
- Train transportations
- Sciences
- Energy and environmental science
- Construction

Pharmaceutical, medical and bioscience The characteristic of this domain is that the translation volume does not vary with economic fluctuations, resulting in a steady market. As translation in this domain requires specialized knowledge, most of the translators have backgrounds in pharmacy, biochemistry and biology.

- Medical equipment
- Pharmaceutical
- Biotechnology
- Cosmetics
- Foods
- Agriculture
- Petroleum

Financial Though the translation volume here is smaller than in the former two domains, this is a rapidly growing domain.

- Economics and finance
- Banking
- Life assurance and insurance
- Stocks
- Business activities
- Governmental

2.2.2 Text types

Text type is the second grouping axis for formal documents handled in the translation services industry. While some of the text types exist only for limited domains, the following text types are fairly common to many of the domains in the industry. Text type is closely related to *sublanguage* as discussed in later subsections.

- Contract documents
- Patent documents
- Regulations
- Compliance documents
- Press releases
- Websites
- Reports

2.3 Machine translation as a translation assisting tool

As described in Section 2.1.4, the upward trend of the translation volume along with the downward trend of the translation cost calls for computerized translation assisting tools.

The computerized translation assisting tool that is widely known in the industry as "translation memory" has proved to improve efficiency and reduce the cost of translation in some domains and text types of formal documents in the translation of both close and distant language pairs. The Wikipedia description of translation memory[2] is as follows.

[2] Translation memory https://en.wikipedia.org/wiki/Translation_memory

A **translation memory (TM)** *is a database that stores "segments," which can be sentences, paragraphs or sentence-like units (headings, titles or elements in a list) that have previously been translated, in order to aid human translators. The translation memory stores the source text and its corresponding translation in language pairs called " translation units." Individual words are handled by terminology bases and are not within the domain of TM.*

According to Wikipedia, translation memory is widely used in companies producing multilingual documentation and a survey reveals that 82.5% out of 874 replies confirmed the use of translation memory. Translation memory is effective for distant language pairs, including Japanese-to-English and Japanese-to-Chinese translation, as long as the terms and expressions occur repeatedly. Although translation memory is a tool that is used in the industry, the improvement in efficiency is steady but limited because the application of translation memory still requires manual intervention by human translation.

The technology of machine translation dates back to the early stages of the development of computers back in the 1950s. This technology is expected to produce further improvements in efficiency compared to translation memory as it generates translated sentences automatically, which is expected to substantially reduce the cost due to utilizing humans in the translation process. Recent advances in the translating mechanism along with hardware improvements and the increased volume of the corpus, have raised the speed and quality of machine translation to a level that is sufficient for browsing through foreign documents in one's own language. However, the translation quality still does not meet the high quality standard required by the translation services industry. Even the machine translated texts in translation between close language pairs requires further development to make it useful for human translators and translation firms. The situation is even worse for distant language pairs, where further improvement in translation quality is necessary. Further research into the technology is expected to meet the requirements of this growing industry.

2.4 Sublanguage for machine translation

Past studies suggest the need to introduce mechanisms for dealing with sublanguage-specific features to improve machine translation for the translation industry (Buchmann, Warwick, and Shann 1984; Luckhardt 1991) [5, 48]. The characteristics of a sublangauge that are relevant to machine translation are listed by Lehrberger (Lehrberger 1982) [46] and further described by Ananiadou (Ananiadou 1990) [2] as follows:

(i) Limited subject matter This means text with domain specific knowledge, e.g. immunology and computer maintenance.

(ii) Lexical, syntactic and semantic restrictions Typical examples of lexical restrictions appear in technical terms that have precise meaning, e.g. "floating point", "eyebolts", "perioxide" etc. Examples of syntactic restrictions include the absence of direct questions in the case of professional translation documents. A typical example of semantic restrictions is that the term "cable" in a particular sublanguage only occurs as a noun and not as a verb.

(iii) Deviant rules of grammar Typical examples of deviant grammar rules are the sentences "the patient presented with influenza" and "the patient presented to the doctor with influenza." This usage of the term "present" in these sentences is not acceptable in general language but is used in common practice in the medical domains.

(iv) High frequency of certain constructions A typical example of this aspect is the high frequency of imperatives, e.g. "check" and "add" in operations manuals.

(v) Particular text structure Text structure is often referred to as document structure. Sublanguage documents often form characteristic text structures in the sense that text belonging to a certain sublanguage tends to possess certain components in certain sequences.

(vi) Use of special symbols A typical example of the use of special symbols is the placement of a colon after a transitional phrase, e.g. "An information processing apparatus comprising: a recording media ... ."

Recent advances in language modeling, such as the language model of statistical machine translation and the sequence-to-sequence model of neural translation, have overcome some of these aspects to a certain extent. In particular, these language model approaches are effective in the lexical selection and the selection of short-range expressions. The characteristics among the above list that have been resolved to a certain extent are (i) *limited subject matter* and (ii) *lexical, syntactic, and semantic restrictions*. The characteristic of (iii) *deviant rules of grammar* is also considerably improved by language models.

The characteristics that are left unsolved by recent advances in machine translation technology are (iv) *high frequency of certain constructions* and (v) *particular text structure*.

2.5 Sublanguage-specific sentence structures (SSSSs)

As described above, I consider the most effective characteristics of sublanguage-specific features left undeveloped to be *high frequency of certain constructions* and *particular text structure*. Among these two characteristics, *particular text structure* requires techniques involving contexts, which will be important in machine translation technology in the near future, but I leave it to future research. For the present I concentrate on sentence-wise technology. Hence, I chose *high frequency of certain constructions* to be my research target among the effective characteristics in sublanguage-specific features.

There are a number of studies dealing with the sentence structures specific to a sublanguage. There is a line of research dealing with the sentence structure specific to the legal document sublanguage (Gotti et al. 2008; Farzindar et al. 2009; Bach et al. 2010; Hung et al. 2012) [23, 16, 3, 4, 30]. There is another line of research dealing with the sentence structure specific to the patent claim sublanguage (Shinmori et al., 2003; Verberne et al. 2010; Fuji et al. 2015; Hu et al. 2016) [65, 18, 70, 28].

2.6 Terminology used in the thesis

The following is a list of terminology used throughout the thesis:

Segments Semantic units comprising a sentence. This is a unit that is used in contrast to syntactic units such as syntactic phrases and clauses, but short segments may coincide with short syntactic units.

Sentence structure The structure within a sentence, made up of segments as defined above. This may be referred to as *intra-sentential structure* in some related works. The antonym of sentence structure, or intra-sentential structure is *inter-sentential structure* which refers to the relationship among a group of sentences.

Sublanguage-specific sentence structure (SSSS) The structure within a sentence that is specific to the given sublanguage.

3. Patent Claim Translation Using Sublanguage-Specific Sentence Structure

3.1 Introduction

Advances in reordering techniques based on syntactic parsing (Isozaki et al., 2010; de Gispert et al., 2015)[34, 11], with growing volumes of parallel patent corpora available, have brought about significant improvements in the performance of statistical machine translation (SMT) for translating patent documents across distant language pairs (Goto et al., 2015)[22]. 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)[5, 48] 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 parts irrespective of language, such as those illustrated in Figure 3 (a), (b) and (c). 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. Hereinafter, I will call the parts constituting a claim sentence, *sentence segments*, or simply *segments*. I 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. The method based on this sublanguage-specific sentence structure (henceforth, *SSSS*) has two major effects. (1) Pre-ordering and SMT are applied for each sentence segment, 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. I manually extracted a set of language independent claim segments. Moreover, using these segments, I constructed a set of synchronous rules for English and Japanese to transfer the SSSS in the source language to the target language.

An experiment demonstrates that my proposed method significantly improves the translation quality in terms of RIBES scores (Isozaki et al., 2010)[32] by over 25 points, in all of the four translation directions, i.e., English-to-Japanese, Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese directions. Alongside the improvement in RIBES scores, improvements of approximately five points in BLEU scores (Papineni et al., 2002)[60] are observed for English-to-Japanese and Japanese-to-English pairs, and that of 1.5 points are observed for Chinese-to-Japanese and Japanese-to-Chinese directions. The number of common segment units required to express English, Chinese and Japanese claim sentence structures is only five, and the number of synchronous rules written with these segments is fewer than ten in all the four translation directions. The substantial gains are obtained with a very small set of segment units and rules.

3.2 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)[10, 63, 39, 67, 27, 6, 22]. To introduce syntactic information to translation of formalized documents forming a sublanguage, it appears necessary to incorporate sublanguage-specific information (Buchmann et al., 1984; Luckhardt, 1991)[5, 48]. Since sublanguage-specific sentences appearing in stylized documents tend to be very long and have characteristic sentence structures, my task is to appropriately deal with sublanguage-specific structures for long input sentences.

Much of the recent work relating to the translation of 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

Segments		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.

(a) English claim

Segments		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。

(b) Japanese claim

Segments		Example strings
Preamble		一種裝置
Transitional phrase		包括：
Body	Element	鉛筆和；
	Element	安裝在所述鉛筆的橡皮擦以及；
	Element	安裝在所述鉛筆的燈泡。

(c) Chinese claim

Figure 3. Example of English, Japanese and Chinese patent claims

and Popescu-Belis, 2012; Meyer et al., 2012)[53, 62, 52, 24, 51] and on resolving the ambiguity of discourse connectives connecting structural segments. Conversely, when dealing with sentence structures across distant language pairs, it is insufficient just to deal with discourse connectives, but to capture the sentence structure of the input sentence and transfer it to the target structure.

A wide range of research has been conducted in this direction. A study by Marcu et al. [49] 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) [44, 72, 37, 68]. 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) [50, 74] attempts to extract the key elements/structure (or skeleton) from the input sentence using a syntactic parser. The divide-and-translate approach (Shinmori et al., 2003; Sudo et al., 2010; Hung et al., 2012) [65, 66, 30] also makes use of syntactically motivated features, such as phrases and clauses, for extracting sub-segments to be translated by SMT.

There are also studies on pattern translation (Xia et al., 2004; Murakami et al., 2009; Murakami et al., 2013) [73, 55, 54] and sentence segmentation (Roh et al. 2008; Xiong et al., 2009; Jin and Liu, 2010) [64, 75, 36] 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.

3.3 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) [71], the fundamental structure of an English claim is that it is a single sentence consisting of three segments:

$$S \rightarrow PREA\ \ TRAN\ \ BODY \tag{1}$$

where S denotes the claim sentence, PREA the preamble, TRAN the transitional phrase and BODY the body. Figure 3 illustrates a typical example of PREA, TRAN and BODY appearing in English, Japanese and Chinese claim sentences. The preamble is an introductory phrase that identifies the category of the invention, the body is the main segment of the claim that describes the elements or purposes of the invention, and the transitional phrase is the segment that connects the preamble and the body. In actual claim documents, the body is expressed either as a series of elements or purposes. Elements are the segments constituting the invention and purposes are the segments expressing the purposes the the invention. Hereinafter, I will denote element as ELEM, and purpose as PURP.

Figure 3 (a) 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 segment comprising the invention. Figure 3 (b) shows the structure of a Japanese claim sentence corresponding to the English claim sentence shown in Figure 3 (a), and Figure 3 (c) shows the structure of a Chinese claim sentence corresponding to the English claim sentence. Note that the sets of segments comprising the claims in the two languages are identical, although the order of segments is different in the three languages.

My proposed method is described as follows. A manual analysis revealed that a claim consists of a fixed set of segments and the set is common to the three languages. It was also found that there are strict generation rules in each language. For example, the English patent claim sentence in Figure 3 (a) is represented by the set of rules in Figure 4 (a). The symbol “＋” denotes a non-null list of the preceding segments. The corresponding Japanese sentence is represented by another set of rules comprising the same segments, as shown in Figure 4 (b).

Having observed a strong regularity in the structure of patent claim sentences across languages, I represent the structural transfer in the form of synchronous context-free grammar (SCFG). For example, I derive the SCFG rules in Figure 4 (c) by connecting the corresponding rules in Figure 4 (a) and (b), where the numeric indices indicate correspondences between non-terminals in both constituent trees. I handcrafted a set of SCFG rules for translating patent claim sentences. The details of the process are presented in Section 3.4.1.

Figure 5 illustrates an example of a bilingual Enligh-to-Japanese claim sen-

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

(a) Generation rules for an English claim sentence

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

(b) Generation rules for a Japanese claim sentence

$$\begin{array}{lll} \text{S} & \rightarrow & \langle PREA_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}},\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ PREA_{\textcircled{1}} \rangle \\ \text{BODY} & \rightarrow & \langle ELEM+,\ ELEM+ \rangle \\ \text{TRAN} & \rightarrow & \langle\ \text{“comprising:”},\ \text{“備える”} \rangle \end{array}$$

(c) SCFG rules derived from English and Japanese generation rules

Figure 4. English and Japanese generation rules, and SCFG rules derived from these generation rules

tence pair. Here, PREA, TRAN and BODY represent the segments of these bilingual sentences, where the order of these segments differs between English and Japanese claim sentences. Hence the figure shows that simple composition of a claim sentence into segments is inadequate, but the segments have to be appropriately reordered to suit the structure of the target language.

[$_{PREA}$ An apparatus] [$_{TRAN}$ comprising:] [$_{BODY}$ a pencil; an eraser attached to the pencil; and a light attached to the pencil.]
[$_{BODY}$ 鉛筆と；鉛筆に取り付けられた消しゴムと；鉛筆に取り付けられたライトと；] [$_{TRAN}$ を備える] [$_{PREA}$ 装置]

Figure 5. An example English-to-Japanese claim pair

3.4 Pipeline for patent claim translation

While patent claim sentences have a distinctive structure, their segments, 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. I 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, my method translates it through the following three-step pipeline (Figure 6).

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 segments 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 segment 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 segment is translated by an SMT system, and the translated segments joined up to form a sentence, with words conjugated and conjunctions added as necessary. Again, as a result of Step 1, shorter segments are input that are easier to translate.

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

3.4.1 SSSS transfer

As described in Section 3.1, 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.

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 segment 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 segment translated by English-to-Japanese SMT

Figure 6. Overview of translation pipeline

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. The rules are created in the following manner. First, the English, Chinese and Japanese claim sentences were analyzed manually in my development set (described in Section 3.5.1) and found that each claim sentence is composed of a fixed set of segments and that the set is common to the three languages. The set of segments U identified is as follows:

$$U = \{PREA, TRAN, BODY, ELEM, PURP\}, \tag{2}$$

where these five items are explained in the previous section, i.e., preamble, transitional phrase, body, element and purpose.

I then constructed a set of generation rules for English and Japanese claims using U as a set of non-terminal symbols, and I handcrafted a set of 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 7 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. Likewise, I constructed a set of generation rules for Chinese claims using U as a set of non-terminal symbols, and obtained sets of SCFG rules for Chinese-to-Japanese and Japanese-to-Chinese translations. See Section A in Appendix for Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese directions. The number of rules for English-to-Japanese translation is eight, that for Japanese-to-English is ten, that for Chinese-to-Japanese is six and that for Japanese-to-Chinese is ten.

The sentences used for constructing the generation rules were taken from the patent claim sentence pairs as described in Section 3.5.1. 500 patent claim sentence pairs were extracted randomly from the training data. However, due to the high regularity of patent claim sublanguage, only about first 50 sentences were sufficient for finding most major rules and another 50 sentences were sufficient for collecting most of the necessary rules. Another 100 sentences were used to verify the rules constructed using the first 100 sentences.

Figure 8 shows an example bilingual sentence pair that may be under subject to SCFG rule R_{ej2} in the SCFG rules set for English-to-Japanese translation. The figure shows that the segments appearing in the order of PREA, TRAN, BODY, TRAN and BODY in the input English claim sentence are reordered to BODY, TRAN, BODY, TRAN and PREA, and as a result, the target Japanese claim sentence has an appropriate sentence structure.

In the SCFG rule set I prepared for my experiment, I designed each of the rules in the rule set to be deterministic, except for the terminal symbols where ambiguity in matching may occur. In the actual implementation of the SCFG rules, I used regular expressions for obtaining a unique match for a terminal symbol, so that the matching process for terminal symbols is also deterministic. For ensuring a unique match for a terminal symbol, I employed the head-directionality of the input language. For example, for head-initial languages such as English and Chinese, I selected the match for TRAN occurring at the position closest to the end of the sentence, while for head-final languages such as Japanese, I selected the match for TRAN occurring at the position closest to the beginning of the sentence.

For example, to analyze input sentences containing more than one occurrence of the string “ comprising: ” I prepared a regular expression to match the first occurrence. This heuristic rule correctly matches the claim string in most cases since writers of English patent claims usually keep in mind to use TRAN to appear toward the beginning of the sentence in practical situations. Figure 9 shows an example of English and Japanese regular expressions written in Perl-like notation, where “+” denotes the longest matching and “+?” denotes the shortest matching.

3.4.2 Pre-ordering

Most of the current pre-ordering techniques may be classified either into phrase structure-based (Isozaki et al. 2010b; Goto et al. 2015; Hoshino, Miyao, Sudoh, Hayashi, and Nagata 2015) [34, 22, 26], or into dependency-based (Yang, Li, Zhang, and Yu 2012; Lerner and Petrov 2013; Jehl, de Gispert, Hopkins, and Byrne 2014; de Gispert et al. 2015) [77, 47, 35, 11] techniques. While my proposed method may employ either of the types, I take phrase structure-based technique as an example in the following explanation.

ID	SCFG rules		
R_{ej1}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}},$ $BODY_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ PREA_{\textcircled{1}}\rangle$
R_{ej2}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{4}}\ BODY_{\textcircled{5}},$ $BODY_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{5}}\ TRAN_{\textcircled{4}}\ PREA_{\textcircled{1}}\rangle$
R_{ej3}	BODY	$\rightarrow$	$\langle ELEM+,\ ELEM+\rangle$
R_{ej4}	BODY	$\rightarrow$	$\langle PURP+,\ PURP+\rangle$
R_{ej5}	TRAN	$\rightarrow$	⟨ “comprising:”, “備えることを特徴とする”⟩
R_{ej6}	TRAN	$\rightarrow$	⟨ “including:”, “備えることを特徴とする”⟩
R_{ej7}	TRAN	$\rightarrow$	⟨ “having:”, “備えることを特徴とする”⟩
R_{ej8}	TRAN	$\rightarrow$	⟨ “wherein:”, “ことを特徴とする”⟩

Figure 7. SCFG rule set for English-to-Japanese translation

For example, when “He likes apples.” is inputted into the English-to-Japanese translation system, it is first parsed as shown in Figure 10. Second, the nodes in the parse tree are reordered using a classifier. In the case of Figure 10, 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) . (.)”

There is a variety of techniques for deciding whether to swap a binary node, such as a technique based on handcrafted rules (Isozaki et al. 2010) [34], and a technique based on statistical method where the decision is made so as to minimize the difference in the word orders between the source sentence and the target sentence by minimizing the value of Kendall's τ as a rank correlation coefficient (Goto et al. 2015; Hoshino et al. 2015) [22, 26]. I adopted the latter statistical technique in my experiment. The detailed setting I used in my experiment are given in Section 3.5.3. However, my proposed method does not depend on any particular pre-ordering methodology.

As described in 3.1, another major issue in patent claim translation is that the extreme lengths cause syntactic parsers to fail with consequent low reordering ac-

[$_{S}$ [$_{PREA}$ An energy management system] [$_{TRAN}$ comprising:] [$_{BODY}$ [$_{ELEM}$ (a) a helmet shell having a bottom edge;] [$_{ELEM}$ (b) a plurality of bell-shaped pockets situated on an inside surface of the helmet shell, each of the bell-shaped pockets having a bottom surface; and [$_{ELEM}$ (c) a bladder positioned inside of each bell-shaped pocket;]] [$_{TRAN}$ wherein] [$_{BODY}$ [$_{PURP}$ the bottom surface of each bell-shaped pocket is configured to allow the bladder to extend beyond the bottom surface of the pocket and beyond the bottom edge of the helmet upon impact.]]]

(a) English claim sentence

[$_{S}$ [$_{BODY}$ [$_{ELEM}$ （ａ）底縁を有するヘルメットシェルと、] [$_{ELEM}$ （ｂ）前記ヘルメットシェルの内面に位置する複数の釣鐘形ポケットであって、各々が底面を有する釣鐘形ポケットと、] [$_{ELEM}$ （ｃ）各釣鐘形ポケットの内部に位置決めされるブラダーとを]] [$_{TRAN}$ 備えることを特徴とする] [$_{BODY}$ [$_{PURP}$ 各釣鐘形ポケットの前記底面が、衝撃時に前記ブラダーを前記ポケットの前記底面を超え、かつ前記ヘルメットの前記底縁を超えて延在させるように構成される、]] [$_{PREA}$ エネルギー制御装置。]]

(b) Japanese claim sentence

Figure 8. Example English-to-Japanese claim sentence pair corresponding to SCFG rule R_{ej2}

```
($prea, $tran, $body) = /^(.+?)(comprising)(.+)$/
($body, $tran, $prea) = /^(.+)(備えることを特徴とする)(.+?)$/
```

Figure 9. Example of perl-like regular expression for English and Japanese

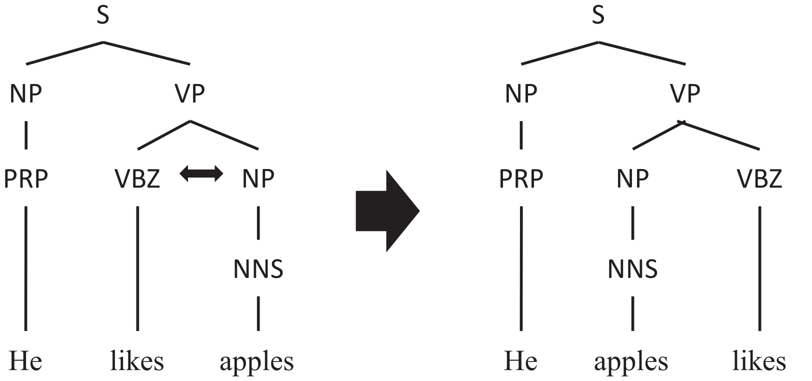


Figure 10. Transformation of the binary structure of input sentence “He likes apples.”

curacy. To evaluate the effect of introducing my SSSS transfer on the translation quality, I also implemented a pre-ordering tool using state-of-the-art techniques.

3.5 Experiments

I evaluated to what extent my SSSS transfer and pre-ordering improved the translation quality. As mentioned in Section 3.4, these methods are implemented as an add-on to off-the-shelf SMT systems. In particular, I used phrase-based SMT (Koehn et al., 2007) [41] as the base system.

3.5.1 Data

I used patent sentence corpora for training SMT. For Chinese-to-Japanese and Japanese-to-Chinese translation, I managed to collect sufficient amount of bilingual corpus just by incorporating patent claim sentences, while for English-to-Japanese and Japanese-to-English translation, I combined a corpus consisting of general patent sentences and a corpus consisting of claim sentences since the corpus consisting of claim sentences alone did not yield sufficient volume for training SMT. Ultimately, I collected an equal number of sentences for each of the four translation directions, namely, English-to-Japanese, Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese directions.

The training data for English-to-Japanese and Japanese-to-English 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-10 Workshop (Goto et al., 2013) [21]. I randomly selected 3.0 million sentence pairs. Henceforth, I 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, I 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) [69]. The concatenation of Corpora A and B was used to train baseline SMT systems, as well as those for my extensions.

The training data for Chinese-to-Japanese and Japanese-to-Chinese SMT consists solely of the bilingual corpus selected from the patent claim sentences of the Japan Patent Office English/Japanese bilingual corpus provided by ALAGIN [3] as linguistic and speech resources. I randomly selected 4.0 million sentence pairs from the Chinese-to-Japanese patent claim bilingual sentences in a similar manner to the previously mentioned Corpus B for English-to-Japanese and Japanese-to-English translation.

Development and test data were constructed separately from the training data in the following manner. First, I 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, I randomly selected 2,000 sentences from the results and asked professional translators specializing in patent translation to translate them into Chinese and 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

[3] ALAGIN: Advanced LAnGuage INformation Forum: http://alagin.jp

test data respectively consisting of 1,000 English-Chinese-Japanese claim sentence pairs.

3.5.2 Systems

In this experiment, I regard the implementation of phrase-based SMT in the Moses toolkit (Koehn et al., 2003) [42] and hierarchical phrase-based SMT (Chiang 2005) [9] as the baseline. I examined each of my SSSS transfer, and pre-ordering modules and their combination over the baseline.

Throughout the experiments, I used KenLM (Heafield et al., 2013) [25] for training language models and SyMGIZA++ (Junczys-Dowmunt and Szał, 2010) [38] for word alignment. I 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) [8] regarding BLEU (Papineni et al., 2002) [60] as the objective. For each system, I performed weight tuning three times and selected for the test the setting that achieved the best BLEU on the development data. For the baseline phrase-based SMT, I carried out evaluation both for distortion limit of six and that of twenty. This is because a distortion limit of six is the default setting of the Moses toolkit, and I also chose the value twenty that is considerably larger than the default setting for comparison purpose.

I did not apply SSSS transfer to the training data for training a model, even in the case where I apply SSSS transfer to the test data. This is for the sake of making fair comparison between different languages, as the availability of patent claim corpora is different for given language pairs, while SSSS transfer requires patent claim sentences as input. However, this setting may result in the reduction in the number of sentences since long sentences not treated by SSSS transfer may be deleted in the training process.

3.5.3 Pre-ordering

Each of the sentence segments outputted by SSSS transfer is pre-ordered using the Berkeley Parser (Petrov et al., 2006) [61] as syntactic parser. This configuration is identical to all four of the translation directions. The training data is parsed into a binary tree structure.

I performed self-learning for domain-adapting the syntactic parser. I first parsed 200,000 patent sentences using the initial parsing model. I then built a patent-adapted (not claim-adapted) parsing model by applying a self-learning procedure (Huang et al., 2009) [29] to the above automatic parses.

The initial parsing model for English was trained on the sentences in the Penn Treebank[4] as well as 3,000 patent sentences manually parsed by the authors. The initial model for Japanese was trained on the EDR Treebank[5] consisting of approximately 200,000 sentences. The initial mode for Chinese was trained CTB-6 (Zhang and Xue 2012) [78]. No patent sentences were used for training Chinese and Japanese models.

The pre-ordering model is trained on a given parallel corpus through the following procedure (de Gispert et al. 2015) [11]:

1. Parse the source sentences of the parallel corpus.

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 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.

I used the open source toolkit, Neural Probabilistic Language Model Toolkit (NPLM)[6] to train a model for predicting whether a given node is SWAP or STRAIGHT. I used the default setting for most of the settings, except for the output layer where I used two outputs corresponding to SWAP and STRAIGHT.

[4] The Penn Treebank Project: http://www.cis.upenn.edu/ treebank/home.html

[5] EDR Corpus: https://www2.nict.go.jp/out-promotion/techtransfer/EDR/JPN/Struct/Struct-CPS.html

[6] Neural Probabilistic Language Model Toolkit: http://nlg.isi.edu/software/nplm/

3.5.4 Evaluation metrics

Each system is evaluated using two metrics: BLEU (Papineni et al., 2002) [60] and RIBES (Isozaki et al., 2010a) [32]. Although my 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 segments 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) [21] and 2nd Workshop on Asian Translation（WAT） (Nakazawa et al. 2015; Isozaki and Kouchi 2015) [57, 33].

Each of the BLEU and RIBES scores are tested for significance (Koehn 2004)[40] against the baseline with the toolkit, MTEval[7], using bootstrapping method with 100 divisions and 1,000 repetitions.

3.6 Results

Tables 1, 2, 3 and 4 show the results of my experiment. In the tables, PB and HPB denote the phrase-based SMT and hierarchical phrase-based SMT of Moses toolkit respectively, and d of PB denotes the value of distortion limit. The numbers in the brackets show the improvement over P1, the vanilla PBSMT system. The scores significantly greater than the baseline at the 5% level are marked with a †, while those significantly greater than the baseline at the 1% level are marked with a ‡.

I used the 1,000 test sentences described in Section 3.5.1 for evaluation. However, I also show for further reference, the results where I selected and used the sentences having less than or equal to 200 words from the 1,000 test sentences. This is to cope with the limitation of Berkeley parser that is unable to parse long sentences in the 1,000 test sentences. Hence, I did not include evaluation using all of the 1,000 sentences for P3 and this is shown with the notation N/A.

[7] MTEval Toolkit https://github.com/odashi/mteval

ID	Settings			Test sentences			
	SSSS transfer	Pre-ordering	SMT	All sentences		Sentences w/ lt 200 tokens (805 sentences)	
				BLEU	RIBES	BLEU	RIBES
P1			PB d=6	23.9	43.9	24.7	42.2
P1'			PB d=20	23.4 (-0.5)	49.1 (+5.2)‡	23.6 (-1.1)	48.3 (+6.1)‡
H1			HPB	24.3 (+0.4)	53.3 (+9.4)‡	25.0 (+0.3)	52.9 (+10.7)‡
P2	√		PB d=6	24.5 (+0.6)†	67.8 (+23.9)‡	25.9 (+1.2)‡	70.7 (+28.5)‡
P3		√	PB d=6	N/A	N/A	25.3 (+0.6)	54.1 (+11.9)‡
P4	√	√	PB d=6	**28.4 (+4.5)**‡	**74.8 (+30.9)**‡	**31.1 (+6.4)**‡	**78.1 (+55.9)**‡

Table 1. Evaluation scores for English-to-Japanese translation

The setting $P4$, a combination of SSSS transfer and pre-ordering, in all of the four translation directions, substantial gains in both BLEU and RIBES scores are observed. Statistical significance test reveals that only $P4$ shows significant improvement at the 1% level over the baseline in both BLEU and RIBES scores and in all of the four translation directions. The settings $P2$ with SSS transfer only and $P3$ with pre-ordering only exhibit large improvement in RIBES scores, however, the improvement in BLEU score is only marginal. Substantial improvement is observed only when SSSS transfer and pre-ordering are used in combination, which may implicate complementary contribution of SSSS transfer and pre-ordering.

The overall tendency for the results with sentences containing less than or equal to 200 words is much the same, however, the scores of all the settings are higher in many cases. The performance of pre-ordering itself is different for different languages, and particular, P3 for Japanese-to-English and Chinese-to-Japanese gains higher values than P4 in some cases. However, steady improvement is observed for P4 regardless the performance of P3.

3.7 Analysis

Experimental results confirm that translation quality can be improved significantly by using SSSS transfer, irrespective of the existence of the pre-ordering process and translation directions. In this section, I first explain how my initial issues, i.e., extreme lengths and sublanguage-specific structures in claim sentences,

ID	Settings			Test sentences			
	SSSS transfer	Pre-ordering	SMT	All sentences		Sentences w/ lt 200 tokens (860 sentences)	
				BLEU	RIBES	BLEU	RIBES
P1			PB d=6	21.5	40.2	22.7	36.7
P1'			PB d=20	22.5 (+1.0)†	46.3 (+6.1)‡	24.2 (+1.5)†	44.0 (+7.3)‡
H1			HPB	23.2 (+1.8)‡	49.6 (+9.4)‡	24.4 (+1.7)†	47.4 (+10.7)‡
P2	√		PB d=6	20.9 (-0.6)	64.0 (+23.8)‡	21.4 (-1.3)	65.7 (+29.0)‡
P3		√	PB d=6	N/A	N/A	**31.9 (+9.2)**‡	**79.0 (+42.3)**‡
P4	√	√	PB d=6	**27.4 (+6.0)**‡	**74.8 (+34.6)**‡	28.9 (+6.0)‡	77.5 (+40.8)‡

Table 2. Evaluation scores for Japanese-to-English translation

ID	Settings			Test sentences			
	SSSS transfer	Pre-ordering	SMT	All sentences		Sentences w/ lt 200 tokens (805 sentences)	
				BLEU	RIBES	BLEU	RIBES
P1			PB d=6	28.4	48.0	27.8	46.6
P1'			PB d=20	28.2 (-0.2)	47.3 (-0.7)	28.8 (+1.0)	45.9 (-0.7)
H1			HPB	28.3 (+0.1)	47.3 (-0.7)	28.9 (+1.1)	45.9 (-0.7)
P2	√		PB d=6	28.8 (+0.4)	72.5 (+24.5)‡	30.4 (+2.6)‡	75.4 (+28.8)‡
P3		√	PB d=6	N/A	N/A	**33.2 (+5.4)**‡	75.4 (+28.8)‡
P4	√	√	PB d=6	**30.2 (+1.8)**‡	**73.8 (+25.8)**‡	32.5 (+4.7)‡	**76.7 (+30.1)**‡

Table 3. Evaluation scores for Chinese-to-Japanese translation

ID	Settings			Test sentences			
	SSSS transfer	Pre-ordering	SMT	All sentences		Sentences w/ lt 200 tokens (860 sentences)	
				BLEU	RIBES	BLEU	RIBES
P1			PB d=6	23.8	48.5	27.1	45.9
P1'			PB d=20	23.5 (-0.3)	50.8 (+2.3)‡	27.3 (+0.2)	48.9 (+3.0)‡
H1			HPB	20.6 (-3.2)	46.5 (-2.0)	23.7 (-3.4)	44.1 (-1.8)
P2	√		PB d=6	22.5 (-1.3)	74.0 (+25.5)‡	26.9 (-0.2)	77.9 (+32.0)‡
P3		√	PB d=6	N/A	N/A	28.0 (+0.9)	48.4 (+2.5)‡
P4	√	√	PB d=6	**25.3 (+1.5)**‡	**75.9 (+27.4)**‡	**30.0 (+2.9)**‡	**79.4 (+33.5)**‡

Table 4. Evaluation scores for Japanese-to-Chinese translation

are resolved by SSSS transfer and pre-ordering. Subsequently, I provide an in-depth analysis of the additional benefit of my SSSS transfer, i.e., making SMT inputs short. Finally, I discuss the different trends of the observed gains in the translation directions.

3.7.1 Complementary contribution of SSSS transfer and pre-ordering

Figure 11 illustrates a typical sequence of example translations generated by the four configurations, P1 to P4, in my Japanese-to-English experiment. Throughout the figure, a labelled bracketing scheme is used to illustrate claim segments. The contributions of SSSS transfer and pre-ordering are summarized as follows.

Contribution of SSSS transfer The order of segments is not changed from the input Japanese sentence in P1. However, in P2, with the introduction of SSSS transfer, the segments 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 segment, P1 and P2 do not seem significantly different. In contrast, I obtain a better translation for the second element in P4 than in P3. This is an evidence that SSSS transfer improves pre-ordering effectively.

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 segments 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 segments, results in the words in each segment being properly reordered.

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

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] [$_{TRAN}$ を備える] [$_{PREA}$ 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,] [$_{PREA}$ well screen system] [$_{ELEM}$ 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 11. Typical example of Japanese-to-English translation

3.7.2 Effects of shortening SMT inputs

As seen above, pre-ordering works better on segments obtained through SSSS transfer rather than on the entire input sentence. To estimate the shortening effect of SSSS transfer, I performed the following two analyses.

First, I evaluated the accuracy of the syntactic parser for varying input sentence lengths. Table 5 shows the sentence-wise accuracy of the English parser invoked by my 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. A parse tree is judged correct if all the constituents are correct in the parse tree, while it is judged incorrect if any incorrect constituent is present in the parse tree. The figure shows that the longer sentences show considerably lower accuracy compared with the shorter sentences. In particular, the parsing accuracy for the long sentences containing over 80 tokens is very low, where only one out of the 16 sentences were parsed correctly.

Second, I compared the distributions of lengths of the processing unit of the succeeding steps, i.e., the entire sentence for P1 and automatically identified claim segments in P2. Figure 12 shows the cumulative ratio of original sentences and identified claim segments in English, Chinese and Japanese, respectively. The syntactic parsing on the input Japanese sentence is identical both for Japanese-to-English and Japanese-to-Chinese translation directions. The figures illustrates that the sentences containing over 80 tokens comprised 31% of all the sentences before SSSS transfer, while it is reduced to 3% after SSSS transfer. Together with the analysis of parsing accuracy for varying sentence lengths, input sentences are expected to be parsed more correctly as a result of SSSS transfer.

For further reference, the number of input sentences and the number of their corresponding segments obtained by SSSS transfer are shown in Figure 6. The number of segments of around 4,000 is obtained for input 1,000 sentences for all three input languages, i.e., an average of four segments are obtained for every input sentence for all the input languages. Table 7 shows the number of correct and incorrect SSSS transfer for 100 input sentences for the three input languages. The high SSSS transfer performace in English and Chinese may be due to the highly formalized nature of the languages. However, the effect of the low performance in Japanese may be negligible, as the translation performance in case

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%
121-140	2	0	0%

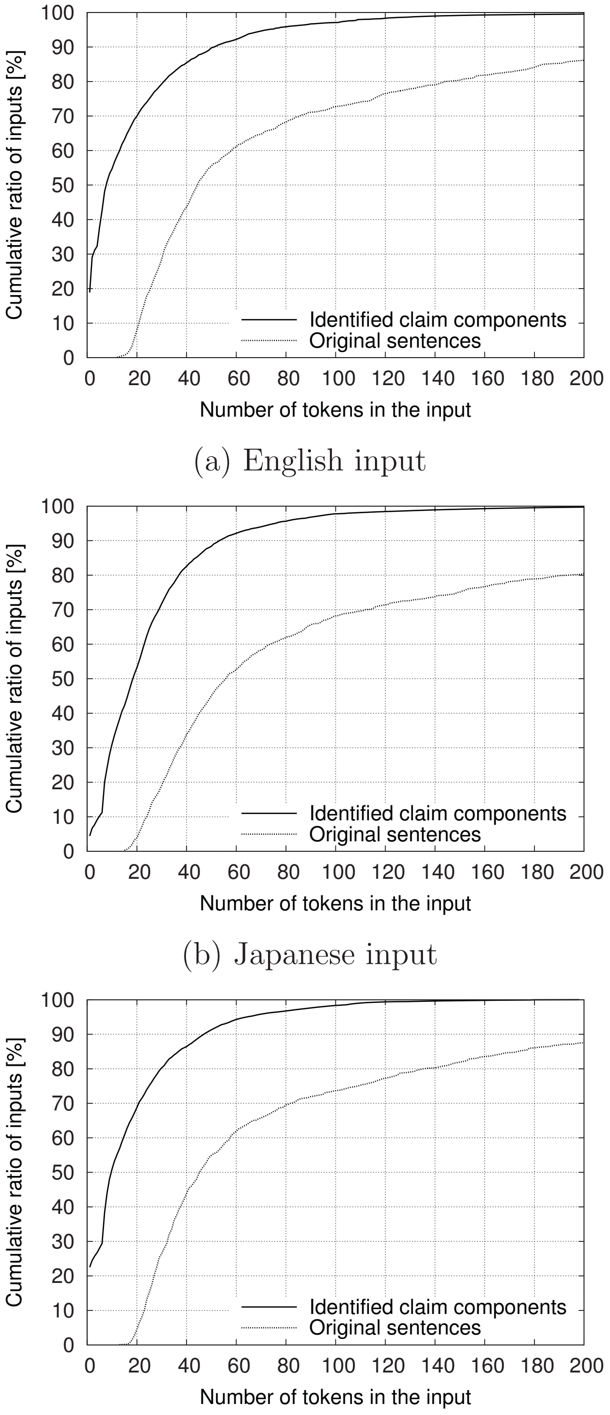
Table 5. Parsing accuracy of English parser used for English-to-Japanese pre-ordering

the source language is Japanese is much the same standard as in case the source language is English or Chinese.

3.7.3 Different trends for translation directions

The experiment showed that substantial gains of over 25 points in RIBES scores were obtained in all of the English-to-Japanese, Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese patent claim translations. I speculate that all these language directions require sentence segments to be reordered for proper translation, and my proposed method successfully realized this reordering in segments resulting in the improvement in RIBES scores.

However, the level of improvement in BLEU scores for English-to-Japanese and Japanese-to-English was larger than that for Chinese-to-Japanese and Japanese-to-Chinese. I speculate that this is because the pre-ordering within each segment in English-to-Japanese and Japanese-to-English translation requires the reordering of attachment directions as well as the reordering of the predicate, while the pre-ordering within each of the segment in Chinese-to-Japanese and Japanese-to-Chinese translation mainly involves reordering of the predicate. Hence the shortening effect of pre-ordering worked more substantially for English-to-Japanese and Japanese-to-English translation.



(a) English input

(b) Japanese input

(c) Chinese input

Figure 12. Cumulative ratio of inputs to SMT with respect to the number of words, with and without SSSS transfer

(a) Number of segments for English inputs

Input	No. of sentences		1,000
After SSSS transfer	No. of segments	PREA	1,000
		ELEM	916
		PURP	1,197
		TRAN	1,134
	Total no. of segments		4,247

(b) Number of segments for Chinese inputs

Input	No. of sentences		1,000
After SSSS transfer	No. of segments	PREA	1,000
		ELEM	465
		PURP	2,697
		TRAN	244
	Total no. of segments		4,406

(c) Number of segments for Japanese inputs

Input	No. of sentences		1,000
After SSSS transfer	No. of segments	PREA	1,000
		ELEM	1,258
		PURP	729
		TRAN	996
	Total no. of segments		3,983

Table 6. Input and number of components after SSSS transfer

(a) English

Correct	95
Incorrect	5
Total	100

(b) Chinese

Correct	89
Incorrect	11
Total	100

(c) Japanese

Correct	97
Incorrect	3
Total	100

Table 7. Number of correct and incorrect SSSS transfer

3.8 Discussion

In this Section, I described a method for transferring sublanguage-specific sentence structure for English-to-Japanese, Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese patent claim translations. The experimental results show that my proposed method, a combination of SSSS transfer and pre-ordering based on syntactic parsing, achieved a substantial gain of more than 25 points in RIBES scores in all four translation directions. In addition, my proposed method achieved five point gains in BLEU scores in English-to-Japanese and Japanese-to-English directions, and 1.5 point gains in BLEU scores in Chinese-to-Japanese and Japanese-to-Chinese translation directions. I achieved these results with only a handful of SCFG rules.

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

My proposed method has demonstrated a successful hybridization of SMT and the human knowledge of the target sublanguage sentence structure, the latter knowledge can only be handled by handcrafted rules currently.

4. Global Pre-ordering for Improving Sublanguage Translation

4.1 Introduction

Formal documents such as legal and technical documents often form sublanguages. Previous studies have highlighted that capturing the sentence structure specific to the sublanguage is extremely necessary for obtaining high-quality translations especially between distant languages [5, 48, 49]. Figure 13 illustrates two pairs of bilingual sentences specific to the sublanguage of patent abstracts. In both sentence pairs, the global sentence structure ABC in the source sentences must be reordered to CBA in the target sentences to produce a structurally appropriate translation. Each of the segments ABC must then be syntactically reordered to complete the reordering.

Various attempts have been made along this line of research. One such method is the skeleton-based statistical machine translation (SMT) which uses a syntactic parser to extract the global sentence structure, or the *skeleton*, from syntactic trees and uses conventional SMT to train global reordering [50, 74]. However, the performance of this method is limited by syntactic parsing, therefore the global reordering has low accuracy where the accuracy of syntactic parsing is low. Another approach involves manually preparing synchronous context-free grammar rules for capturing the global sentence structure of the target sublanguage [18]. However, this method requires manual preparation of rules. Both methods are unsuitable for formal documents such as patent abstracts, because they fail to adapt to sentences with various expressions, for which manual preparation of rules is complex.

This section describes a novel global reordering method for capturing sublanguage-specific global sentence structure to supplement the performance of conventional syntactic reordering. The method learns a global pre-ordering model from non-annotated corpora without using syntactic parsing and uses this model to perform global pre-ordering on newly inputted sentences. As the global pre-ordering method does not rely on syntactic parsing, it is not affected by the degradation of parsing accuracy, and is readily applicable to new sublanguages. Glob-

Pair 1	Japanese	[[$_A$ アンテナ資源を有効に活用して信頼性の高い通信を行うことができる][$_B$ 通信装置を][$_C$ 提供すること。]]
	Japanese (word-for-word translation)	[[$_A$ Antenna resources effectively utilizing reliability high communication perform capable][$_B$ communication apparatus][$_C$ to provide.]]
	English	[[$_C$ To provide][$_B$ a communication apparatus][$_A$ capable of performing highly reliable communication by effectively utilizing antenna resources.]]
Pair 2	Japanese	[[$_A$ 高画質な画像を形成できる][$_B$ 画像形成装置を][$_C$ 提供する。]]
	Japanese (word-for-word translation)	[[$_A$ High quality images form enable][$_B$ image formation device][$_C$ to provide.]]
	English	[[$_C$ To provide][$_B$ an image formation device][$_A$ which enables high quality images to be formed.]]

Figure 13. Example of sublanguage-specific bilingual sentences requiring global reordering. A, B, C are the sentence segments constituting global sentence structures.

ally pre-ordered sentence segments are then syntactically reordered before being translated by SMT.

In this empirical study on the patent abstract sublanguage in Japanese-to-English and English-to-Japanese translations, the translation quality of the sublanguage was improved when global pre-ordering was combined with syntactic pre-ordering. A statistically significant improvement was observed against the syntactic pre-ordering alone, and a substantial gain of more than 25 points in RIBES score against the baseline was observed for both Japanese-to-English and English-to-Japanese translations, and the BLEU scores remained comparable.

4.2 Related work

The hierarchical phrase-based method [9] is one of the early attempts at reordering for SMT. In this method, reordering rules are automatically extracted from non-annotated text corpora during the training phase, and the reordering rules are applied in decoding. As the method does not require syntactic parsing and learns from raw text corpora, it is highly portable. However, this method does not specifically capture global sentence structures.

The tree-to-string and string-to-tree SMTs are the methods which employ syntactic parsing, whenever it is available, either for the source or for the target language to improve the translation of the language pair [76, 1]. However, these methods too are not specifically designed for capturing global sentence structures.

The skeleton-based SMT is a method particularly focusing on the reordering of global sentence structure [50, 74]. It uses a syntactic parser to extract the global sentence structure, or the *skeleton*, from syntactic trees, and uses conventional SMT to train global reordering. Another related approach is the reordering method based on predicate-argument structure [43]. However, the performance of sentence structure extraction tends to be low when the accuracy of the syntactic parsing is low.

The syntactic pre-ordering is the state-of-the-art method which has substantially improved reordering accuracy, and hence the translation quality [34, 22, 11, 26]. However, the adaptation of this method to a new domain requires manually parsed corpora for the target domains. In addition, the method does not have a specific function for capturing global sentence structure. Thus, I apply here my proposed global reordering model as a preprocessor to this syntactic reordering method to ensure the capturing of global sentence structures.

4.3 Global pre-ordering method

I propose a novel global reordering method for capturing sublanguage-specific global sentence structure. On the basis of the finding that sublanguage-specific global structures can be detected using relatively shallow analysis of sentences [5], I extract from the training set the n-grams frequently occurring in sentences involving global reordering and use these n-grams to detect the global structure of newly inputted sentences.

For example, Figure 13 shows two sentence pairs in the training set that contain global reordering. In this dissertation, I will call the semantic units comprising a sentence, *segments*. In the figure, the segments ABC in the source sentence must be reordered globally to CBA in the target sentence to obtain structurally appropriate translations. With segment boundaries represented by the symbol ”|”, the extraction of unigrams on both sides of the two segment boundaries of sentence E1 of Figure 13 yields

$$\{provide, |, a\} \ \ \{apparatus, |, capable\}.$$

When I input the sentence *"To provide a heating apparatus capable of maintaining the temperature,"* this is matched against the above unigrams. Thus, the segment boundary positions are detected as *"To provide | a heating apparatus | capable of maintaining the temperature"*. The end-of-sentence marker "." is excluded in the global reordering process, and it is restored after all the reordering and translation processes. The detected segments are then reordered globally to yield the sentence *"Capable of maintaining the temperature | a heating apparatus | to provide,"* which has the appropriate global sentence structure for the target Japanese sentence. Each segment is then syntactically reordered before inputting to English-to-Japanese SMT.

The method consists of two steps. Step (i): Extract sentence pairs containing global reordering from the training corpus. Hereinafter, I shall call this subset of the training corpus the *global reordering corpus*. Step (ii): Extract features from the source sentences of the global reordering corpus, and use these features to detect the segments of newly inputted sentences. Then reorder these detected segments globally. In step (ii), I experiment with a detection method based on heuristics, as well as a method based on machine learning. Steps (i) and (ii) are described in the following subsections.

4.3.1 Extraction of sentence pairs containing global reordering

I consider that a sentence pair contains *global reordering* if the segments in the target sentence appear in swap orientation in phrase-based sense [20] to the source segments within the alignment table, when the sentences are divided into two or three segments each. The number of source segments must equal that of target segments. Figure 14 shows an example of a sentence pair involving global reordering with the sentence divided into three segments. The sentence pair in Figure 14 meets the requirement and is regarded as containing global reordering.

Although in theory, my proposed method can be applied to sentence pairs containing more than three segments, I have limited the number of segments

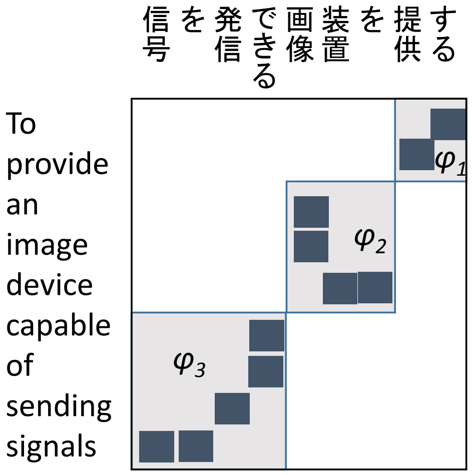


Figure 14. An example of segments arranged in swap orientations for English to Japanese translation

to two and three based on my empirical observation that particular n-grams frequently appear around segment boundaries only when an input sentences is divided into two or three segments.

The steps are as follows:

1. By regarding that all positions between two adjacent words can be a segment boundary both for the source and target sentence, all candidate segments starting at all possible word positions are created. The number of segments both for the source and target is restricted to two and three, and the number of source segment must equal that of target segments. Also overlapping candidates are removed. Here, a sentence pair consisting of K segments is represented as $(\phi_1, \phi_2 \cdots \phi_K)$, where ϕ_k consists of the k^{th} phrase of the source sentence and ${\alpha_k}^{th}$ phrase of the target sentence.
2. Out of all the candidates in step 1, all the candidates with the segments in swap orientation are extracted. The source and target phrases of ϕ_k are considered to be in swap orientation if $\alpha_k = \alpha_{k+1} + 1$.
3. If there is more than one candidate, the segment candidate are selected based on the head directionality of the source sentence. In head-initial languages such as English, the most important two or three segments tend to

Cand. 1

ϕ_1	*To provide*
ϕ_2	*an outlet facilitating assembly work*
ϕ_3	*for employees .*

Cand. 2

ϕ_1	*To provide*
ϕ_2	*an outlet*
ϕ_3	*facilitating assembly work for employees .*

Figure 15. Ranking of segmenting candidates according to head-directionality

appear near the beginning of the sentence, whereas in head-final languages, the most important segments tend to appear toward the end of the sentence. For a head-initial language, such as English, I select the candidate for which ϕ_K has the largest length. For a head-final language, such as Japanese, I select the candidate for which ϕ_1 has the largest length. Figure 15 shows two example English segment candidates in which the value of K is three. In this case, I select candidate 1 since the length of ϕ_3 is *six* which is larger than that of candidate 2 whose length is *three*.

The sentences containing global reordering are extracted from the automatically aligned training corpus and stored in the global reordering corpus which is subsequently used for training and prediction.

4.3.2 Training and prediction of global reordering

4.3.3 Heuristics-based method

In the heuristics-based method, I extract n-grams from the source sentences of the global reordering corpus and match these n-grams against a newly inputted sentence to perform global reordering. I call this method *heuristics-based*, because automatic learning is not used for optimizing the extraction and matching processes of the n-grams, but rather, I heuristically find the optimal setting for the given training data. Below, I describe the extraction and matching processes.

N-gram extraction I extract n-grams occurring on both sides of the segment boundary between adjacent segments ϕ_k and ϕ_{k+1}. In the heuristic-based method, n can assume different values in the left- and right-hand sides of the segment boundary. Let B be the index of the first word in ϕ_{k+1}, and f be the source sentence. Then the range of n-grams extracted on the left-hand side of f is as follows where n_L is the value n of the n-gram.

$$(f_{B-n_L}, f_{B-n_L+1} \cdots f_{B-1}) \tag{3}$$

Likewise, the range of n-grams extracted from the right-hand side of f is as follows where n_R denotes the value n of the n-gram.

$$(f_B \cdots f_{B+n_R-2}, f_{B+n_R-1}) \tag{4}$$

Decoding The decoding process of my global reordering is based on n-gram matching. I hypothesize that the matching candidate is more reliable (i) when the length of the n-gram matching is larger and/or (ii) when the occurrence frequency of the n-grams is higher. Thus, I heuristically determine the following score where *len* denotes the length of n-gram matching and $freq$ denotes the occurrence frequency of the n-grams. I calculate the score for all matching candidates and select the candidate that has the highest score.

$$\log(freq) \times len \tag{5}$$

An example of the decoding process for sentences containing two segments is presented in Figures 16. Examples of the decoding process for sentences containing two or three segments are presented in Figures 16 and 17 respectively. When a candidate matches the n-grams for both two and three segments, I use those with three segments for segment detection in accordance with my hypothesis that long n-gram matching is more reliable than short matching.

Figure 16 shows an example of the decoding process for an input sentence containing two segments, i.e., $K = 2$, with one segment boundary. $m1$ through $m4$ are the n-grams matching the input sentence "*To prevent imperfect coating and painting,*" where "|" denotes the position of the segment boundary. The matching length is indicated by *len* which is the sum of n_L

ID	n-grams	len	freq
m1	*prevent,* |	1	2217
m2	*To, prevent,* |	2	1002
m3	*To, prevent,* |, *imperfect*	3	120
m4	*To, prevent,* |, *imperfect, coating*	4	18

Figure 16. Example of n-gram matching against an input sentence containing two segments. The input sentence is "*To prevent imperfect coating and painting.*"

and n_R on both sides of the segment boundary. For example, for $m3$, the occurrence frequency is given as 120 and len is calculated such that $len = n_L + n_R = 2 + 1 = 3$. A score is calculated using equation 5 for all candidates, $m1$ through $m4$, and the candidate obtaining the highest score is used to determine the segment boundary.

Figure 17 shows an example of n-gram matching for a sentence containing three segments with two segment boundaries. $n1$ through $n5$ are the n-grams matching the input sentence "*To provide a household heating device capable of maintaining the room temperature*," where "|" denotes the positions of the two segment boundaries. Here, len is the sum of len_1 and len_2 where len_1 is the matching length for the first segment boundary and len_2 is that for the second boundary. The matching length of the first boundary len_1 is calculated as $len_1 = n_L + n_R = 2+1 = 3$, whereas that for the second boundary len_2 is calculated as $len_2 = n_L + n_R = 1 + 1 = 2$, which yields $len = len_1 + len_2 = 3 + 2 = 5$. Consequently, the score for $n3$ is calculated with $len = 5$ and $freq = 112$. A score is calculated using equation 5 for all candidates, $n1$ through $n5$, and the candidate with the highest score is used to determine the segment boundary.

4.3.4 Machine learning–based method

As the heuristic method involves intuitive determination of settings, which makes it difficult to optimize the performance of the system, I introduce machine learning to facilitate the optimization of segment detection. I regard segment boundary prediction as a binary classification task and use support vector machine (SVM)

ID	n-grams for 1st boundary	n-grams for 2nd boundary	len	freq
n1	*To, provide, |*	*|, capable*	3	334
n2	*To, provide, |, a*	*|, capable*	3	254
n3	*To, provide, |, a*	*device, |, capable*	5	112
n4	*To, provide, |, a*	*device, |, capable*	6	94
n5	*To, provide, |, a*	*device, |, capable, of, maintaining*	7	3

Figure 17. Matching n-grams against input sentence for sentences containing three segments. The input sentence is "*To provide a household heating device capable of maintaining the room temperature*".

models to perform training and prediction. I train an SVM model to predict whether each of the word positions in the input sentence is a segment boundary by providing the features relating to the word in question. I use two types of features, as described below, for SVMs, both for training and prediction.

- **N-grams**: Here, n-grams are extracted from both sides of the word under training/prediction. In contrast to the heuristics-based method, for simplicity, I use here the same value of n for n-grams in the left- and right-hand sides of the examined word. The n-grams used are as follows, where f is the sentence, i is the index of the word in question, and n is the value n of n-grams.
$$(f_{i-n+1}, f_{i-n+2} \cdots f_i \cdots f_{i+n-1}, f_{i+n}) \tag{6}$$
- **Position in the sentence**: The position of the word under training/prediction is provided as a feature. This feature is introduced to differentiate multiple occurrences of identical n-grams within the same sentence. The position value is calculated as the position of the word counted from the beginning of the sentence divided by the number of words contained in the sentence. This is shown as follows, where i denotes the index of the word in question and F is the number of words contained in the sentence.
$$\frac{i}{F} \tag{7}$$

In the prediction process, I extract the features corresponding to the word position i and then input these features to the SVM model to make a prediction for i. By repeating this prediction process for every i in the sentence, I obtain a sentence with each position i marked either as a segment boundary or as *not* a segment boundary. These predicted segments are then reordered globally to produce the global sentence structure of the target language.

As I do not control the number of segment boundaries within a sentence during training, an arbitrary number of segments is produced for each input sentence. Therefore, I then limit the number of segments to a maximum of three as the global reordering corpus contains only sentences with two and three segments. I limit the number of segments according to the head directionality of the source language. For head-initial languages such as English I select segment boundaries from the beginning of the sentence, whereas for head-final languages such Japanese I select segment boundaries from the end of the sentence.

4.4 Experiments

I conducted experiments to illustrate the effect of introducing global pre-ordering. In this section, I first describe the reordering configuration for depicting the effect of global pre-ordering. I then describe the primary preparation of global reordering, followed by a description of the settings used in my translation experiment.

4.4.1 Reordering configuration

To illustrate the effect of introducing global pre-ordering, I evaluate the following four reordering configurations: **(T1)** Baseline SMT without any reordering; **(T2)** T1 with global pre-ordering only. The input sentence is globally pre-ordered, and this reordered sentence is translated and evaluated; **(T3)** T1 with conventional syntactic pre-ordering [22]. The input sentence is pre-ordered using conventional syntactic pre-ordering and the reordered sentence is translated and evaluated; and **(T4)** T1 with a combination of syntactic and global pre-ordering. The input sentence is globally pre-ordered, each segment is reordered using syntactic pre-ordering and the reordered sentence is translated and evaluated.

4.4.2 Preparation of global reordering

As described in Section 4.3.2, I experiment two methods for training global reordering, namely, heuristics-based and machine learning-base methods. For the heuristics-base method, the optimal setting for the given data set is heuristically found and hence no quantitative evaluation of the performance of this method on its own is performed. On the other hand, for the machine learning-based method, I calibrated the performance of the learning tools by varying the parameters and setting of the given data prior to the experiment. I describe the calibration procedure of the machine learning-based method in this section.

As described in Section 4.3.4, I use n-grams and the position in the sentence as the features. In preparation for global pre-ordering, I calibrated the machine learning-based detection to determine the optimal feature set for detecting segments. To determine the optimal feature set, I plotted the prediction accuracy with respect to the size of the global reordering corpus and value n of n-grams. As my support vector machines, I used liblinear 1.94 [15] for training and prediction. I used the default settings for executing liblinear since no obvious improvement over the default settings was observed by performing scaling with the svm-scale tool, or by performing grid searching using the grid.py tool.

Figure 18 shows the variation in the prediction accuracy with respect to the size of the global reordering corpus and the order of an n-gram for Japanese input. Figure 19 shows the same for English input. The *accuracy* is the average accuracy of a ten-fold cross-validation for the global reordering corpus. From the calibration shown in the tables, I select the settings producing the highest prediction accuracy, namely, a value of $five$ for the n of n-grams and a size of $100k$ for the global reordering corpus, for both Japanese and English inputs. However, overfitting may be taking place when the size of the global reordering corpus is 100,000 sentence, since the prediction performance of the machine degrades as the size of n-gram increases as shown in the figure.

4.4.3 Translation experiment setup

Data As my experimental data, I use the Patent Abstracts of Japan (PAJ), the English translations of Japanese patent abstracts. I automatically align [69] PAJ with the corresponding original Japanese abstracts, from which I

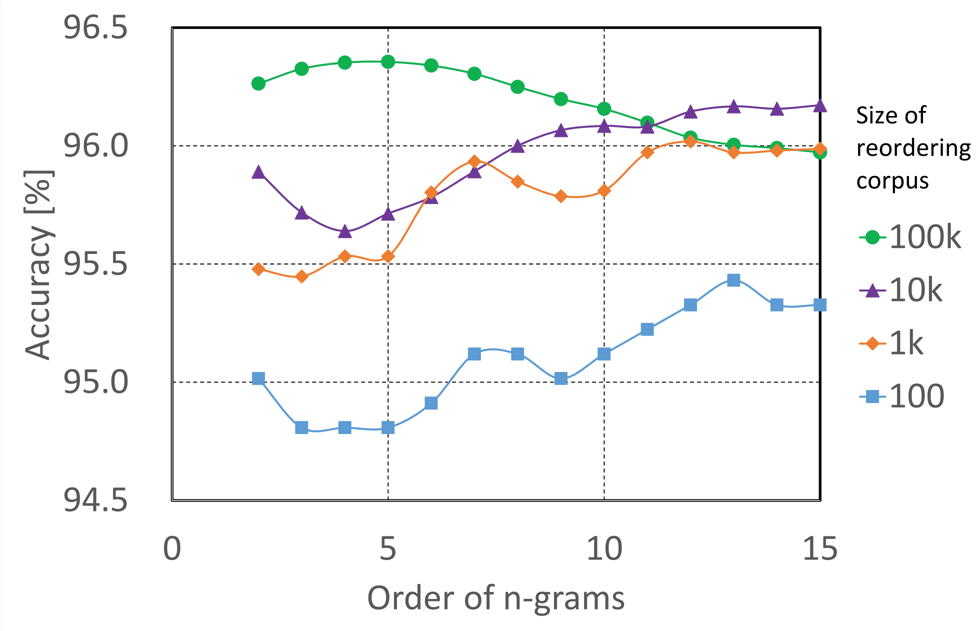


Figure 18. Accuracy of Japanese segment boundary detection using an SVM model for various values of n for an n-gram with various sizes of global reordering corpus. The legend on right-hand side shows the size of the global reordering corpus.

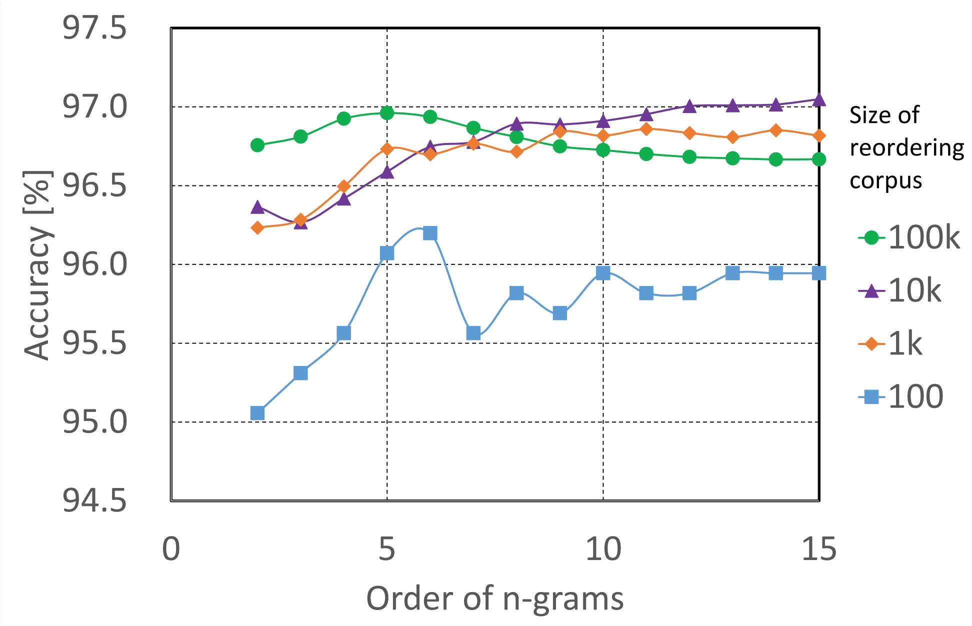


Figure 19. Accuracy of English segment boundary detection with an SVM for various values of n for an n-gram with various sizes of global reordering corpus. The legend on right-hand side shows the size of the global reordering corpus.

randomly select 1,000,000 sentence pairs for training, 1,000 for development and 1,000 for testing. This training data for the translation experiment are also used for training global reordering as described in the previous subsection. Out of the 1,000 sentences in the test set, I extract the sentences that show any matching with the n-grams and use these sentences for my evaluation. In my experiments, the number of sentences actually used for evaluation is 300.

Out of the 100,000 sentence pairs used for training global reordering, those sentence pairs containing global reordering amounted to 38,194 pairs. For the heuristic-based method, the 2- to 5-grams extracted from the above sentences amounted to 381,311 n-grams which are subsequently stored in the global reordering corpus.

Baseline SMT The baseline system for my experiment is Moses phrase-based SMT [41] with the default distortion limit of six. I use KenLM [25] for training language models and SyMGIZA++ [38] for word alignment. The weights of the models are tuned with the n-best batch MIRA [8] regarding BLEU [60]as the objective. For each system, I performed weight tuning three times and selected the setting that achieved the best BLEU on the development data.

As variants of the baseline, I also evaluate the translation output of the Moses phrase-based SMT with a distortion limit of 20, as well as that of the Moses hierarchical phrase-based [9] SMT with the default maximum chart span of ten. I also evaluate the translation output of Travatar (Neubig 2013) [58], a tree-to-string SMT. For training and decoding using Travatar, I used Ckylark (Oda, Neubig, Sakti, Toda, and Nakamura 2015) [59] as the syntactic parser which is the recommended setting of Travatar.

Conventional syntactic pre-ordering Syntactic pre-ordering is implemented on the Berkeley Parser. The input sentences are parsed using the Berkeley Parser, and the binary nodes are swapped by the classifier [22]. This basic pre-ordering setting is identical both for Japanese-to-English and English-to-Japanese translation directions.

The Berkeley Parser is domain-adapted by self-learning. The initial parsing model for English was trained on the sentences in the Penn Treebank[8] as well as 3,000 patent sentences manually parsed by the authors. The initial model for Japanese was trained on the EDR Treebank[9] consisting of approximately 200,000 sentences. No patent sentences were used for training Japanese models.

As a variant of conventional reordering, I also use a reordering model based on the top-down bracketing transducer grammar (TDBTG) [10](Nakagawa 2015) [56]. I use the output of mkcls and SyMGIZA++ obtained during the preparation of the baseline SMT for training TDBTG-based reordering.

Global pre-ordering Global pre-ordering consists of the detection of segment boundaries and the reordering of the detected segments. Out of the 1,000,000 phrase-aligned sentence pairs in the training set for SMT, I use the first 100,000 sentence pairs for extracting the sentence pairs containing global reordering. I only use a portion of the SMT training data due to the slow execution speed of the current implementation of the software program for extracting sentence pairs containing global reordering. Since the alignment table described in Section 4.3.1 contains a certain amount of erroneous alignments especially those alignments involving punctuation marks, the alignments in the alignment table that are aligned with punctuation marks are removed prior to the experiment. I evaluate both the heuristic and the machine learning-based methods for comparison.

Evaluation metrics I use the RIBES [32] and the BLEU [60] scores as evaluation metrics. I use both metrics because n-gram-based metrics such as BLEU alone cannot fully illustrate the effects of global reordering. RIBES is an evaluation metric based on rank correlation which measures long-range relationships and is reported to show much higher correlation with human evaluation than BLEU for evaluating document translations between distant languages [33].

[8] The Penn Treebank Project: http://www.cis.upenn.edu/ treebank/home.html

[9] EDR Corpus: https://www2.nict.go.jp/out-promotion/techtransfer/EDR/JPN/Struct/Struct-CPS.html

[10] Top-Down BTG-based Preordering https://github.com/google/topdown-btg-preordering

Table 8. Evaluation of Japanese-to-English translation where *glob-pre* denotes global pre-ordering and *pre* denotes conventional syntactic pre-ordering, *dl* denotes distortion limit, HPB denotes hierarchical phrase-based SMT and TDBTG denotes reordering based on top-down bracketing transduction grammar.

	Reordering config		Settings			Results	
	glob-pre	pre	SMT	glob-pre	pre	RIBES	BLEU
T1			PB dl=6			44.9	17.9
			PB dl=20			53.7 (+8.8)	21.3 (+3.4)
			HPB			54.9 (+10.0)	23.1 (+5.2)
T2	√		PB dl=6	heuristic		61.7 (+16.8)	19.6 (+1.7)
			PB dl=6	SVM		61.0 (+16.1)	19.3 (+1.4)
T3		√	PB dl=6		TDBTG	64.6 (+19.7)	22.3 (+4.4)
			PB dl=6		syntactic	64.9 (+20.0)	**25.5 (+7.6)**
T4	√	√	PB dl=6	heuristic	syntactic	**71.3 (+26.4)**	**25.3 (+7.4)**
			PB dl=6	SVM	syntactic	**72.1 (+27.2)**	**25.6 (+7.7)**

Table 9. Evaluation of English-to-Japanese translation

	Reordering config		Settings			Results	
	glob-pre	pre	SMT	glob-pre	pre	RIBES	BLEU
T1			PB dl=6			43.2	27.9
			PB dl=20			54.4 (+11.1)	29.0 (+1.1)
			HPB			59.1 (+15.8)	32.1 (+4.2)
T2	√		PB dl=6	heuristic		59.5 (+16.2)	28.4 (+0.5)
			PB dl=6	SVM		65.3 (+22.1)	29.1 (+1.2)
T3		√	PB dl=6		TDBTG	**77.7 (+34.5)**	34.9 (+7.0)
			PB dl=6		syntactic	76.1 (+32.8)	**36.9 (+9.0)**
T4	√	√	PB dl=6	heuristic	syntactic	**77.3 (+34.1)**	**36.5 (+8.6)**
			PB dl=6	SVM	syntactic	**77.7 (+34.5)**	**36.5 (+8.6)**

4.5 Results

The evaluation results based on the present translation experiment are shown in Tables 8 and 9 for Japanese-to-English and English-to-Japanese translations respectively, listing the RIBES and BLEU scores computed for each of the four reordering configurations. The numbers in the brackets refer to the improvement over the baseline phrase-based SMT with a distortion limit of six. *glob-pre* denotes global pre-ordering and *pre* denotes conventional syntactic pre-ordering, *dl* denotes distortion limit, HPB denotes hierarchical phrase-based SMT and TDBTG denotes reordering based on top-down bracketing transduction grammar. The bold numbers indicate a statistically insignificant difference from the best system performance according to the bootstrap resampling method at $p = 0.05$.

A substantial gain of more than 25 points in the RIBES scores compared to the baseline is observed for both Japanese-to-English and English-to-Japanese translations, when global pre-ordering is used in conjunction with conventional pre-ordering. Also, the combination of global syntactic pre-ordering performs significantly better than syntactic pre-ordering alone. The BLEU score is not as sensitive to the introduction of global reordering, probably because the improvement is mainly concerned with long-distance reordering. I will further discuss the matter of evaluation metrics in the following section.

4.6 Analysis

4.6.1 Correctness of global sentence structure

I evaluated to find out the extent my proposed method succeeded in producing structurally correct target sentences as a result of global pre-ordering. I consider this evaluation important as the target sentence must have the correct structure in the first place for obtaining high quality translation. I counted a target sentence to have the correct structure if the target sentence satisfy the following requirements.

- In case the input structure ABC has to be translated as CBA in the target sentence, the sentence is actually translated as above

- All the segments in the input sentence are all present in the target sentence, and the segments are arranged in the correct order in the target sentence

Table 10. Number of sentences with correctly identified structures for Japanese-to-English translation (Out of 100 sentences)

	Reordering configuration		Settings			No. of sents with correct structure (100 sents)	
	glob-pre	pre	SMT	glob-pre	pre	No. of sents	Rate (%)
T1			PB dl=6			4	4%
			PB dl=20			12	12%
			HPB			12	12%
			Travatar			38	38%
T2	√		PB dl=6	heuristic		30	30%
			PB dl=6	SVM		31	31%
T3		√	PB dl=6		TDBTG	21	21%
			PB dl=6		syntactic	27	27%
T4	√	√	PB dl=6	heuristic	syntactic	46	46%
			PB dl=6	SVM	syntactic	58	58%

Table 11. Number of sentences with correctly identified structures for English-to-Japanese translation (Out of 100 sentences)

	Reordering configuration		Settings			No. of sents with correct structure (Out of 100 sentences)	
	glob-pre	pre	SMT	glob-pre	pre	No. of sents	Rate (%)
T1			PB dl=6			6	6%
			PB dl=20			15	15%
			HPB			28	28%
			Travatar			52	52%
T2	√		PB dl=6	heuristic		23	23%
			PB dl=6	SVM		26	26%
T3		√	PB dl=6		TDBTG	59	59%
			PB dl=6		syntactic	67	67%
T4	√	√	PB dl=6	heuristic	syntactic	68	68%
			PB dl=6	SVM	syntactic	63	63%

- All the words constituting a segment must appear consecutively without any gaps in th target sentence

Table 10 shows the number of sentences that have the correct sentence structure out of the 100 target sentences for Japanese-to-English translation. Table 11 shows that for English-to-Japanese translation. In both tables, T4, a combination of global reordering and syntactic reordering, produces substantially more sentences with correct structures compared with T1 and T2. For Japanese-to-English translation, T4 shows substantially better performance compared with T3, syntactic reordering alone. However, for English-to-Japanese, the performance of T4 is not so obvious as the syntactic reordering for English-to-Japanese already performs well.

Even though, T2 and T4 are both expected to improve the sentence structure, our analysis reveals that the correctness of sentence structure for T4 is much higher than that for T2. I speculate this is because the effect of global reordering alone as in T2 may yield undesirable word orders in the vicinity of segment boundaries, while this undesirable word orders is alleviated by the syntactic reordering performed alongside the global reordering.

Even though the BLEU score for HPBSMT as is considerably higher than the baseline in Tables 8 and 9, this improvement is not obvious in Tables 10 and 11. I speculate that this is because the improvement by HPBSMT is relatively localized and as a result, contributes the improvement in BLEU score, while it does not contribute to producing correct structures.

4.6.2 Human evaluation

I performed human evaluation for each of the translations. This is for determining the more human-friendly evaluation metrics because the tendency of BLEU scores and that of RIBES considerably differed in Tables 8 and 9. For example, in Table 8, the BLEU score for T4 is not significantly better than T1 and T3, while the RIBES score for T4 is significantly better than T1 and T3.

As human evaluation incurs time and cost, I selected the settings that obtained the highest score in each of T1 and T2. I used all the settings for T3 and T4 as there was not notable difference in the performance. Sentences were evaluated by one evaluator who possesses technical knowledge of the field of the test sentences. Sentences to be evaluated are shuffled so that the evaluator can not determine the source data. For evaluation, I used the first 100 sentences out of the 1,000 test sentences I used for automatic evaluation.

Table 12. Human evaluation for Japanese-to-English translation (Out of 100 sents

	Reordering configuration		Settings			No. of SABC evaluated sents (100 sents)	
	glob-pre	pre	SMT	glob-pre	pre	No. of sents	Rate (%)
T1			PB dl=6				
			PB dl=20				
			HPB				
			Travatar			55	55%
T2	√		PB dl=6	heuristic			
			PB dl=6	SVM		33	33%
T3		√	PB dl=6		TDBTG	27	27%
			PB dl=6		syntactic	36	36%
T4	√	√	PB dl=6	heuristic	syntactic	60	60%
			PB dl=6	SVM	syntactic	65	65%

Table 13. Human evaluation for English-to-Japanese translation (Out of 100 sents

	Reordering configuration		Settings			No. of SABC evaluated sents (100 sents)	
	glob-pre	pre	SMT	glob-pre	pre	No. of sents	Rate (%)
T1			PB dl=6				
			PB dl=20				
			HPB				
			Travatar			40	40%
T2	√		PB dl=6	heuristic			
			PB dl=6	SVM		9	9%
T3		√	PB dl=6		TDBTG	37	37%
			PB dl=6		syntactic	55	55%
T4	√	√	PB dl=6	heuristic	syntactic	50	50%
			PB dl=6	SVM	syntactic	55	55%

The evaluation scale used for the human evaluation comprises five levels, namely, S (Native level), A (Good), B (Fair), C (Acceptable) and D (Nonsense). Although conventional human evaluation often measures the quality of translated sentences in terms of adequacy and fluency (Denkowski and Lavie 2010) [12], studies on human evaluation metrics reveal that adequacy and fluency are often difficult to differentiate as adequacy and fluency demonstrate high correlation (Callison-Burch, Fordyce, Koehn, Monz, and Schroeder 2007) [7]. Hence, I used a single five-level scale not differentiating adequacy and fluency.

Table 12 show the human evaluation for Japanese-to-English translation, and Table 13 show that for English-to-Japanese translation. In the evaluation of Japanese-to-English translation shown in Table 12, T4 obtained considerably higher evaluation scores compared with T1, T2 and T3, and hence the RIBES score in Table 8 may be regarded to resemble human evaluation then BLEU score. On the other hand, in English-to-Japanese translation shown in Table 13, T4 is considerably higher than T1 and T2, but the comparison with T3 varies according to the syntactic pre-ordering method. However, the RIBES score seems to exhibit similar tendency with the human evaluation.

4.6.3 Typical translations

Figure 20 shows typical translations for each of the reference, and the four settings T1, T2, T3 and T4, demonstrating how T4, our proposed method, proves effective especially for Japanese-to-English translations.

T1 T1, the baseline, lacks segment A in the target sentence when compared with the reference, and segment B and C are not arranged in the correct order. Also, the words in each of the segments are not arranged in the appropriate order.

T2 T2, the baseline with global reordering alone, gives all the input segments in the the target sentence, and the segments are arranged in the correct order. However, the segment-wise translation is not improved as the word order within each segment is not appropriate.

T3 T3, the baseline with syntactic reordering alone, the words within each segments are appropriately ordered. However, the segments are not arranged in the correct order.

Input sentence	[$_C$ 固体含有量の多い高いトナーケーク層を生成し、静電印刷エンジンで作動可能な] [$_B$ トナーケーク層形成装置を] [$_A$ 提供する。]
Reference	[$_A$ To provide] [$_B$ a toner cake layer forming apparatus] [$_C$ which forms a toner cake layer having a high solid content and which can be actuated by an electrostatic printing engine.]
T1	[$_C$ Solid content of high toner cake layer for generating an electrostatic print engine operates in] [$_B$ a toner cake layer forming device.]
T2	[$_A$ To provide] [$_B$ toner cake layer forming apparatus] [$_C$ of the solid content of high toner cake layer for generating an electrostatic print engine can be operated.]
T3	[$_C$ For generating toner cake layer having a high solids content and] [$_A$ to provide] [$_B$ a toner cake layer forming device] [$_C$ which can be operated by an electrostatic printing engine.]
T4	[$_A$ To provide] [$_B$ a toner cake layer forming device] [$_C$ for generating toner cake layer having a high solid content, and operable by an electrostatic printing engine.]

Figure 20. Typical translations for Japanese-to-English translation

Input sentence	[$_A$ 投影面の中心部をその周辺部より高い光 強度で照明しうる] [$_B$ 光学投影装置を] [$_C$ 提供する。]
Reference	[$_C$ To provide] [$_B$ an optical projection system] [$_A$ which can illuminate a central part of a projection plane with light intensity higher than that of a peripheral section.]

Identified segments	投影面の中心部をその周辺部より高い光 強度で照明しうる
	光学投影装置を
	提供する
Result of global reordering	提供する
	光学投影装置を
	投影面の中心部をその周辺部より高い光 強度で照明しうる
Result of syntactic reordering to each seg.	する 提供
	光学 投影 装置
	中心部 の 投影 面 うる し 照明 で 光 強度 高い より その 周辺 部
SMT output	To provide an optical projection device center part of the projection plane can be illuminated by a light intensity higher than the peripheral part.

Table 14. An erroneous Japanese-to-English translation and intermediate stages

T4 T4, the baseline with global reordering together with syntactic reordering produces a translation with all the segments are arranged in correct order and the words within each segments are arranged in appropriate order.

Since only 65% of the sentences translated with T4 scores better than D in Japanese-to-English direction and 55% in English-to-Japanese direction, I carried out manual inspection of the sentences scoring a D. There are two stages where the translated sentences score D. First, the translated sentences score a D in most cases if the global sentence structure is not correctly identified in the first place. Second, the translated sentences score a D even when the global sentence structure is identified correctly.

Table 14 shows a typical erroneous translation corresponding to the latter case. While the steps up to the global reordering are performed correctly, the syntactic reordering for the third segment is not performed correctly. Specifically, while the token "うる" meaning "which can" must be placed at the beginning of the segment as a result of syntactic reordering, it is erroneously placed as the fifth word in the segment and instead, the token "中心部" meaning "center part" is placed at the beginning of segment. This error causes SMT to produce a translation where the translated segment "an optical projection device" is erroneously succeeded by "center part", rather than "which can", resulting in a string "an optical projection device center part". As a result, the evaluator fails to identify the segment "an optical projection device" yielding a low evaluation score. This erroneous reordering may be due to the training process that assumes sentence-wise reordering rather than segment-wise reordering.

4.6.4 Different trends for translation directions

Through the above mentioned experiment and analysis, the proposed method gives translation quality that is significantly better than conventional reordering for English-to-Japanese translation direction, while the translation quality does not outperform conventional reordering for Japanese-to-English translation direction. I speculate this is due to the difference in the readiness of reordering based on syntactic parsing for different translation directions. Since English sentences have more rigid syntactic structure compared with Japanese sentences, it is relatively simple to parse an English sentence and use this rigid syntactic structure to generate a Japanese sentence, than generating an English sentence from a less rigid Japanese syntax structure. It can be thought that the proposed method aids the recognition of long distance relationship in the Japanese-to-English translation direction and substantially improves the translation quality.

4.7 Discussion

In this Section, I proposed a global pre-ordering method that supplements conventional syntactic pre-ordering and improves translation quality for sublanguages. The proposed method learns global reordering models without syntactic parsing from a non-annotated corpus. The experimental results on the patent abstract sublanguage show substantial gains of more than 25 points in RIBES and comparable BLEU scores when compared with baseline SMT for Japanese-to-English and English-to-Japanese translations. Comparison with conventional syntactic reordering gives the results that the proposed method substantially improves Japanese-to-English translation direction, while the method does not outperform the conventional syntactic reordering in English-to-Japanese translation direction.

5. Conclusion

5.1 Summary

Many of the documents to be translated in the translation services industry are said to form sublanguages in the sense that the vocabulary, sentence structure, and expressions used in each translation domain and application are considerably different from those of general documents. It is this sublanguage that makes machine translation difficult. On the other hand, the key to improving machine translation for the translation services industry is to devise methods for incorporating information specific to the sublanguages into the translation mechanism. As the sentences comprising these formal documents forming sublanguages are often very long and complex, characteristic writing styles have been devised for each sublanguage in daily practice among writers so that readers can easily comprehend the documents.

This paper presents methods for incorporating features specific to each sublanguage into the translation mechanism to recognize the sentence structure correctly and improve translation quality. The correct recognition of sentence structure is particularly important for translating long sentences between distant language pairs because not only the syntactic order but also the sentence structure is different between these language pairs. This paper empirically demonstrated the following points:

1. The effectiveness of incorporating SSSSs into the mechanism of SMT is shown empirically.

2. The effectiveness of incorporating handcrafted rules for recognizing SSSSs is shown when high regularity in the writing style is present in the formal document in question. An experiment that demonstrates the effect of this method is shown in Section 3.

3. The effectiveness of incorporating an automatic detection method for recognizing SSSSs is shown when moderate regularity in the writing style is present in the formal document in question. An experiment that demonstrates the effect of this method is shown in Section 4.

A brief summary of each of the experiments is described in Section 3 and Section 4 as follows.

Section 3 describes translation experiments for patent claim sentences that are extremely long but exhibit exceptionally high regularity in the writing style. The experimental results show that my proposed method, a combination of SSSS transfer and pre-ordering based on syntactic parsing, achieved a substantial gain of more than 25 points in the RIBES scores in all four translation directions. In addition, my proposed method achieved five-point gains in BLEU scores in English-to-Japanese and Japanese-to-English translations, and 1.5 point gains in BLEU scores in Chinese-to-Japanese and Japanese-to-Chinese translations. These results were achieved with only a handful of SCFG rules. My proposed method successfully improved the translation of patent claims with a quality comparable to that of the other parts of patent documents. My proposed method has demonstrated a successful hybridization of SMT and human knowledge of the target SSSS; the latter knowledge can only be handled by handcrafted rules currently.

Section 4 describes a method for capturing the sentence structure with moderate regularity of writing style and higher occurrence frequency compared with patent claim sentences. A substantial improvement in translation quality was observed by incorporating global reordering along with conventional reordering. The proposed method learns global reordering models without syntactic parsing from a non-annotated corpus. The experimental results on the patent abstract sublanguage show substantial gains of more than 25 points in RIBES and comparable BLEU scores for Japanese-to-English and English-to-Japanese translations.

5.2 Discussion

Variation in quality of handcrafted rules for different rule writers As the generation rules and SCFG rules in Section 3 are constructed by human rule writers, there is inevitably variation in the quality of the rules constructed. However, because the regularity of the patent claim sublanguage is considerably high, the variation in the constructed rules is expected to be minimal provided the rule writers have sufficient knowledge of the patent claim sublanguage. To construct sublanguage-specific rules, the writers are required to have up-to-date knowledge of the common practice of patent claims, but are not required to have specific knowledge of the individual patent domain in question. The writers are required to update their knowledge periodically, because the common practice in patent claims is considerably affected by major patent cases. For example, constructions such as "*XXX method comprising: a step for AAAing ...; a step for BBBing ...; and a step for CCCing ... .*" has become less popular recently following judicial precedents that a patent claim expressed in this writing style is regarded to explicitly specify the order in which these steps are executed. Constructions such as "*XXX method comprising: AAAing ...; BBBing ...; and CCCing ... .*" have been more favored recently in cases where the inventor does not wish to specify the order of the steps.

Evaluators of experiment in Section 4 For the evaluation of the translation quality of formal documents by humans, evaluators with high language skills as well as high sublanguage knowledge are needed. It was found through provisional human evaluation, that a complete understanding of both the target domain and sublanguage is necessary to make appropriate judgments. For example, to evaluate the patent abstract sublanguage, it was found essential for the evaluators to grasp the exact content of the invention in question, including an understanding of the components of the invented apparatus or method and that of the exact configuration and action created by each of the components. This is especially important for judging the appropriateness of the sentence structure, where the ability to judge syntactic appropriateness is not sufficient to arrive at the correct judgment.

5.3 Future directions

Comparison of Section 3 method with previous methods The paper has compared the translation quality resulting from the proposed method with the baseline SMT systems as well as conventional syntactic reordering. However, the proposed method should also be compared with the range of methods that focus on sentence structures. These conventional methods include the RST-based approach (Kurohashi and Nagao, 1994; Wu and Fung, 2009; Joty et al., 2013; Tu et al., 2013) [44, 72, 37, 68], the skeleton-based approach (Mellebeek et al., 2006; Xiao, 2014) [50, 74], the divide-and-translate approach (Shinmori et al., 2003; Sudo et al., 2010; Hung et al., 2012) [65, 66, 30], the pattern-based approach (Xia et al., 2004; Murakami et al., 2009; Murakami et al., 2013) [73, 55, 54], and the method based on sentence segmentation (Roh et al. 2008; Xiong et al., 2009; Jin and Liu, 2010) [64, 75, 36].

Testing effectiveness of proposed methods on other sublanguages Although it is generally accepted that formal documents tend to form sublanguages, the degree of effectiveness of the proposed method has to be experimentally evaluated for other sublanguages, because the degree and extent of sublanguage-specific characteristics may vary from sublanguage to sublanguage. For example, a wide variety of work has been carried out to improve the translation quality of legal dcouments because legal documents tend to possess sublanguage-specific characteristics (Gotti et al. 2008; Farzindar et al. 2009; Bach et al. 2010; Hung et al. 2012) [23, 16, 3, 4, 30], and hence it is expected that my proposed method will be effective to a certain extent. However, evaluation experiments are required to prove this expectation.

Comparison of translation by the two methods To compare the methodological aspects of the two experiments, the translation when applying the Section 4 method to Section 3 text data should be evaluated. However, this evaluation is not possible using the Section 4 system in its current form. First, the current Section 4 system only deals with the global reordering of input sentence structure ABC into CBA in the output sentence structure, whereas the Section 3 text data often requires the transfer of ABC sentence structure into ACB structures. Second, the Section 3 text data when the source language is Japanese sometimes requires a unification process to operate on repeated segments, such as the transfer of ABCA into CBA, which corresponds to rule R_{je2} of Figure 21 with an example sentence pair shown in Figure 24. The first transfer pattern is applicable to a wide range of sublanguages, whereas the second transfer pattern is specific to patent claim sublanguages. It is planned to add these transfer patterns to the Section 4 method, which can be achieved by just extending the current method.

Measuring usefulness of translation using the proposed method The method of Section 4 was evaluated with automatic and human evaluation, and it was found that the RIBES score seems to be a measure that is close to human evaluation. However, further consideration will be necessary to devise some evaluation metrics because there are some discrepancies between RIBES scores and human scores.

In addition, further evaluation will be required to estimate the usefulness of the translated outputs in practical usage. The further evaluation consists of two aspects of practical usage, i.e., for *assimilation* and *dissemination*. The former refers to the use of machine translation for translating foreign texts with the object of obtaining the gist of the text, and the latter refers to the use of machine translation for producing automatic translation to be post-edited by human translators. A range of evaluation methods has been proposed both for assimilation (Fuji et al. 1999; Fuji et al. 2001; Doherty et al. 2012) [17, 19, 14] and dissemination (Läubli et al. 2013) [45].

Choosing the appropriate method for a given input sentence The proposed method described in Section 3 is designed to be effective for sublanguages consisting of very long sentences with high regularity, whereas the method described in Section 4 is designed to be effective for sublanguages consisting of moderately long sentences with moderate regularity. Currently, the choice of method to be used for a given sublanguage, which depends on the degree of regularity of the sublanguage, is left to the intuition of the system user. It is anticipated, however, that the choice of methods will be automated. An idea for achieving this semi-automatic judgment would be application of the method in Section 4 to all the newly incoming sublanguage sentences and the arrival of a judgment from the repeatability of the n-grams occurring in the sentences. The sublanguages containing highly repeated n-grams may be suitable for the method of Section 3. Some judgment criterion must be devised to develop this semi-automatic judgment.

It must also be pointed out that there are other aspects of judgment that influence the selection of the methods, such as the relationship between the improvement of the handcrafting method and the cost of constructing human rules.

Introduction of inter-sentential structures All the structures employed in this research are what I call "sentence structures" that are structures within each individual sentence. Before dealing with sublanguage-specific documents or sentences, it will become necessary to determine the sublanguage of the document or sentence in question, by using inter-sentential information.

Incorporation into neural machine translation On the grounds that neural-network-based machine translation (*"neural machine translation"*) has recently been performing comparably to or even outperforming SMT, it is natural to incorporate the proposed method into neural network-based machine translation. Provisional manual comparison of a few patent claim sentences between the output of the proposed method of Section 3 and that of neural machine translation shows that the sentence structure that is appropriately handled by the proposed method is not handled appropriately by neural machine translation. In many cases some of the segments present in the input sentence are missing in the output sentence in the case of neural machine translation. Therefore, it can be concluded that the method of Section 3 performs better for the patent claim sublanguage in terms of sentence structure. It follows from this observation that the combination of SSSS transfer and neural machine translation may further improve the results obtained in Section 3. The most straightforward way of combining SSSS transfer and neural machine translation would be to use the pipeline of Section 3.4 by just replacing each reordering and SMT with neural machine translation, though the capability of the neural machine translation to create fluent translation may be somewhat impaired. More elaborate ways of combining the two methods and maximizing the capability of each method are expected to be developed in future.

References

[1] Vamshi Ambati and Wei Chen. *Cross Lingual Syntax Projection for Resource-Poor Languages*. CMU, 2007.

[2] Effie Ananiadou. The use of sublanguages in machine translation. In *Proceedings of a workshop on machine translation, UMIST, Manchester*, page Unpaged. Speech and Language Technology Club, Department of Trade and Industry, London, 1990.

[3] Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. Recognition of requisite part and effectuation part in law sentences. In *Proceedings of (ICCPOL)*, pages 29–34, 2010.

[4] Ngo Xuan Bach, NL Minh, and Akira Shimazu. Exploring contributions of words to recognition of requisite part and effectuation part in law sentences. *Proceedings of JURISIN*, pages 121–132, 2010.

[5] Beat Buchmann, Susan Warwick-Armstrong, and Patrick Shane. Design of a machine translation system for a sublanguage. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, Proceedings of COLING '84, July 2-6, 1984, Stanford University, California, USA.*, pages 334–337, 1984.

[6] Jingsheng Cai, Masao Utiyama, Eiichiro Sumita, and Yujie Zhang. Dependency-based pre-ordering for Chinese-English machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 155–160, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[7] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics, 2007.

[8] Colin Cherry and George Foster. 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*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[9] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[10] Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 531–540, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[11] Adrià de Gispert, Gonzalo Iglesias, and William Byrne. 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 (NAACL HLT 2015)*, June 2015.

[12] Michael Denkowski and Alon Lavie. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*. AMTA, 2010.

[13] Donald A. DePalma, Hélène Pielmeier, Stephen Henderson, and Robert G. Stewart. *The Language Services Market: 2016*. Common Sense Advisory, Inc., 2016.

[14] Stephen Doherty, Dorothy Kenny, and Andy Way. A user-based usability assessment of raw machine translated technical instructions. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translations in the Americas (AMTA 2012)*, 2012.

[15] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[16] Atefeh Farzindar and Guy Lapalme. Machine translation of legal information and its evaluation. In *Canadian Conference on Artificial Intelligence*, pages 64–73. Springer, 2009.

[17] Masaru Fuji. Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of MT Summit VII*, pages 285–289, 1999.

[18] Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita, and Yuji Matsumoto. Patent claim translation based on sublanguage-specific sentence structure. In *Proceedings of the 15th Machine Translation Summit*, pages 1–16, 2015.

[19] Masaru Fuji, N Hatanaka, E Ito, S Kamei, H Kumai, T Sukehiro, T Yoshimi, and Hitoshi Isahara. Evaluation method for determining groups of users who find mt useful. In *MT Summit VIII: Machine Translation in the Information Age*, pages 103–108, 2001.

[20] Michel Galley and Christopher D. Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[21] Isao Goto, Ka-Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, June 18-21, 2013*, 2013.

[22] Isao Goto, Masao Utiyama, Eiichiro Sumita, and Sadao Kurohashi. Preordering using a target-language parser via cross-language syntactic projection for statistical machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 14(3):13:1–13:23, June 2015.

[23] Fabrizio Gotti, Atefeh Farzindar, Guy Lapalme, and Elliott Macklovitch. Automatic translation of court judgments. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai*, pages 370–379, 2008.

[24] Najeh Hajlaoui and Andrei Popescu-Belis. Translating English discourse connectives into Arabic: A corpus-based analysis and an evaluation metric. In *Fourth Workshop on Computational Approaches to Arabic Script-based Languages at Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012.

[25] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *ACL (2)*, pages 690–696, 2013.

[26] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, Katsuhiko Hayashi, and Masaaki Nagata. Discriminative preordering meets Kendall's τ maximization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 139–144, Beijing, China, July 2015. Association for Computational Linguistics.

[27] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Two-stage pre-ordering for Japanese-to-English statistical machine translation. In *IJCNLP*, pages 1062–1066, 2013.

[28] Mengke Hu, David Cinciruk, and John MacLaren Walsh. Improving automated patent claim parsing: Dataset, system, and experiments. *CoRR*, abs/1605.01744, 2016.

[29] Zhongqiang Huang and Mary Harper. Self-training PCFG grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 832–841, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[30] Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. Divide and translate legal text sentence by using its logical structure. In *Proceedings of the 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems*, KICSS '12, pages 18–23, Washington, DC, USA, 2012. IEEE Computer Society.

[31] IBISWorld. *Translation Services in the US: Market Research Report*. IBISWorld, Inc., 2016.

[32] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[33] Hideki Isozaki and Natsume Kouchi. Dependency analysis of scrambled references for better evaluation of Japanese translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 450–456, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[34] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 244–251, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[35] Laura Jehl, Adrià de Gispert, Mark Hopkins, and Bill Byrne. Source-side preordering for translation using logistic regression and depth-first branch-and-bound search. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 239–248, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[36] Y. Jin and Z. Liu. Improving Chinese-English patent machine translation using sentence segmentation. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010)*, pages 1–6, Aug 2010.

[37] Shafiq R Joty, Giuseppe Carenini, Raymond T Ng, and Yashar Mehdad. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of ACL (1)*, pages 486–496, 2013.

[38] Marcin Junczys-Dowmunt and Arkadiusz Szal. SyMGiza++: A tool for parallel computation of symmetrized word alignment models. In *Proceedings of the International Multiconference on Computer Science and Information Technology - IMCSIT 2010, Wisla, Poland, 18-20 October 2010, Proceedings*, pages 397–401, 2010.

[39] Jason Katz-Brown and Michael Collins. Syntactic reordering in preprocessing for Japanese → English translation: MIT System description for NTCIR-7 patent translation task. In *NTCIR*, 2008.

[40] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395, 2004.

[41] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[42] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[43] Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *IWSLT*, pages 77–82. Citeseer, 2006.

[44] Sadao Kurohashi and Makoto Nagao. Automatic detection of discourse structure by checking surface information in sentences. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pages 1123–1127, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[45] Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 83–91, 2013.

[46] John Lehrberger. *Automatic Translation and the Concept of Sublanguage.* De Gruyter, Berlin, 1982.

[47] Uri Lerner and Slav Petrov. Source-side classifier preordering for machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, 2013.

[48] Heinz-Dirk Luckhardt. Sublanguages in machine translation. In *EACL 1991, 5th Conference of the European Chapter of the Association for Computational Linguistics, April 9-11, 1991, Congress Hall, Alexanderplatz, Berlin, Germany*, pages 306–308, 1991.

[49] Daniel Marcu, Lynn Carlson, and Maki Watanabe. The automatic translation of discourse structures. In *ANLP*, pages 9–17, 2000.

[50] Bart Mellebeek, Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. A syntactic skeleton for statistical machine translation. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, pages 195–202, 2006.

[51] Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012.

[52] Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, pages 194–203, 2011.

[53] Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005), Barcelona, Spain, December*, 2005.

[54] Jin'ichi Murakami, Isamu Fujiwara, and Masato Tokuhisa. Pattern-based statistical machine translation for NTCIR-10 PatentMT. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, June 18-21, 2013*, 2013.

[55] Jin'ichi Murakami, Masato Tokuhisa, and Satoru Ikehara. Statistical machine translation adding pattern-based machine translation in Chinese-English translation. In *Proceedings of 2009 International Workshop on Spoken Language Translation, IWSLT 2009, Tokyo, Japan, December 1-2, 2009*, pages 107–112, 2009.

[56] Tetsuji Nakagawa. Efficient top-down BTG parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 208–218, Beijing, China, July 2015. Association for Computational Linguistics.

[57] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation*, pages 1–28, 2015.

[58] Graham Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proceedings of the ACL Demonstration Track*, Sofia, Bulgaria, August 2013.

[59] Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Ckylark: A more robust PCFG-LA parser. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 41–45, Denver, Colorado, June 2015. Association for Computational Linguistics.

[60] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[61] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 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*, ACL-44, pages 433–440, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[62] Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[63] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[64] Yoon-Hyung Roh, Ki-Young Lee, Sung-Kwon Choi, Oh-Woog Kwon, and Young-Gil Kim. Recognizing coordinate structures for machine translation of English patent documents. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 460–466, The University of the Philippines Visayas Cebu College, Cebu City, Philippines, November 2008. De La Salle University, Manila, Philippines.

[65] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. Patent claim processing for readability: Structure analysis and term explanation.

In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, pages 56–65. Association for Computational Linguistics, 2003.

[66] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. 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. Association for Computational Linguistics, 2010.

[67] Katsuhito Sudoh, Jun Suzuki, Hajime Tsukada, Masaaki Nagata, Sho Hoshino, and Yusuke Miyao. NTT-NII statistical machine translation for NTCIR-10 PatentMT. In *NTCIR*, 2013.

[68] Mei Tu, Yu Zhou, and Chengqing Zong. A novel translation framework based on rhetorical structure theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–374, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[69] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In *Proceedings of the Eleventh Machine Translation Summit*, pages 475–482, 2007.

[70] Suzan Verberne, Eva D'hondt, Nelleke Oostdijk, and Cornelis Koster. Quantifying the challenges in parsing patent claims. *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010)*, 2010.

[71] The World Intellectual Property Organization (WIPO). *WIPO Patent Drafting Manual*. The World Intellectual Property Organization (WIPO), 2014.

[72] Dekai Wu and Pascale Fung. 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. Association for Computational Linguistics, 2009.

[73] Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[74] Tong Xiao, Jingbo Zhu, and Chunliang Zhang. A hybrid approach to skeleton-based translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 563–568, 2014.

[75] Hao Xiong, Wenwen Xu, Haitao Mi, Yang Liu, and Qun Liu. Sub-sentence division for tree-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 137–140, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[76] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2001.

[77] Nan Yang, Mu Li, Dongdong Zhang, and Nenghai Yu. A ranking-based approach to word reordering for statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 912–920, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[78] Xiuhong Zhang and Nianwen Xue. Extending and scaling up the Chinese treebank annotation. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 27–34, Tianjin, China, December 2012. Association for Computational Linguistics.

List of Publications

Journal Papers

1. Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita and Yuji Matsumoto. Patent claim translation based on sublanguage-specific sentence structure. *Journal of Natural Language Processing*, Vol. 23 No. 5, pp. 407-435, December 2016. (In Japanese)

2. Masaru Fuji, Masao Utiyama, Eiichiro Sumita and Yuji Matsumoto. Global preordering for improving sublanguage translation. *Journal of Natural Language Processing*, Vol. 24 No. 3, June 2017. (In Japanese; to appear)

International Conference and Workshop Papers

1. Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita and Yuji Matsumoto. Patent claim translation based on sublanguage-specific sentence structure. In *Proceedings of the Machine Translation Summit XV (MT Summit XV)*, Vol. 1: MT Researchers' Track, pp.1-16, Miami, FL, USA, October 2015.

2. Masaru Fuji, Masao Utiyama, Eiichiro Sumita, and Yuji Matsumoto. Global pre-ordering for improving sublanguage translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT 2016)*, pp. 84-93, Osaka, Japan, October 2016.

Other Publications and Talks

1. Masaru Fuji, Atsushi Fujita, Masao Utiyama and Eiichiro Sumita. Patent claim translation using bilingual patterns. In *Proceedings of the 21st Annual Meetings of the Association for Natural Language Processing*, pp. 1020-1023, March 2015. (In Japanese)

2. Masaru Fuji, Masao Utiyama and Eiichiro Sumita. Machine translation apparatus and model learning apparatus for machine translation. *Patent application to Japan Patent Office*, P2016-164707A, March 2015. (In Japanese)

3. Masaru Fuji and Masao Utiyama. Pattern learning and preprocessing apparatus for machine translation and computer program thereof. *Patent application to Japan Patent Office*, P2016-227583A, November 2016. (In Japanese)

4. Masaru Fuji. Patent Claim translation based on sublanguage-specific sentence structure. *Invited keynote talk for the AAMT/Japio Patent Translation Study Group*, December 2016, AAMT/Japio Special Interest Group on Patent Information. (In Japanese)

Appendix

A. SCFG rules for the experiments of Section 3

Figures 21, 22 and 23 illustrate the SCFG rule sets for Japanese-to-English, Chinese-to-Japanese and Japanese-to-Chinese translation directions respectively. The SCFG rule set for English-to-Japanese is shown in Figure 7 of Section 3.4.1.

ID	SCFG rules		
R_{je1}	S	$\rightarrow$	$\langle BODY_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ PREA_{\textcircled{3}},\ PREA_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{1}}\rangle$
R_{je2}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{3}}\ PREA_{\textcircled{1}},\ PREA_{\textcircled{1}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{2}}\rangle$
R_{je3}	S	$\rightarrow$	$\langle BODY_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{4}}\ PREA_{\textcircled{5}},$
			$PREA_{\textcircled{5}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{1}}\ TRAN_{\textcircled{4}}\ BODY_{\textcircled{3}}\rangle$
R_{je4}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{4}}\ TRAN_{\textcircled{5}}\ PREA_{\textcircled{1}},$
			$PREA_{\textcircled{1}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{5}}\ BODY_{\textcircled{4}}\rangle$
R_{je5}	BODY	$\rightarrow$	$\langle ELEM+,\ ELEM+\rangle$
R_{je6}	BODY	$\rightarrow$	$\langle PURP+,\ PURP+\rangle$
R_{je7}	TRAN	$\rightarrow$	⟨ “備えることを特徴とする”, “comprising:” ⟩
R_{je8}	TRAN	$\rightarrow$	⟨ “備える”, “comprising:” ⟩
R_{je9}	TRAN	$\rightarrow$	⟨ “ことを特徴とする”, “wherein:”⟩
R_{je10}	TRAN	$\rightarrow$	⟨ “する”, “wherein:”⟩
R_{je11}	TRAN	$\rightarrow$	⟨ “であって、”, “wherein:”⟩

Figure 21. SCFG rule set for Japanese-to-English translation

ID	SCFG rules		
R_{cj1}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}},\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ PREA_{\textcircled{1}}\rangle$
R_{cj2}	S	$\rightarrow$	$\langle PREA_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{4}}\ BODY_{\textcircled{5}},$
			$BODY_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{5}}\ TRAN_{\textcircled{4}}\ PREA_{\textcircled{1}}\rangle$
R_{cj3}	BODY	$\rightarrow$	$\langle ELEM+,\ ELEM+\rangle$
R_{cj4}	BODY	$\rightarrow$	$\langle PURP+,\ PURP+\rangle$
R_{cj5}	TRAN	$\rightarrow$	⟨ “包括：”, “備えることを特徴とする”⟩
R_{cj6}	TRAN	$\rightarrow$	⟨ “其中：”, “ことを特徴とする”⟩

Figure 22. SCFG rule set for Chinese-to-Japanese translation

ID	SCFG rules
R_{jc1}	S $\rightarrow$ $\langle BODY_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ PREA_{\textcircled{3}},\ PREA_{\textcircled{3}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{1}}\rangle$
R_{jc2}	S $\rightarrow$ $\langle PREA_{\textcircled{1}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{3}}\ PREA_{\textcircled{1}},\ PREA_{\textcircled{1}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{2}}\rangle$
R_{jc3}	S $\rightarrow$ $\langle BODY_{\textcircled{1}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{3}}\ TRAN_{\textcircled{4}}\ PREA_{\textcircled{5}},$ $PREA_{\textcircled{5}}\ TRAN_{\textcircled{2}}\ BODY_{\textcircled{1}}\ TRAN_{\textcircled{4}}\ BODY_{\textcircled{3}}\rangle$
R_{jc4}	S $\rightarrow$ $\langle PREA_{\textcircled{1}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{4}}\ TRAN_{\textcircled{5}}\ PREA_{\textcircled{1}},$ $PREA_{\textcircled{1}}\ TRAN_{\textcircled{3}}\ BODY_{\textcircled{2}}\ TRAN_{\textcircled{5}}\ BODY_{\textcircled{4}}\rangle$
R_{jc5}	BODY $\rightarrow$ $\langle ELEM+,\ ELEM+\rangle$
R_{jc6}	BODY $\rightarrow$ $\langle PURP+,\ PURP+\rangle$
R_{jc7}	TRAN $\rightarrow$ ⟨ “備えることを特徴とする”, “包括：” ⟩
R_{jc8}	TRAN $\rightarrow$ ⟨ “備える”, “包括：” ⟩
R_{jc9}	TRAN $\rightarrow$ ⟨ “ことを特徴とする”, “其中’⟩
R_{jc10}	TRAN $\rightarrow$ ⟨ “する”, “其中”⟩

Figure 23. SCFG rule set for Japanese-to-Chinese translation

B. Example claim sentence pair corresponding to SCFG rules

Figure 24 shows an example Japanese-to-English translation sentence pair matching Rule R_{je2} of the rule set in Figure 21. This example sentence pair illustrates an instance where the PREA segment appearing twice in the source Japanese sentence is reduced to a single occurrence of PREA segment in the target English sentence.

[$_S$ [$_{PREA}$ 安全ヘルメット] [$_{TRAN}$ であって、][$_{BODY}$ [$_{PURP}$ 接合手段が、一組の防振要素を互いに連結する単一構造を含み、] [$_{PURP}$ 前記単一構造が蜘蛛の形態であり、その蜘蛛の頭部が上部防振要素に固定され、かつその蜘蛛の各脚が、前記上部防振要素と周囲防振要素との間をつなぎ合わせるように働き、] [$_{PURP}$ 前記上部および周囲防振要素が前記単一構造にオーバーモールドすることにより得られることを特徴とする、]] [$_{PREA}$ 安全ヘルメット。]]

(a) Japanese claim sentence

[$_S$ [$_{PREA}$ A safety helmet,] [$_{TRAN}$ wherein:] [$_{BODY}$ [$_{PURP}$ the joining means comprises a single structure connecting the set of damping elements to one another,] [$_{PURP}$ wherein the single structure is in the form of a spider, the head of which is fixed to a top damping element and each leg of which performs joining between the top damping element and a peripheral damping element, and] [$_{PURP}$ wherein the top and peripheral damping elements are obtained by overmolding on the single structure.]]]

(b) Corresponding English claim sentence

Figure 24. Example Japanese-to-English claim sentence pair corresponding to SCFG rule R_{je2}