Japanese-to-English Patent Translation System based on Domain-adapted Word Segmentation and Post-ordering

Katsuhito Sudoh sudoh.katsuhito@lab.ntt.co.jp
NTT Communication Science Laboratories, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan /
Graduate School of Informatics, Kyoto University, Yoshida-hommachi, Sakyo-ku, Kyoto
606-8501, Japan

Masaaki Nagata nagata.masaaki@lab.ntt.co.jp
NTT Communication Science Laboratories, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

Shinsuke Mori mori@ar.media.kyoto-u.ac.jp
Tatsuya Kawahara kawahara@i.kyoto-u.ac.jp
Academic Center for Computing and Media Studies, Kyoto University, Yoshida-hommachi, Sakyo-ku, Kyoto 606-8501, Japan

Abstract

This paper presents a Japanese-to-English statistical machine translation system specialized for patent translation. Patents are practically useful technical documents, but their translation needs different efforts from general-purpose translation. There are two important problems in the Japanese-to-English patent translation: long distance reordering and lexical translation of many domain-specific terms. We integrated novel lexical translation of domain-specific terms with a syntax-based post-ordering framework that divides the machine translation problem into lexical translation and reordering explicitly for efficient syntax-based translation. The proposed lexical translation consists of a domain-adapted word segmentation and an unknown word transliteration. Experimental results show our system achieves better translation accuracy in BLEU and TER compared to the baseline methods.

1 Introduction

Machine translation (MT) is now widely used for various languages and fields. One typical and important use of MT is *assimilation*, understanding the content of documents written in foreign languages. Among such documents, patents are very important technical documents written in official languages of the countries to which they are filed. MT of the patents is beneficial for industrial use, such as technical surveys in other countries.

One practical problem in the patent MT is its domain dependence. Patents usually include long sentences and many technical terms in various fields, which are distinct difference from general-purpose MT. This work focuses on statistical MT (SMT) for patents, from Japanese to English. It is more difficult than English-to-Japanese in: 1) long distance reordering, and 2) word segmentation and lexical translation of domain-specific terms. The first problem is due to large syntactic differences between Japanese and English. Although the reordering in the English-to-Japanese direction can be solved effectively by very simple heuristics called Head

Finalization (Isozaki et al., 2010), the reordering in Japanese-to-English is not so straightforward. The second problem results from the Japanese orthography in which there are no explicit word boundaries. General-purpose word segmenters often fail to segment the domain-specific words and those words are translated incorrectly or remain untranslated. Some domain-specific words cannot be translated as unknown words even if they are segmented correctly, due to limited SMT training data. The first problem has been addressed by a syntax-based approach (Yamada and Knight, 2002; Galley et al., 2004; Zollmann and Venugopal, 2006), while most previous studies did not deal with the second problem. Since the lexical translation also affects the reordering based on a lexicalized reordering model and an n-gram language model, considering both problems is important for an overall SMT system. The goal of this work is to improve the Japanese-to-English patent SMT by tackling both problems at the same time. The domain-specific words have important roles in the patents and should be translated carefully for meaningful translations. We propose a novel domain adaptation method for the word segmentation, using effective features derived from a large-scale patent corpus. We also incorporate machine transliteration for the unknown Japanese words written in katakana (Japanese phonograms), bootstrapped from the parallel corpus (Sajjad et al., 2012; Sudoh et al., 2013a).

Our SMT system integrates these techniques with a post-ordering framework (Sudoh et al., 2013b), which divides the SMT problem explicitly into two sub-problems of the lexical translation and the reordering. In the post-ordering framework, the lexical translation precedes the reordering, different from pre-ordering in which the reordering precedes the lexical translation (Xia and McCord, 2004; Isozaki et al., 2010). An advantage of the post-ordering is that it is easy to integrate the domain-adapted word segmentation and the unknown word transliteration in its lexical translation step and that the reordering can use the improved lexical translation results. If we are to do the same thing in the pre-ordering, we need domain adaptation of its Japanese syntactic parser in addition to the word segmenter, and have to integrate the transliteration process with the SMT decoder as Durrani et al. (2014). Our system shows better translation accuracy in BLEU and TER than baseline methods in Japanese-to-English patent translation experiments.

2 Related Work

The patent MT between Japanese and English has been studied actively on shared tasks in NTCIR (Fujii et al., 2008, 2010; Goto et al., 2011, 2013). Recent important achievements in these studies are on the reordering problem especially in English-to-Japanese direction. Isozaki et al. (2010) proposed a very simple but effective rule-based syntactic pre-ordering method called Head Finalization. It is very effective for the long distance reordering. On the other hand, the Japanese-to-English direction is more difficult due to the lack of such simple rules. Hoshino et al. (2013) proposed an effective rule-based syntactic pre-ordering based on predicate-argument structures. Sudoh et al. (2013b) proposed a different approach called post-ordering for the Japanese-to-English patent MT, and achieved high translation performance by an efficient syntax-based translation. Our system uses the latter approach based on English syntax rather than the former one based on Japanese syntax. This is because our word segmentation adaptation can be applied directly to it without the Japanese parser adaptation as described earlier. General-purpose Japanese parsers do not work well in the patent domain, and their domain adaptation is not easy without a treebank in the patent domain. In this work, we use an English parser Enju[1] that includes a parsing model for biomedical articles and works relatively well in the patent domain. The domain adaptation of syntactic parsers is another important problem for further studies, but it is beyond the scope of this paper.

The problem of word segmentation has not been addressed in previous studies on the patent MT, and general-purpose Japanese morphological analyzers have been used in common. In Chi-

[1] http://www.nactem.ac.uk/tsujii/enju/index.html

nese, domain adaptation of the word segmentation to the patent domain has been studied (Guo et al., 2012). Their domain adaptation introduces features extracted from a large number of unlabeled patent data into a supervised word segmentation framework based on a labeled corpora in a different (newspaper) domain. They reported the improvement in word segmentation, and did not report its effect on the patent MT. Their work can be seen an application of a semi-supervised learning method (Sun and Xu, 2011) to the domain adaptation. Such an approach is appropriate for the patent domain where a huge number of patent documents are publicly available. We extend their domain adaptation by more effective and easy-to-use features, and also incorporate additional Japanese-oriented features to improve the Japanese word segmentation in the patent domain.

With respect to the relation between word segmentation and SMT, Chang et al. (2008) reported consistency and granularity of word segmentation is important in Chinese-to-English MT and modified their Chinese word segmenter to optimize the translation performance. Dyer et al. (2008) and Zhang et al. (2008) used multiple word segmentation results to overcome the problem of different word segmentation standards. Xu et al. (2008) optimized Chinese word segmentation for Chinese-to-English SMT using an extended Bayesian word segmentation method with bilingual correspondence. These studies aim to optimize word segmentation using bilingual correspondence and are different from the domain adaptation.

Machine transliteration is an important problem for translating names and other imported words (Knight and Graehl, 1998). Conventional methods need to prepare parallel transliteration pairs for training. Sajjad et al. (2012) proposed an unsupervised transliteration mining from standard parallel corpora for bootstrapping machine transliteration from the parallel corpora. Durrani et al. (2014) integrated it with a SMT framework. The transliteration process in our SMT system is a character-based SMT basically same as Durrani et al. (2014), but uses an extended transliteration mining method for Japanese compound words (Sudoh et al., 2013a).

3 System Overview

Our Japanese-to-English patent SMT is based on large-scale language resources in the patent domain. This work uses NTCIR PatentMT dataset (Goto et al., 2011, 2013) including a Japanese-English parallel corpus of 3.2 million sentences and monolingual corpora of more than 300 million sentences of Japanese and English. The parallel corpus was developed by an automatic sentence alignment over patent documents in the Japan Patent Office and the United States Patent and Trademark Office (Utiyama and Isahara, 2007). The workflow of our SMT system is illustrated in Figure 1. The translation is divided into the following four processes.

1. Japanese word segmentation using a patent-adapted word segmentation model
2. Translation into an intermediate language, Head Final English (HFE), by a monotone phrase-based SMT
3. Transliteration of untranslated Japanese katakana words (i.e. unknown words in the previous process) into English words, by a monotone phrase-based SMT in the character level
4. Post-ordering into English by a syntax-based SMT

Here, HFE is Japanese-ordered English, which was proposed by Isozaki et al. (2010) for English-to-Japanese translation. Figure 2 shows an example of a HFE sentence and corresponding English and Japanese sentences. The word order of the HFE sentence is almost the same as the Japanese one, while that of the original English sentence is different from them largely in the verb phrase. Using HFE as the intermediate language, Japanese-to-English SMT is decomposed into two sub-problems: monotone lexical translation and reordering. For these two sub-problems, we basically follow the work by Sudoh et al. (2013b). They used phrase-based

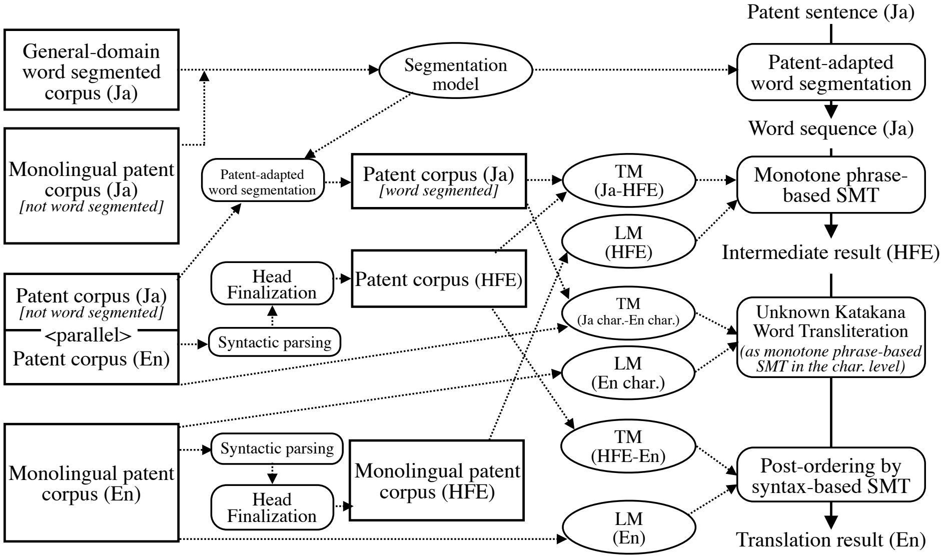
Patent sentence (Ja)
General-domain word segmented corpus (Ja)
Segmentation model
Patent-adapted word segmentation
Word sequence (Ja)
Monolingual patent corpus (Ja) [not word segmented]
Patent-adapted word segmentation
Patent corpus (Ja) [word segmented]
TM (Ja-HFE)
Monotone phrase-based SMT
LM (HFE)
Intermediate result (HFE)
Head Finalization
Patent corpus (HFE)
Patent corpus (Ja) [not word segmented]
<parallel>
Patent corpus (En)
Syntactic parsing
TM (Ja char.-En char.)
Unknown Katakana Word Transliteration (as monotone phrase-based SMT in the char. level)
LM (En char.)
Monolingual patent corpus (En)
Syntactic parsing
Head Finalization
Monolingual patent corpus (HFE)
TM (HFE-En)
Post-ordering by syntax-based SMT
LM (En)
Translation result (En)

Figure 1: Training and translation workflow by our patent-oriented Japanese-to-English SMT. HFE stands for Head Final English, Japanese-ordered English obtained by Head Finalization (Isozaki et al., 2010). TM and LM stand for a translation model and a language model.

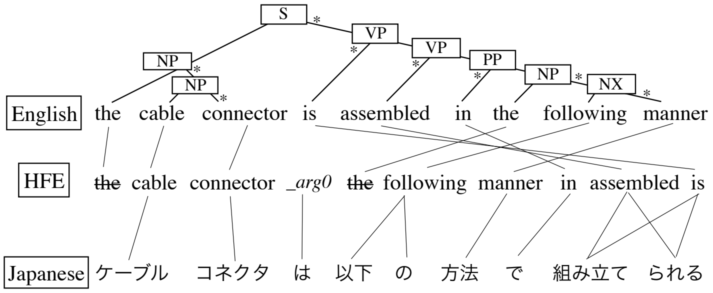
S
NP
NP
VP
VP
PP
NP
NX
English
the cable connector is assembled in the following manner
HFE
~~the~~ cable connector _arg0 ~~the~~ following manner in assembled is
Japanese
ケーブル コネクタ は 以下 の 方法 で 組み立て られる

Figure 2: Example of Head Final English (HFE) and word correspondences among English, Japanese, and HFE. Asterisk (*) indicates syntactic heads. *_arg0* is a pseudo-word for Japanese topic marker "は", inserted after subject phrases by Head Finalization rules (Isozaki et al., 2010). English articles "the" do not correspond to any Japanese words and deleted by the rules.

SMT for the first lexical translation problem to utilize its advantage on the use of phrasal contexts, and a syntax-based SMT with target language syntax for the second reordering problem to utilize its advantage on the long distance reordering. The transliteration process gives lexical translation of the unknown katakana words as the monotone character-level SMT. Katakana is often used for transcribing imported words (especially from Western languages) in Japanese, so the transliteration helps to reduce unknown words in the following HFE-to-English SMT.

The models are trained as follows. The word segmentation model is trained using a general domain labeled (word segmented) corpus and a patent unlabeled (not word segmented) corpus, by a semi-supervised learning described later in the next section. The transliteration model is trained using transliteration pairs mined from the parallel corpus by the method of Sudoh et al. (2013a). The translation models used in the monotone phrase-based SMT and syntax-based SMT are trained using the parallel corpora. Since HFE can be generated automatically by the Head Finalization rules, we can easily obtain the parallel corpus of three languages: Japanese, English, and HFE. The language models are trained using the monolingual corpora.

4 Domain Adaptation of Japanese Word Segmentation for Patents

We aim to improve the Japanese-to-English translation performance further by the word segmentation adaptation for patent-specific words and technical terms. We use the large-scale monolingual Japanese patent corpora for the domain adaptation, by the semi-supervised approach as Sun and Xu (2011) and Guo et al. (2012). There are also active learning-based supervised domain adaptation (Tsuboi et al., 2008; Neubig et al., 2011) and unsupervised word segmentation (Kempe, 1999; Kubota Ando and Lee, 2003; Goldwater et al., 2006, and many others) approaches, but the semi-supervised approach is expected to be effective; the active learning method is not easy to utilize for such large-scale corpora and the unsupervised method is not so accurate as existing supervised word segmenters.

4.1 Baseline Word Segmentation based on Conditional Random Fields

We use a character-based word segmenter based on CRFs (Peng et al., 2004; Tseng et al., 2005). It solves a character-based sequential labeling problem. In this work we employ four classes `B`, `M`, `E` (beginning/middle/end of a word), and `S` (single-character word)[2], as Sun and Xu (2011).

Our baseline features follow the work of Japanese word segmentation by Neubig et al. (2011): label bigrams, character n-grams (n=$1, 2$), and character type n-grams (n=$1, 2, 3$). We use the n-gram features within [i-2, i+2] for classifying the word at the position i. The character types are *kanji*, *katakana*, *hiragana*, digits, roman characters, and others.

4.2 Conventional Method: Word Segmentation Adaptation using Accessor Variety

Sun and Xu (2011) and Guo et al. (2012) used Accessor Variety (AV) (Feng et al., 2004) derived from unlabeled corpora as word segmentation features. AV is a word extraction criterion from un-segmented corpora, focusing on the number of distinct characters appearing around a string. The AV of a string $\boldsymbol{x}_n$ is defined as

$$AV(\boldsymbol{x}_n) = \min \left\{ AV_L(\boldsymbol{x}_n), AV_R(\boldsymbol{x}_n) \right\},$$

where $AV_L(\boldsymbol{x}_n)$ is the left AV (the number of distinct predecessor characters) and $AV_R(\boldsymbol{x}_n)$ is the right AV (the number of distinct successor characters). The AV-based word extraction is based on an intuitive assumption; *a word appears in many different context so that there is a large variation of its accessor characters.* Intuitively this assumption seems true. Figure 3 shows an example of the AV calculation for a character “前”. If the character is a word by itself, it is expected to appear in many different context so that the AV values become large by different accessor characters. This kind of information benefits the supervised word segmentation because large-scale corpora derive word boundary clues for many different character sequences that are not included in the labeled training corpus. Many technical terms in the patent domain do not appear in the general-domain labeled corpus but can be segmented using this kind of information as features in the CRF-based word segmentation. To obtain reliable AV values for many different technical terms, we need to use unlabeled corpora in the patent domain as large as possible. Here note that the AV values are frequency-based and proportional to the corpus size in general. Previous studies use several frequency classes with corresponding threshold values tuned according to the corpus, but it is not straightforward to determine appropriate classes and threshold values.

Sun and Xu (2011) used the following features based on the left and right AVs of character n-grams for classifying x_i, which imply word boundaries around x_i, as illustrated in Figure 4.

- Left AV of n-gram starting from x_i: $AV_L(x_i, ..., x_{i+n-1})$

[2] Guo et al. (2012) used six classes including `B2`, `B3` (second and third character in a word) proposed by Zhao et al. (2006) for Chinese word segmentation. This paper uses the four classes, because the six classes did not improve the word segmentation accuracy in our pilot test.

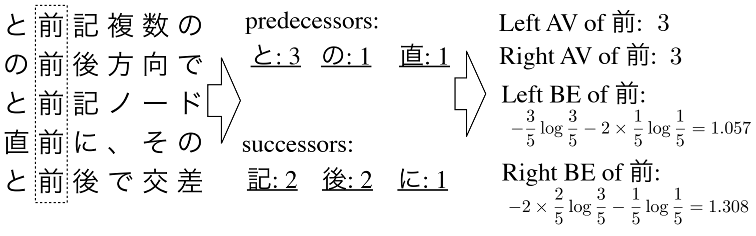

Figure 3: Example of accessor variety (AV) and branching entropy (BE) for a character “前”.

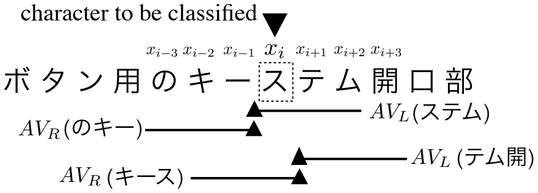
character to be classified
x_{i-3} x_{i-2} x_{i-1} x_i x_{i+1} x_{i+2} x_{i+3}
ボタン用のキーステム開口部
AV_L(ステム)
AV_R(のキー)
AV_L (テム開)
AV_R (キース)

Figure 4: Accessor variety features on the character x_i.

- Right AV of n-gram ending with x_{i-1}: $AV_R(x_{i-n}, ..., x_{i-1})$
- Left AV of n-gram starting from x_{i+1}: $AV_L(x_{i+1}, ..., x_{i+n})$
- Right AV of n-gram ending with x_i: $AV_R(x_{i-n+1}, ..., x_i)$

Sun and Xu (2011) classified the AV values into frequency classes by frequency thresholds, and used them as binary bucket features. Guo et al. (2012) used different relative frequency classes: H, M, L for top 5%, between top 5% and 20%, below top 20%. This kind of frequency-based grouping quantizes the AV values. The thresholds and the number of the classes are tuned using some held-out data (Sun and Xu, 2011) or chosen empirically (Guo et al., 2012). Such a tuning is not easy in general, especially with a large number of the classes. We follow the relative frequency classes of Guo et al. (2012) in the following experiments.

4.3 Proposed Word Segmentation Adaptation Method

We propose a word segmentation adaptation method using two additional novel types of features: branching entropy (BE) features and pseudo-dictionary (PD) features. Our semi-supervised learning framework is the same as Sun and Xu (2011) and Guo et al. (2012). The BE features are practically useful because of the probabilistic attribute of the BE, and the PD features reflect characteristics of Japanese compound words.

4.3.1 Branching Entropy Features

The BE (Jin and Tanaka-Ishii, 2006) is a different word boundary clue based on probabilistic uncertainty of accessor characters. Jin and Tanaka-Ishii (2006) used the BE for unsupervised Chinese word segmentation. Their approach is based on an intuitive assumption; *the uncertainty of successive characters is large at a word boundary.* The uncertainty of the successive character X after a given string $\boldsymbol{x}_n = x_1...x_n$ of the length n can be measured by the BE as the local conditional entropy of X with X_n instantiated:

$$H(X|\boldsymbol{X}_n = \boldsymbol{x}_n) = -\sum_{x \in V_x} P(x|\boldsymbol{x}_n) \log P(x|\boldsymbol{x}_n),$$

where $\boldsymbol{X}_n$ is the context of the length n, V_x is a set of characters. Jin and Tanaka-Ishii (2006) used the BE around character n-grams: left BE $H_L(\boldsymbol{x}_n)$ for predecessor characters and right

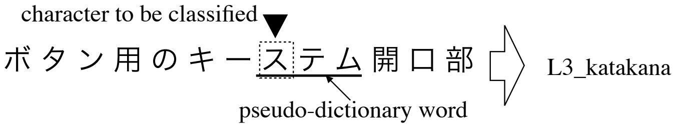

Figure 5: Example of pseudo-dictionary features.

BE $H_R(\boldsymbol{x}_n)$ for successor characters. Figure 3 also shows an example of the BE calculation. The left and right BE values are slightly different due to the different distributions of predecessor and successor characters. Even if the number of distinct accessor characters is large, the probabilistic certainty varies with their variance and is not necessarily large. Another important advantage of the BE is its probabilistic attribute. The uncertainty of accessor characters represented by a certain BE value is basically the same even for different corpus sizes, while the AV values increase with the corpus size in general.

The BE features are binary bucket features based on rounded integer values of the left and right BEs of character n-grams, similarly defined as the AV features illustrated in Figure 4. This simple quantization is motivated by the probabilistic attribute of the BE.

- Left BE of n-gram starting from x_i: $H_L(x_i, ..., x_{i+n-1})$
- Right BE of n-gram ending with x_{i-1}: $H_R(x_{i-n}, ..., x_{i-1})$
- Left BE of n-gram starting from x_{i+1}: $H_L(x_{i+1}, ..., x_{i+n})$
- Right BE of n-gram ending with x_i: $H_R(x_{i-n+1}, ..., x_i)$

4.3.2 Pseudo-dictionary Features

We additionally use Japanese-oriented heuristic word boundary clues, based on characteristics of Japanese compound words. Compound words in Japanese patents are usually written in kanji (for Japanese- or Chinese-origin words) or katakana (for imported words from Western languages). Most of their component words are also used individually and in different compound words. For example, a katakana word "ステム" (stem) is used in many compound words such as "キーステム" (key stem) and "ステムセル" (stem cell). Appearance of a distinct katakana sequence "ステム" implies word boundaries between "キー" and "ステム" and between "ステム" and "セル". Such a word boundary clue may help to identify component words appearing in different contexts. The motivation of these intuitive word boundary clues based on the character type is similar to the use of punctuations as reliable word boundaries by Sun and Xu (2011).

To include such information, we use distinct kanji and katakana sequences as pseudo-dictionary entries. The definition of the pseudo-dictionary features follows the dictionary word features used in Japanese morphological analyzer KyTea[3]: whether or not the character is in the beginning/middle/end of one of the dictionary words of a certain length. An example of the pseudo-dictionary features is shown in Figure 5. The character "ス" in the example has a feature "L3_katakana", representing the character is located at the leftmost position of a matched katakana pseudo-dictionary word of the length of three characters "ステム". We use two distinguished pseudo-dictionaries for kanji and katakana for the PD features. Short sequences whose length is shorter than 2 characters for kanji and 3 characters for katakana are excluded from the pseudo-dictionaries. The length of the long pseudo-dictionary words exceeding five characters is labeled as "5+" to mitigate feature sparseness.

[3] http://www.phontron.com/kytea/method.html

Dataset	Type	#sentences	#Ja characters	#words
BCCWJ	Labeled (Training)	53,899	1,810,675	1,242,137
BCCWJ$_{UL}$	Unlabeled	6,017,627	185,289,168	n/a
NTCIR	Unlabeled	3,191,228	214,963,715	n/a
PatentJP	Unlabeled	537,494,485	42,175,165,488	n/a
Test$_{Patent}$	Labeled (Test)	2,000	127,825	81,481
Test$_{BCCWJ}$	Labeled (Test)	6,406	201,080	135,664

Table 1: Corpus statistics for word segmentation experiments.

5 Evaluation

We conducted experiments using the NTCIR PatentMT data to investigate the performance of our Japanese-to-English patent SMT system. We evaluated the word segmentation itself in addition to the overall Japanese-to-English translation, to see the effect of the proposed word segmentation adaptation on the translation results.

5.1 Evaluation of Word Segmentation

We evaluated the word segmentation accuracy by the proposed word segmentation adaptation and compared it with those by other methods.

5.1.1 Setup

We implemented a word segmenter using the features described in the previous section, with CRFsuite[4] and its default hyperparameters. We used CORE data of Balanced Corpus of Comtemporary Written Japanese (Maekawa, 2007) as the labeled general domain corpus for training the word segmentation model[5], and split them for training (BCCWJ) and test (Test$_{BCCWJ}$) sets by about 9:1. For unlabeled corpus, we used its non-CORE portion (BCCWJ$_{UL}$), the Japanese portion of NTCIR-9 PatentMT (Goto et al., 2011) Japanese-English bitext (NTCIR), and Japanese monolingual patent corpus provided for NTCIR-9 PatentMT (PatentJP). The test set in the patent domain (Test$_{Patent}$) was in-house 2,000 sentences in which the word segmentation was manually annotated by the same word segmentation standards as the other labeled data. Corpus statistics are shown in Table 1.

5.1.2 Compared Methods

We compared the following word segmentation features in the word segmentation experiments.

- Baseline: only the baseline features described in 4.1
- +AV: the AV (n=2,3,4,5) and baseline features
- +BE: the BE (n=1,2,3,4,5) and baseline features
- +PD: the PD and baseline features
- +BE +PD: the BE (n=1,2,3,4,5), PD, and baseline features

To investigate the impact of the unlabeled corpus size in the semi-supervised approach, we compared two different conditions, mid-scale and large-scale; BCCWJ$_{UL}$ and NTCIR were used in the mid-scale condition, and BCCWJ$_{UL}$ and PatentJP[6] in the large-scale condition. Here, the pseudo-dictionaries of kanji and katakana sequences were composed by kanji and katakana sequences found in the unlabeled data. We also compared the word segmenters with a publicly available word segmenter MeCab[7] with a Japanese dictionary UniDic[8], for reference.

[4] http://www.chokkan.org/software/crfsuite/

[5] We replaced kanji numbers with digits for consistency with the patent corpus.

[6] The patent sentences in NTCIR is also included in PatentJP.

[7] https://code.google.com/p/mecab/

[8] http://sourceforge.jp/projects/unidic/

Condition	Feature	Patent		BCCWJ
		F-measure	OOV Recall	F-measure
Labeled	Baseline	96.87	87.94	97.85
Unlabeled (Mid-scale)	+AV	^{L}98.08	91.25	98.27
	+BE	^{A}98.25	91.58	98.38
	+PD	^{L}97.85	91.18	98.08
	+BE +PD	^{A,B}98.32	92.09	98.39
Unlabeled (Large-scale)	+AV	^{P}97.80	90.79	98.26
	+BE	^{A,P,M}98.34	91.62	98.33
	+PD	97.12	89.32	98.36
	+BE +PD	^{A,P}98.32	92.33	98.37
	+BE +PD$_m$	^{A,P,M}**98.36**	**92.61**	98.37
MeCab		97.73	86.94	98.35

Table 2: Word segmentation F-measures (%) for the patent and original domains and OOV recalls (%) in the patent domain. ^{A}, ^{B}, and ^{P} indicate significantly better results than +AV, +BE, +PD (in the same group), ^{M} and ^{L} indicate significantly better results than mid- and large-scale. PD$_m$ means the PD features derived from the mid-scale unlabeled corpora.

5.1.3 Results

Table 2 shows word segmentation results in F-measures in the patent and general (BCCWJ) domains, and recalls of out-of-vocabulary words (OOV recall) in the patent domain focusing on domain-specific words not included in the general domain corpus. All the additional features showed better results in the patent domain than the baseline features and MeCab, which were statistically significant (p=0.05) by bootstrap resampling tests.

The AV and BE features helped to outperform MeCab in the patent domain especially in the OOV recall while the baseline performance was much worse. The BE features worked consistently with the different corpus sizes. The AV features with the large-scale data showed obviously worse results than with the mid-scale data; this indicates instability of the AV features with different corpus sizes. The PD features showed good performance especially in OOV recall, but those from the large-scale corpora did not work so well. This is possibly due to inappropriate pseudo-dictionary entries extracted around typographical errors, which sometimes occur between characters with similar type faces. Thus we additionally tested the combination of the PD features from the mid-scale data and the large-scale BE features, and that showed the best results. This indicates our domain adaptation is very effective for domain-specific words.

5.2 Evaluation of Translation

We finally conducted MT experiments to investigate the performance of our patent SMT system and the effect of each technique.

5.2.1 Setup

We used Japanese-to-English patent translation dataset used in NTCIR-9 (Goto et al., 2011) and NTCIR-10 (Goto et al., 2013) PatentMT. They shared the same training and development sets and used different test sets. Corpus statistics are shown in Table 3. English sentences were tokenized and parsed by an English syntactic parser Enju with its “GENIA” models for biomedical articles, and then lowercased. Japanese sentences were tokenized by the different tokenizers described later. Here in the training set, long sentences exceeding 64 words in either Japanese or English were filtered out. For the long sentence filtering, we used the segmentation results by KyTea because it is based on a short word unit and resulted in the largest number of segmented words. Note that the sentence set was the same for all Japanese segmenters.

Tokenizer	Training (2,862,022 sents.)	Development (2,000 sents.)	Test9 (2,000 sents.)	Test10 (2,300 sents.)
Proposed	95,465,533	75,020	75,962	101,309
Baseline	94,914,460	74,627	75,504	100,589
KyTea	101,718,532	80.025	80,842	107,405
MeCab	93,030,977	73,263	74,066	99,163
JUMAN	91,052,206	71,707	72,515	97,205
English	88,192,234	68,854	69,806	94,906

Table 3: Corpus statistics in the number of words for translation experiments.

The Japanese-to-HFE monotone PBMT was implemented with Moses and trained using Japanese-HFE parallel sentences. The distortion limit of the PBMT was set to zero, but a standard lexicalized reordering model (wbe-msd-bidirectional-fe) was used to constrain adjacent phrase translations. The HFE-to-English SAMT was implemented with Moses-chart and trained using the HFE sentences and the corresponding English parse trees. Its reordering parameter max-chart-span was set to 200 to allow arbitrary distance reordering for accurate Japanse-to-English translation[9]. The search space parameter cube-pruning-pop-limit was set to 32 for efficiency, according to Sudoh et al. (2013b). Their language models were word 6-gram models trained using a large-scale English patent corpus with more than 300 million sentences. Model weights were optimized in BLEU (Papineni et al., 2002) using Minimum Error Rate Training (MERT) (Och, 2003). We chose the best weights among ten individual runs of MERT.

The katakana transliteration was implemented as a Moses-based monotone PBMT in the character level, trained using transliteration pairs mined from the Japanese-English phrase table entries whose Japanese part consisted of katakana only. Its character-level language model was character 9-gram models trained using the large-scale English patent corpus which is used for the word-level language models described above. It was used to replace katakana words remained in the intermediate results in HFE with their transliteration results.

5.2.2 Compared Methods

We compared following segmenters for the translation experiments.

- Baseline: the baseline word segmenter same as the word segmentation experiments above
- Proposed: the patent-adapted segmenter using the labeled general-domain corpus and the large-scale unlabeled patent corpus with the BE and PD features
- KyTea, MeCab, and JUMAN[10]: publicly available Japanese morphological analyzers

We also compared the results by the post-ordering with those by standard SAMT and PBMT. The search space parameters of the standard SAMT were set to the same value as the HFE-to-English SAMT, to compare the performance with similar computation time[11].

5.2.3 Results and Discussion

Table 4 shows the translation performance in BLEU and TER (Snover et al., 2006) with the results of statistical significance tests (p=0.05) by bootstrap resampling (Koehn, 2004), in which our overall system resulted in the best. The table also shows the results of intermediate Japanese-to-HFE translation. The advantage of our system can be attributed to three techniques included in the system: domain adaption of word segmentation, katakana unknown word transliteration, and post-ordering.

[9] It exceeded the maximum sentence length in the development and test sets.

[10] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[11] Actually the post-ordering needs the time for the first monotone PBMT but it ran very fast and did not affect so much (Sudoh et al., 2013b).

System	Ja word	Test9		Test10	
	segmenter	BLEU (%)	TER (%)	BLEU (%)	TER (%)
Overall system (Ja-HFE monotone PBMT + transliteration + HFE-En SAMT)	Proposed	**34.77**	$^{+}$**51.86**	**35.75**	**50.71**
	Baseline	$^{+}$*34.29	$^{+}$*52.16	$^{+}$*35.21	$^{+}$50.90
	KyTea	$^{+}$*34.42	$^{+}$*52.30	*35.37	*51.38
	MeCab	$^{+}$*34.52	$^{+}$*52.21	$^{+}$*35.41	$^{+}$*50.92
	JUMAN	$^{+}$34.59	$^{+}$52.10	$^{+}$*35.41	$^{+}$*51.00
Post-ordering (Ja-HFE monotone PBMT + HFE-En SAMT)	Proposed	**34.75**	**51.90**	**35.71**	**50.71**
	Baseline	*34.18	*52.30	*35.14	50.96
	KyTea	*34.32	*52.40	*35.33	*51.41
	MeCab	*34.33	*52.35	*35.28	*51.03
	JUMAN	*34.50	52.19	*35.35	*51.07
SAMT (efficiency-oriented)	MeCab	*33.11	*53.26	*33.67	*52.19
PBMT (distortion limit=12)	MeCab	*31.96	*55.04	*33.06	*53.77
Ja-HFE monotone PBMT	Proposed	**35.87**	**49.05**	**37.11**	**48.19**
	Baseline	*35.30	*49.46	*36.46	*48.58
	KyTea	*35.50	*49.77	*36.66	*48.86
	MeCab	*35.45	*49.57	*36.71	*48.54
	JUMAN	35.70	*49.42	*36.65	*48.73

Table 4: Results of overall Japanese-to-English translation and intermediate Japanese-to-HFE translation in BLEU and TER. $^{+}$ indicates the difference from the results without transliteration is statistically significant. * indicates the difference from Proposed in the same group is statistically significant.

First, the post-ordering contributed the largest and significant improvements compared with the standard SAMT and PBMT, by about 1-2 points in BLEU and 2-3 points in TER. They basically followed the results by Sudoh et al. (2013b).

Second, the proposed word segmentation showed significant improvements in most cases, by the better intermediate translation results shown at the bottom of Table 4. Although the absolute improvement was not so large, the domain adaptation worked consistently. These results suggest that the domain adaptation of word segmentation actually worked for the patent SMT. We also analyzed the advantage of the patent-adapted word segmentation by the number of unknown words in translation. Table 5 shows the numbers of unknown kanji and katakana words that were not translated in the monotone PBMT, by the five word segmenters in the experiments. These values reflect the consistency and granularity problem in word segmentation (Chang et al., 2008). If the word segmentation is consistent and have relatively small granularity (choosing shorter words), the number of the unknown words becomes small. The granularity is closely related to the problem of compound words in this work; the translation of compound words becomes easy if they are segmented to short and appropriate component words. MeCab and JUMAN are dictionary-based word segmenters that have an advantage on precise segmentation of in-vocabulary words. JUMAN used a large-scale dictionary collected from web texts covering many domain-specific words, and resulted in a smaller number of unknown words than MeCab. KyTea and this paper's segmenter are character-based ones that have an advantage on identifying out-of-vocabulary words (as shown in Table 2). KyTea worked well on kanji words, but derived a large number of katakana unknown words. It was probably due to the difference of embedded information between ideogram (kanji) and phonogram (katakana). Katakana compound words are usually difficult to segment only by their poor character-based information. The proposed method used reliable word boundary clues derived from the large-scale corpora and achieved consistent word segmentation of katakana compound words with

Ja word segmenter	test9		test10	
	kanji	katakana	kanji	katakana
Proposed	18 (18)	30 (20)	29 (23)	34 (20)
Baseline	54 (43)	87 (59)	98 (71)	101 (59)
KyTea	10 (10)	108 (79)	14 (14)	132 (78)
MeCab	48 (39)	68 (50)	100 (73)	87 (55)
JUMAN	2 (2)	48 (41)	9 (9)	71 (45)

Table 5: Statistics of unknown kanji and katakana words (non-translated words by monotone PBMT). The numbers in parentheses are the number of unique unknown words.

Test9	Test10
53.33 (16/30)	59.37 (19/32)

Table 6: Transliteration accuracy in sample-wise correctness (ACC) in the proposed system.

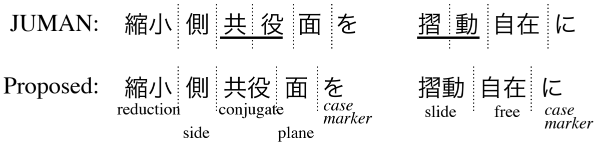

Figure 6: Examples of small granularity segmentation for out-of-vocabulary words by JUMAN.

more appropriate granularity than others, as suggested by the smallest number of katakana unknown words in Table 5. Such an advantage was not found in kanji words compared to KyTea and JUMAN. However, JUMAN tended to choose small granularity segmentations for out-of-vocabulary words as shown in the examples in Figure 6, so these results may not indicate directly the disadvantage of the proposed method.

Finally, the transliteration itself did not improve BLEU and TER significantly in our system, although some significant improvements were found in the results by the other segmenters because of their many unknown katakana words. Its effect was limited only on the unknown katakana words and their context words (related to the word n-gram language model and the post-ordering) and did not contribute well to BLEU and TER with a small number of the unknown katakana words. We analyzed the transliteration accuracy in the intermediate HFE results with the transliteration as shown in Table 6. About a half of the unknown katakana words were transliterated correctly. This improvement is practically important for the assimilation.

6 Conclusion

This paper presented our Japanese-to-English SMT system specialized for patent translation, including the effective word segmentation by our domain adaptation method, the unknown katakana word transliteration, and the efficient syntax-based post-ordering. We achieved better translation performance by the system than by other existing Japanese word segmenters and standard SMT methods.

Domain adaptation is also expected to be effective in other components such as syntactic parsing, translation models, and language models. The application of monolingual and bilingual knowledge of other domains to the patent domain especially for named entities, and the use of patent MT knowledge in MT for other domains are also practically important.

References

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio. Association for Computational Linguistics.

Durrani, N., Sajjad, H., Hoang, H., and Koehn, P. (2014). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden. Association for Computational Linguistics.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing Word Lattice Translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio. Association for Computational Linguistics.

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1):75–93.

Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T. (2008). Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of the NTCIR-7 Workshop Meeting*, pages 389–400.

Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., and Shimohata, S. (2010). Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proceedings of the NTCIR-8 Workshop Meeting*, pages 371–376.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA. Association for Computational Linguistics.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual Dependencies in Unsupervised Word Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia. Association for Computational Linguistics.

Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K. (2011). Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 559–578.

Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K. (2013). Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260–286.

Guo, Z., Zhang, Y., Su, C., and Xu, J. (2012). Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation. In *Proceedings of the 1st CCF Conference on Natural Language Processing & Chinese Computing*, pages 121–131.

Hoshino, S., Miyao, Y., Sudoh, K., and Nagata, M. (2013). Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1062–1066, Nagoya, Japan. Asian Federation of Natural Language Processing.

Isozaki, H., Sudoh, K., Tsukada, H., and Duh, K. (2010). Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251, Uppsala, Sweden. Association for Computational Linguistics.

Jin, Z. and Tanaka-Ishii, K. (2006). Unsupervised Segmentation of Chinese Text by Use of Branching Entropy. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 428–435, Sydney, Australia. Association for Computational Linguistics.

Kempe, A. (1999). Experiments in Unsupervised Entropy-Based Corpus Segmentation. In *Proceedings of Computational Natural Language Learning (CoNLL-99)*.

Knight, K. and Graehl, J. (1998). Machine Transliteration. *Computational Linguistics*, 24(4):599–612.

Koehn, P. (2004). Statistical Significance Test for Machine Translation Evaluation. In *Proceedinsg of EMNLP*, pages 388–395.

Kubota Ando, R. and Lee, L. (2003). Mostly-unsupervised statistical segmentation of Japanese kanji sequences. *Natural Language Engineering*, 9(2):127–149.

Maekawa, K. (2007). Design of a Balanced Corpus of Contemporary Written Japanese. In *Proceedings of Symposium on Large-Scale Knowledge Resources (LKR2007)*, pages 55–58.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

Peng, F., Feng, F., and McCallum, A. (2004). Chinese Segmentation and New Word Detection using Conditional Random Fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland. COLING.

Sajjad, H., Fraser, A., and Schmid, H. (2012). A Statistical Model for Unsupervised and Semi-supervised Transliteration Mining. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 469–477, Jeju Island, Korea. Association for Computational Linguistics.

Snover, M., Dorr, B., and Schwartz, R. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*.

Sudoh, K., Mori, S., and Nagata, M. (2013a). Noise-Aware Character Alignment for Bootstrapping Statistical Machine Transliteration from Bilingual Corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 204–209, Seattle, Washington, USA. Association for Computational Linguistics.

Sudoh, K., Wu, X., Duh, K., Tsukada, H., and Nagata, M. (2013b). Syntax-Based Post-Ordering for Efficient Japanese-to-English Translation. *ACM Transactions on Asian Language Information Processing*, 12(3).

Sun, W. and Xu, J. (2011). Enhancing Chinese Word Segmentation Using Unlabeled Data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 970–979, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., and Matsumoto, Y. (2008). Training Conditional Random Fields Using Incomplete Annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 897–904, Manchester, UK. Coling 2008 Organizing Committee.

Utiyama, M. and Isahara, H. (2007). A Japanese-English Patent Parallel Corpus. In *Proceedings of the 11th Machine Translation Summit*, pages 475–482.

Xia, F. and McCord, M. (2004). Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland. COLING.

Xu, J., Gao, J., Toutanova, K., and Ney, H. (2008). Bayesian Semi-Supervised Chinese Word Segmentation for Statistical Machine Translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK. Coling 2008 Organizing Committee.

Yamada, K. and Knight, K. (2002). A Decoder for Syntax-based Statistical MT. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 303–310, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhang, R., Yasuda, K., and Sumita, E. (2008). Improved Statistical Machine Translation by Multiple Chinese Word Segmentation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 216–223, Columbus, Ohio. Association for Computational Linguistics.

Zhao, H., Huang, C.-N., Li, M., and Lu, B.-L. (2006). Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 87–94.

Zollmann, A. and Venugopal, A. (2006). Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City. Association for Computational Linguistics.