TRANSLATING LEGAL SENTENCE BY SEGMENTATION AND RULE SELECTION

Bui Thanh Hung[1], Nguyen Le Minh[2] and Akira Shimazu[3]

Graduate School of Information Science,
Japan Advanced Institute of Science and Technology, Japan
[1] hungbt@jaist.ac.jp ,[2] nguyenml@jaist.ac.jp ,[3] shimazu@jaist.ac.jp

ABSTRACT

A legal text usually long and complicated, it has some characteristic that make it different from other daily-use texts.Then, translating a legal text is generally considered to be difficult. This paper introduces an approach to split a legal sentence based on its logical structure and presents selecting appropriate translation rules to improve phrase reordering of legal translation. We use a method which divides a English legal sentence based on its logical structure to segment the legal sentence. New features with rich linguistic and contextual information of split sentences are proposed to rule selection. We apply maximum entropy to combine rich linguistic and contextual information and integrate the features of split sentences into the legal translation, tree-based SMT system. We obtain improvements in performance for legal translation from English to Japanese over Moses and Moses-chart systems.

KEYWORDS

Legal Translation, Logical Structure of a Legal Sentence, Phrase Reordering, CRFs, Maximum Entropy Model, Linguistic and Contextual Information

1. INTRODUCTION

Legal translation is the task of how to translate texts within the field of law. Translating legal texts automatically is one of the difficult tasks because legal translation requires exact precision, authenticity and a deep understanding of law systems. Because of the meticulous nature of the composition (by experts), sentences in legal texts are usually long and complicated. When we translate long sentences, parsing accuracy will be lower as the length of sentence grows. It will inevitably hurt the translation quality and decoding on long sentences will be time consuming, especially for forest approaches. So splitting long sentences into sub-sentences becomes a natural way to improve machine translation quality.

A legal sentence represents a requisite and its effectuation [10], [17], [23]. Dividing a sentence into shorter parts and translating them has a possibility to improve the quality of translation. For a legal sentence with the requisite-effectuation structure (logical structure), dividing a sentence into requisite-and-effectuation parts is simpler than dividing the sentence into its clauses because such legal sentences have specific linguistic expressions that are useful for dividing. We first recognize the logical structure of a legal sentence using statistical learning model with linguistic information. Then we segment a legal sentence into parts of its structure and apply rule selection to translate them with statistical machine translation (SMT) models.

In the phrase-based model [11], phrase reordering is a great problem because the target phrase order differs significantly from the source phrase order for several language pairs such as English-Japanese. Linguistic and contextual information have been widely used to improve translation

DOI : 10.5121/ijnlc.2013.2403

performance. It is helpful to reduce ambiguity, thus guiding the decoder to choose correct translation for a source text on phrase reordering. In this paper, we focus on selecting appropriate translation rules to improve phrase reordering for the tree-based statistical machine translation, the model operates on synchronous context free grammars (SCFG) (Chiang [4], [5]). SCFG rules for translation are represented by using terminal (words or phrases), non-terminals and structural information. SCFG consists of a left-hand-side (LHS) and a right-hand-side (RHS). Generally, there are cases that a source sentence pattern-matches with multiple rules which produce quite different phrase reordering as the following example:

$$R \rightarrow (at\ X_1\ of X_2 \mid X_2\ の\ X_1)$$
$$R \rightarrow (at X_1\ of X_2 \mid で X_1\ の\ X_2)$$
$$R \rightarrow (at X_1\ of X_2 \mid に X_1\ の X_2)$$

During decoding, without considering linguistic and contextual information for both nonterminals and terminals, the decoder may make errors on phrase reordering caused by inappropriate translation rules. So rule selection is important to tree-based statistical machine translation systems. This is because a rule contains not only terminals (words or phrases), but also nonterminals and structural information. During decoding, when a rule is selected and applied to a source text, both lexical translations (for terminals) and reorderings (for nonterminals) are determined. Therefore, rule selection affects both lexical translation and phrase reorderings.
We propose translating split sentence based on the logical structure of a legal sentence and rule selection for legal translation specifically:

- We divide a legal sentence based on its logical structure into the first step
- We apply a statistical learning method - Conditional Random Fields (CRFs) with rich linguistic information to recognize the logical structure of a legal sentence and the logical structure of a legal sentence is adopted to divide the sentence.
- We use a maximum entropy-based rule selection model for tree-based English-Japanese statistical machine translation in legal domain. The maximum entropy-based rule selection model combines local contextual information around rules and information of sub-trees covered by variables in rules.
- We propose using rich linguistic and contextual information of split sentences for both non-terminals and terminals to select appropriate translation rules.
- We obtain substantial improvements by BLEU over the Moses and Moses-chart baseline system.

This paper is organized as follows. Section II gives related works. Section III describes our Method : how to segment legal sentence to the logical structure and how to use rich linguistic and contextual information of split sentences for rule selection. The experiment results are discussed in Section IV, and the conclusion and future work are followed in Section V.

2. PREVIOUS WORK

Machine translation can work well for simple sentences but a machine translation system faces difficulty while translating long sentences, as a result the performance of the system degrades.
Most legal sentences are long and complex, the translation model has a higher probability to fail in the analysis, and produces poor translation results. One possible way to overcome this problem is to divide long sentences to smaller units which can be translated separately. There are several approaches on splitting long sentences into smaller segments in order to improve the translation.

Xiong et al., [25] used Maximum Entropy Markov Models to learn the translation boundaries based on word alignments in hierarchical trees. They integrated soft constraints with beginning

and ending translation boundaries into the log linear model decoder. They proposed a new feature: translation boundary violation counting to prefer the hypotheses that are consistent with the translation boundaries.
Sudoh et al., [22] proposed dividing a source sentence into smaller clauses using a syntactic parser. They used a non-terminal symbol served as a place-holder for a relative clause and trained a clause translation model with these symbols. They proposed a clause alignment method using a graph-based method to build the non-terminal corpus. Their model can perform short and long distance reordering simultaneously.
Goh et al., [7] proposed rule-based method for splitting the long sentences using linguistic information and translated the sentences with the split boundaries. They used two types of Constraints : split condition and block condition with “zone” and “wall” markers in Moses.

Each of these approaches has its strength and weakness in application to sentence partitioning. However, in order to develop a system for splitting legal sentences, dividing a legal sentence based on its logical structure is preferable. Dividing a sentence into requisite-and-effectuation parts (logical structure) is simpler than dividing the sentence into its clauses because such legal sentences have specific linguistic expressions that are useful for dividing.

Our approach is different from those of previous works. We apply the logical structure of a legal sentence to split legal sentences. We use characteristics and linguistic information of legal texts to split legal sentences into logical structures. Bach et al., [1] used Conditional Random Fields (CRFs) to recognize the logical structure of a Japanese legal sentence. We use the same way as in [1], [9] to recognize the logical structure of a English legal sentence. We propose new features to recognize its logical structure. The logical structure of a legal sentence by the recognition task will be used to split long sentences. Our approach is useful for legal translation.It will reserve a legal sentence structure, reduce the analysis in deciding the correct syntactic structure of a sentence, remove ambiguous cases in advanced and promise results.

Linguistic and contextual information were used previously. They are very helpful for SMT system. There are many works using them to solve the selection problem in SMT.Carpuat and Wu, [2] integrated word-sense-disambiguation (WSD) and phrase-sensedisambiguation (PSD) into a phrased-based SMT system to solve the lexical ambiguity problem.Chan et al., [3] incorporated a WSD system into the hierarchical SMT system, focusing on solving ambiguity for terminals of translation rules. He et al., Liu et al., extended WSD like the approach proposed in [2] to hierarchical decoders and incorporated the MERS model into a state-of-the-art syntax-based SMT model, the tree-to-string alignment template model [8], [16].Chiang et al. [6], used 11,001 features for statistical machine translation.

In this paper, we propose using rich linguistic and contextual information for English-Japanese legal translation, specifically:

- We recognize the logical structure of a legal sentence and divide the legal sentence based on its logical structure as the first step.
- We use rich linguistic and contextual information for both non-terminals and terminals.Linguistic and contextual information around terminals have never been used before,we see that these new features are very useful for selecting appropriate translation rules if we integrate them with the features of non-terminals.
- We propose a simple and sufficient algorithm for extracting features in rule selection.
- We classify features by using maximum entropy-based rule selection model and incorporate this model into a state-of-the-art syntax-based SMT model, the tree-based model (Moses-chart).

- Our proposed method can achieve better results for English-Japanese legal translation based on the BLEU scores.

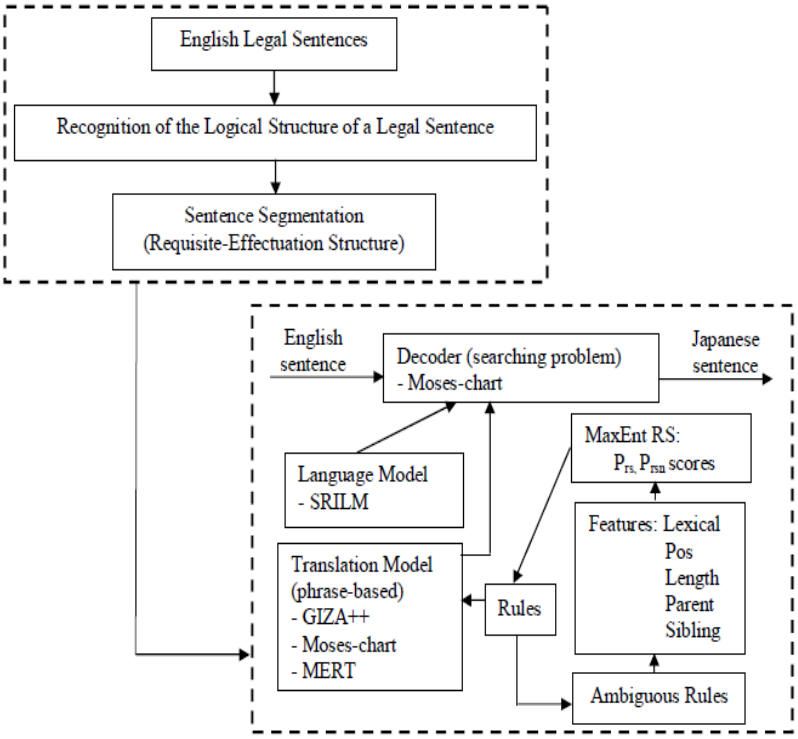


Figure 1. The diagram of our proposed method

3. PROPOSED METHOD

The method of translating legal sentence by segmentation and rule selection follows in two steps:

- Legal sentence segmentation
- Rule selection for legal translation

The diagram of our proposed method is shown in Fig. 1.

3.1. Legal sentence segmentation

To segment legal sentence to its structure, at the first we recognize the logical structure of legal sentence. Most law sentences are the implication and the logical structure of a sentence defining a term is the equivalence type. An implication law sentence consists of a law requisite part and a law effectuation part which designate the legal logical structure described in [10], [17], [23]. Structures of a sentence in terms of these parts are shown in Fig. 2.

The requisite part and the effectuation part of a legal sentence are generally composed from three parts: a topic part, an antecedent part and a consequent part. In a legal sentence, the partusually describes a law provision, and the antecedent part describes cases in which the law provision can be applied. The topic part describes a subject which is related to the law provision.There are four cases (illustrated in Fig. 2) basing on where the topic part depends on: case 0 (no

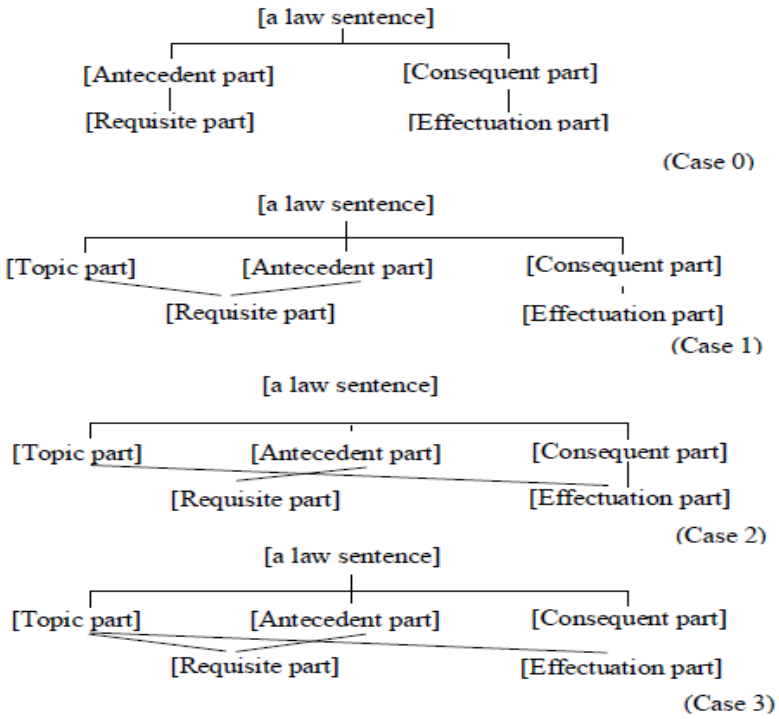


Figure 2. Four cases of the logical structure of a legal sentence

topic part), case 1 (the topic part depends on the antecedent part), case 2 (the topic part depends on the consequent part), and case 3 (the topic part depends on both the antecedent and the consequent parts). Let us show examples of four cases of the logical structure of a legal sentence. The annotation in these examples and in the test corpus was carried out by a person who was an officer of the Japanese government, and persons who were students of a graduate law school and a law school.

- Case 0:

<A> When a period of an insured is calculated, </A>
<C> it is based on a month. </C>

- Case 1:

<T1> For the person </T1>
<A> who is qualified for the insured after s/he was disqualified, </A>
<C> the terms of the insured are added up together. </C>

- Case 2:

<T2> For the amount of the pension by this law, </T2>
<A> when there is a remarkable change in the living standard of the nation of the other situation, </A>
<C> a revision of the amount of the pension must be taken action promptly to meet the situations. </C>

- Case 3:

<T3> For the Government, </T3>
<A> when it makes a present state and a perspective of the finance, </A>
<C> it must announce it officially without delay. </C>

In these examples, A refers to the antecedent part, C refers to the consequent part, and T1, T2, T3 refer to the topic parts which correspond to case 1, case 2, and case 3.
We use sequence learning model described in [13], [15] to recognize the logical structure of a legal sentence. We model the structure recognition task as a sequence labeling problem, in which each sentence is a sequence of words. We consider implication types of legal sentences,and five kinds of logical parts for the recognition of the logical structure of a legal sentence, as follows:

- Antecedent part (A)
- Consequent part (C)
- Topic part T1 (correspond to case 1)
- Topic part T2 (correspond to case 2)
- Topic part T3 (correspond to case 3)

in the IOB notation [13], [15], we will have 11 kinds of tags: B-A, I-A, B-C, I-C, B-T1, I-T1,BT2, I-T2, B-T3, I-T3 and O (used for an element not included in any part). For example, an element with tag B-A begins an antecedent part, while an element with tag B-C begins a consequent part.
We use Conditional Random Fields (CRFs) [13], [15] as a learning method because the recognition task of the logical structure of a legal sentence can be considered as a sequence learning problem, and CRFs is an efficient and powerful framework for sequence learning tasks.
We propose some new features to recognize the logical structure of a English legal sentence based on its characteristics and linguistic information. We designed a set of features:

- Word form: phonological or orthographic appearance of a word in a sentence.
- Chunking tag: tag of syntactically correlated parts of words in a sentence.
- Part-of-Speech features: POS tags of the words in a sentence
- The number of particular linguistic elements which appear in a sentence as follows:
  + Relative pronouns (e.g, where, who, whom, whose, that)
  + Punctuation marks (. , ; :)
  + Verb phrase chunks
  + Relative phrase chunks
  + Quotes

We parse the individual English sentences by Stanford parser [20] and use CRFs tool [13] for sequence learning tasks. Experiments were conducted in the English-Japanese translation corpus. We collected the corpus using Japanese Law Translation Database System (available at http://www.japaneselawtranslation.go.jp/). The corpus contains 516 sentences pairs. Table 1 shows statistics on the number of logical parts of each type.

We divided the corpus into 10 sets, and conducted 10-fold cross-validation tests for recognizing logical structures of the sentences in the corpus. We evaluated the performance by *precision*, *recall*, and F1 scores as:

$$precision = \frac{\#\ correct\ \ parts}{\#\ predicted\ \ parts}$$

$$recall = \frac{\#\ correct\ \ parts}{\#\ actual\ \ parts}$$

$$F_1 = \frac{2\ *\ precision\ \ *\ recall}{precision\ \ +\ recall}$$

Experimental results on the corpus are described in Table 2.

Table 1. Statistics on logical parts of the corpus

Logical Part					
C	A	T1	T2	T3	Total
576	561	0	4	130	1271

Table 2. Experimental results for recognition of the logical structure of a legal sentence

	Precision (%)	Recall (%)	F1 (%)
C	84.89	82.85	83.86
A	86.03	85.03	85.56
T2	89.10	85.64	87.34
T3	78.08	56.12	65.30
Overall	84.64	83.24	83.93

✓ Case 0:
<A> When a period of an insured is calculated, </A>
<C> it is based on a month. </C>

The sentence will be split as:
[When a period of an insured is calculated]
[it is based on a month.]

✓ Case 1:
<T1> For the person </T1>
<A> who is qualified for the insured after s/he was disqualified, </A>
<C> the terms of the insured are added up together. </C>

The sentence will be split as:
[For the person, who is qualified for the insured after s/he was disqualified]
[the terms of the insured are added up together.]

✓ Case 2:
<T2> For the amount of the pension by this law, </T2>
<A> when there is a remarkable change in the living standard of the nation of the other situation, </A>
<C> a revision of the amount of the pension must be taken action promptly to meet the situations. </C>

The sentence will be split as:
[When there is a remarkable change in the living standard of the nation of the other situation]
[For the amount of the pension by this law, a revision of the amount of the pension must be taken action promptly to meet the situations.]

✓ Case 3:
<T3> For the Government, </T3>
<A> when it makes a present state and a perspective of the finance, </A>
<C> it must announce it officially without delay. </C>

The sentence will be split as:
[For the Government, when it makes a present state and a perspective of the finance]
[For the Government, it must announce it officially without delay.]

Figure 3. Examples of sentence segmentation

After recognizing the logical structure of a legal sentence, we segment a sentence to its structure.

According to the logical structure of a legal sentence (Fig. 1), a sentence of each case is divided as follows:

- Case 0:
  Requisite part: [A]
  Effectuation part: [C]

- Case 1:
  Requisite part: [T1 A]
  Effectuation part: [C]

- Case 2:
  Requisite part: [A]
  Effectuation part: [T2 C]

- Case 3:
  Requisite part: [T3 A]
  Effectuation part: [T3 C]
  The examples of the sentences in section 3.1 are separated as shown in Fig 3.

3.2. Rule Selection for Legal Translation

Rule selection is important to tree-based statistical machine translation systems. This is because a rule contains not only terminals (words or phrases), but also nonterminals and structural information. During decoding, when a rule is selected and applied to a source text, both lexical translations (for terminals) and reorderings (for nonterminals) are determined. Therefore, rule selection affects both lexical translation and phrase reorderings. However, most of the current tree-based systems ignore contextual information when they select rules during decoding, especially the information covered by nonterminals. This makes the decoder hardly to distinguishrules. Intuitively, information covered by nonterminals as well as contextualinformation of rules is believed to be helpful for rule selection.Linguistic and contextual information have been widely used to improve translation performance. It is helpful to reduce ambiguity, thus guiding the decoder to choose correct translation for a source text on phrase reordering. In our research, we integrate dividing a legal sentence based on its logical structure into the first step of the rule selection. We propose a maximum entropy-based rule selection model for tree-based English-Japanese statistical machine translation in legal domain. The maximum entropy-based rule selection model combines local contextual information around rules and information of sub-trees covered by variables in rules. Therefore, the nice properties of maximum entropy model (lexical and syntax for rule selection) are helpful for rule selection methods better.

3.2.1. Maximum Entropy based rule selection model (MaxEnt RS model)

The rule selection task can be considered as a multi-class classification task. For a source-side, each corresponding target-side is a label. The maximum entropy approach (Berger et al., 1996) is known to be well suited to solve the classification problem. Therefore, we build a maximum entropy-based rule selection (MaxEnt RS) model for each ambiguous hierarchical LHS (left-hand side).
Following [4], [5] we use (α, γ) to represent a SCFG rule extracted from the training corpus, where α and γ are source and target strings, respectively. The nonterminal in α and γ are represented by Xk , where k is an index indicating one-one correspondence between nonterminal in source and target sides. Let us use $e(Xk)$ to represent the source text covered by Xk and $f(Xk)$

to represent the translation of *e(Xk)*. Let *C(α)* be the context information of source text matched

Table 3. Lexical features of nonterminals

Side	Type	Name	Description
Source-side	Lexical features	$W_{\alpha-1}$	The source word immediately to the left of α
		$W_{\alpha+1}$	The source word immediately to the right of α
		$WL_{e(X_k)}$	The first word of $e(X_k)$
		$WR_{e(X_k)}$	The last word of $e(X_k)$
	Pos features	$P_{\alpha-1}$	POS of $W_{\alpha-1}$
		$P_{\alpha+1}$	POS of $W_{\alpha+1}$
		$PL_{e(X_k)}$	POS of $WL_{e(X_k)}$
		$PR_{e(X_k)}$	POS of $WR_{e(X_k)}$
	Length feature	$LEN_{e(X_k)}$	Length of source subphrase $e(X_k)$
Target-side	Lexical features	$WL_{f(X_k)}$	The first word of $f(X_k)$
		$WR_{f(X_k)}$	The last word of $f(X_k)$
	Length feature	$LEN_{f(X_k)}$	Length of target subphrase $f(X_k)$

by α and $C(\alpha)$ be the context information of source text matched by γ . Under the MaxEnt model, we have:

$$P_{rs}(\gamma \mid \alpha, e(X_k), f(X_k)) = \frac{\exp[\sum_i \lambda_i h_i(C(\gamma), C(\alpha), e(X_k), f(X_k))]}{\sum_{\gamma'} \exp[\sum_i \lambda_i h_i(C(\gamma'), C(\alpha), e(X_k), f(X_k))]}$$

Where h_i a binary feature function, λ_i the feature weight of h_i. The MaxEnt RS model combines rich context information of grammar rules, as well as information of the subphrases which will be reduced to nonterminal *X* during decoding. However, these information is ignored by Chiang's hierarchical model.

We design five kinds of features for a rule (α, γ): Lexical, Parts-of-speech (POS), Length, Parent and Sibling features.

3.2.2. Linguistic and Contextual Information For Rule Selection

A. Lexical Features of Nonterminal

In the each hierarchical rules, there are nonterminals. Features of nonterminal consist of Lexical features, Parts-of-speech features and Length features:
Lexical features, which are the words immediately to the left and right of α, and boundary words of subphrase $e(X_k)$ and $f(X_k)$;

Parts-of-speech (POS) features, which are POS tags of the source words defined in lexical features.
Length features, which are the length of subphrases $e(X_k)$ and $f(X_k)$.
Table 3 shows lexical features of nonterminals. For example, we have a rule, source phrase, source sentence and alignment as following:
Rule:

$$X \rightarrow (\, X_1 \text{ it officially } X_2 \mid X_2 \text{ これを } X_1 \,)$$

Table 4. Lexical features of nonterminal of the example

Type	Features
Lexical Features	$W_{\alpha-1}$ = it
	$WL_{e(X_1)}$ = must $WR_{e(X_1)}$ = announce $WL_{e(X_2)}$ =without $WR_{e(X2)}$ = delay
	$WL_{f(X_1)}$ =公表 $WR_{f(X_1)}$ = しなければならない $WL_{f(X_2)}$ =遅滞 $WR_{f(X_2)}$ = なく
POS Features	$P_{\alpha-1}$ =NP
	$PL_{e(X_1)}$ = V $PR_{e(X_1)}$ = V $PL_{e(X_2)}$ = P $PR_{e(X_2)}$ = N
Length Features	$LEN_{e(X_1)}$ =2 $LEN_{e(X_2)}$ =2 $LEN_{f(X_1)}$ =2 $LEN_{f(X_2)}$ =2

Table 5. Lexical features around nonterminal

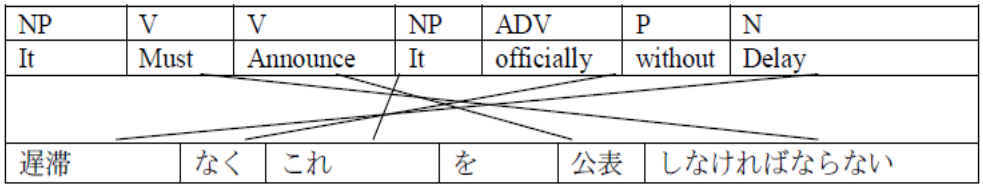
Side	Type	Name	Description
Source-side	Lexical feature	$WL_{e(X_k)-1}$	The first word immediately to the left of $e(X_k)$
		$WR_{e(X_k)-1}$	The first word immediately to the right of $e(X_k)$
	POS Features	$PL_{e(X_k)-1}$	POS of $WL_{e(X_k)-1}$
		$PR_{e(X_k)+1}$	POS of $WR_{e(X_k)+1}$
Target-side	Lexical features	$WL_{f(X_k)-1}$	The first word of $f(X_k)-1$
		$WR_{f(X_k)+1}$	The last word of $f(X_k)+1$

Source Phrase:
must announce it officially without delay
遅滞なくこれを公表しなければならない

X_1 must announce
X_2 without delay
X_1 公表しなければならない
X_2 遅滞なく

Source sentence:
It must announce it officially without delay
遅滞なくこれを公表しなければならない

Alignment of English-Japanese sentence pair:



Features of this example are shown in Table 4.

b. Lexical features around nonterminals

These features are same meaning as features of nonterminal.

Lexical features, which are the words immediately to the left and right of subphrase $e(X_k)$ and $f(X_k)$;

Table 6. Lexical features around nonterminal of the example

Side	Type	Features
Source-side	Lexical feature	$WL_{e(X_1)-1}$ = remarkable $WR_{e(X_1)+1}$ = in
		$WL_{e(X_2)-1}$ = standard $WR_{e(X_2)-1}$ = the
	POS Features	$PL_{e(X_1)-1}$ = ADJ $PR_{e(X_1)+1}$ = P
		$PL_{e(X_2)-1}$ = N $PR_{e(X_2)+1}$ = DET
Target-side	Lexical features	$WL_{f(X_1)-1}$ =著しい
		$WL_{f(X_2)-1}$ = 国民 $WR_{f(X_2)+1}$ =生活

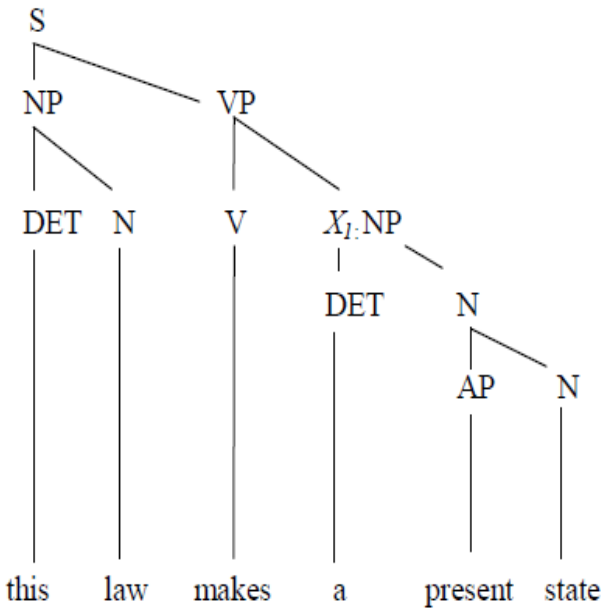


Figure 4. Sub-tree covers nonterminal *X1*

Parts-of-speech (POS) features, which are POS tags of the source words defined in lexical features.

Table 5 shows lexical features around nonterminal.

Example: with a rule:

$X \rightarrow$ (a remarkable X_1 in the living standard X_2 the nation |
国民 X_2 生活に水準の諸事情に著しい X_1)

We have lexical features around nonterminal shown in Table 6.

c. Syntax features

Let $R \rightarrow <\alpha, \gamma\sim>$ is a translation rule and $e(\alpha)$ is source phrase covered by α.
X_k is nonterminal in α, $T(X_k)$ is sub-tree covering X_k.

Parent feature (PF):

The parent node of $T(X_k)$ in the parse tree of source sentence. The same sub-tree may have different parent nodes in different training examples. Therefore, this feature may provide information for distinguishing source sub-trees

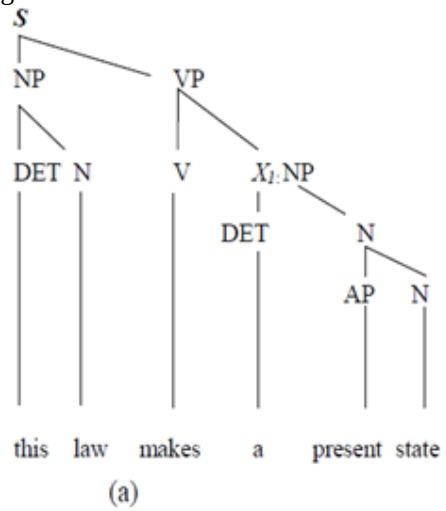
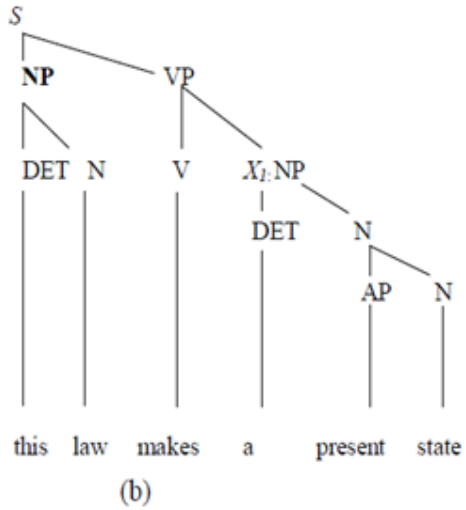


Figure 5. (a) *S*: Parent feature of sub-tree covers nonterminal *X1*



(b) NP: Sibling feature of sub-tree covers nonterminal *X1*

$R=\{R_i\}=\{set\ of\ Rules\}$, $P=\{P_j,P'_j\}=\{set\ of\ English\text{-}Japanese\ phrase\ alignments\}$,
$S=\{S_l,E_l\}=\{set\ of\ sentence\ pairs\}$, $S'=\{S_l\}=\{set\ of\ tagged\ English\ sentences\}$,
$S''=\{S''_l\}=\{set\ of\ parsed\ English\ split\ sentences\}$.

Input: Rules, English-Japanese phrase alignments, sentence pairs, tagged English sentences, parsed English split sentences.

Output: Features of non-terminals; Features around non-terminals and Syntax features

```
For i = 1 to n do
  Xk  = Non-terminal of LHS of Ri
  X'k = Non-terminal of RHS of Ri
  Y= LHS of Ri
  Z= RHS of Ri
  For  j = 1 to m do
     If  Y ∈Pj, Z∈ P'j then
      Xk  = phrase        (English phrase)
      X'k = phrase'       (Japanese phrase)
        For  l = 1 to v do
            features of non-terminal
            features around non-terminal
            syntax features
          endfor
       endif
   endfor
endfor
```

Figure 6. Algorithm for extracting features

Sibling feature (SBF)

The sibling features of the root of *T(Xk)*. This feature considers neighboring nodes which share the same parent node.Fig. 4 shows the subtree covers non terminal *X1*, Fig. 5(a) shows *S* node is the Parent feature ofsubtree covering *X1* and *NP* node is the Sibling feature shown in Fig. 5(b). Those features:Lexical feature, Parts-of-speech features, Length features, Parent features and Sibling featuresmake use rich of information around a rule, including the contextual information of a rule andthe information of sub-trees covered by non terminals. These features can be gathered accordingto Chiang's rule extraction method. We use Moses-chart [12] to extract phrases and rules,Stanford Tagger toolkits and Cabocha [14] to tag, tokenize English and Japanese sourcesentence, Stanford parser [20] to parse English test sentence, after that we use algorithm in Fig.6 to extract features.

In Moses-chart, the number of nonterminal of a rule are limited up to 2. Thus a rule may have 36 features at most. After extracting features from training corpus, we use the toolkit implemented in [24] to train a MaxEnt RS model for each ambiguous hierarchical LHS.

3.2.3. Integrating Maxent RS Model Into Tree-Based Model

We integrate the MaxEnt RS model into the tree-based model during the translation of each source sentence. Thus the MaxEnt RS models can help the decoder perform context-dependent rule selection during decoding.

In Chiang, [4] the log-linear model combines 8 features: the translation probabilities $P(\gamma \mid \alpha)$ And $P(\alpha \mid \gamma)$, the lexical weights $Pw(\gamma \mid \alpha)$ and $Pw(\alpha \mid \gamma)$, the language model, the word penalty, thephrase penalty, and the glue rule penalty. For integration, we add two new features:

(1) $P_{rs}(\gamma \mid \alpha, e(X_k), f(X_k))$.

This feature is computed by the MaxEnt RS model, which gives a probability that the model selecting a target-side γ given an ambiguous source-side α, considering context information.

(2) *Prsn = exp(1).*

This feature is similar to phrase penalty feature. In our experiment, we find that some sourcesides are not ambiguous, and correspond to only one target-side. However, if a source-side α' is not ambiguous, the first features *Prs* will be set to *1.0*. In fact, these rules are not reliable since they usually occur only once in the training corpus. Therefore, we use this feature to reward the ambiguous source-side. During decoding, if an LHS has multiple translations, this feature is set to *exp(1)*, otherwise it is set to *exp(0)*.

Chiang [5] used the CKY (Cocke-Kasami-Younger) algorithm with a cube pruning method for decoding. This method can significantly reduce the search space by efficiently computing the top-n items rather than all possible items at a node, using the k-best algorithms of Huang and Chiang (2005) to speed up the computation. In cube pruning, the translation model is treated as the monotonic backbone of the search space, while the language model score is a nonmonotonic cost that distorts the search space. Similarly, in the MaxEnt RS model, source-side features form a monotonic score while target-side features constitute a non-monotonic cost that can be seen as part of the language model.

For translating a source sentence E^J_I, the decoder adopts a bottom-up strategy. All derivations are stored in a chart structure. Each cell *c[i, j]* of the chart contains all partial derivations which correspond to the source phrase e^j_i . For translating a source-side span *[i , j]*, we first select all possible rules from the rule table. Meanwhile, we can obtain features of the MaxEnt RS model which are defined on the source-side since they are fixed before decoding. During decoding, for a source phrase $e^j i$, suppose the rule $X \rightarrow (e^k_i X1\ e^j_t, f^{k'}_{i'} X_1 f^{j'}_{t'})$ is selected by the decoder, where $i \le k < t \le j$ and $k+1 < t$, then we can gather features which are defined on the target-side of the subphrase X_1 from the ancestor chart cell *c[k+1, t-1]* since the span *[k+1, t-1]* has already been covered. Then the new feature scores P_{rs} and P_{rsn} can be computed. Therefore, the cost of derivation can be obtained. Finally, the decoding is completed. When the whole sentence is

Table 7. Statistical table of train and test corpus

Corpus	**#words**		**#sentences**
Training corpus	English	990,011	40,000
	Japanese	935,467	
Development corpus	English	45,150	1,400
	Japanese	45,020	
Test corpus	English	17,475	516
	Japanese	17,753	

Table 8. Statistics of the test corpus

Name of Law	Number of sentences
Act on General Rule for Application of Law	78
Act on Land and Building Leases	120
Administrative Procedure Act	99
Foreign Exchange and Foreign Trade Act	219
Total	516

Table 9. Number of requisite part, effectuation part in the test data

Sentence	516
#of requisite part	436
#of effectuation part	513
#of segment	949

covered, and the best derivation of the source sentence E^J_I is the item with the lowest cost in cell *c[I,J]*.

The advantage of our integration is that we need not change the main decoding algorithm of a SMT system. Furthermore, the weights of the new features can be trained together with other features of the translation model.

4. EXPERIMENTS

We conducted the experiments on the English-Japanese translation corpus provided by Japanese Law Translation Database System. The training corpus consisted of 40,000 English-Japanese original sentence pairs, the development and test set consisted of 1,400 and 516 sentence pairs, respectively. The statistics of the corpus is shown in Table 7. We tested on 516 English- Japanese sentence pairs. Table 8 shows statistics of the test corpus. The test set is divided by the method described in Section 3.1. Table 9 shows the number of sentences, the statistics of the requisite parts, the effectuation parts and the logical parts after splitting in the test set. Then, we applied rule selection for the split sentence in the test set.

To run decoder, we share the same pruning setting with Moses, Moses-chart [12] baseline systems. To train the translation model, we first run GIZA++ [18] to obtain word alignment in both translation directions. Then we use Moses-chart to extract SCFG grammar rules. We use Stanford Tagger [20] and Cabocha [14] toolkits to token and tag English and Japanese sentences. We parse the split English test sentence by Stanford parser [20] and gather lexical and syntax features for training the MaxEnt RS models. The maximum initial phrase length is set to 7 and the maximum rule length of the source side is set to 5.

Lex= Lexical Features, POS= POS Features, Len= Length Feature, Parent= Parent Features, Sibling = Sibling Features.

Table 10. BLEU-4 scores (case-insensitive) on English-Japanese corpus.

System	BLEU
MM	0.287
MC	0.306
MS	0.318
MR (MaxEnt RS)	
Lexical features of nonterminal (Lex+POS+Len)	0.326
Lexical features around nonterminal (Pos+Lex)	0.320
Syntax features (Parent and sibling)	0.325
Lexical features of nonterminal + syntax features	0.327
All features	0.329

We use SRI Language modeling toolkit [21] to train language models. We use minimum error rate training integrated in Moses-chart to tune the feature weights for the log-linear model. The translation quality is evaluated by BLEU metric [19], as calculated by mteval-v12.pl with case insensitive matching of n-grams, where n=4. After using Moses-chart to extract rules, we have a rule-table, then we insert two new scores into the rules. We evaluate both original test sentence and split test sentence with Maxent RS model. We compare the results of four systems: Moses using original test sentence (MM), Moses-chart using original test sentence (MC), Moses-chart using split test sentence (MS) and Moses-chart using split test sentence and applying rule selection or our system (MR). The results are shown in Table 10.

As we described, we add two new features to integrate the Maxent RS models into the Moses chart: $P_{rs}(\gamma \mid \alpha, e(X_k), f(X_k))$ and P_{rsn} .We do not need to change the main decoding algorithm of a SMT system and the weights of the new features can be trained together with other features of the translation model.

In Table 10, Moses system using original test sentence (MM) got 0.287 BLEU scores, Moses chart system using original test sentence (MC) got 0.306 BLEU scores, Moses-chart system using split sentence (MS) got 0.318 BLEU scores, using all features defined to train the MaxEnt RS models for Moses-chart using split test sentence our system got 0.329 BLEU scores, with an absolute improvement 4.2 over MM system, 2.3 over MC system and 1.1 over MS system. In order to explore the utility of the context features, we train the MaxEnt RS models on different features sets. We find that lexical features of non terminal and syntax features are the most useful features since they can generalize over all training examples. Moreover, lexical features around non terminal also yields improvement. However, these features are never used in the baseline. When we used MS system to extract rule, we got the rules as shown in Table 11. Table 12 shows the number of source-sides of SCFG rules for English-Japanese corpus. After extracting grammar rules from the training corpus, there are 12,148 source-sides match the split test sentence, they are hierarchical LHS's (H-LHS, the LHS which contains non terminals). For the hierarchical LHS's, 52.22% is ambiguous (AH-LHS, the H-LHS which has multiple translations). This indicates that the decoder will face serious rule selection problem during decoding. We also noted the number of the source-sides of the best translation for the split test sentence. However, by incorporating MaxEnt RS models, that proportion increases to 67.36%, since the number of AH-LHS increases.

The reason is that, we use the feature Prsn to reward ambiguous hierarchical LHS's. This has some advantages. On one hand, H-LHS can capture phrase reordering

Table 11. Statistical table of rules

Name	Number
The number of rules	1,480,741
The number of rules contain nonterminal	1,126,440
The number of rules don't contain nonterminal	354,298
The number of glue grammar rules	3
The number of rules match test	12,148

Table 12. Number of possible source-sides of SCFG rule for English-Japanese corpus and number of source-sides of the best translation.
H-LHS = Hierarchical LHS, AH-LHS = Ambiguous hierarchical LHS

System	Rule	NO of H-LHS	NO of AH-LHS
MS	12,148	6,541	3,416
Our system (MR, all features)	12,148	7,741	5,214

Table 13. Translation examples of test sentences in Case 3 in MS and our systems (MR, all features)The Japanese sentence in Japanese-English translation is the original sentence. The English sentence in English-Japanese translation is the reference translation in the government web page

Sentence	<C> Notwithstanding the preceding paragraph, </C> <T3> the formalities </T3> <A> that comply with the law of the place where said act was done </A> <C> shall be valid. </C>
Split Sentence	the formalities notwithstanding the preceding paragraph, shall be valid. the formalities that comply with the law of the place where said act was done
MS	同項の規定にかかわらず、手続きは、有効なものでなければならない。行為が行われていたと述べた場所の法律を遵守手続き
Our system (MR, all features)	前項の規定にかかわらず、方式は、有効でなければならない 方式は、当該行為が行われた場所の法律の遵守します。

phrase reorderings. On the other hand, AH-LHS is more reliable than non-ambiguous LHS, since most non-ambiguous LHS occurred only once in the training corpus. In order to know how the MaxEnt RS models improve the performance of the SMT system, we study the best translation of MS and our systems. We find that the MaxEnt RS models improve translation quality in two ways:

Better Phrase reordering

Since the SCFG rules which contain nonterminals can capture reordering of phrases, better rule selection will produce better phrase reordering.
Table 13 shows translation examples of test sentences in Case 3 in MS and our systems, our system gets better result than MS system in phrase reordering.

Better Lexical Translation

The MaxEnt RS models can also help the decoder perform better lexical translation than MSsystem. This is because the SCFG rules contain terminals. When the decoder selects a rule for a source-side, it also determines the translations of the source terminals. The examples of our system get better result than MS system in lexical translation shown in the underlined parts of Table 13.

The advantage of the proposed method arises from the translation model based on the logical structure of a legal sentence where the decoder searches over shortened inputs. Because we use the logical structure of a legal sentence to split sentence, the split sentence reserves its structure and the average length of split sentence is much smaller than those of no split sentence. They are expected to help realize an efficient statistical machine translation search.

The syntax features and lexical features of non-terminals are the useful features since they canbe generalized over all training examples. However, the lexical features around non-terminals also yield improvement because translation rules contain terminals (words or phrases), nonterminals and structural information. Terminals indicate lexical translation, and non-terminal and structural information can capture short or long-distance reordering. Therefore, rich lexical and contextual information can help decoder capture reordering of phrases. Since the rules which contain non-terminals can capture reordering of phrases, better rule selection will produce better phrase reordering. The Maximum entropy models can also help the decoder perform better lexical translation than the baseline. This is because the rules contain terminals, when the decoder selects a rule for a source side, it also determines the translations of the source terminals.

5. CONCLUSION AND FUTURE WORK

We show in this paper that translating legal sentence by segmentation and rule selection can help improving legal translation. We divided a legal sentence based on its logical structure and applying the split sentence into rule selection. We use rich linguistic and contextual information for both non-terminals and terminals and integrate them into maximum entropy-based ruleselection to solve each ambiguous rule among the translation rules to help decoder know which rules are suitable. Experiment shows that this approach to legal translation achieves improvements over tree-based SMT (Moses-chart) and Moses systems.

In the future, we will investigate more sophisticated features to improve legal sentencesegmentation and the maximum entropy-based rule selection model. We will apply our proposed method into training and test the performance on a large scale corpus.

REFERENCES

[1] Bach, N., X., Minh, and N., L., Shimazu, A., (2010). “PRE Task: The Task of Recognition of Requisite Part and Effectuation Part in Law Sentences”, in International Journal of Computer Processing of Languages (IJCPOL), Volume 23, Number 2.

[2] Carpuat, M., and Wu, W., (2007). “Improving statistical machine translation using word sense disambiguation”. In Proceedings of EMNLP-CoNLL, pp. 61–72.

[3] Chan, Seng, Y., Tou, N., H., and Chiang, D., (2007). "Word sense disambiguation improves statistical machine translation". In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 33-40.

[4] Chiang, D., (2005). “A hierarchical phrase-based model for statistical machine translation”. In Processing of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 263-270.

[5] Chiang, D., (2007). “Hierarchical phrase-based translation”. Computational Linguistics, pp. 33(2):201–228.

[6] Chiang, D., Knight, K., and Wang, W., (2009). “11,001 new features for statistical machine translation”. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

[7] Goh, C., Onishi, T., and Sumita, E., (2011). "Rule-based Reordering Constraints for Phrase-based SMT", EAMT.

[8] He, Z., Liu, Q., and Lin, S., (2008). “Improving statistical machine translation using lexicalized rule selection”. Proceedings of the 22nd International Conference on Computational Linguistics, August, pp. 321-328.

[9] Hung, B., T., Minh, N., L., and Shimazu, A., (2012). “Divide and Translate Legal Text Sentence by Using its Logical Structure”, in Proceedings of 7th International Conference on Knowledge, Information and Creativity Support Systems, pp. 18-23.

[10] Katayama, K., (2007). “Legal Engineering-An Engineering Approach to Laws in E-society Age”, in Proc. of the 1st International Workshop on JURISIN.

[11] Koehn, P., (2010). “Statistical machine translation”, 488 pages, Cambridge press.

[12] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E., (2007). “Moses: Open Source Toolkit for Statistical Machine Translation”, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June.

[13] Kudo, T., (2003). “CRF++: Yet Another CRF toolkit”, http://crfpp.sourceforge.net/

[14] Kudo, T., (2003). “Yet Another Japanese Dependency Structure Analyzer”, http://chasen.org/taku/software/cabocha/

[15] Lafferty, J., McCallum, A., and Pereira, F., (2001). “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, in Proceedings of ICML, pp. 282–289

[16] Liu, Q., He, Z., Liu, Y., and Lin, S., (2008). "Maximum Entropy based Rule Selection Model for Syntax-based Statistical Machine Translation". Proceedings of EMNLP, Honolulu, Hawaii.

[17] Nakamura, M., Nobuoka, S., and Shimazu, A., (2007). “Towards Translation of Legal Sentences into Logical Forms”, in Proceedings of the 1st International Workshop on JURISIN.

[18] Och, F., and Ney, H., (2000). “Improved Statistical Alignment Models”, Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, China, pp. 440-447, October.

[19] Papineni, Kishore, Roukos, S., Ward, T., and Zhu, W., (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation”, in Proceedings of the 40th Annual Meeting of the ACL, pp.311-318.

[20] Socher, R., Bauer, J., Manning, C., D., and Andrew, Y., Ng., (2013). "Parsing With Compositional Vector Grammars". Proceedings of ACL.

[21] Stolcke, (2002). “SRILM-An Extensible Language Modeling Toolkit”, in Proceedings of International Conference on Spoken Language Processing, vol. 2, (Denver, CO), pp. 901-904.

[22] Sudoh, Katsuhito, Duh, K., Tsukada, H., Hirao, T., and Nagata, M., (2010). “Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation”, in Proceedings of the Joint 5th Workshop on SMT and MEtricsMATR, pp 418-427.

[23] Tanaka, K., Kawazoe, I., and Narita, H., (1993). “Standard Structure of Legal Provisions-for the Legal Knowledge Processing by Natural Language-(in Japanese)”, in IPSJ Research Report on Natural Language Processing, pp.79-86.

[24] Tsuruoka, Y., (2011). “A Simple C++ Library for Maximum Entropy Classification”. http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/maxent/.

[25] Xiong, H., Xu, W., Mi, H., Liu, Y., and Lu, Q., (2009). “Sub-Sentence Division for Tree-Based Machine Translation”, in Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP, Short Papers, Singapore, pp. 137-140

Authors

Bui Thanh Hung

He received his M.S. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2010. He is a Ph.D. research student. His research interest includes natural language processing and machine translation.

Nguyen Le Minh

He received his M.S. degree from Vietnam National University in 2001 andPh.D. degree from Japan Advanced Institute of Science and Technology in 2004.He is currently a Associate Professor at the Graduate School of InformationScience, Japan Advanced Institute of Science and Technology (JAIST). His main research interests are Natural Language Processing, Text Summarization,Machine Translation, and Natural Language Understanding

Akira Shimazu

He received his M.S. and Ph.D. degrees from Kyushu University in 1973 and in1991. He is currently a Professor at the Graduate School of Information Science,Japan Advanced Institute of Science and Technology (JAIST). His main researchinterests are Natural Language Processing, Dialog Processing, MachineTranslation, and Legal Engineering.