# A joint inference of deep case analysis and zero subject generation for Japanese-to-English statistical machine translation

工藤 拓　市川 宙　中川 哲治　賀沢 秀人
Google Japan
{taku,ichikawa,tnaka,kazawa}@google.com

## 1 Introduction

Japanese to English translation is known to be one of the most difficult language pair for statistical machine translation (SMT). It has been widely believed for years that the difference of word orders, i.e., Japanese is an SOV language, while English is an SVO language, makes the English-to-Japanese and Japanese-to-English translation difficult. However, simple, yet powerful pre-ordering techniques have made this argument a thing of the past (Isozaki et al., 2010b; Komachi et al., 2006; Fei and Michael, 2004; Lerner and Petrov, 2013; Wu et al., 2011; Katz-Brown and Collins, 2008; Neubig et al., 2012; Hoshino et al., 2013).

While many successes of English-to-Japanese translation have been reported recently, the quality improvement of Japanese-to-English translation is still small even with the help of pre-ordering (Goto et al., 2013). We found that there are two major issues that make Japanese-to-English translation difficult. One is that Japanese subject and object cannot easily be identified compared to English, while their detections are the key process to generate correct English word orders. Japanese surface syntactic structures are not always corresponding to their deep structures, i.e., semantic roles. The other is that Japanese is a pro-drop language in which certain classes of pronouns may be omitted when they are pragmatically inferable. In Japanese-to-English translation, these omitted pronouns have to be generated properly.

There are several researches that focused on the pre-ordering with Japanese deep syntactic analysis (Komachi et al., 2006; Hoshino et al., 2013) and zero pronoun generation (Taira et al., 2012) for Japanese-to-English translation. However, these two issues have been considered independently, while they heavily rely on one another.

In this paper, we propose a simple joint inference which handles both Japanese deep structure analysis and zero pronoun generation. To the best of our knowledge, this is the first study that addresses these two issues at the same time.

This paper is organized as follows. First, we describe why Japanese-to-English translation is difficult. Second, we show the basic idea of this work and its implementation based on pointwise probabilistic models and a global inference with an integer linear programming (ILP). Several experiments are employed to confirm that our new model can improve the Japanese to English translation quality.

## 2 What makes Japanese-to-English translation difficult?

Japanese syntactic relations between arguments and predicates are usually specified by particles. There are several types of particles, but we focus on が (*ga*), を (*wo*) and は (*wa*) for the sake of simplicity [1].

Table 1: An example of difficult sentence for parsing

| | |
|---|---|
| Sentence: | 今日 は お酒 が 飲める. |
| Gloss: | today $wa_{\text{TOP}}$ liquor $ga_{\text{NOM}}$ can_drink. |
| Translation: | (I) can drink liquor today. |

- *ga* is usually a subject marker. However, it becomes an object marker if the predicate has a potential voice type, which is usually translated into *can, be able to, want to, or would like to*.
- *wo* is an object marker.
- *wa* is a topic case marker. The topic can be anything that a speaker wants to talk about. It can be a subject, object, location, time or any other grammatical elements.

We cannot always identify Japanese subject and object only by seeing the surface case markers *ga*, *wo* and *wa*. Especially the topic case marker is problematic, since there is no concept of topic in English. It is necessary to get a deep interpretation of topic case markers in order to develop accurate Japanese-to-English SMT systems.

Another big issue is that Japanese subject (or even an object) can be omitted when they can pragmatically be inferable from the context. Such a pronoun-dropping is not an unique phenomenon in Japanese actually. For instance, Spanish also allows to omit pronouns. However, the inflectional suffix of Spanish verbs include a hint for the person of the subject. On the other hand, inferring Japanese subjects is more difficult than Spanish, since Japanese verbs usually do not have any grammatical cues to tell the subject person.

Table 1 shows an example Japanese sentence which cannot be parsed only with the surface structures. The second token *wa* specifies the relation between 今日 (*today*) and 飲める (*can drink*). Human can easily tell that the relation of them is not a subject but an adverb (time). The topic case marker *wa* implies that the time when the speaker drinks liquor is the focus of this sentence. 4th token *ga* indicates the relation between お酒 (*liquor*) and 飲める (*can drink*). Since the predicate has a potential voice (*can drink*), the *ga* particle should be interpreted as an object here. In this sentence, the subject is omitted. In general, it is unknown who speaks this sentence, but the first person is a natural interpretation in this context.

Another tricky phenomenon is that detecting voice type is not always deterministic. There are several ways to generate a potential voice in Japanese, but we usually put the suffix word れる (*reru*) or られる (*rareru*) after the predicates. However, these suffix words are also used for a passive voice.

In summary, we can see that the following four factors are the potential causes that make the Japanese parsing difficult.

[1] Other case markers are less often than these three markers

- Japanese voice type detection is not straightforward. *reru* or *rareru* are used either for passive or potential voice.
- surface case *ga* changes its interpretation from subject to object when the predicate has a potential voice.
- topic case marker *wa* is used as a topic case marker which doesn't exist in English. Topic is either subject, object or any grammatical elements depending on the context.
- Japanese subject is often omitted when it is inferable from the context. There is no cue to tell the subject person in verb suffix (inflection) like in Spanish verbs.

We should note that they are not always independent issues. For instance, the deep case detection helps to tell the voice type, and vice versa.

Another note is that they are unique issues observed only in Japanese-to-English translation. In English-to-Japanese translation, it is acceptable to generate Japanese sentences that don't use Japanese topic markers *wa*. Also, generating Japanese pronoun from English pronoun is acceptable, although it sounds redundant and unnatural for native speakers.

# 3 A joint inference of deep case analysis and zero subject generation

## 3.1 A Probabilistic model over predicate-argument structures

Our deep parser runs on the top of a dependency parse tree. First, it extracts all predicates and their arguments from a dependency tree by using manual rules over POS tags. Since our pre-ordering system generates the final word orders from a labeled dependency tree, we formalize our deep parsing task as a simple labeling problem over dependency links, where the label indicates the deep syntactic roles between a head and modifier.

We here define a joint probability over a predicate and its arguments as follows:

$$P(p, z, v, A, S, D), \tag{1}$$

where

- $p$: a predicate
- $z$: zero subject candidate for $p$.
  $z \in Z = \{$*I, you, we, it, he/she, imperative, already_exists*$\}$
- $v$: voice type of the predicate $p$.
  $v \in V = \{active, passive, potential\}$
- $a_k \in A$: $k$-th argument which modifies or is modified by the predicate[2].
- $d_k \in D$: deep case label which represents a deep relation between $a_k$ and $p$.
  $d \in D = \{$ *subject, object, other* $\}$, where *other* means that deep case is neither subject nor object.
- $s_k \in S$: surface relation (surface case marker) between $a_k$ and $p$.

We assume that a predicate $p$ is independent from other predicates in a sentence. This assumption allows us to estimate the deep structures of $p$ separately, with no regard to which decisions are made in other predicates.

[2]Generally, an argument modifies a predicate, but in relative clauses, a predicate modifies an argument

An optimal zero subject label $\hat{z}$, deep cases $\hat{D}$, and voice type $\hat{v}$ for a given predicate $p$ can be solved as the following optimization problem.

$$\langle \hat{z}, \hat{v}, \hat{D} \rangle = \underset{z,v,D}{argmax}\, P(p, z, v, A, S, D)$$

Since the inference of this joint probability is difficult, we decompose $P(p, z, v, A, S, D)$ into small independent sub models:

$$P(p, z, v, A, S, D) \approx$$
$$P_z(z|p, A, S) P_v(v|p, A, S) P_d(D|p, v, A, S) P(p, A, S). \tag{2}$$

We don't take the last term $P(p, A, S)$ into consideration, since it is constant for the optimization. In the next sections, we describe how these probabilities $P_z$, $P_d$, and $P_v$ are computed.

### 3.1.1 Zero subject model: $P_z(z|p, A, S)$

This model estimates the syntactic zero subject[3] of the predicate $p$. For instance, z= *I* means that the subject of $p$ is omitted and its type is first person. z=*imperative* is internally handled as an invisible subject. z=*already_exists* means that a subject already appears in the sentence. A maximum entropy classifier is used in our zero subject model, which takes the features extracted from $p$, $A$, and $S$.

### 3.1.2 Voice type model: $P_v(v|p, A, S)$

This model estimates the voice type of a predicate. We also use a maximum entropy classifier for this model. This classifier is used only when the predicate has the ambiguous suffix *reru* or *rareru*. If the predicate is a potential verb[4], this model returns *potential* with a very high probability.

### 3.1.3 Deep case model: $P_d(D|p, v, A, S)$

This model estimates the deep syntactic role between a predicate $p$ and its arguments $A$. This model helps to resolve the deep cases when their surface cases are topic. We define $P_d$ as follows with an independent assumption on predicate-argument structures:

$$P_d(D|p, v, A, S) \approx \prod_i [\max(p(d_i|a_i, p) - m(s_i, d_i, v),\ \delta)].$$

$p(d|a, p)$ models the deep relation between $p$ and $a$. We use a maximum likelihood estimation for $p(d|a, p)$:

$$\begin{aligned} p(d = subj|a, p) &= \frac{freq(s = ga, a, \text{active form of p})}{freq(a, \text{active form of p})} \\ p(d = obj|a, p) &= \frac{freq(s = wo, a, \text{active form of p})}{freq(a, \text{active form of p})}, \end{aligned}$$

where $freq(s = ga, a, \text{active form of p})$ is the frequency of how often an argument $a$ and $p$ appears with the surface case *ga*. The frequencies are aggregated only when the predicate appears in an active voice. If the voice type is active, we can safely assume that the surface cases *ga* and *wo* correspond to subject and object respectively (Kawahara and Kurohashi,

[3]Here *syntactic subject* means the subject which takes the voice type into account.
[4]飲む *(drink)* → 飲める *(can_drink)*

2006). We compute the frequencies from a large amount of auto-parsed data.

$m(s,d,v)$ is a non-zero penalty variable describing how the deep case $d$ generates the surface case $s$ depending on the voice type $v$. Since the number of possible surface cases, deep cases, and voice types are small, we define this penalty manually by reference to the Japanese grammar book (日本語記述文法研究会, 2009). We use these manually defined penalties in order to put more importance on syntactic preferences rather than those of semantics. Even if a predicate-augment structure is semantically irrelevant, we take this structure as long as it is syntactically correct [5].

$\delta$ is a very small positive constant to avoid zero probability.

### 3.2 Joint inference with linguistic constraints

Our initial model (2) assumes that zero subjects and deep cases are generated independently. However, this assumption does not always capture real linguistic phenomena. English is a subject-prominent language in which almost all sentences (or predicates) must have a subject. This implies that it is more reasonable to introduce strong linguistic constraints to the final solution for pre-ordering, which are described as follows:

- Subject is a mandatory role. A subject must be inferred either by zero subject or deep case model [6]. When the voice type is passive, an object role in $D$ is considered as a syntactic subject.
- A predicate can not have multiple subjects and objects respectively.

These two constraints avoid the model from inferring syntactically irrelevant solutions.

In order to find the result with the constraints above, we formalize our model as an integer linear programming, ILP. Let $\{x_1,,...,x_n\}$ be binary variables, i.e., $x_i \in \{0,1\}$. $x_i$ corresponds to the binary decisions in our model, e.g., $x_k = 1$ if $d_i =$ subj and $v =$ active. Let $\{p_1,...,p_n\}$ be probability vector corresponding to the binary decisions. ILP can be formalized as a mathematical problem, in which the objective function and the constraints are linear:

$$\{\hat{x}_1,...,\hat{x}_n\} = argmax \sum_i \log(p_i)x_i$$
$$\text{s.t. linear constraints over } \{x_1,..,x_n\}.$$

After taking the log of (2), our optimization model can be converted into an ILP. Also, the constraints described above can be represented as linear equations over binary variables $X$. We leave the details of the representations to (Punyakanok et al., 2004; Iida and Poesio, 2011).

### 3.3 Japanese pre-ordering with deep parser

We use a simple rule-based approach to make pre-ordered Japanese sentences from our deep parse trees, which is similar to the algorithms described in (Komachi et al., 2006; Katz-Brown and Collins, 2008; Hoshino et al., 2013). First, we naively reverse all the *bunsetsu*-chunks [7]. Then, we move a subject chunk just before its predicate. This process converts SOV to SVO. When the subject is omitted, we generate a subject with our deep parser and insert it to a subject position in the source sentence. There are three different ways to generate a subject.

Table 2: Examples of deep parser output

| |
|---|
| 今日は_{d=other} 酒が_{d=obj} 飲める_{v=potential, z=I} |
| ニュースが_{d=subj} 伝えられた_{v=passive, z=already} |
| ラーメンは_{d=obj} 食べましたか_{v=active, z=you} |
| あなたは_{d=subj} 食べましたか_{v=active, z=already} |

1. Generate real Japanese words. (Insert 私 *(I)*, あなた *(you)*.. etc)
2. Generate virtual seed words. (Insert *1st_person*, *2nd_person*..., which are not in the Japanese lexicon)
3. Generate only a single virtual seed word regardless of the subject type. (Insert *zero_subject*)

1) is the most aggressive method, but it causes completely incorrect translations if the detection of subject type fails. 2) and 3) are rather conservative, since they leave SMT to generate English pronouns.

We decided to use the following hybrid approach, since it shows the best performance in our preliminary experiments.

- In the training of SMT, use 3).
- In decoding, use 1) if the input sentence only has one predicate. Otherwise, use 3).

For example, "今日 は 酒 が 飲める" is reordered into "*zero_subject* 飲める が 酒 は 今日" in the training.

### 3.4 Examples of parsing results

Table 2 shows examples of our deep parser output. It can been seen that our parser can correctly identify the deep case of topic case marker *wa*.

## 4 Experiments

### 4.1 Experimental settings

We carried out all our experiments using a state-of-the-art phrase-based statistical Japanese-to-English machine translation system (Och, 2003) with pre-ordering. During the decoding, we use the reordering window (distortion limit) to 4 words. For parallel training data, we use an in-house collection of parallel sentences. These come from various sources with a substantial portion coming from the web. We trained our system on about 300M source words. Our test set contains about 10,000 sentences randomly sampled from the web.

The dependency parser we apply is an implementation of a shift-reduce dependency parser which uses a *bunsetsu*-chunk as a basic unit for parsing (Kudo and Matsumoto, 2002).

The zero subject and voice type models were trained with about 20,000 and 5,000 manually annotated web sentences respectively. In order to simplify the rating tasks for our annota-

[5] We want to regard *apple* as the subject of the sentence "りんごが食べる (apple *ga* eat)", although it is semantically irrelevant.

[6] *imperative* is also handled as an invisible subject

[7] *bunsetsu* is a basic Japanese grammatical unit consisting of one content word and functional words.

Table 3: Examples of training data

| Task | Input ({{}} is the annotation part.) | Output |
|---|---|---|
| Zero | 彼の優しさに {{ 感動した }} | I |
| Zero | 病院で診察を {{ 受けるべき }} | You |
| Voice | USB が {{ 認識されない }} | Passive |
| Voice | 返品には {{ 応じられません }} | Potential |

Table 4: Results for different reordering methods

| System | BLEU | RIBES |
|---|---|---|
| baseline (no reordering) | 16.15 | 52.67 |
| surface reordering | 19.39 | 60.30 |
| independent deep reordering | 19.68 | 61.27 |
| independent deep reordering + zero subject | 19.81 | 61.67 |
| joint deep reordering | 19.76 | 61.43 |
| joint deep reordering + zero subject | 19.90 | 61.89 |

tors, we extracted only one candidate predicate from a sentence for annotations. Table 3 shows examples of our training data [8].

We tested the following six systems.

- **baseline**: no pre-ordering.
- **surface reordering** : pre-ordering only with surface dependency relations.
- **independent deep reordering**: pre-ordering using deep parser without global linguistic constraints.
- **independent deep reordering + zero subject**: pre-ordering using deep parser and zero subject generation without global linguistic constraints.
- **joint deep reordering**: pre-ordering using our new deep parser with global linguistic constraints.
- **joint deep reordering + zero-subject**: pre-ordering using deep parser and zero subject generation with global linguistic constraints.

As translation metrics, we used BLEU (Papineni et al., 2002), as well as RIBES (Isozaki et al., 2010a), which is designed for measuring the quality of distant language pairs in terms of word orders.

### 4.2 Results

Table 4 shows the experimental results for six pre-reordering systems. It can be seen that the proposed method with deep parser outperforms baseline and naive reordering with surface syntactic trees. The zero subject generation can also improve both BLEU and RIBES scores, but the improvements are smaller than those with reordering. Also, joint inference with global linguistics constraints outperforms the model which solves deep syntactic relations and zero subject generation independently.

## 5 Conclusions

In this paper, we proposed a simple joint inference of deep case analysis and zero subject generation for Japanese-to-English SMT. Our parser consists of pointwise probabilistic models and a global inference with linguistic constraints. We applied our new deep parser to pre-ordering in Japanese-to-English SMT system and showed substantial improvements in automatic evaluations.

Our future work is to enhance our deep parser so that it can handle other linguistic phenomena, including causative voice, coordinations, and object ellipse. Also, the current system is built on the top of a dependency parser. The final output of our deep parser is highly influenced by the parsing errors. It would be interesting to develop a full joint inference of dependency parsing and deep syntactic analysis.

[8]This format allows us to use a spreadsheet to perform annotations.

## References

Xia Fei and McCord Michael. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proc. of ACL*.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proc. of NTCIR*.

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Two-stage pre-ordering for japanese-to-english statistical machine translation. In *Proc. IJCNLP*.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ilp solution to zero anaphora resolution. In *Proc. of ACL*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*. Association for Computational Linguistics.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head finalization: A simple reordering rule for sov languages. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Jason Katz-Brown and Michael Collins. 2008. Syntactic reordering in preprocessing for japanese → english translation: Mit system description for ntcir-7 patent translation task. In *Proc. of the NTCIR-7 Workshop Meeting*.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proc. of NAACL*.

Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2006. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proc. of the International Workshop on Spoken Language Translation*.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of CoNLL*.

Uri Lerner and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *Proc. of EMNLP*.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proc. of EMNLP*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proc. of ACL*.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of je translation. In *Proc. of Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Extracting pre-ordering rules from predicate-argument structures. In *Proc. of IJCNLP*.

日本語記述文法研究会. 2009. **現代日本語文法 *2* 第 *3* 部格と構文 第 *4* 部ヴォイス**. くろしお出版.