Language independent null subject prediction for statistical machine translation

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

1 Introduction

There are groups of languages whose grammar permits an independent clause to lack an explicit subject. These languages are called **null subject languages (NSLs)**. When we translate NSLs into non-NSLs including English and French, we need to restore the missing subjects from source sentences to generate syntactically correct target sentences.

Null subject prediction in statistical machine translation (SMT) is challenging for the following two reasons. First, SMT is not well designed to generate target phrases that are not aligned to source sentences. A special mechanism is required to predict the missing subjects accurately. Second, the null subject is not a rare phenomenon that happens only in a limited number of languages. In fact, a considerable number of languages exhibit null subjects. They include most Romance languages (Italian, Spanish, Portuguese, Romanian), Balto-Slavic languages (Czech, Polish, Russian, Croatian), Arabic, Greek, Hungarian, Turkish, Korean, Chinese, Thai and Japanese. The method for null subject predictions should ideally be language independent to handle these many NSLs with minimum effort.

There are several researches that focus on null subject predictions (Mori et al., 1999; Le Nagard and Koehn, 2010; Taira et al., 2012; Kudo et al., 2014; Russo et al., 2012; Kopeć, 2014). Most of previous work has been in the context of monolingual anaphora resolutions. They have presented null subject analyzers that predict omitted subjects only from a monolingual source sentence. Although source side analysis can easily be integrated into SMT system as preprocessing, it is not trivial to develop such source side analyzers for every NSL individually. Because of this limitation, these previous works mainly focused on specific languages and no cross-language analysis has been employed in machine translation literature.

In this paper, we propose a weakly supervised and source-language independent method for null subject prediction. Our method can be applicable to any NSLs to English translations, since it does not require any deep source side analyzers, such as morphological analyzers, parsers, nor anaphora resolvers. Our method is based on a post-editing of SMT, which inserts or corrects missing or incorrect pronouns respectively. The null subject predictors are trained from a parallel corpus and monolingual text in target language.

This paper is organized as follows. First, we describe null subject problems in machine translation with some real examples. Second, we show the basic idea of this work and its implementation based on post-editing. We conduct machine translation experiments with 18 language pairs (all pairs are "to English") to show substantial improvements in human evaluations.

2 Null subjects in machine translation

The NSLs can roughly be classified into two groups. One is **consistent NSL** and the other is **discourse-related NSL** (Camacho, 2013).

Most Romance languages and Balto-Slavic languages are consistent NSLs. In these languages, person, number and/or gender are explicitly marked in the verb, which has rich information to determine or restore the missing pronouns. In the following Italian sentence, the pronoun *lui/lei* is omitted, because the verb *vuole* encodes the person of the subject.

(1) Non vuole mangiare.
Not want to-eat.
'He/She does not want to eat.'

Languages like Chinese and Japanese are classified as discourse-related NSLs. These languages do not have any explicit morphology that can identify a null subject. In these languages, the pronoun type can only be identified from discourse-level information, including tense, mood, and aspect. Null subject detection in discourse-related NSLs tends to be more difficult than that of consistent NSLs. The following is an example of Japanese sentence without a subject.

(2) 京都 に 行き たい.
Kyoto to go want.
'I want to go to Kyoto'

Unlike the Italian example (1), the person of the subject is not encoded in the verb "行く" *(go)*. However, first person is a natural interpretation of this sentence, since this sentence expresses a subjective statement of the speaker with the auxiliary verb "たい" (*want*)

Although the mechanisms of null subject generation are different in these language groups, it can be seen that the clues to tell the missing pronouns are kept in the *local* contexts around the verb in different forms.

Standard phrase-based SMT systems e.g.,(Och, 2003) do not implement any treatments for null subject predictions. When we translate the above Italian sentence (1) with our in-house phrase-based SMT system, we can obtain the following incorrect translations in the N-best list.

(3) a. Do not want to eat.
b. I do not want to eat.

In (3-a), the correct pronoun "He/She" is not generated. Since the negative construction of English and Italian are different, we extract a word alignment from *vuole* to *want*, which loses the person type encoded in the source word. In (3-b), an incorrect pronoun "I" is inserted although the pronoun corresponding "I" does not exist in the source sentence. In this paper, we call the error (3-a) a **missing pronoun** and (3-b) an **incorrect**

pronoun respectively. According to the survey by (Russo et al., 2012), such missing and incorrect pronouns in target sentence often happen both in rule-based and statistical machine translations. They reported that about 48% and 33% of subject pronouns are not correctly translated in their Spanish to French rule-based and SMT (Moses) systems respectively.

There are several studies that addressed the null subject predictions in machine translation. These previous works are mainly implemented as a preprocessor that takes an input sentence and outputs the same sentence with the dropped subject pronouns restored. (Russo et al., 2012) proposed a rule-based preprocessor that uses a POS tagger and a list of personal and impersonal verbs. (Kudo et al., 2014) presented a joint model for null subject insertion and predicate-argument analysis. Although the preprocessing is the most straightforward and effective way to deal with the null subject problem, it is not easily applicable to other languages since it requires language dependent resources, such as verb lexicon, POS taggers and predicate-argument analyzers.

3 Null subject predictions with Post editing

3.1 Basic idea

The null subject prediction can be divided into the following two sub tasks.

- Null subject identification
- Null subject type estimation

Null subject identification is a task to detect whether the verb in a source sentence has a missing subject or not. We need a *global* sentence analyzer for this task, which can output predicate-argument structures of the sentence. Ignoring imperative sentences, it is reasonable to suppose that a subject is omitted if no subject role exists in the predicate-argument structures.

Once we can know that a subject is missing, we need to determine which pronouns should be augmented. This estimation task does not require predicate-argument structures, since, as described in the previous section, the null subject type is usually determined by *local* contexts around the verb, including verb suffixes and auxiliary verbs around the verb.

The basic idea of our null subject prediction is to perform the first null subject identification task in the target (English) side where accurate sentence analyzers are available. This target side null subject identification allows us to apply our method to many languages where source side dependency parsers are not available.

3.2 Null subject prediction algorithm

Here we describe our null subject prediction algorithm in detail. We first focus on the model for missing pronoun, and later present how to treat the incorrect pronoun with the same model. Our algorithm is implemented as a post editor of an underlying SMT decoder. First, we parse the one-best translation with a dependency parser of the target language to detect a missing subject. In this stage, we apply a simple rule-based detection algorithm. i.e., if a verb has no subject argument in its predicate-argument structures, we consider that a subject should be augmented for this verb. For instance, the sentence (3-a) is triggered by this rule based algorithm, as (3-a) has no subjects in its main verb *(want)*. One note in this step is that we may see many false-positive detections since the original source sentence may be imperative. We handle this issue in the following type estimation task by introducing a special person type "imperative". This imperative type is used when the source sentence is imperative. In addition, by choosing the imperative type, we can recover the false positive errors happened in the first identification phase.

The second null subject type estimation task is formalized as the following optimization problem:

$$\hat{z} = \underset{z \in Z}{argmax}\, S(z, f, e, t_e, v_e, v_f)$$

where

- z: a null subject candidate.
  $z \in Z = \{$*I, you, we, it, they, he/she, imperative*$\}$ [1].
- f: source sentence.
- e: target sentence (one-best translation of f).
- t_e: dependency parse tree of e.
- v_e: target verb that has a missing subject.
- v_f: source words that aligned to the target verb v_e. We assume that the baseline SMT decoder can output word alignments.
- S: scoring function over z, f, e, t_e, v_e, and v_f.

In order to combine different signals and training data to compute S, we decompose S into sub-components and infer the final result with the weighted sum of the scores that individual sub components output:

$$\begin{aligned} S(z, f, e, t_e, v_e, v_f) \quad \stackrel{\text{def}}{=} \quad & w_f \cdot S_f(z, f, v_f) + \\ & w_e \cdot S_e(z, t_e, v_e) + \\ & w_{LM} \cdot S_{LM}(z, t_e, v_e) + \\ & b_z, \end{aligned}$$

where

- $S_f(z, f, v_f)$: source side score
- $S_e(z, t_e, v_e)$: target side score
- $S_{LM}(z, t_e, v_e)$: language model score
- $\{w_f, w_e, w_{LM}\} \in \mathbb{R}^3$: confidence weights for each component
- $b_z \in \mathbb{R}$: bias term for each pronoun z. They work as default scores when no component scores are available.

In the next sections, we describe how these scores S_e, S_f, and S_{LM} are computed and how the weights w_f, w_e, w_{LM} and bias b_z are optimized.

3.2.1 Source side score: $S_f(z, f, v_f)$

This scoring function estimates how likely the target pronoun z can be generated from the source verb v_f in f. We

[1] We did not distinguish the gender when training the model, since its detection is turned out to be extremely difficult. In our experiments, we used "he" when making the final translation for our own convenience, but we can use different methods depending on the domains in which the MT system is used. For instance, we can use the singular "they".

formalize it as a simple multi-class classification task as follows:

$$S_f(z, f, v_f) = \log(P_f(z|f, v_f)),$$

where the conditional probability $P_f(z|f, v_f)$ is modeled by a maximum entropy classifier. We use contextual features extracted from f and v_f, such as verb suffix and word n-grams.

The training data for this classifier is obtained from a parallel corpus with automatic word alignments. The idea here is similar to the parser projection (Jiang and Liu, 2009) in the sense that we project the subject-verb relations from the target to the source sentence. First, we parse the target sentence. If the target sentence has a pronoun-verb dependency, we extract the pair of source context and target pronoun as a training instance. If the target sentence is imperative, we consider the target pronoun z as "imperative". In order to emulate the same decoding environment, we remove the source pronouns from the source context in this training phase. This treatment avoids the classifier from capturing the obvious relations from the source pronouns to the target pronouns.

3.2.2 Target side score: $S_e(z, t_e, v_e)$

This scoring function estimates how likely the target pronoun z can be generated from the target verb v_e and the parse tree t_e. We also formalize it as a multi-class classification task as follows:

$$S_e(z, t_e, v_e) = \log(P_e(z|t_e, v_e)).$$

We also use a maximum entropy classifier to model the conditional probability $P_e(z|t_e, v_e)$. The features of this classifier are word/POS n-grams surrounding v_e and children of v_e.

In order to train the classifier, we make artificial training data from gold English parse trees. Given a gold parse tree with a pronoun-verb relation, we remove the pronoun from the parse tree. The classifier is trained such that it can restore the removed pronoun from the parse tree without the pronoun.

3.2.3 Language model score: $S_{LM}(z, t_e, v_e)$

This scoring function computes how the language model score changes after inserting the pronoun z as a subject of v_e. In order to ignore the effect of sentence length, we do not use the raw language model score, but use the difference after inserting the pronoun z. We apply a naive rule-based algorithm to determine the position to which the pronoun z is inserted[2]. In addition, we change the verb suffixes so that the pronoun z can agree with the main verb. e.g., *He want* $\rightarrow$ *He want***s**.

3.2.4 Weight and Bias optimization with MERT

We make a small development data to optimize the component weights w_f, w_e, w_{LM} and bias b_z. First, we make an artificial parallel corpus by translating a large amount of source sentences with an underlying SMT system. Second, we use the rule-based null subject identification algorithm to sample parallel sentences that have no subject in the target part. We finally generate seven translation hypotheses by inserting different pronouns $z \in Z$ and employ a human evaluation to compare which hypothesis is valid as a translation. Each evaluator can rate each translation from 0 (very bad) to 6 (very good). The final weights and bias are optimized through MERT-like (Och, 2003) optimization algorithm so that the average of the evaluation scores is maximized.

Since this optimization procedure requires human evaluations, our algorithm cannot be considered as a fully unsupervised algorithm. However, the human evaluation task is much easier than making gold null subject annotations or reference sentences.

[2]The pronoun is inserted just before the main verb. If there are auxiliary verbs, we insert the pronoun before them.

3.3 Model for incorrect pronouns

We make the null subject predictor for incorrect pronouns as follows.

1. In rule-based null subject detection, we select instances where the target pronoun is aligned to null or source words that are not pronouns.
2. We remove the target pronoun from the parse tree t_e when computing the target side score. For instance, we remove "I" from (3-b). In the language model computation, we replace the existing pronoun with z.
3. We use different weight and bias parameters that are trained from different training data extracted with the detection algorithm described in 1.

4 Experiments

4.1 Experimental settings

We conducted subjective evaluations to compare two MT systems, 1) baseline phrase-based SMT system (Och, 2003), and 2) the same baseline SMT with our null subject predictors. We first translated a large amount of source sentences sampled from the web with the two systems. From these initial translations, we randomly sampled 500 sentences for each language pair, which were translated differently with the two systems. We then conducted subjective evaluations in which human raters were asked to judge translations on a scale from 0 (very bad) to 6 (very good).

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. The source side classifier is trained with about 1M parallel sentences sampled from the same in-house collection. The target side classifier was trained with our in-house English treebank. We made 800 development data from the web for MERT-based parameter tuning.

4.2 Results and Discussions

Table 1 shows the experimental results of our models for missing and incorrect pronouns respectively. One note in this table is that the test sets handled by these two models are disjoint. The end-to-end evaluation scores can be obtained by aggregating these two results.

We can see that our models substantially improve the translation qualities regardless of the NSL groups. The improvement on Japanese to English translation is remarkable with respect to AES and CR.

The affected sentences with our model are generally small (around 1% for all languages). We found that the ratio of sentences that contain null subjects highly rely on the domain of

Table 1: Results of subjective evaluations (Target is all English)

	missing pronouns		incorrect pronouns	
Source Lang.	CR %	Δ AES	CR %	Δ AES
Spanish	2.72	+0.12	0.61	−0.60
Italian	1.01	+0.19*	0.48	+0.02
Portuguese	1.28	+0.02	0.30	−0.89
Romanian	0.83	+0.16*	0.58	+0.12*
Polish	1.94	+0.20*	0.73	−0.06
Czech	0.92	+0.38*	0.68	+0.34*
Russian	0.75	+0.21*	0.38	−0.22
Slovak	0.79	−0.02	0.57	+0.22*
Slovenian	1.01	+0.20*	0.30	−0.48
Bulgarian	0.96	+0.19*	0.62	+0.30*
Hungarian	0.59	+0.19*	0.73	+0.22*
Greek	0.75	+0.38*	0.67	+0.31*
Arabic	0.57	+0.10	0.77	−0.29
Turkish	1.22	+0.26*	0.83	+0.09
Japanese	1.55	+0.19*	5.48	+0.40*
Korean	1.29	+0.08*	0.66	+0.04
Chinese	0.76	+0.36*	0.30	+0.19*
Thai	0.10	+0.26*	1.92	+0.13*

CR (Change Rate): Ratio of affected translations to all translations.
Δ AES (Averaged Eval Score): The diff of averaged subjective evaluation score against the baseline system.
Asterisked results are significantly improved with bootstrapping resampling on the test corpus (Koehn, 2004) ($p < 0.05$)

Table 2: Accuracies of classifiers and MERT weights (missing pros.)

Source Lang.	Acc. (S_f)	Acc. (S_e)	w_f	w_e	w_{LM}
Spanish	0.80	0.67	0.85	0.03	0.13
Italian	0.81	0.67	0.83	0.01	0.15
Portuguese	0.80	0.67	0.70	0.00	0.30
Romanian	0.78	0.67	0.61	0.14	0.25
Polish	0.81	0.67	0.77	0.01	0.21
Czech	0.84	0.67	0.74	0.03	0.23
Russian	0.75	0.67	0.66	0.03	0.31
Slovak	0.82	0.67	0.85	0.01	0.14
Slovenian	0.76	0.67	0.61	0.15	0.24
Bulgarian	0.83	0.67	0.70	0.03	0.27
Hungarian	0.73	0.67	0.64	0.07	0.29
Greek	0.80	0.67	0.60	0.04	0.35
Arabic	0.81	0.67	0.69	0.09	0.22
Turkish	0.72	0.67	0.64	0.14	0.21
Japanese	0.56	0.67	0.44	0.14	0.42
Korean	0.51	0.67	0.62	0.20	0.18
Chinese	0.52	0.67	0.61	0.10	0.29
Thai	0.49	0.67	0.28	0.23	0.49

Note: Accuracies of target score S_e are the same for all language pairs, since the target language is all English and we used the same model.

the corpus. When we apply our model to other corpus that contain short conversational texts, the change ratios are almost doubled while keeping almost the same qualities.

Our error analysis reveals that parsing and POS tagging errors often cause incorrect predictions. There are a lot of words in English that can have multiple POS interpretations, e.g., *saw, rose, hope*. When these words are incorrectly analyzed as a verb, our model may insert a pronoun by mistake.

Table 2 summarizes the accuracies of source and target classifiers[3], and normalized MERT weights assigned to source, target, and language model components. It can be seen in general that consistent NSLs, e.g., Spanish, Italian, and Polish, rely more on source side information, while discourse-related NSLs, e.g., Japanese, Chinese, Korean, and Thai, rely more on target side information. These results support our linguistic analysis described in section 2. They also imply that null subjects in consistent NSLs can solely be identified with source side information, but that is not the case in discourse-related NSLs and combining source and target side analysis is necessary.

5 Conclusions

In this paper, we proposed a language independent null subject prediction algorithm for machine translation. Our method is based on post-editing of SMT, which inserts or corrects missing or incorrect pronouns respectively. We conducted machine translation experiments with 18 language pairs to show substantial improvements in human evaluations.

Our future work includes two main directions. One short-term direction is to enhance the target-side English parser so that it can analyze machine translated sentences correctly, which often contain many ungrammatical expressions. Second long-term direction is the integration of our model into the decoder. As we described, the combination of source and target side information is necessary especially for discourse-related NSLs. It is more natural to solve the null subject issues not as a post-editing but in a decoder for these NSLs.

[3]The accuracies here are the macro averages of F-scores computed with a cross validation

References

José A. Camacho. 2013. *Null Subjects*. Cambridge Studies in Linguistics 137.

Wenbin Jiang and Qun Liu. 2009. Automatic Adaptation of Annotation Standards for Dependency Parsing: Using Projected Treebank As Source Corpus. In *Proc. of IWPT*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation.

Mateusz Kopeć. 2014. Zero subject detection for Polish. In *Proc. of EACL*.

Taku Kudo, Hiroshi Ichikawa, and Hideto Kazawa. 2014. A joint inference of deep case analysis and zero subject generation for Japanese-to-English statistical machine translation. In *Proc. of ACL*.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation*.

Tatsunori Mori, Mamoru Matsuo, and Hiroshi Nakagawa. 1999. Zero-subject Resolution Using Linguistic Constraints and Defaults: The Case of Japanese Instruction Manuals. *Machine Translation*, 14:231–245.

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

Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. 2012. Improving machine translation of null subjects in Italian and Spanish. In *Proc. of EACL*.

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*.