POSTECH's Statistical Machine Translation Systems for NTCIR-9 PatentMT Task (English-to-Japanese)

Hwidong Na
Pohang University of Science and Technology (POSTECH),
leona@postech.ac.kr

Jin-Ji Li *
Pohang University of Science and Technology (POSTECH),
ljj@postech.ac.kr

Se-Jong Kim
Pohang University of Science and Technology(POSTECH),
sejong@postech.ac.kr

Jong-Hyeok Lee
Pohang University of Science and Technology(POSTECH),
jhlee@postech.ac.kr

ABSTRACT

We present a two-stage statistical machine translation (SMT) framework as proposed in [10]. In the first stage, it resolves structural differences using a phrase-based SMT with syntax-aided preprocessing (SMT1). In the second stage, it resolves lexical differences using a phrase-based SMT (SMT2). For morpho-syntactically divergent language pairs such as English-Japanese, this framework strengthens the structural transfer of phrase-based SMT whose capability for lexical transfer has already been well established.

Translation from a morphologically-poor language (isolating language) to a morphologically-rich one (agglutinative language) is more difficult than the converse. Our proposed approach fills morpho-syntactic gaps with the transferred syntactic roles. It facilitates the generation of adequate case markers that appear only in the target languages.

In addition, we take into consideration word order differences between English and Japanese. Our proposed approach moves modality-bearing words to the end of a sentence as Japanese is a verb-final language.

Finally, as they are complementary, we combine the two above-mentioned approaches in a cascaded model to perform a more generalized structural transfer. The input sentences are syntactically reordered, and the thematic divergences of the *subject* and *object* relations of the reordered sentences are then resolved, and vice versa (transfer and reorder).

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Natural Language Processing*

General Terms

Experimentation, Languages

Keywords

Phrase-based SMT, reordering, morpho-syntatic, preprocessing

Team Name

[KLE][POSTECH]

*This work was mainly done during her Ph.D course.

Subtasks/Languages

[Patent Translation][English-to-Japanese]

External Resources Used

[Stanford Parser][CaboCha][Mecab][GIZA++][Moses]

1. INTRODUCTION

Resolving lexical and structural ambiguities both within a language (monolingual ambiguity) and between two languages (bilingual ambiguity) are major problems in all machine translation (MT) systems. In most MT systems, monolingual ambiguity is usually resolved in the analysis phase of the source language. Bilingual ambiguity refers to the translational (transfer) ambiguity caused by lexical and structural differences between languages. Therefore, an effective lexical and structural transfer directly impacts the performance of an MT system.

Many different techniques have been developed with the goal of enabling statistical machine translation (SMT) systems to resolve transfer ambiguities. The capability of modern SMT systems for lexical transfer has been unequivocally established in phrase-based SMT (PBSMT) system, but the structural transfer of these systems is still poor. Augmenting SMT systems with the capability for resolving structural differences has become the main focus area in current research e.g. syntax-aided PBSMT and syntax-based SMT.

Syntax-aided PBSMT refers to independent sub-components such as pre- or post-processing approaches. These approaches augment a phrase-based system for the structural transfer. Syntax-aided phrase-based SMT is efficient in encoding the linguistically motivated features because independent sub-components do not introduce additional complexity to the decoder. Syntax-based SMT increases decoding complexity to a greater extent than syntax-aided methods because it directly embeds the syntax in the translation model. Thus, syntax-aided PBSMT is a loosely-coupled method, while syntax-based SMT is a tightly-coupled one. In this study, we focus on the syntax-aided preprocessing methods, proposed in [10]. Syntax-aided preprocessing methods provide the corpora for intermediate languages used in two-stage SMT systems.

English-Japanese is a morpho-syntactically divergent language pair. In this instance, the direction of translation is

from a morphologically-poor language to a morphologically-rich one. In addition, syntactic roles are implicitly expressed by word order in English, while in Japanese they are explicitly expressed by case markers. These syntactic roles are frequently transferred into other syntactic roles during translating. Among various kinds of structural differences, we focus on the thematic divergences of syntactic roles such as *subject* and *object* between source and target languages. Our proposed approach fills the morpho-syntactic gaps with the transferred syntactic roles in order to resolve thematic divergences.

From the viewpoint of word order typology, English is a subject-verb-object (SVO) language with rigid word order while Japanese belongs to a SOV language with flexible word order. This difference in verb positioning causes difficulty in generating correct verbal phrases for target languages. For syntactic reordering, we move modality-bearing words to the end of sentences as Japanese is a verb-final language.

2. RELATED WORKS

2.1 Morpho-syntactic reconstruction

The purpose of mopho-syntactic reconstruction is two-fold. One is to decrease the morpho-syntactic differences between the source and target languages, and the other is to access syntactic information at the word level.

For morpho-syntactically divergent language pairs, the granularity of lexical units and the representation methods of syntax are completely different. Some researchers insert pseudo words such as functional words unique in the target language [15], or syntactic relations [7, 8] into source sentences to fill the morpho-syntactic gaps between two languages.

Other techniques include the use of supertags [6] or micro-tags [1] to enrich the word with syntactic information. In this approach, each word is supertagged using Lexicalized Tree-Adjoining Grammar (LTAG) or Combinatory Categorial Grammar (CCG) supertag sets or enriched with micro-tags i.e. per-word projections of chunk labels.

2.2 Syntactic reordering

One of the major weaknesses of phrase-based SMT is long-distance reordering. Syntactic reordering restructures the source sentences into a more target-like word order using syntactic information as a guide. That is, as pre-processing of SMT, the input are first syntactically analyzed, and then reordered according to reordering rules. The reordering rules can be hand-crafted [2, 8, 12, 13, 14] or automatically generated [3, 4, 9, 11, 16].

Syntactic reordering effectively compensates the low long-distance reordering power of phrase-based SMT without introducing additional complexity to the decoding process. We can also turn on the distortion models to capture local reordering not captured in the preprocessing stage.

Most syntactic reordering methods deterministically reorder input sentences. As a result, once there is faulty reordering, the mistakes cannot be recovered. To resolve this problem, word lattice reordering (in which the preference is encoded as the path probability in the lattice) has been proposed as input instead of syntactic reordering [3, 9, 16]. However, we do not take that approach in this paper.

3. METHOD

Table 1: 6 Japanese case markers

Japanese grammatical functions	Case markers
Subject; Object	が(ga)
Object; Path	を(wo)
Genitive; Subject	の(no)
Dative object; Location	に(ni), には(niwa)
Topic	は(wa)

We present a two-stage framework that first resolves structural differences using a phrase-based SMT with syntax-aided preprocessing (SMT1), and then resolves lexical differences using a phrase-based SMT (SMT2), as proposed in [10]. SMT1 and SMT2 train E-E' and E'-J corpora, respectively, where E' is a corpus in the intermediate language.

3.1 Transferring syntactic roles

In the case of English-Japanese, SVO patterns retain structural transfer ambiguities such as thematic divergences during translating. We propose a preprocessing method that transfers the syntactic roles of SVO patterns. As a result of our proposed method, the transferred syntactic roles promote the generation of correct case markers in the target languages. This transfer phase is realized in SMT1. The process is similar to the structural transfer phase of a traditional transfer-based machine translation but without the lexical transfer. We leave the lexical transfer to the SMT decoder as it is one of the greatest strengths of a phrase-based SMT system.

To transfer syntactic roles of SVO patterns, we identify grammatical relations in the source languages. More specifically, we adopt grammatical relations that are produced by the Stanford English typed dependency parsers. The previous work provides 7 grammatical roles that are related to *subject* and *object* in English. [1].

Training data for SMT1 is automatically constructed using a word-aligned and dependency-parsed English-Japanese bilingual corpora. In other words, we generate the gold standard data of intermediate sentences (E'). More specifically, for each word with a *subject* or *object* relation in the source sentences, a case marker of the target language is assigned via the word-alignment information.

This process should take into consideration the following: (1) Which case marker in the target language can be observed through the structural transfer? (2) How many case markers should we consider? If there are too many case markers to predict, it will decrease the prediction accuracy and the overall translation performance. Therefore, we focus on only certain kinds of relations.

For each word with a *subject* or *object* relation in the source sentence, if the aligned word is a content morpheme, we find the Japanese *bunsetsu* that contains it, and the corresponding postpositions. If the corresponding postpositions does not belong to one of the case markers listed in Table 1, then we set this case to 'null'.

3.2 Syntactic reordering

[1] *nsubj* (nominal subjects), *nsubjpass* (passive nominal subject), *scubj* (clausal subject), *csubjpaa* (passive clausal subject), *dobj* (direct object), *iobj* (indirect object), and *cop* (copular)

Algorithm 1 English syntactic reordering rules for predicates

Input: *L_Children*, *R_Children* of a Predicate P
Output: *L_Advcl*, *L_Other*, *L_FromRight*, *L_Modal*, *R_Modal*, *R_Other*

```
for node N in L_Children do
  if dep.relation of N ∈ {aux, auxpass, neg, cop} then
    L_Modal ← L_Modal ∪ {N}
  else
    L_Other ← L_Other ∪ {N}
  end if
end for
for node N in R_Children do
  if dep.relation of N ∈ {prt} then
    R_Modal ← R_Modal ∪ {N}
  else if dep.relation of N ∈ {advcl} then
    L_Advcl ← L_Advcl ∪ {N}
  else if dep.relation of N ∈ {conj, cc, punct} then
    R_Other ← R_Other ∪ {N}
  else
    L_FromRight ← L_FromRight ∪ {N}
  end if
end for
```

Since Japanese is a head-final language, all to other elements should take pre-verbal positions in Japanese sentences. In [14], they used verb precedence to organize a verb group and move it to the end of the sentence. Although they did not use the term "modality-bearing word", the elements that they grouped are closely related to that such as phrasal verb particle, auxiliary verb, passive auxiliary verb, and negation.

We place modality-bearing words [2] close to their verbal heads. Every predicate in an English D-tree consists of left children (*L_Children*) and right children (*R_Children*).

From the left children, the modality-bearing words (*L_Modal*) are relocated near the predicate, while the other elements (*L_Other*) remain on the left side of the predicate. For the right children, the process is slightly different. Modality-bearing words(*R_Modal*) are relocated near the predicate (just like the *L_Modal*). However, most right children will be moved to the left side of the predicate (*L_FromRight*) since Japanese is a head-final language. A right child belonging to *R_Other* is either a coordination conjunction or a punctuation. For the adverbial clause modifier, we move it to the leftmost position (*L_Advcl*) of the given predicate considering it is usually translated at the beginning of Japanese sentences. After then, we obtain the reordered sentence by traversing in the following order: L_Advcl, L_Other, $L\_FromRight$, L_Modal, R_Modal, P, $and R\_Other$ for each predicate P. Algorithm 1 gives more the details on the procedure.

3.3 Structural transfer as preprocessing

Syntactic reordering and resolving of thematic divergences are complementary operations. Therefore, we combine the two approaches to perform a more generalized structural transfer. We simply cascade the two approaches by first syntactically reordering the input, then resolving the thematic divergence of *subject* and *object* relations of the reordered sentences, and vice versa (transfer and reorder).

[2]The followings are a set of dependency relations defined in the Stanford English typed dependency parser: aux (auxiliary), auxpass (passive auxiliary), neg (negation modifier), cop (copular), prt (phrasal verb particle), advcl (adverbial clause modifier), conj (conjunction), cc (coordination), and punct (punctuation)

Table 2: System description of our runs. Transfer-Reorder and Reorder-Transfer are the cascaded methods.

Run ID	E' Generation	E-E'	E'-J
KLE-01	Transfer	PBSMT	Hiero
KLE-02	Transfer	PBSMT	PBSMT
KLE-03	Transfer-Reorder	PBSMT	PBSMT
KLE-04	Reorder-Transfer	PBSMT	PBSMT
KLE-05	Hiero E-J		

Table 3: Official evaluation results for the primary run

Run ID	Adequacy	Pairwise	Tie
KLE-01	2.3533	0.4342	0.3095
TOP	3.67	0.6947	0.1977

4. RESULT

We submitted five formal runs as follows. For of each runs, except the baseline (KLE-05), we built a pair of SMT systems for E-E' and E'-J. We trained each pair of systems using E-E' and E'-J parallel corpora, where E' was generated by our proposed methods. At the decoding stage, we allowed unlimited distortion (distortion-limit = -1) for the PBSMT systems. KLE-05 is a hierarchical PBSMT (Hiero) system. We used "moses" and "moses-chart" decoders as PBSMT and Hiero systems, respectively. Table 2 gives the systems.

Table 3 and 4 show our official and automatic evaluation results of ours as well as the top scorer, respectively. Please consult [5] for the evaluation metrics.

5. DISCUSSION AND FUTURE WORK

Table 5 shows various example sentences from our runs that required resolution of thematic divergences and global reordering.

- In the first example, the object "states" and its predicate "show" were moved to the end of the Japanese sentence. Our proposed method correctly inserted the object case marker "_を" in the first stage following which Hiero reordered the object and predicate with a pseudo word. In contrast, when Hiero alone was applied from start to finish, it generated the wrong case marker "は", and failed to reorder the object and predicate.

- In the second example, the subject "The image forming apparatus" has the adverbial case marker "として(as)" in Japanese. Although our proposed method wrongly inserted the case marker "の" that represents either subject or genitive, the results appeared to be better than that obtained using Hiero alone.

- The predicate "comprises" was translated into "からなり" in our proposed method. However, wehn only Hiero was applied, it was translated as "であり".

Table 5: Selected examples that required resolution of thematic divergences

ID	20040623_2004184512=20050621_11157283-135
Source	FIGS . 6 and 7 show states in which the switch S1 is connected to the output terminal .
E'	FIGS . 6 and 7 show states _を in which the switch S1 is connected to the output terminal .
KLE-01	図６及び図７は、スイッチＳ１の出力端子に接続されている状態を示す。
Reference	図６，図７は、スイッチ部Ｓ１が接続された状態を示している。
KLE-05	図６及び図７に示す状態では、スイッチＳ１が出力端子に接続されている。
ID	20041124_2004339567=20050624_11166995-604
Source	The image forming apparatus will be exemplified by an electrophotographic copying machine .
E'	The image forming apparatus _の will be exemplified by an electrophotographic copying machine .
KLE-01	画像形成装置の 電子写真複写機を例示する。
Reference	本実施の形態では画像形成装置として 電子写真方式の複写機を例に挙げて説明する。
KLE-05	この画像形成装置は 電子写真複写機について説明する。
ID	20040617_2004179699=20050615_11154415-853
Source	The spring 25 comprises a coil spring , and inserted into the valve control chamber 23 in a compressed state .
E'	The spring 25 _はcomprises a coil spring , and inserted into the valve control chamber 23 in a compressed state .
KLE-01	バネ２５はコイルスプリングからなり 、圧縮状態で弁制御室２３に挿入される。
Reference	バネ２５はコイルスプリングからなり 、圧縮状態で弁制御室２３に挿入配置される。
KLE-05	バネ２５はコイルバネであり 、弁制御室２３内に挿入され、圧縮された状態にある。
ID	20041122_2004337309=20051006_11244083-242
Source	First , as shown in FIG . 3A , a lead frame 52 in which an interconnect pattern is formed is placed on a sealing tape 21 .
E'	First , as shown in FIG . 3A , a lead frame 52 in which an interconnect pattern _をis formed is placed on a sealing tape 21 .
KLE-01	まず、図３（ａ）に示すように、リードフレーム５２を封止テープ２１の上に載置する。形成された配線パターンを
Reference	まず、図３（ａ）に示す工程で、配線パターンが形成されたリード フレーム５２を 封止テープ２１の上に載置する。
KLE-05	まず、図３（ａ）に示すように、封止テープ２１上に載置されたリードフレーム５２には、配線パターンが形成されている。
ID	20040618_2004181735=20050617_11154625-675
Source	This causes transition from the ECC decoding state EDST to the self-refresh state SRST .
E'	This _にcauses transition from the ECC decoding state EDST to the self-refresh state SRST .
KLE-01	これにより、ＥＣＣデコード状態ＥＤＳＴからセルフリフレッシュ状態ＳＲＳＴに遷移する。
Reference	これにより、ＥＣＣデコード状態ＥＤＳＴからセルフリフレッシュ状態ＳＲＳＴに遷移する。
KLE-05	これにより、ＥＣＣデコード状態ＥＤＳＴからセルフリフレッシュ状態ＳＲＳＴに遷移する。

Table 4: Automatic evaluation results for the submitted formal runs

Run ID	BLEU	NIST	RIBES
KLE-01	0.3403	8.2467	0.690476
KLE-02	0.2982	7.8441	0.645376
KLE-03	0.2851	7.6125	0.640937
KLE-04	0.2839	7.6761	0.641663
KLE-05	0.3510	8.2846	0.742908
TOP	0.3948	8.7134	0.78129

- The clause “an interconnect pattern is formed” modifiers the subject “a lead frame 52”, and the subject becomes the object in Japanese. Our proposed method inserted the object case marker “_を” (the first stage), but Hiero subsequently failed to reorder the clause correctly (the second stage). On the other and, standalone Hiero both failed to insert the correct case marker and reorder the clause.

- The subject “This” has the adverbial case marker “により(by)”in Japanese. Even though our proposed method inserted the wrong case marker “_に”, Hiero subsequently translated the case marker “により(by)” correctly. Thus, it appears that there exist the phrase pair 〈 “This _にcause”, “これにより、” 〉. In this case, thematic divergence was resolved automatically within the translation rule.

Although we got high scores in automatic evaluation, the official result shows that our primary run was ranked the 10th in terms of adequacy, and 7th in terms of acceptability. We actually conducted the experiment in a manner that was different from how we had intended to conduct it for the cascades systems. Since reordering of PBSMT is one of the major weaknesses, it would have been impractical to let PBSMT deal with global reordering during decoding. The differences of automatic evaluation scores between KLE-01 and KLE-02 also reveal that PBSMT is less effective at global reordering than Hiero which facilitates formally syntatic reordering. A more reasonable reordering methods would be to syntactically reordering at decoding stage as well as training. That is, the correct organization of two cascaded systems is as follows:

- Transfer-Reorder: E -> reorder(E') -> J
- Reorder-Transfer: reorder(E) -> E' -> J

We will examine these methods in future work.

Acknowledgements

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST No. 2011-0003029), in part by the Korea Ministry of Knowledge Economy (MKE) under Grant No.2009-S-034-01, and in part by the BK 21 Project in 2011.

6. REFERENCES

[1] M. Cettolo, M. Federico, D. Pighin, and N. Bertoldi. Shallow-syntax phrase-based translation: Joint versus factored string-to-chunk models. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawaii, October 2008.

[2] M. Collins, P. Koehn, and I. Kucerova. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] J. Elming. Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 46–54, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[4] D. Genzel. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 376–384, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[5] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, June 2011.

[6] H. Hassan, K. Sima'an, and A. Way. Supertagged phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 288–295, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[7] G. Hong, S.-W. Lee, and H.-C. Rim. Bridging morpho-syntactic gap between source and target sentences for english-korean statistical machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 233–236, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[8] Y.-S. Lee, B. Zhao, and X. Luo. Constituent reordering and syntax models for english-to-japanese statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 626–634, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[9] C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 720–727, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[10] J.-J. Li, J. Kim, and J.-H. Lee. Resolving thematic divergences of subject and object relations for smt. *Intl J. of Computer Processing of Oriental Languages*, (submitted), 2010.

[11] K. Visweswariah, J. Navratil, J. Sorensen, V. Chenthamarakshan, and N. Kambhatla. Syntax based reordering with automatically derived rules for improved statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1119–1127, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[12] C. Wang, M. Collins, and P. Koehn. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[13] F. Xia and M. McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

[14] P. Xu, J. Kang, M. Ringgaard, and F. Och. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[15] Y. Xu and S. Seneff. Two-stage translation: A combined linguistic and statistical machine translation framework. In *Proceedings of the Association for Machine Translation in the Americas 2008*, pages 222–231, Waikiki, Hawaii, USA, Oct 21–Oct 25 2008. AMTA.

[16] Y. Zhang, R. Zens, and H. Ney. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, SSST '07, pages 1–8, Morristown, NJ, USA, 2007. Association for Computational Linguistics.