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

Hwidong Na, Jin-Ji Li, Se-Jong Kim, and Jong-Hyeok Lee
Pohang university of science and technology (POSTECH), Pohang, Korea

Introduction

Two-stage SMT (E → E' → J): 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). SMT1 and SMT2 train E-E' and E'-J corpora, respectively, where E' is a corpus in the intermediate language.

For morphological divergence: English is a **morphologically-poor** language while Japanese is 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**.

For word order difference: English is a subject-verb-object (**SVO**) language **with rigid word order** while Japanese belongs to a **SOV** language **with flexible word order**. The modality-bearing words are located at end of sentences as Japanese is a **head-final** language.

Transferring syntactic roles

For thematic divergence: In the case of English-Japanese, syntactic roles such as ***subject*** and ***object*** are frequently transferred into other syntactic roles during translating.

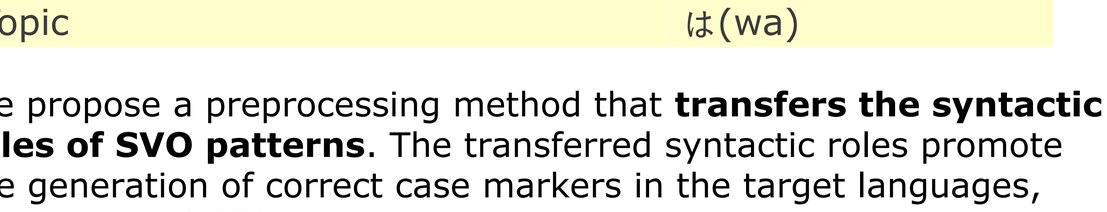
Japanese grammatical functions	Case markers
Subject; **Object**	が(ga)
Object; Path	を(wo)
Genitive; **Subject**	の(no)
Dative object; Location	に(ni), には(niwa)
Topic	は(wa)

We propose a preprocessing method that **transfers the syntactic roles of SVO patterns**. The transferred syntactic roles promote the generation of correct case markers in the target languages, **realized in SMT1**.

E' is automatically constructed using a word-aligned and dependency-parsed English-Japanese bilingual corpora. 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.

Syntactic reordering

For predicate: Since Japanese is a head-final language, **modality-bearing words** such as phrasal verb particle, auxiliary verb, passive auxiliary verb, and negation should **take pre-verbal positions** in Japanese sentences.



	Dependency label
LA	advcl (adverbial clause modifier)
LO	- (default)
LR	- (default)
LM	aux (auxiliary), auxpass (passive aux), neg (negation modifier) , cop (copular)
RM	prt (phrasal verb particle)
RO	conj (conjuction), cc (coordination), punt (punctuation)

E' is automatically constructed using dependency-parsed English monolingual corpus. For each predicate, we obtain the reordered sentence.

Cascade methods

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 (KLE-04), and vice versa (KLE-03).

Result

Run ID	Adequacy	Pairwise	Tie	BLEU	NIST	RIBES
KLE-01	2.3533	0.4342	0.3095	0.3404	8.2467	0.690476
TOP	3.67	0.6947	0.1977	0.3948	8.7134	0.78129

Run ID	E' Generation	E-E'	E'-J	BLEU	NIST	RIBES
KLE-01	Transfer	PBSMT	Hiero	0.3404	8.2467	0.690476
KLE-02	Transfer	PBSMT	PBSMT	0.2982	7.8411	0.645376
KLE-03	Transfer-Reorder	PBSMT	PBSMT	0.2851	7.6125	0.640937
KLE-04	Reorder-Transfer	PBSMT	PBSMT	0.2839	7.6761	0.641663
KLE-05	Hiero E-J			0.3501	8.2846	0.740298

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.

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**.

Discussion

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。
Ref	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。
Ref	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。
Ref	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	... a lead frame 52 in which an interconnect pattern is formed is placed on a sealing tape 21 .
E'	... a lead frame 52 in which an interconnect pattern を is formed is placed on a sealing tape 21 .
KLE-01	... 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
Ref	... 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	... 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。
Ref	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。

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 syntactic 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

However, it **also fail** to improve the translation quality (BLEU).

TEMPLATE DESIGN © 2008
www.PosterPresentations.com