Integrating Empty Category Detection into Preordering Machine Translation

Shunsuke TAKENO, Masaaki NAGATA, Kazuhide YAMAMOTO

Abstract:

We propose a method for integrating Japanese empty category detection into the preordering process of Japanese-to-English statistical machine translation.
First, we apply machine-learning-based empty category detection to estimate the position and the type of empty categories in the constituent tree of the source sentence.
Then, we apply discriminative preordering to the augmented constituent tree in which empty categories are treated as if they are normal lexical symbols.
We find that it is effective to filter empty categories based on the confidence of estimation.
Our experiments show that, for the IWSLT dataset consisting of short travel conversations, the insertion of empty categories alone improves the BLEU score from 33.2 to 34.3 and the RIBES score from 76.3 to 78.7, which imply that reordering has improved For the KFTT dataset consisting of Wikipedia sentences, the proposed preordering method considering empty categories improves the BLEU score from 19.9 to 20.2 and the RIBES score from 66.2 to 66.3, which shows both translation and reordering have improved slightly.

	BLEU		RIBES	
	w/o EC	w/ EC	w/o EC	w/ EC
BASELINE	33.1	33.6 (+0.5)	74.2	75.7 (+1.5)
REORDERING(C)	33.2 (+0.1)	**34.3** (+1.2)	76.3 (+2.1)	**78.8** (+4.6)
REORDERING(H)	33.8	34.1	76.8	78.6

Proposal:

家 に は 早く 帰る ほう が よい 。

↓ EC detection (Takeno+2015)

$(pro)_1$ $(pro)_2$家 に は 早く 帰る ほう が よい 。

↓ Reordering (Hoshino+2015)

$(pro)_1$ よい が ほう $(pro)_2$ 帰る 早く は に 家 。

It 's better if you come home early.

- **Preordering model** alleviate the word order problem w/ EC
  Plain insertion of EC slightly improve due to ***word order*** problem including ECs
  Word alignments about EC are needed for building the model
- **Elimination of unreliable ECs** refines EC detection
  Accuracy of structural parse is insufficient for practical usage
  Cutting lower confidence of ECs alleviate the problem