Source Phrase Segmentation and Translation for Japanese to English Translation Using Dependency Structure

Junki Matsuo, Kenichi Ohwada, and Mamoru Komachi

Graduate School of System Design, Tokyo Metropolitan University, Japan

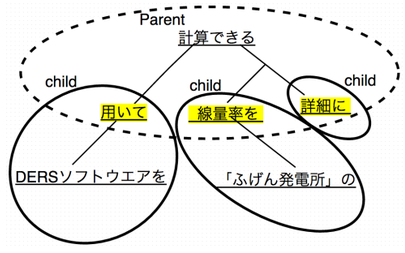
① Abstract

We propose a linguistically motivated approach based on segmenting a source phrase using dependency structure and translating each phrase with PBSMT. This work presents the results of our method on Japanese-English translation and discusses potential improvements.

② Source Phrase Segmentation and Translation Using Dependency Structure

Algorithm

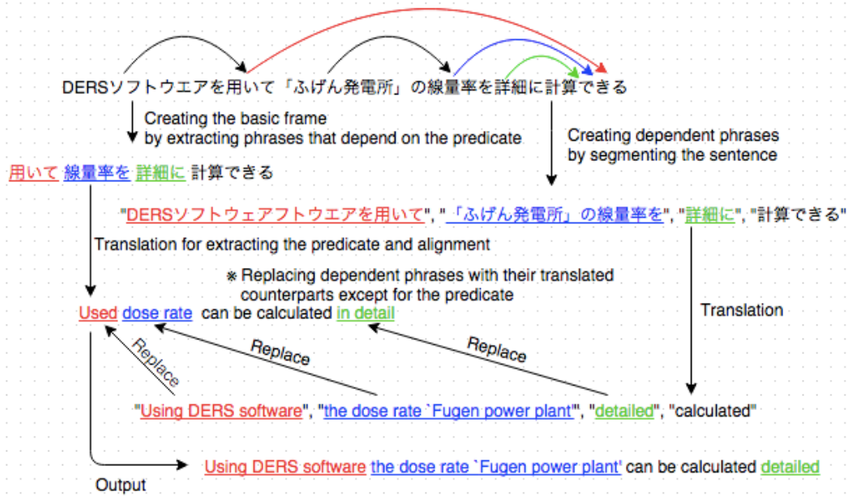
1. Our method creates a basic frame and dependent phrases. (Basic frame : broken line circle, Dependent phrases : solid line circle.)
2. The basic frame and the dependent phrases are translated by a decoder.
3. The anchor words of the basic frame are replaced with the translations of the corresponding dependent phrases. (Anchor words : the yellow under-lined words)



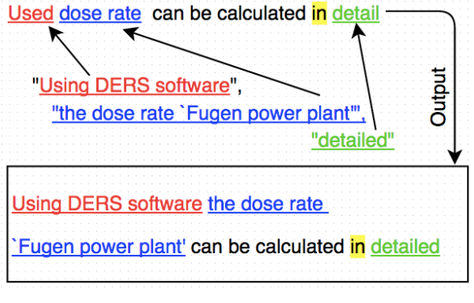
The difference between the first and second methods is not to replace a preposition that follows the predicate in the basic frame.

Because the translation of dependent phrases in our first method uses the language and translation models optimized for a sentence, our first method might not be able to use a model optimized for translating phrases.

An example of the first method



An example of the second method



③ Experimental Settings

Decoder : Moses (2.1.1) default settings
Corpus : The ASPEC (3 million parallel sentences)
Segmentation : JUMAN (7.0)
Alignment : GIZA++ (1.0.7)
MERT : Performed on the full dev-set
Dependency : CaboCha(0.68)
Baseline : Standard output

Our method only uses CaboCha for creating a basic frame and dependent phrases. (Not segmentation)

④ Experimental Results

	BLEU	RIBES
no-seg&trans	18.32	0.641456
seg&trans	15.85	0.628897
seg&trans (w/o prep)	15.72	0.628463

no-seg&trans ... Standard Moses output
seg&trans ... The first method
seg&trans ... The second method

⑤ Error Analysis (in 100 sentences)

We investigate the reason why BLEU and RIBES fall off. The number of the pairs of input and reference that have different voices (active or passive) is 35.

We observed that translation of dependent phrases is the most frequent error type. It is because the language and the translation models are not optimized for translating phrases.

error types	frequency
Dependency parsing	3
Translation of a basic frame	18
Translation of dependent phrases	46
Total (Each error may overlap)	57

Error Example

・Translation of the basic frame
“雷撃比は等しい”
↓
“The equal to lightning stroke ratio”
→ *Ungrammatical Sentence*

・Translation of the dependent phrases
“解決を” “DERSに”
↓ ↓
“We solve” “In DERS”
→ *Not an NP* → *Missing context*

⑥ Conclusion & Future Work

- Our finding is that our proposed methods have three problems: a dependency parsing, translation of a basic frame and translation of dependent phrases.
- We plan to optimize the language and the translation models for translating phrases.