Domain Adaptation and Attention-Based Unknown Word Replacement in Chinese-to-Japanese Neural Machine Translation

Kazuma Hashimoto, Akiko Eriguchi, and Yoshimasa Tsuruoka

The University of Tokyo

The UT-KAY system

Overview: The UT-KAY System for Chinese-to-Japanese Machine Translation

有关Yukon和西北领域、Hudson和James湾、北部魁北克、拉布拉多、Greenland的污染物质的信息从文献、组织、研究者方面进行了大范围的收集。

↓

NMT (Luong et al., 2015) + **Domain adaptation** (Watanabe et al., 2016)

↓

UNKと北西分野、UNKとUNK湾、北部のUNK、UNK、UNKの汚染物質の情報について文献、組織、研究者から広範囲の収集を行った。

↓

Attention-based unknown word (UNK) replacement (Jean et al. 2015)

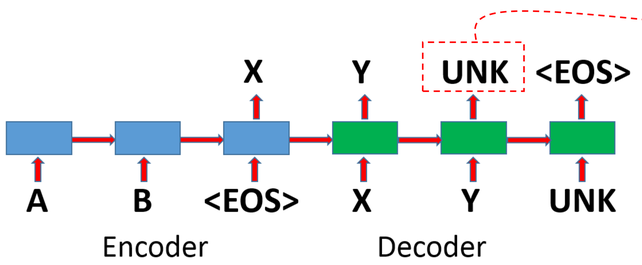
↓

Yukonと北西分野、**Hudson**と**James**湾、北部の**魁北克**、**拉布拉多**、**Greenland**の汚染物質の情報について文献、組織、研究者から広範囲の収集を行った。

Method	Dev. data BLEU	Dev. data RIBES	Test data BLEU	Test data RIBES
(1) ANMT	38.09	83.67	-	-
(2) ANMT w/ UNK replacement	39.05	83.98	39.06	84.23
(3) ANMT w/ domain adaptation	38.28	83.83	-	-
(4) ANMT w/ domain adaptation and UNK replacement	39.24	84.20	39.07	84.21
(5) Ensemble of (1) and (3)	40.66	84.91	-	-
(6) Ensemble of (1) and (3) w/ UNK replacement	41.72	85.25	41.81	85.47
The best system at WAT 2015 (Neubig et al., 2015)	-	-	42.95	84.77
The best system at WAT 2016 (Kyoto-U, NMT)	-	-	46.70	87.29

Selected as one of the 3-best systems in the subtask

Model: Attention-Based Neural Machine Translation (NMT) with Multi Domain Adaptation



Word-based, 512-dimensional, and single layer
LSTM encoder-decoder model with an attention mechanism

Cost function

$$-\log p(y_j|y_1, y_2, \ldots, y_{j-1}, \boldsymbol{x})$$

$$p(y_j|y_1, y_2, \ldots, y_{j-1}, \boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_p \tilde{\boldsymbol{t}}_j + \boldsymbol{b}_p)$$

$$(2\boldsymbol{W}_p^{\mathcal{G}}, 2\boldsymbol{b}_p^{\mathcal{G}}) \qquad (2\overline{\boldsymbol{W}}_p^{\mathcal{D}_1}, 2\overline{\boldsymbol{b}}_p^{\mathcal{D}_1})$$

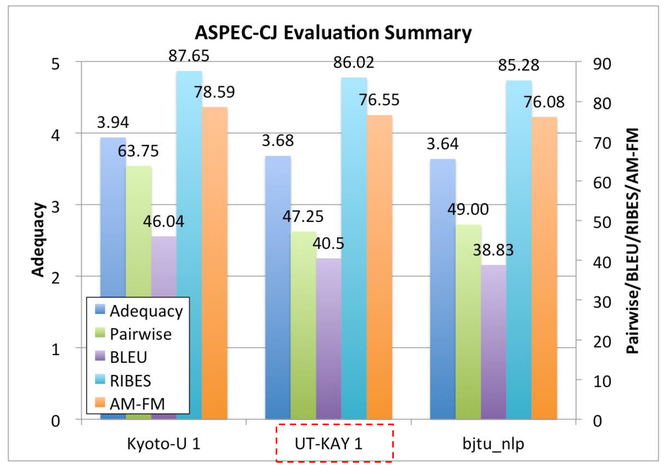
$$-\frac{1}{2}\log p^{\mathcal{G}}(y_j|y_1, y_2, \ldots, y_{j-1}, \boldsymbol{x}) - \frac{1}{2}\log p^{\mathcal{D}_1}(y_j|y_1, y_2, \ldots, y_{j-1}, \boldsymbol{x})$$

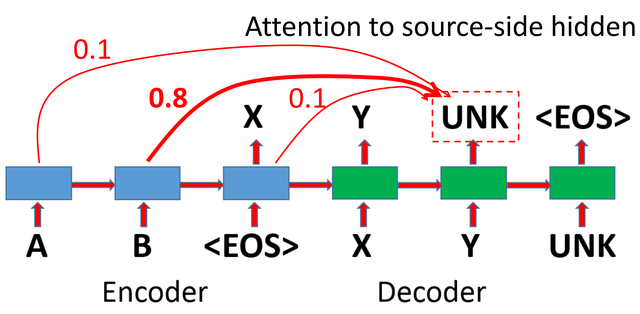
Test time: $\boldsymbol{W}_p^{\mathcal{D}_1} = \boldsymbol{W}_p^{\mathcal{G}} + \overline{\boldsymbol{W}}_p^{\mathcal{D}_1}, \quad \boldsymbol{b}_p^{\mathcal{D}_1} = \boldsymbol{b}_p^{\mathcal{G}} + \overline{\boldsymbol{b}}_p^{\mathcal{D}_1}$

Originally, the domain adaptation method (Watanabe et al., 2016) was proposed for two (source and target) domain settings

↓

Multiple domain adaptation





Postprocessing

Output: X Y UNK <EOS>

0.8

Selecting the source-side word with **the highest attention score**

X Y B <EOS>

Chinese and Japanese share many Chinese characters (Kanji)

		BIO	CHEM	ENE	ENVI	INFO	MATE
BLEU	Method (2)	35.27	37.24	39.74	36.21	41.91	34.92
	Method (4)	34.86	33.96	40.37	37.16	41.58	37.80
	Method (6)	37.84	42.77	43.64	39.29	44.17	38.65
# of samples in the development data		216	19	37	804	982	32

There are no clear trends on the results,
but **ensemble of the two different models** with different objective functions boosted the **BLEU score by 2.7 point**

Analysis: Manual Evaluation on Attention-Based UNK Replacement

More than 70% of the UNK replacement find relevant positions

Type	Count	Ratio
(A) Correct	76	30.4%
(B) Acceptable	5	2.0%
(C) Correct with word translation	104	41.6%
(D) Partially correct	50	20.0%
(E) Incorrect	15	6.0%
Total	250	100.0%

Most of the errors are caused by word segmentation

The six different unknown words are correctly replaced

Input: Chinese
有关Yukon和西北领域、Hudson和James湾、北部魁北克、拉布拉多、Greenland的污染物质的信息从文献、组织、研究者方面进行了大范围的收集。

Output: Japanese
UNKと北西分野、UNKとUNK湾、北部のUNK、UNK、UNKの汚染物質の情報について文献、組織、研究者から広範囲の収集を行った。

(A) (A) (A) (C) (C) (B)
Yukonと北西分野、**Hudson**と**James**湾、北部の**魁北克**、**拉布拉多**、**Greenland**の汚染物質の情報について文献、組織、研究者から広範囲の収集を行った。

"グリーンランド" in the human translation

Input: Chinese
高尾山的环境保护与京王的社会贡献

Output: Japanese
高UNKの環境保全とUNKの社会貢献

(A) (D)
高尾山の環境保全と京の社会貢献

This should be a single word, but the two characters are split by a word segmentation tool

Incorrect segmentation