System Description I

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

University of Tokyo

The **UT-KAY** system

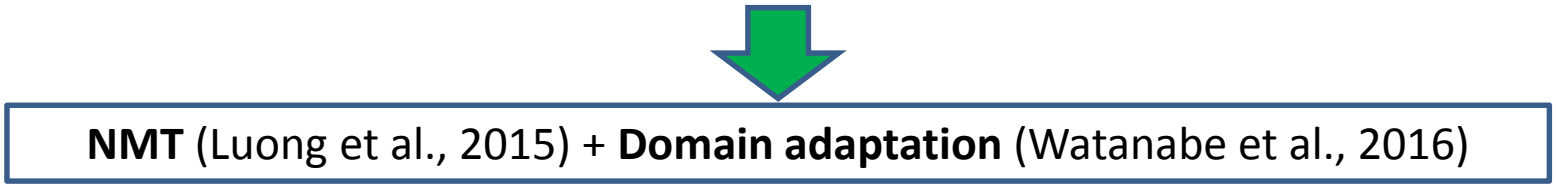
Kazuma Hashimoto
Akiko Eriguchi
Yoshimasa Tsuruoka

12/12/2016 WAT 2016 in Osaka, Japan

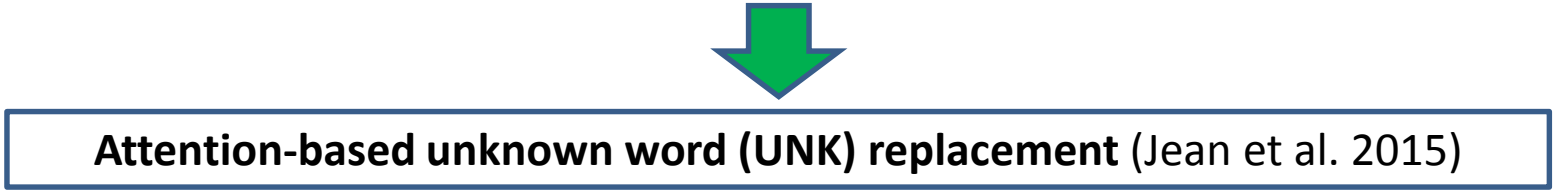
The UT-KAY System

- Chinese-to-Japanese **N**eural **M**achine **T**ranslation (NMT)

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。



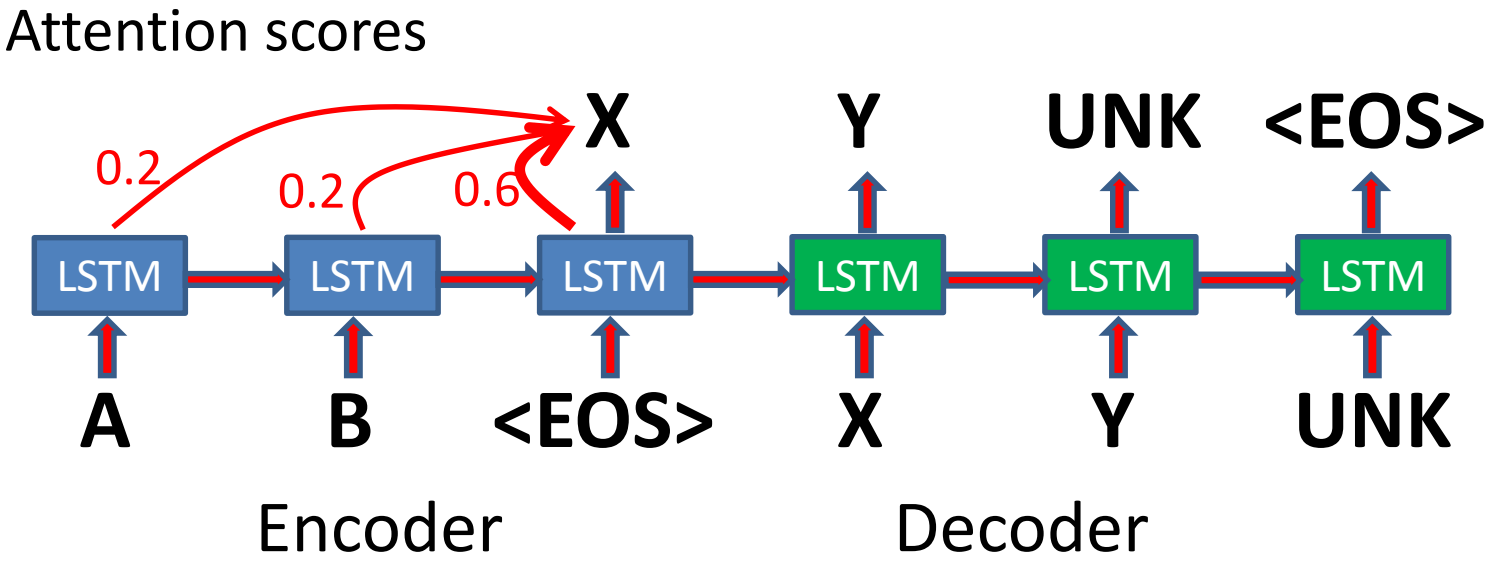
UNKと北西分野、UNKとUNK湾、北部のUNK、UNK、UNKの汚染物質の情報について文献、組織、研究者から広範囲の収集を行った。



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。

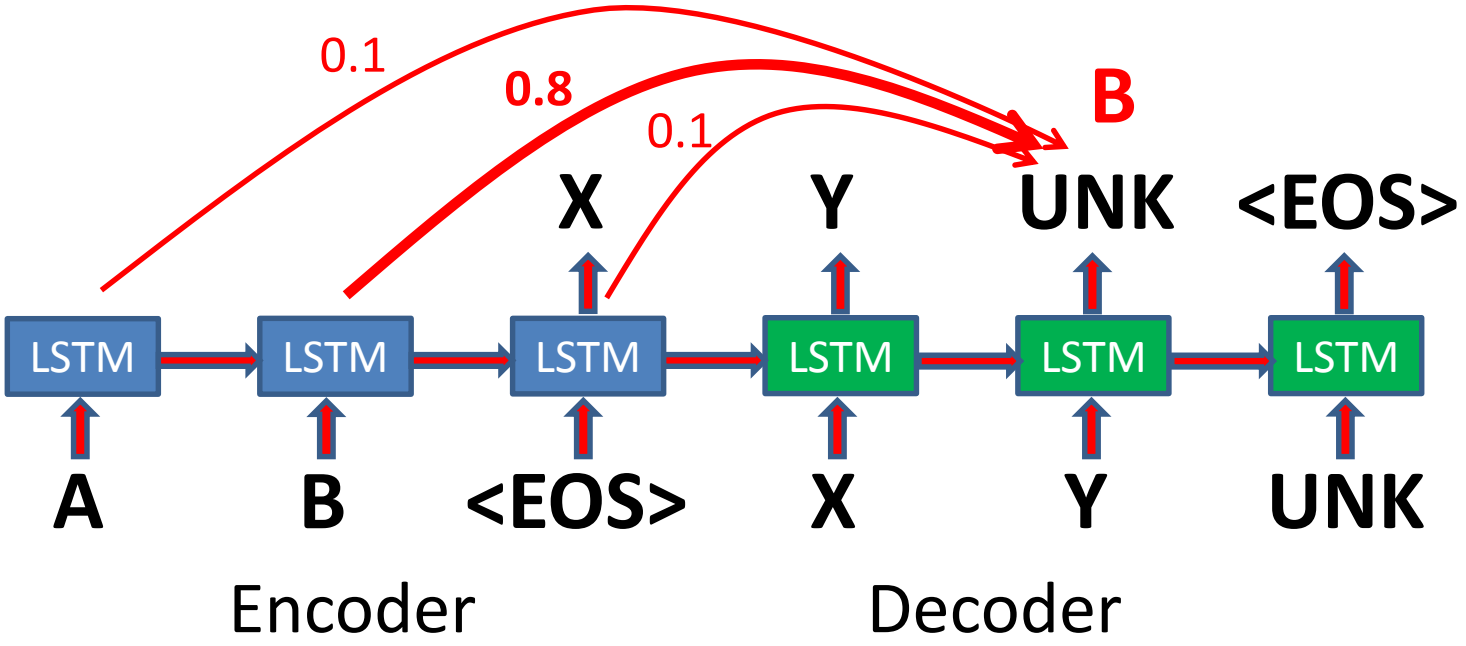
The UT-KAY System

- Attention-based NMT (ANMT) (Luong et al., 2015)
  - Domain adaptation objective (Watanabe et al., 2016)
    - Applied to multiple domain settings
  - Attention-based UNK replacement (Jean et al., 2015)



The UT-KAY System

- Attention-based NMT (ANMT) (Luong et al., 2015)
  - Domain adaptation objective (Watanabe et al., 2016)
    - Applied to multiple domain settings
  - Attention-based UNK replacement (Jean et al., 2015)



Summary of Our Results

- Attention-based UNK replacement improves the results
- Domain adaptation does not improve the results

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 top 3 systems in the subtask

How Accurate?

- Manual check for the replacement results of 250 cases in 132 sentences

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

Example 1

- Six different unknown words are replaced correctly

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。

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%

“グリーンランド” in the human translation

Example 2

- Word segmentation should be improved

Input: Chinese

高尾山的环境保护与京王的社会贡献

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

Output: Japanese

高UNKの環境保全とUNKの社会貢献

Incorrect segmentation

(A) (D)

高尾山の環境保全と京の社会貢献

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%

Summary

- Attention-based unknown word replacement is effective in Chinese-to-Japanese neural machine translation
  - There is still room for improvement by using high quality word-level dictionaries

- For more details, please come to see the poster!