Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Decoder Models

Shun Kiyono[1] Sho Takase[2] Jun Suzuki[1,2,4] Naoaki Okazaki[3] Kentaro Inui[1,4] Masaaki Nagata[2]

[1]Tohoku University
[2]NTT Communication Science Laboratories
[3]Tokyo Institute of Technology
[4]RIKEN AIP

Summary

- Attention matrix is **not an optimal choice** for interpreting the output of Encoder-Decoder-based models (e.g. low alignment acc.)
- Proposed method, Unsupervised Alignment Module (UAM), models **token-wise alignment** between the source and target
- UAM provides **better interpretability of Encoder-Decoder-based models** through alignments

Task: Headline Generation

Input $\boldsymbol{X}$: 1st Sentence of an Article → EncDec → Output $\boldsymbol{Y}$: Headline

Taiwan on Tuesday bowed to calls to make the island more international by easing its immigration rules .

Taiwan eases immigration rules

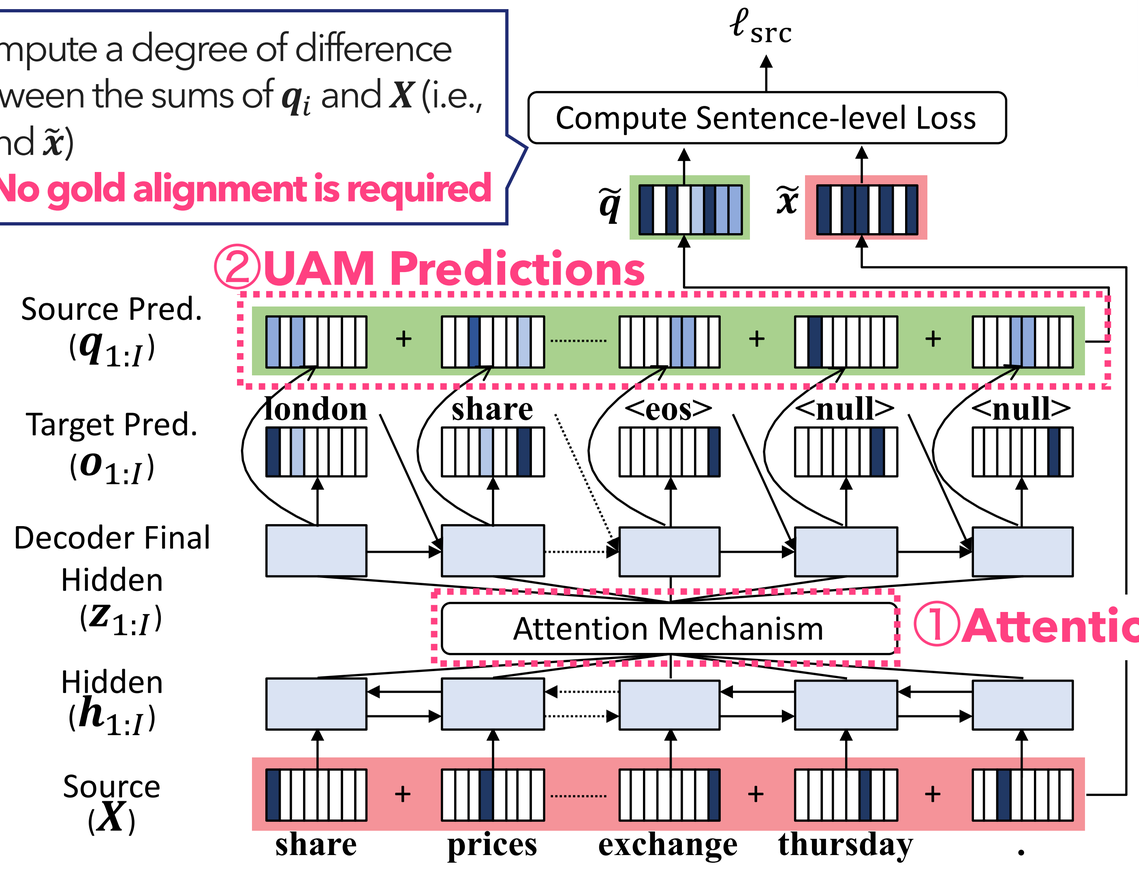
4M sentence pairs are available from Gigaword corpus

Proposed Method: UAM

Key Idea: Predict Source-side Tokens

$$G_1(\theta,\gamma) = \frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{D}} \Big(\underbrace{\ell_{\mathrm{trg}}(\boldsymbol{Y}',\boldsymbol{X},\theta)}_{\textbf{EncDec Loss}} + \underbrace{\ell_{\mathrm{src}}(\tilde{\boldsymbol{x}},\boldsymbol{X},\boldsymbol{Y}',\gamma,\theta)}_{\textbf{UAM Loss}}\Big)$$

Compute a degree of difference between the sums of $\boldsymbol{q}_i$ and $\boldsymbol{X}$ (i.e., $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{x}}$)
→ **No gold alignment is required**

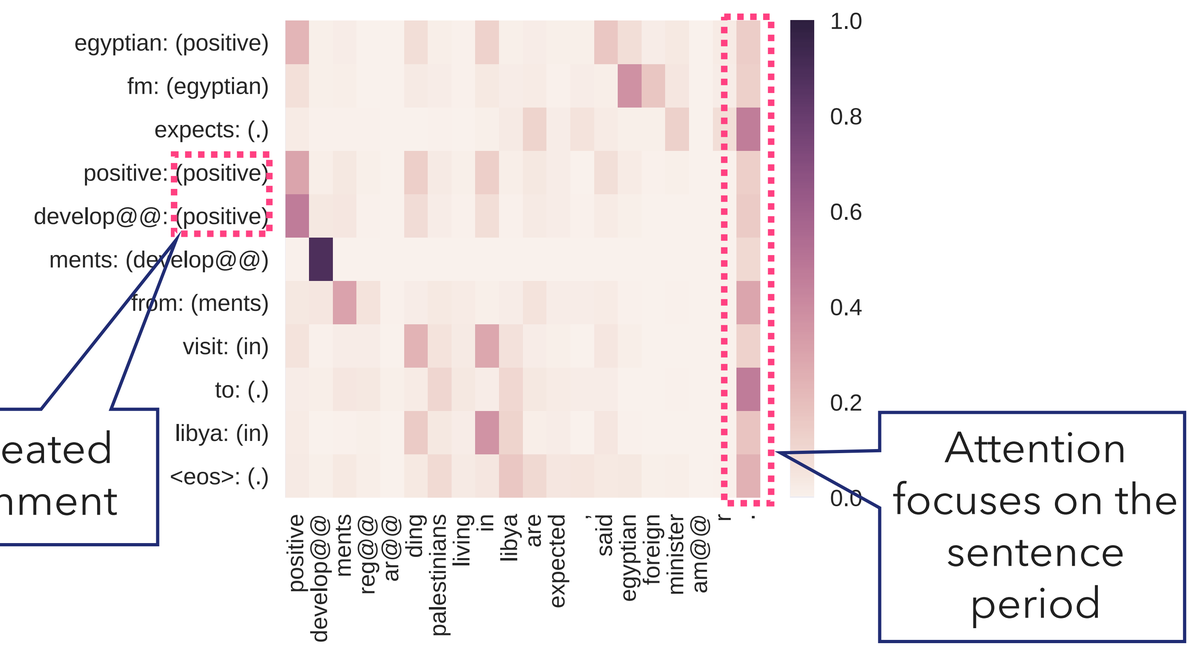


Analysis: Attention vs. Uam

Qualitative Analysis

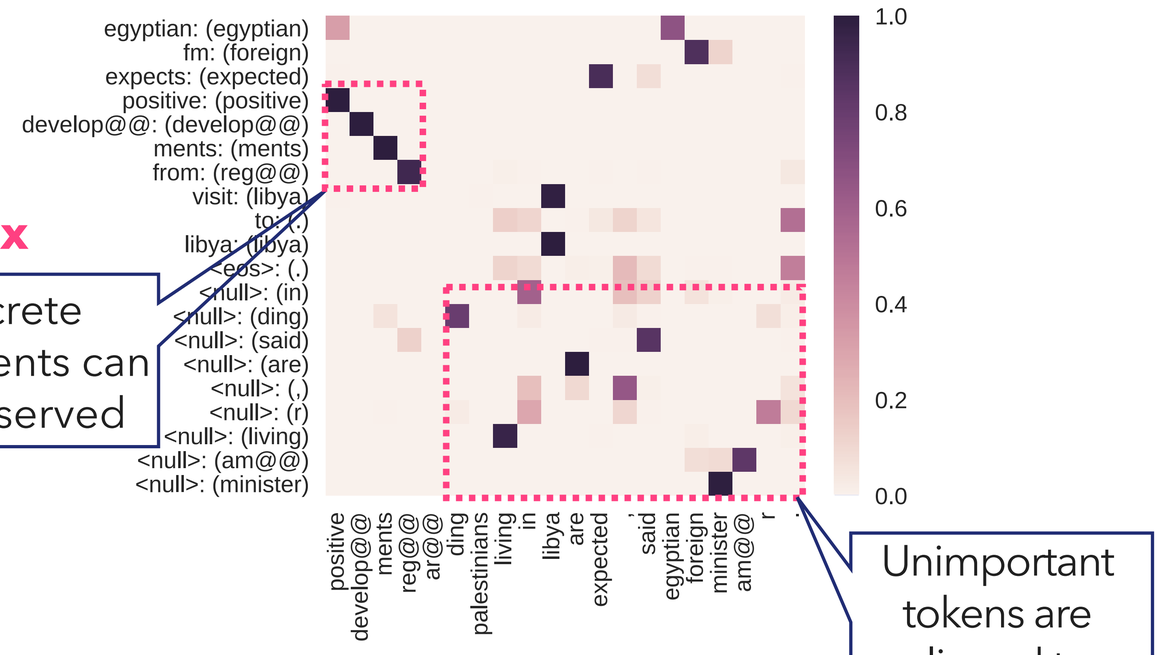
① Attention Matrix

😩 Repeated alignments to same source-side tokens
😩 Attention values are distributed to several tokens



② UAM Predictions

😀 Repeated alignments are rarely observed
😀 Alignments are more discrete than attention



Quantitative Analysis

Table: Alignment Accuracy

Model	Test (Ours)	Test (Zhou)
Baseline (EncDec)	8.60	5.97
Baseline+UAM	**52.52**	**50.91**

UAM alignments are significantly more accurate than that of the attention matrix
→ **better interpretation of the model**

Results (ROUGE Scores)

Table: ROUGE F1 Evaluation Results

Model	Test (Ours)			Test (Zhou)		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
Baseline (EncDec)	46.80	24.48	43.74	46.79	24.75	43.62
Baseline +UAM	**46.91**	**24.86**	**43.87**	**46.89**	**24.93**	**43.68**

UAM improves ROUGE scores
→**<null> alignment handles selection of unimportant information?**