Content Word Aware Neural Machine Translation

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita

National Institute of Information and Communications Technology (NICT), Kyoto, Japan

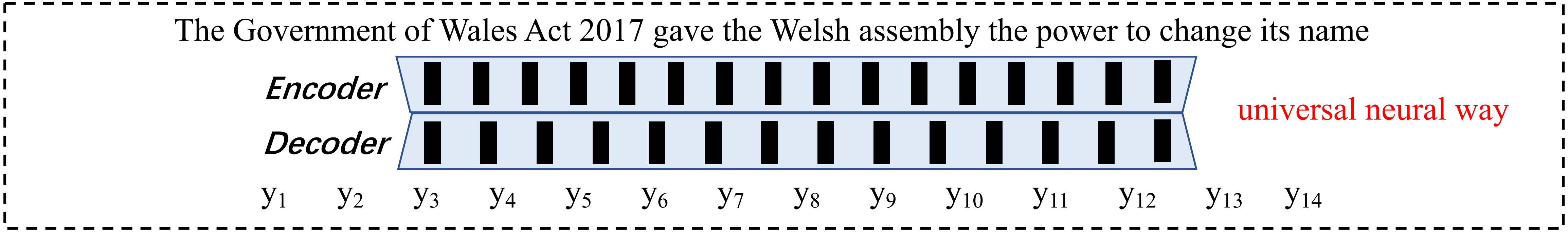
Outline

- Motivation
- Preliminary Experiment
- Content Word Recognition
- Proposed NMT models
- Experiments
- Conclusion

Motivation

Existing Encoder-Decoder NMT model

- Encoding and generating all source/target words in a universal neural way

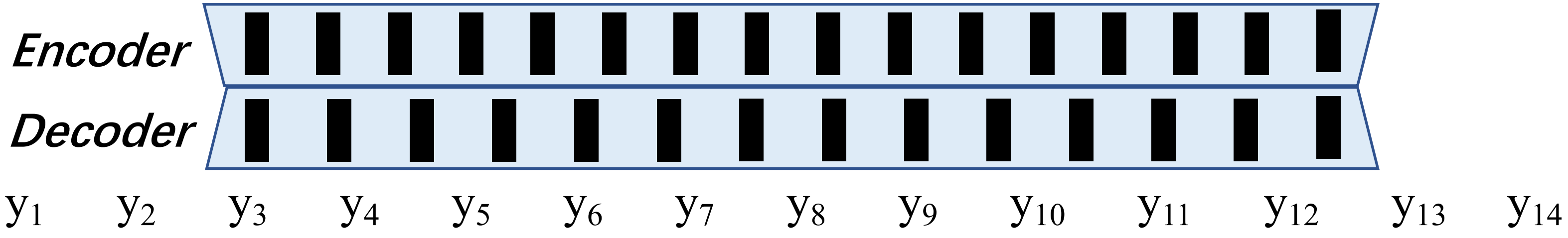


Motivation

Existing Encoder-Decoder NMT model

- Encoding and generating all source/target words in a universal neural way

The Government of Wales Act 2017 gave the Welsh assembly the power to change its name



- Not considering the importance of word in the sentence meaning

Function words: The Government of Wales Act 2017 gave the Welsh assembly the power to change its name

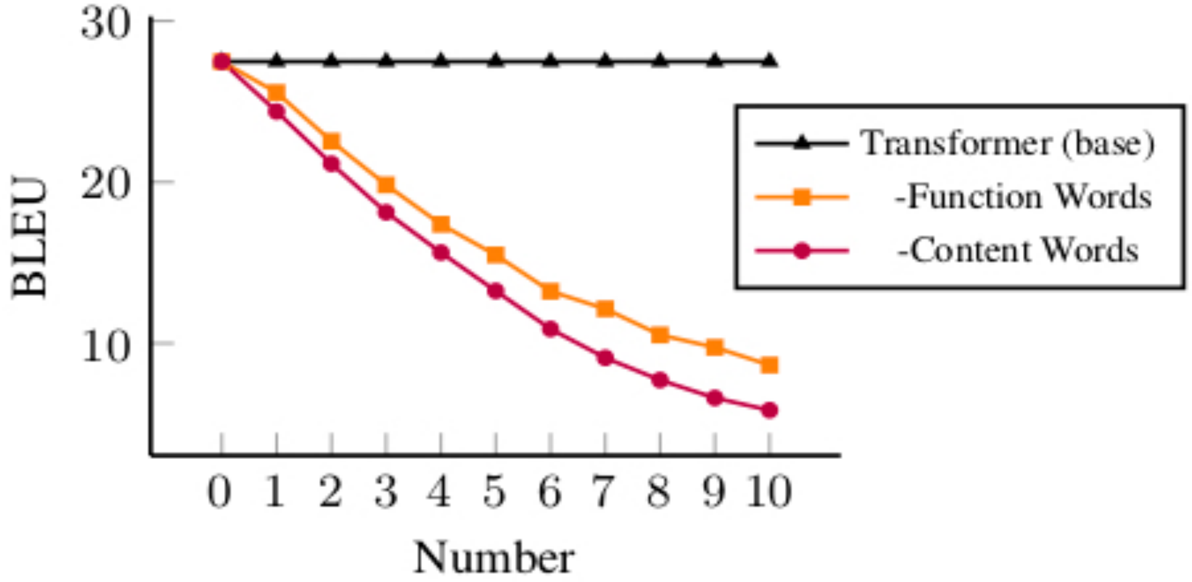
Content words: The Government of Wales Act 2017 gave the Welsh assembly the power to change its name

Content words express more important meaning than *function words* in the sentence meaning.

Preliminary Experiment

Preliminary

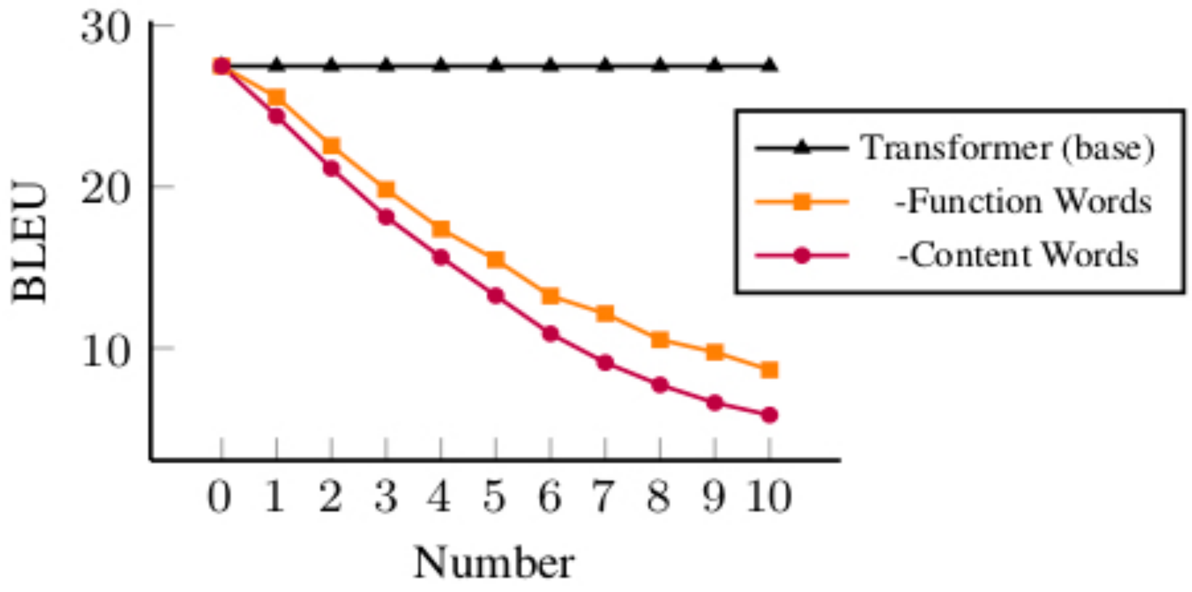
- ✓ Given a trained Transformer-based NMT model for the WMT14 English-German translation task
- ✓ Randomly masked content ("-Content Words") or function words ("-Function Words") with UNK in a source sentence
- ✓ The trained NMT model decodes these masked test set



Preliminary Experiment

Preliminary

- ✓ Given a trained Transformer-based NMT model for the WMT14 English-German translation task
- ✓ Randomly masked content (“-Content Words”) or function words (“-Function Words”) with UNK in a source sentence
- ✓ The trained NMT model decodes these masked test set



Findings

- ➢ Content words have a greater effect on modeling translation between a language pair
- ➢ NMT should pay more attention to content words in a sentence

Content Word Recognition

The Government of Wales Act 2017 gave the Welsh assembly the power to change its name

term frequency-inverse document frequency (TF-IDF)

$$TI_j = \frac{k_{j,m}}{J_m} \times \log \frac{|M|}{1 + |m : d_j \in D_m|}$$

$k_{j,m}$: *the number of occurrences of the j-th word in the input sentence* d_t*;*
$|M|$: *the total number of sentences in the monolingual data;*
$|m{:}d_j \in D_m|$: *the number of sentences including word* d_j *in the monolingual data;*
J_m: *an input sentence of length;*
D_m: *a input sentence.*

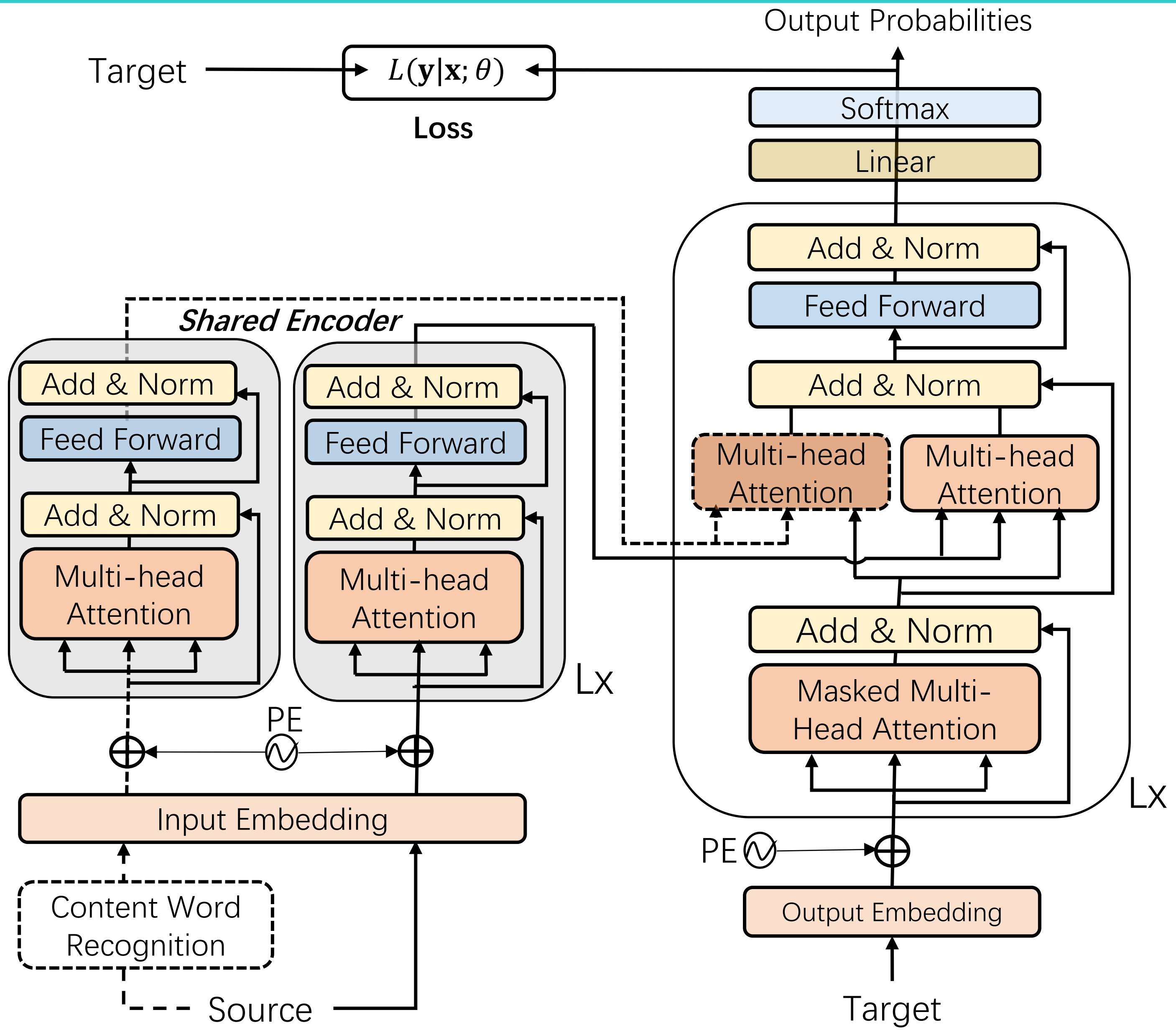
Government Wales Act 2017 gave Welsh assembly power change its name

Selecting a sequence of fixed percent content words

Government Wales gave Welsh power change name

Typically, TF-IDF is often used to recognize topics in a document.
Intuitively, the recognized topics can often summary the main content of this document.
Therefore, we regarded a sentence as a document.

Proposed NMT models

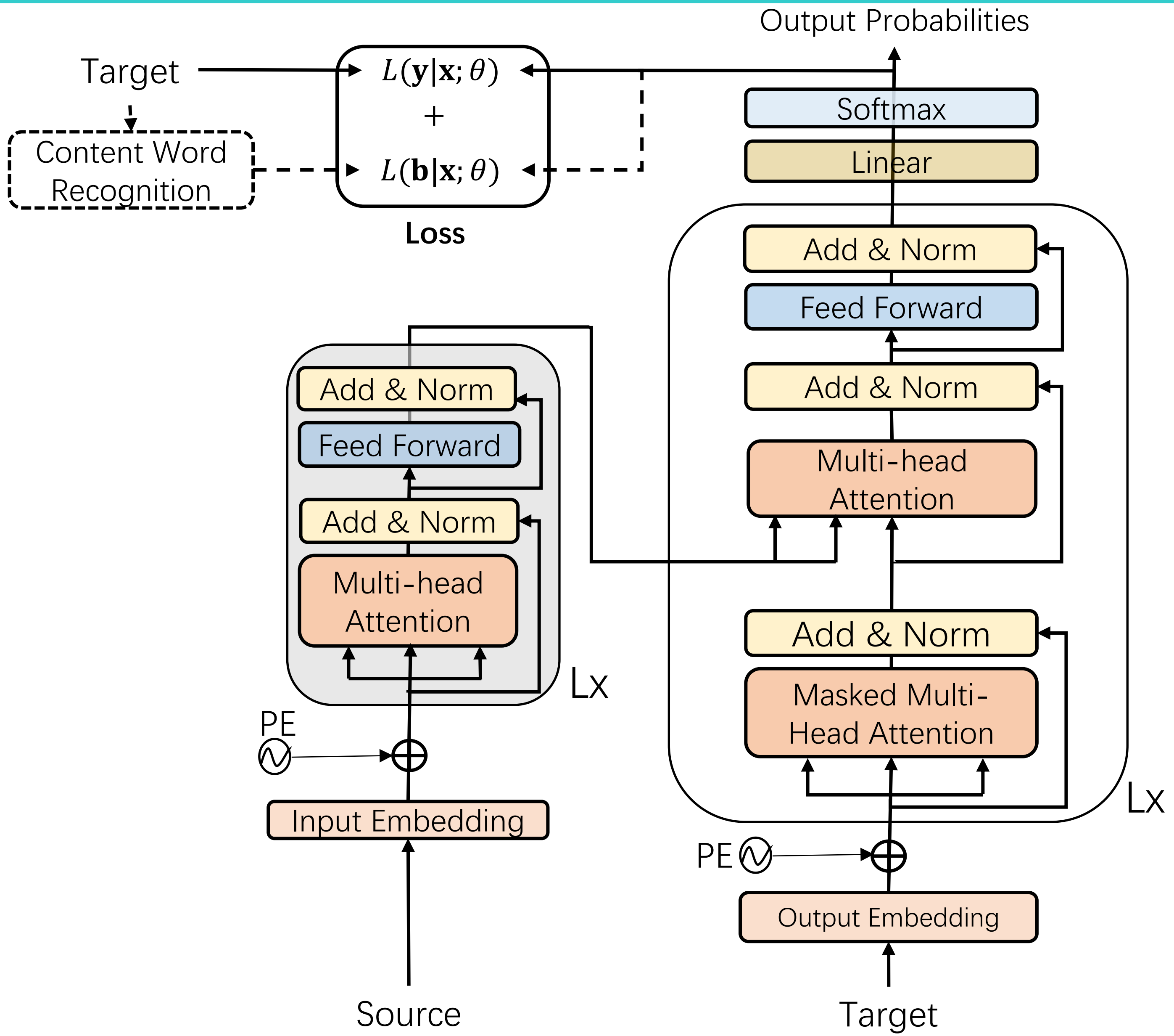


SCWAContext: Based on the sequence of the recognized source content words, we use a shared encoder to learn its representation, and thereby obtain an additional context vector to improve the prediction of target word:

$$\begin{aligned}
\overline{\boldsymbol{\mathcal{S}}}_i^l &= \mathrm{LN}(\mathrm{ATT}_d^l(\mathbf{Q}_i^{l-1}, \mathbf{K}_i^{l-1}, \mathbf{V}_i^{l-1}) + \boldsymbol{\mathcal{S}}_i^{l-1}), \\
\mathbf{C}_i^l &= \mathrm{LN}(\mathrm{ATT}_c^l(\overline{\boldsymbol{\mathcal{S}}}_i^l, \mathbf{K}_e^L, \mathbf{V}_e^L) + \overline{\boldsymbol{\mathcal{S}}}_i^l), \\
\boldsymbol{\mathcal{C}}_i^l &= \mathrm{LN}(\mathrm{ATT}_y^l(\overline{\boldsymbol{\mathcal{S}}}_i^l, \boldsymbol{\mathcal{K}}_e^L, \boldsymbol{\mathcal{V}}_e^L) + \overline{\boldsymbol{\mathcal{S}}}_i^l), \\
\boldsymbol{\mathcal{S}}_i^l &= \mathrm{LN}(\mathrm{FFN}_d^l(\mathbf{C}_i^l + \boldsymbol{\mathcal{C}}_i^l) + \mathbf{C}_i^l),
\end{aligned}$$

$$P(y_i|y_{<i}, \mathbf{x}) \propto \exp(\boldsymbol{W}_o \tanh(\boldsymbol{W}_w \boldsymbol{\mathcal{S}}_i^L))$$

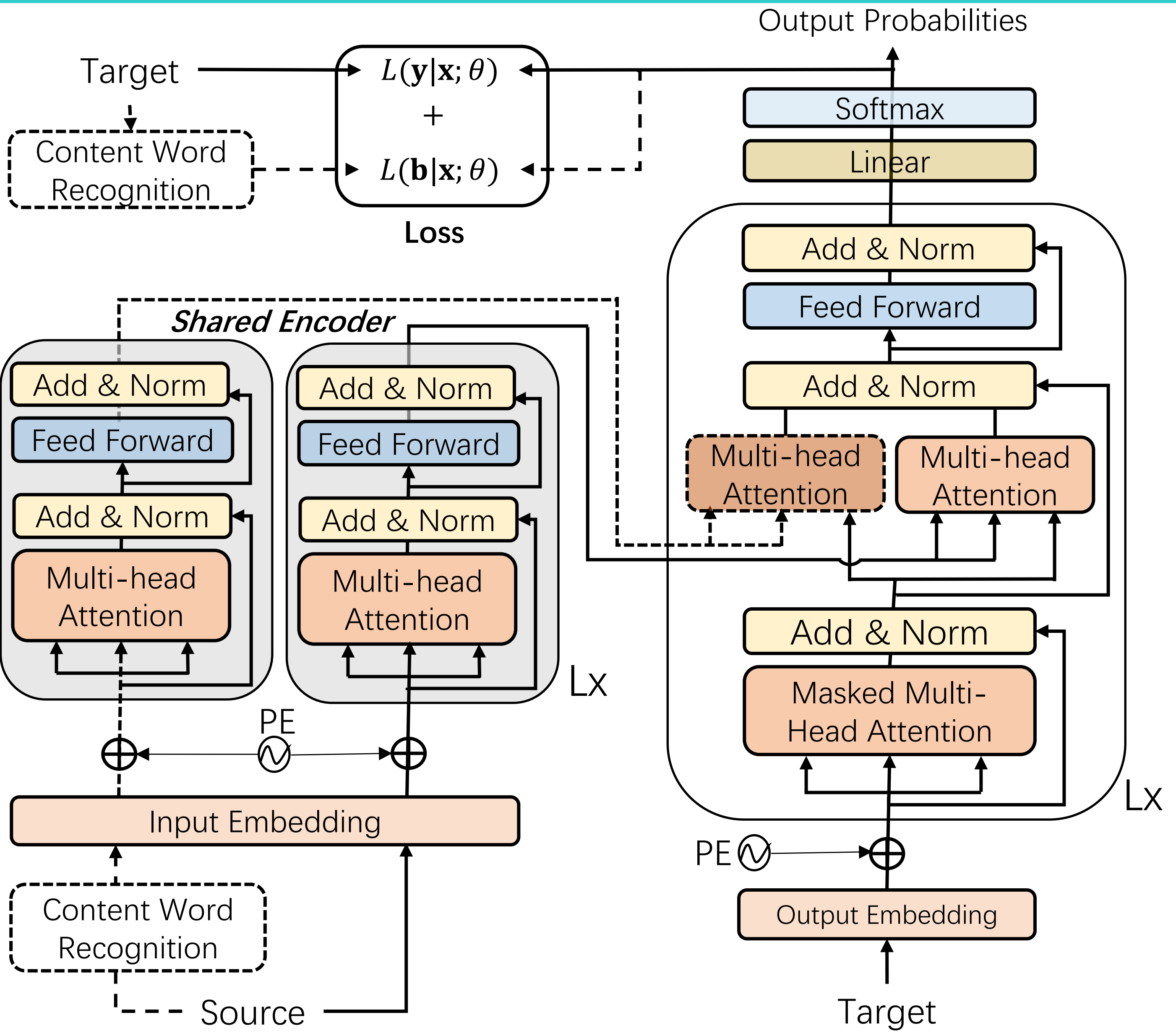
Proposed NMT models



TCWALoss: Based on the sequence of the recognized target content words, we utilize it to compute an additional loss to guide the training of the translation model:

$$\mathcal{J}(\theta) = \arg\max_{\theta}\{P(\mathbf{y}|\mathbf{x};\theta) + \lambda * P(\mathbf{b}|\mathbf{x};\theta)\}$$

Proposed NMT models



BCWAContLoss: It captures the content words of both the source and the target sentences to further improve translation performance.

Experiments

❖ Main Results

Systems	EN-DE			ZH-EN		EN-FR	
	BLEU	#Speed	#Param	BLEU	#Param	BLEU	#Param
Existing NMT systems							
Trans.base (Vaswani et al., 2017)	27.3	N/A	65.0M	N/A	N/A	38.1	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.26	N/A	106.9M	24.67	126.8M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.18	N/A	88.0M	24.80	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.21	N/A	97.4M	24.70	107.3M	N/A	N/A
Trans.big (Vaswani et al., 2017)	28.4	N/A	213.0M	N/A	N/A	41.0	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.89	N/A	339.6M	24.56	379.4M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.74	N/A	339.6M	25.01	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.98	N/A	333.5M	25.10	373.3M	N/A	N/A
Our NMT systems							
Trans.base	27.48	13.2K	66.5M	24.28	74.7M	38.32	66.9M
+SCWAContext	28.28+	12.1K	72.8M	24.79+	81.0M	39.41+	73.2M
+TCWALoss	27.94+	13.3K	66.5M	24.65	74.7M	38.89+	66.9M
+BCWAContLoss	28.51+	12.1K	72.8M	24.94+	81.0M	39.56+	73.2M
Trans.big	28.45	11.2K	221.1M	24.55	237.5M	41.21	222.9M
+BCWAContLoss	29.14+	10.1K	246.3M	25.12+	262.7M	42.57+	247.1M

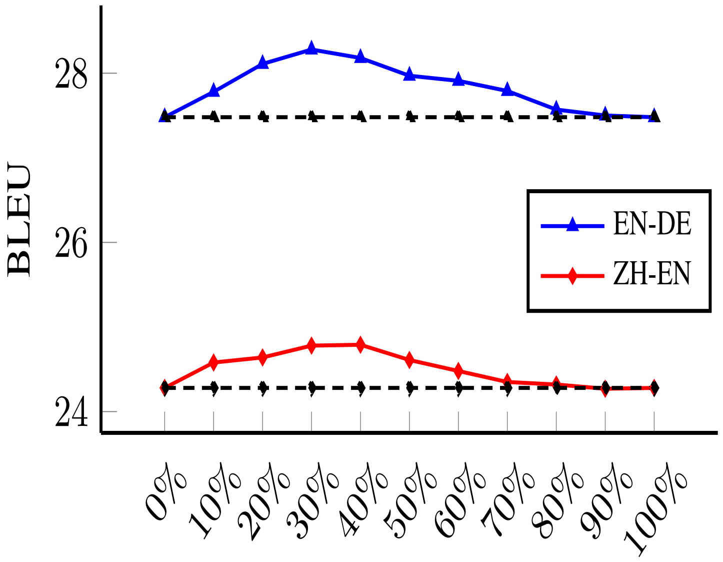
Experiments

❖ Main Results

Systems	EN-DE			ZH-EN		EN-FR	
	BLEU	#Speed	#Param	BLEU	#Param	BLEU	#Param
Existing NMT systems							
Trans.base (Vaswani et al., 2017)	27.3	N/A	65.0M	N/A	N/A	38.1	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.26	N/A	106.9M	24.67	126.8M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.18	N/A	88.0M	24.80	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.21	N/A	97.4M	24.70	107.3M	N/A	N/A
Trans.big (Vaswani et al., 2017)	28.4	N/A	213.0M	N/A	N/A	41.0	N/A
+Context-Aware SANs (Yang et al., 2019a)	28.89	N/A	339.6M	24.56	379.4M	N/A	N/A
+Convolutional SANs (Yang et al., 2019b)	28.74	N/A	339.6M	25.01	N/A	N/A	N/A
+BIARN (Hao et al., 2019)	28.98	N/A	333.5M	25.10	373.3M	N/A	N/A
Our NMT systems							
Trans.base	27.48	13.2K	66.5M	24.28	74.7M	38.32	66.9M
+SCWAContext	28.28+	12.1K	72.8M	24.79+	81.0M	39.41+	73.2M
+TCWALoss	27.94+	13.3K	66.5M	24.65	74.7M	38.89+	66.9M
+BCWAContLoss	28.51+	12.1K	72.8M	24.94+	81.0M	39.56+	73.2M
Trans.big	28.45	11.2K	221.1M	24.55	237.5M	41.21	222.9M
+BCWAContLoss	29.14+	10.1K	246.3M	25.12+	262.7M	42.57+	247.1M

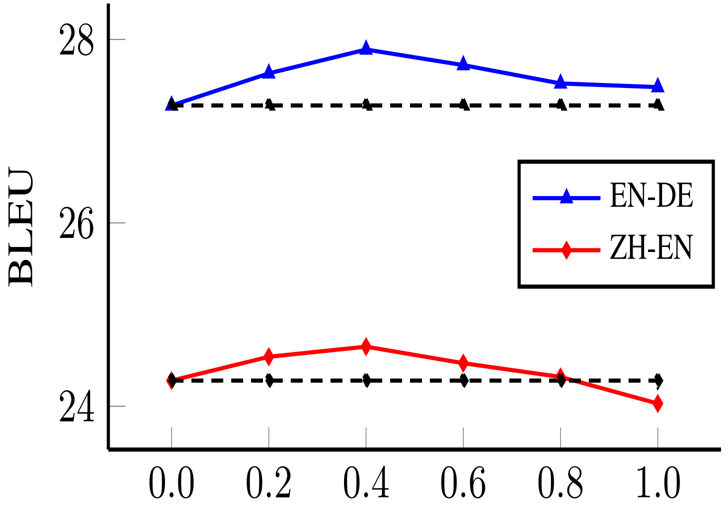
Experiments

❖ Evaluating Content Word Recognition



The percent of N for +SCWACont model

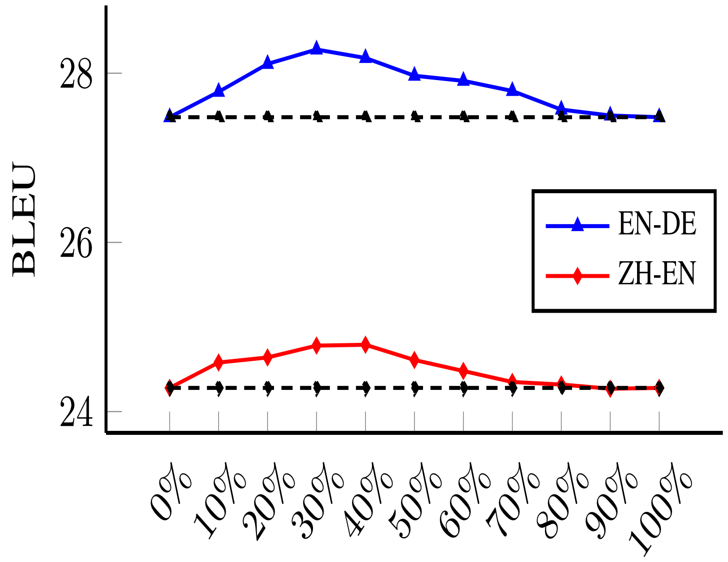
❖ Effect of Content Word-Aware Loss



λ for +TCWALosst model

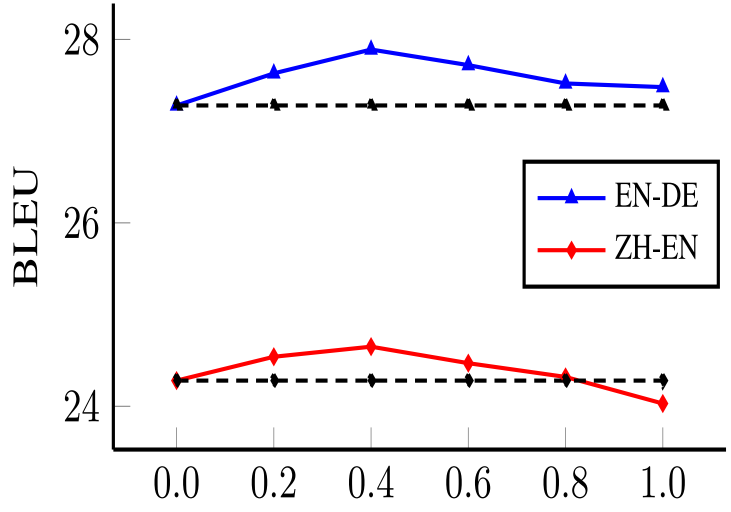
Experiments

❖ Evaluating Content Word Recognition



The percent of N for +SCWACont model

❖ Effect of Content Word-Aware Loss



λ for +TCWALosst model

❖ **Evaluating Generation of Content Words**

System	EN-DE	ZH-EN
Trans.base	51.0%	53.8%
+SCWAContext	51.9%	54.6%
+TCWALoss	51.5%	54.2%
+BCWAContLoss	52.1%	54.7%

Table 2: Accuracy of unigram content words on the EN-DE and ZH-EN test sets

Conclusion

- Explored the importance of word in a sentence for NMT
- Recognized content words through statistical word frequency information
- Simple and efficient, not much time and space cost, and introduced to the training and inference