A Machine Translation

A.1 Detailed Experimental Settings

For Zh-En, we segmented Chinese data using ICTCLAS[1]. We limited the maximum sentence length to 50 tokens. For En-De and En-Ro, we did not filter out the sentence length for En-De and En-Ro. We applied byte pair encoding (Sennrich et al., 2016, BPE) to segment all sentences with merge operations of 32K. All out-of-vocabulary words were mapped to a distinct token `<UNK>`.

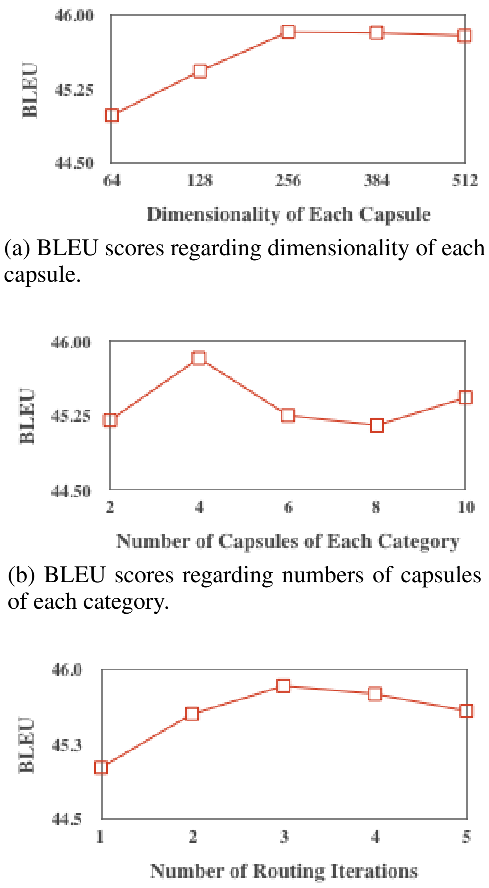
We used the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. We used the same learning rate schedule strategy as (Vaswani et al., 2017) with 4,000 warmup steps. The training batch consisted of approximately 25,000 source tokens and 25,000 source and target tokens. Label smoothing of the value of 0.1 (Szegedy et al., 2016) was used for training. We trained our models for 100k steps on single GTX 1080ti GPU.

For evaluation, we used beam search with a width of 4 with length penalty of 0.6 (Wu et al., 2016). We did not apply checkpoint averaging (Vaswani et al., 2017) on the parameters for evaluation. The translation evaluation metric is case-insensitive BLEU (Papineni et al., 2002) for Zh-En[2], and case-sensitive BLEU for En-De and En-Ro[3], which are consistent with previous work.

A.2 Effect of Hyperparameters

To examine the effects of different hyperparameters of the proposed model, we list the results of different settings of the dimension and the number of the capsules, and the number of routing iteration in Table 1.

We observe that increasing the dimension or the number of capsules does not bring better performance. We attribute these results to the sufficient expressive capacity of medium scales of the capsules. Likewise, the BLEU score goes up with the increase of the number of iterations, while it turns to decrease after the performance climb to the peak at the best setting of 3 iterations. The number of routing iteration affects the estimation of the agreement between two capsules. Hence, redundant iterations may lead to over-estimation of the agreement, which has also been revealed in Dou et al. (2019).



(a) BLEU scores regarding dimensionality of each capsule.

(b) BLEU scores regarding numbers of capsules of each category.

(c) BLEU scores regarding routing iterations of GDR

Figure 1: BLEU scores in terms of different hyperparameters.

A.3 Visualization

We show two more examples of visualization of the assignment probabilities of the guided dynamic routing mechanism in Figure 2.

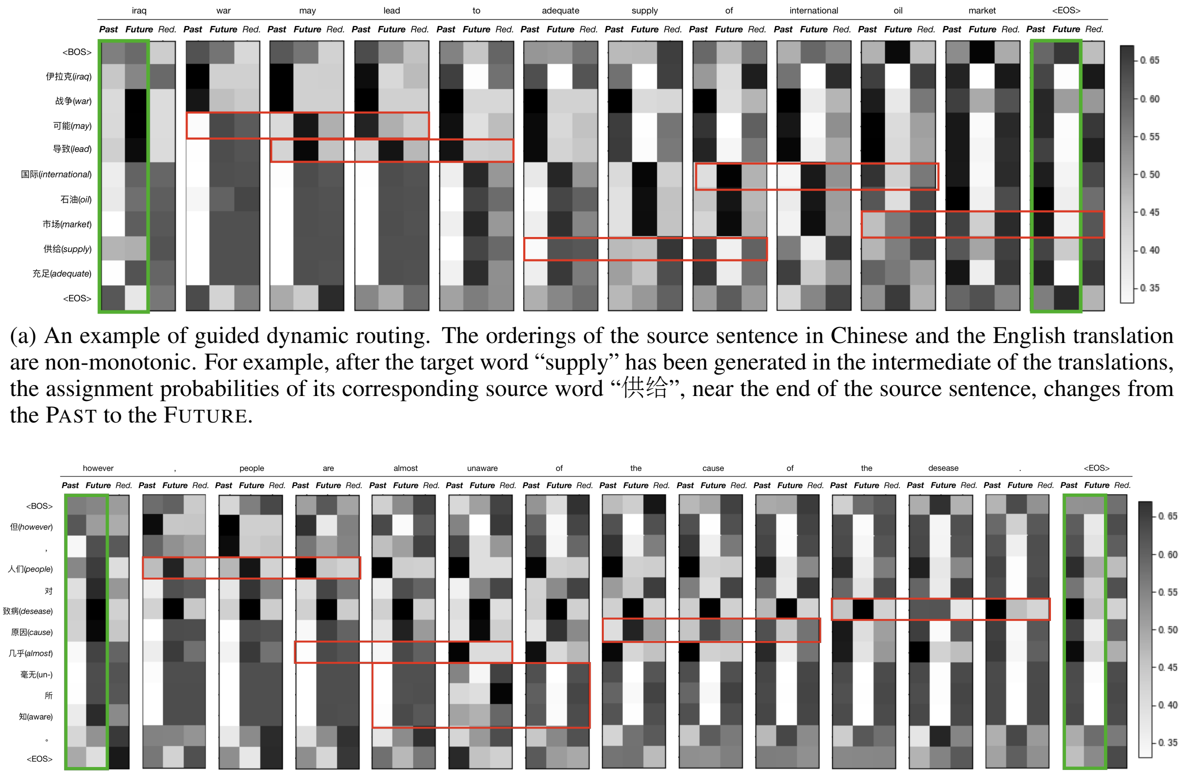
B Abstract Summarization

Abstract summarization is another prevalent sequence-to-sequence task, which aims to find a short, fluent and effective summary from a long text article (Chang et al., 2018). Intuitively, discarding redundant contents in the source article is important for abstract summarization, which could be achieved by our proposed redundant capsules. Moreover, figuring out which parts of the source article have been summarized completely and which have not yet should be also explicitly modeled for abstract summarization.

[1] `http://ictclas.nlpir.org/`

[2] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl`

[3] `https://github.com/awslabs/sockeye/tree/master/contrib/sacrebleu`



(a) An example of guided dynamic routing. The orderings of the source sentence in Chinese and the English translation are non-monotonic. For example, after the target word "supply" has been generated in the intermediate of the translations, the assignment probabilities of its corresponding source word "供给", near the end of the source sentence, changes from the PAST to the FUTURE.

(b) Another example. An interesting case is related to "unware", which has a source counterpart of three Chinese words ("毫无", "所", and "知"). When it has been generated, the assignment probabilities of the words in its counterpart phrase change from PAST to the FUTUREsimultaneously.

Figure 2: Visualization of the assignment probabilities of the iterative guided dynamic routing. For each translation word, the three columns represent the probability assigning to PAST, FUTURE or redundant capsule, respectively. Green frames indicate that the assignment probabilities of source words change from PAST at the beginning to FUTURE in the end. Red frames highlight the changes before and after a specific word's generation.

B.1 Experimental Settings

Dataset We conduct experiments on the LCSTS dataset (Hu et al., 2015) to evaluate our proposed model for abstract summarization. This dataset contains a large number of short Chinese news articles with their headlines as the short summaries collected from Sina Weibo, a Twitter-like microblogging website in China. As shown in Table 1, this dataset is composed of three parts. Part I contains a large number of 2,400,591 pairs of (article, summary). Part II and III contain not only text data but also manually rated scores from 1 to 5 for the quality of summaries in terms of their relevance to the source articles. We follow Hu et al. (2015) to use Part I as the training set, and the subset scored from 3 to 5 of Part II and Part III as the development set and testing set.

Training and evaluation Following Hu et al. (2015), we conducted experiments based on character. We set the vocabulary size to 4,000 for both source and target sides. We used the Transformer as our baseline system. Model configurations and other training hyperparameters are the same as machine translation tasks. For evaluation, we used beam search with a width of 10 without length penalty. We report three variants of recall-based ROUGE (Lin, 2004), namely, ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest-common substring).

	Part I	Part II	Part III
#pairs	2,400,591	10,666	1,106
#pairs (score $\geq$ 3)	-	8,685	725

Table 1: Statistics of the LCSTS dataset.

B.2 Results

Table 2 shows the results of existing systems and our proposed model. We observe that our proposed model outperforms the Transformer baseline system by a significant margin. As we expected, the redundant capsules are vital for abstract summarization task, in which summary is

Model	R-1	R-2	R-L
RNN+context (Hu et al., 2015)	29.9	17.4	27.2
CopyNet (Gu et al., 2016)	34.4	21.6	31.3
Distraction (Chen et al., 2016)	35.2	22.6	32.5
DGRD (Li et al., 2017)	36.99	24.15	34.21
MRT (Ayana et al., 2016)	37.87	25.43	35.33
WEAN (Ma et al., 2018)	37.80	25.60	35.20
AC-ABS (Li et al., 2018)	37.51	24.68	35.02
Transformer (Chang et al., 2018)	40.49	26.83	37.32
+HWC (Chang et al., 2018)	44.38	32.26	41.35
Transformer	40.18	25.76	35.69
OURS	**43.85**	**29.57**	**39.10**
- *redundant capsules*	42.43	28.17	37.66

Table 2: ROUGE scores on LCSTS abstract summarization task.

required to be short and concise by discarding tons of less important contents from the original article. Our model also beats most of the previous approaches except the model of Chang et al. (2018), which benefit from a hybrid word-character vocabularies of max size (more than 900k entries) and data cleaning. We expect our model to benefit from these improvements as well. In addition, our model does not rely on any extraction-based methods, which are designed for abstract summarization task to extract relevant parts from the source article directly (e.g., CopyNet (Gu et al., 2016)). Our method may achieve further gains by incorporating these task-specific mechanisms.

References

Ayana, Shiqi Shen, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with minimum risk training. *CoRR*, abs/1604.01904.

Chieh-Teng Chang, Chi-Chia Huang, and Jane Yung jen Hsu. 2018. A hybrid word-character model for abstractive summarization. *CoRR*, abs/1802.09968.

Qian Chen, Xiao-Dan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for modeling document. In *IJCAI*.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Longyue Wang, Shuming Shi, and Tong Zhang. 2019. Dynamic layer aggregation for neural machine translation with routing-by-agreement. *arXiv preprint arXiv:1902.05770*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL 2016*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR 2014*.

Piji Li, Lidong Bing, and Wai Lam. 2018. Actor-critic based training framework for abstractive summarization. *CoRR*, abs/1803.11070.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *EMNLP*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *NAACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Computer Science*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.