A Supplementary

Token-independent attention In the vanilla self-attention model, the attention weight is based on the pairwise token interactions. We try a latest trick, in which the attention weights do not depend on any input tokens. Instead, the attention weights are initialized to random values. These values can then either be trainable or kept fixed (Tay et al., 2020). We combine this token-independent weight with the vanilla weight matrix in Transformer and use the following attention calculation equation instead:

$$\text{Attention}(Q, K, V) = (\text{softmax}(\frac{QK^T}{\sqrt{d_k}}) + G)V. \tag{1}$$

where G is a random initialized global weight matrix and trained with the model. We implement this idea in the graph attention network, and the ROUGE F score is 41.90/19.27/38.94.

Entity filter module In order to avoid redundant information in our system, we try to introduce the entity filter module before Context2Entity. We tried two method. In the first method, we simply apply the dense connection to the entity vector to get the salient score for each entity. In the second method, we apply attention mechanism between entities and decode state $\mathbf{s}_t$, this derive $\text{score}_{i,t} = \mathbf{g}_i^T \mathbf{s}_t$. Therefore we get the salient score $\text{salience}_i = \sum_t \mathbf{s}_t$ for each entity. Then select the top-k entities to go on with Context2Entity module. However, the performance is below expectations, so we apply the attention mechanism to let the model decide which entity should be paid more attention.

Usage of graph information There is another way to utilize the graph information in the prediction process. We use the hidden state as query, the graph node representation as key and value. This derive the final graph-enhanced representation $\mathbf{s}^g$:

$$\mathbf{s}^g = \text{Transformer}(\mathbf{s}, \mathbf{g}, \mathbf{g}). \tag{2}$$

But this method performs slightly worse than our Context2Entity attention.

Entity-to-context attention Entity-to-context attention aims to detect which decoder state has the closest similarity with each entity. It should be noticed that current state is unaware of subsequent states, thus we modify the initial BIDAF by masking the followed state for each state. We obtain the attention weights by $b_t = softmax(max_{col}(\mathbf{M}_{:t,:}))$, the weighted state vector is $\tilde{\mathbf{s}}_t = \sum_{i=1}^{t} b_i \mathbf{s}_i$.

Copy mechanism In the prediction process, we also try to apply copy salient words or entities from the input document but the results are not well. We infer that the input of the summarization task is often long documents, implementing copy mechanism directly may not get satisfying results. Gehrmann et al. (2018) use mask and modify the copy attention distribution to only include tokens identified by the content selector. Inspired by this idea, we join the tokenized entities with the [SEP] token, which produces an entity text C similar to the input document D. Then we use the document encoder to get embedding c_i of every token in this text.

For a given token representation c_i in t-th decode process, we concatenate c_i with current decoder state $\mathbf{s}_t^g$, and calculate the similarity score of these two vector:

$$\beta_t^i = W_c[\mathbf{s}_t^g; c_i]. \tag{3}$$

We adopt the basic pointer generator structure which combines two probabilities, one comes from the pointer network $P_{copy}(c_i) = \frac{e^{\beta_t^i}}{\sum_j e^{\beta_t^j}}$ and the other is the normal generation probability P_{vocab}. Then a switch probability $p_{gen} = \sigma(w_g \mathbf{s}_t^g + w_c \mathbf{c} + b_{gen})$ combine these two distribution and determine the final probabilities of each word in vocabulary. But the results is still not up to our expectations. We left this part to the future work.

Significant test The size of the CNNDM test set is 11,490. We sample 3000 samples from the test set with replacement, and this process is repeated 100 times. And We draw 1000 samples from the nyt50 test set which has 3,452 examples. We mention in the table 1 of our paper that we adopt the method proposed in the paper "An Empirical Investigation of Statistical Significance in NLP". And the calculated p value is 0. In other words, our model performs significantly better than other models with $p < 0.05$. We admit that the improved score is not too much, compared with BERTSUMABS. But it is worth noting that Berg-Kirkpatrick et al. (2012) also propose that systems whose outputs are highly correlated will achieve higher confidence at lower metric gains.

Triple filtering The extracted entities include not only the noun phrase but some structures that modifies the major part such as articles and preposition structure. The modification parts are noisy sometimes. Some restricted parts are too lengthy and interfere with the model's understanding of the main part. We use a simple trick to alleviate this problem, in which we remove any triple whose argument (subject or object) has more than 10 words. The rouge score is 41.85/19.15/38.92. The decrease in score may be that some important triples are dropped when the entity length is limited to 10. We will continue to explore the effect of the entity length limitation in the future.

Manual evaluation of extracted triples It is worth noting that this article replaces all relations with "have relation with" in order to simplify the model. We show the extracted triples and golden summary for reference in Table 1:

References

Berg-Kirkpatrick T, Burkett D, Klein D (2012). An empirical investigation of statistical significance in NLP. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 995–1005. Association for Computational Linguistics, Jeju Island, Korea.

Gehrmann S, Deng Y, Rush A (2018). Bottom-up abstractive summarization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4098–4109. Association for Computational Linguistics, Brussels, Belgium.

Tay Y, Bahri D, Metzler D, Juan DC, Zhao Z, Zheng C (2020). Synthesizer: Rethinking self-attention in transformer models. *arXiv preprint arXiv:2005.00743*.

Table 1: Manual evaluation of extracted triples, with golden summarization as reference.

extrated triples
(’david hawkins’, ’have relation with’, "the organization ’s commission on the use of standardized tests") (’david hawkins’, ’have relation with’, ’baylor university’) (’david hawkins’, ’have relation with’, ’america’) (’baylor university’, ’have relation with’, ’america’) (’baylor university’, ’have relation with’, ’the sat’) (’baylor university’, ’have relation with’, ’enough students’) (’the sat’, ’have relation with’, ’enough students’) (’act’, ’have relation with’, ’the sat and act scores’) (’act’, ’have relation with’, ’act scores’) (’the sat and act scores’, ’have relation with’, ’act scores’) (’u.s. news’, ’have relation with’, ’u.s. news and world report’) (’u.s. news’, ’have relation with’, ’world report’) (’u.s. news’, ’have relation with’, ’standardized tests play an integral role in the college admissions process’) (’u.s. news and world report’, ’have relation with’, ’world report’) (’u.s. news and world report’, ’have relation with’, ’standardized tests play an integral role in the college admissions process’) (’world report’, ’have relation with’, ’standardized tests play an integral role in the college admissions process’) (’the rankings’, ’have relation with’, ’which colleges’)
golden summary
david hawkins : admission tests are wrongly used to rank college quality. hawkins says baylor university ’s incentives for test scores are a mistake. grades are much more important than test scores in admissions decisions , he says. hawkins : u.s. news should drop sat and act scores in rankings.
extrated triples
(’martin robinson a home guard’, ’have relation with’, ’britain’) (’martin robinson a home guard’, ’have relation with’, ’the government’) (’martin robinson a home guard’, ’have relation with’, ’cabinet office minister francis maude’) (’britain’, ’have relation with’, ’the government’) (’britain’, ’have relation with’, ’cabinet office minister francis maude’) (’the government’, ’have relation with’, ’cabinet office minister francis maude’) (’security services’, ’have relation with’, ’nick clegg’) (’mrs may’, ’have relation with’, ’lib dem leader nick clegg’) (’mrs may’, ’have relation with’, ’communications data bill’) (’mrs may’, ’have relation with’, ’victims of serious crime , terrorism and child sex offences in the eye’) (’lib dem leader nick clegg’, ’have relation with’, ’communications data bill’) (’communications data bill’, ’have relation with’, ’victims of serious crime , terrorism and child sex offences in the eye’) (’this bill’, ’have relation with’, ’victims of crime , police and the public’)
golden summary
home guard of cyber experts will protect uk business and armed forces. 9 out of 10 companies suffered online attack in 2011 - costing £ 250,000 a time.warning that cyber terrorists are targeting utility firms to disrupt supplies of gas , electricity and water. home secretary says opponents of her communications data bill must look victims of terrorism ’ in the eye ’. deputy pm nick clegg wants ’ snooper ’s charter ’ delayed until 2014.
extrated triples
(’high school kindergarten mr smyth’, ’have relation with’, ’parents’) (’high school kindergarten mr smyth’, ’have relation with’, ’school’) (’parents’, ’have relation with’, ’school’) (’mr smyth’, ’have relation with’, ’most people’) (’mr smyth’, ’have relation with’, ’parents of new kindergarten pupils’) (’mr smyth’, ’have relation with’, ’high school’) (’mr smyth’, ’have relation with’, ’parents should never underestimate the old fashioned note in a lunch-box’) (’mr smyth’, ’have relation with’, ’parents of new students’) (’mr smyth’, ’have relation with’, ’their child’) (’parents of new kindergarten pupils’, ’have relation with’, ’high school’) (’parents of new kindergarten pupils’, ’have relation with’, ’parents should never underestimate the old fashioned note in a lunch-box’) (’high school’, ’have relation with’, ’parents should never underestimate the old fashioned note in a lunch-box’) (’high school’, ’have relation with’, ’parents of new students’) (’high school’, ’have relation with’, ’their child’) (’parents of new students’, ’have relation with’, ’their child’) (’their child’, ’have relation with’, ’the teacher’)
golden summary
teacher and education expert ciaran smyth , has revealed his top tips for making the transition from holidays to education more bearable. for high school students , he recommends creating a routine , picking a perfect study space and the occasional positive note in the lunch box. he warned parents of new kindergarten pupils to make their goodbyes quick and unemotional and label everything.