maries extracted by RNES are of higher quality than summaries produced by previous works.

Table 2: Performance comparison on CNN/Daily Mail test set, evaluated with full-length F1 ROUGE scores (%). All scores of RNES are statistically significant using 95% confidence interval with respect to previous best models.

Model	R-1	R-2	R-L
Lead-3	39.2	15.7	35.5
(Nallapati et al. 2016)	35.4	13.3	32.6
(Nallapati et al. 2017)	39.6	16.2	35.3
(See et al. 2017)	39.53	17.28	35.38
NES	37.75	17.04	33.92
RNES w/o coherence	**41.25**	**18.87**	**37.75**
RNES w/ coherence	40.95	18.63	37.41

Though RNES with the coherence reward achieves higher ROUGE scores than baselines, there is a small gap between its score and that of RNES trained without coherence model. This is because that the coherence objective and ROUGE score do not always agree with each other. Since ROUGE is simply computed based on n-grams or longest common subsequence, it is ignorant of the coherence between sentences. Therefore, enhancing coherence may lead to a drop of ROUGE. However, the 95% confidence intervals of the two RNES models overlap heavily, indicating that their difference in ROUGE is insignificant.

Table 3: Comparison of human evaluation in terms of informativeness(Inf), coherence(Coh) and overall ranking. Lower is better.

Model	Inf	Coh	Overall
RNES w/o coherence	1.183	1.325	1.492
RNES w/ coherence	**1.125**	**1.092**	**1.209**

We also conduct a qualitative evaluation to find out whether the introduction of coherence reward improves the coherence of the output summaries. We randomly sample 50 documents from the test set and ask three volunteers to evaluate the summaries extracted by RNES trained with or without coherence as the reward. They are asked to compare and rank the outputs of two models regarding three aspects: informativeness, coherence and overall quality. The better one will be given rank 1, while the other will be given rank 2 if it is worse. In some cases, if the two outputs are identical or have the same quality, the ranks could be tied, i.e., both of them are given rank 1. Table 3 shows the results of human evaluation. RNES model trained with coherence reward is better than RNES model without coherence reward in all three aspects, especially in the coherence. The result indicates that the introduction of coherence effectively improves the coherence of extracted summaries, as well as the overall quality. It is surprising that summaries produced by RNES with coherence are also more informative than RNES without coherence, indicating that ROUGE might not be the gold standard to evaluate informativeness as well.

Table 4 shows a pair of summary produced by RNES with or without coherence. The summary produced by RNES without coherence starts with pronoun 'That' which is referring to a previously mentioned fact, and hence it may lead to confusion. In contrast, the output of RNES trained with coherence reward includes the sentence "The earthquake disaster . . . " before referring to this fact in the second sentence, and therefore is more coherent and readable. This is because the coherence model gives a higher score to the second sentence if it can form a coherent sentence pair with the first sentence. In REINFORCE training, if the second sentence receives a high coherence score, the action of extracting the first sentence before the second one will be strengthened. This example shows that coherence model is indeed effective in changing the behavior of RNES towards extracting summaries that are more coherent.

Table 4: Examples of extracted summary.

Reference: Peter Spinks from the Sydney Morning Herald reported on Amasia. Within 200 million years, he said the new supercontinent will form. One researcher recently travelled to Nepal to gather further information. He spotted that India, Eurasia and other plates are slowly moving together.
RNES w/o coherence: That's according to one researcher who travelled to the country to study how the Indian and Eurasian plates are moving together. And using new techniques, researchers can now start examining the changes due to take place over the next tens of millions of years like never before. Earth's continents are slowly moving together, and in 50 to 200 million years they are expected to form a new supercontinent called Amasia. In 2012 a study suggested this may be centered on the North Pole. The idea that Earth is set to form a new supercontinent-dubbed Amasia - is not new.
RNES w/ coherence: **The earthquake disaster in Nepal has highlighted how Earth's land masses are already in the process of forming a new supercontinent. That**'s according to one researcher who travelled to the country to study how the Indian and Eurasian plates are moving together. And using new techniques, researchers can now start examining the changes due to take place over the next tens of millions of years like never before. Earth's continents are slowly moving together, and in 50 to 200 million years they are expected to form a new supercontinent called Amasia.

Conclusion

In this paper, we proposed a Reinforced Neural Extractive Summarization model to extract a coherent and informative summary from a single document. Empirical results show that the proposed RNES model can balance between the cross-sentence coherence and importance of the sentences effectively, and achieve state-of-the-art performance on the benchmark dataset. For future work, we will focus on improving the performance of our neural coherence model and introducing human knowledge into the RNES.

Acknowledgments

This work is supported by grants from WeChat-HKUST Joint Lab on Artificial Intelligence Technology (WHAT Lab). Baotian Hu acknowledges partial support from the University of Massachusetts Medical School.