

Exploiting Discourse-Level Segmentation for Extractive Summarization

Zhengyuan Liu, Nancy F. Chen

Institute for Infocomm Research, A*STAR

{liu.zhengyuan, nfychen}@i2r.a-star.edu.sg

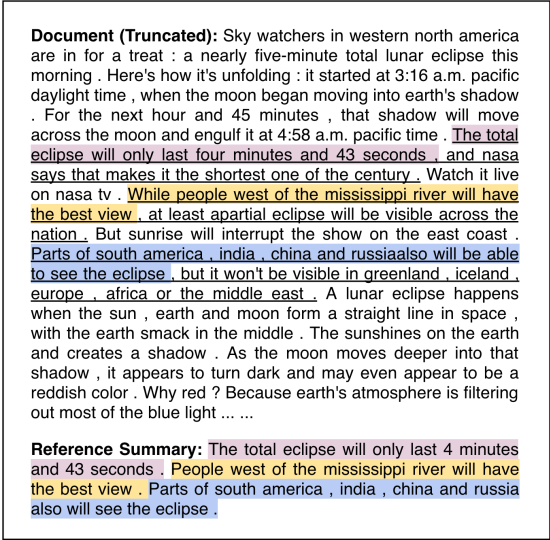
Abstract

Extractive summarization selects and concatenates the most essential text spans in a document. Most, if not all, neural approaches use sentences as the elementary unit to select content for summarization. However, semantic segments containing supplementary information or descriptive details are often nonessential in the generated summaries. In this work, we propose to exploit discourse-level segmentation as a finer-grained means to more precisely pinpoint the core content in a document. We investigate how the sub-sentential segmentation improves extractive summarization performance when content selection is modeled through two basic neural network architectures and a deep bi-directional transformer. Experiment results on the CNN/Daily Mail dataset show that discourse-level segmentation is effective in both cases. In particular, we achieve state-of-the-art performance when discourse-level segmentation is combined with our adapted contextual representation model.

1 Introduction

Document summarization is a core task in natural language processing, targeting to automatically generate a shorter version of one or multiple documents while retaining the most important information. As a straightforward and effective method, extractive summarization creates a summary by selecting and subsequently concatenating the most salient semantic units in a document; much effort has been devoted to this area. While traditional approaches rely heavily on human-engineered features, which is time-consuming and difficult to expand to massive data, neural networks can be trained in an end-to-end manner with fewer linguistic annotation, achieving favorable improvements on large-scale benchmarks (Hermann et al., 2015; Cheng and Lapata, 2016; Zhou et al., 2018).

However, the selected content in current neural approaches is often not succinct enough. As



Document (Truncated): Sky watchers in western north america are in for a treat : a nearly five-minute total lunar eclipse this morning . Here's how it's unfolding : it started at 3:16 a.m. pacific daylight time , when the moon began moving into earth's shadow . For the next hour and 45 minutes , that shadow will move across the moon and engulf it at 4:58 a.m. pacific time . The total eclipse will only last four minutes and 43 seconds . and nasa says that makes it the shortest one of the century . Watch it live on nasa tv . While people west of the mississippi river will have the best view . at least a partial eclipse will be visible across the nation . But sunrise will interrupt the show on the east coast . Parts of south america , india , china and russia also will be able to see the eclipse . but it won't be visible in greenland , iceland , europe , africa or the middle east . A lunar eclipse happens when the sun , earth and moon form a straight line in space , with the earth smack in the middle . The sunshines on the earth and creates a shadow . As the moon moves deeper into that shadow , it appears to turn dark and may even appear to be a reddish color . Why red ? Because earth's atmosphere is filtering out most of the blue light

Reference Summary: The total eclipse will only last 4 minutes and 43 seconds . People west of the mississippi river will have the best view . Parts of south america , india , china and russia also will see the eclipse .

Figure 1: An example of news summarization. Colored spans are salient segments selected to form a summary, and their corresponding sentences are underlined.

shown in Figure 1, human editors tend to further distill the selected sentences by removing nonessential phrases or clauses to compose more concise summaries. While the extracted sentences often contain the main points of the document, such sentences are usually embellished with more clauses or segments of background knowledge to give the readers more context, descriptive details to paint a more colorful picture, supplementary information to make the content more comprehensive, or subtle nuances to give a more polished touch. Therefore, sentence-level extraction might dilute the density of the key information in the summary.

To tackle this problem, we postulate that content selection can benefit from finer-grained text segmentation. Inspired by the rhetorical structure theory (RST) (Mann and Thompson, 1988), we propose to split documents to sub-sentential segments following its discourse structure, as RST provides a coherent and well-organized representation of documents and suggests discourse-level

segmentation can help model semantic information with more refined granularity. This can help us more precisely pinpoint the key information when we subsequently use neural models to select content for summarization. We empirically compare two different selector architectures: a multi-layer recurrent neural network (RNN) and a Transformer network, as they each have their own model assumptions and knowledge representations (Liu et al., 2019), and we further fine-tune a contextualized language model based on the deep bi-directional Transformer. Our experiments on the CNN/Daily Mail dataset demonstrate that discourse-level segmentation is effective, achieving state-of-the-art performance when combined with an adapted large-scale pre-trained model of contextualized language representation.

2 In Relation to Other Work

Content selection plays a key role for both extractive and abstractive paradigms of text summarization (Nallapati et al., 2017; Zhou et al., 2018; Gehrmann et al., 2018; Hsu et al., 2018). While traditional approaches utilize human-engineered linguistic features (Jones, 2007; Shen et al., 2007), neural network approaches learn the features in a data-driven manner, with components such as semantic vector representation of words (Pennington et al., 2014), contextual representation with various neural structures (Schuster and Paliwal, 1997; Kalchbrenner et al., 2014), attention mechanism and hierarchical document modeling (Cheng and Lapata, 2016). Despite the achievement of sophisticated neural extractive models (Kedzie et al., 2018), sentences are the default elementary semantic unit, potentially leading to low density of key information in the summary. Thus, we target to introduce a finer-grained segmentation scheme.

Discourse structure has proved effective for analyzing and extracting important spans in a document (Louis et al., 2010; Hirao et al., 2015). Utilizing the elementary unit segmentation for extractive summarization has been studied via traditional feature-based approaches (Li et al., 2016). However, to the best of our knowledge, it has not been adopted in the recent neural approaches for summarization. While discourse analysis contains unit segmentation, nucleus-satellite recognition and relation classification (Carlson et al., 2001), segmentation has the highest accuracy (Joty et al., 2013; Heilman and Sagae, 2015), thus making it a more

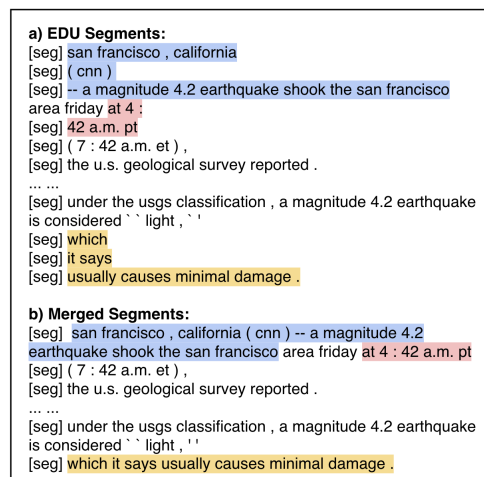


Figure 2: Examples of discourse-level segmentation. a) spans in blue and yellow are the EDUs with semantically fragmented information and spans in red are the inaccurate EDU splits; b) the sub-sentential segments after merging.

mature pre-processing task to be integrated with downstream tasks such as summarization.

3 Discourse-Level Segmentation

For discourse-level segmentation for content selection, our target is to split a document into sub-sentential segments that preserve congruently semantic information.

In the RST discourse framework, a document is split into elementary discourse units (EDUs) that are contiguous token spans similar to independent clauses, and re-organized in a binary tree structure. EDU pairs are assigned to specific discourse relations like elaboration, condition, and contrast, ensuring the semantic coherence and integrity of the entire structure. Therefore, we followed the conventions annotated in the RST Discourse Treebank¹ (Carlson et al., 2001), which contains discourse tree annotations for 385 WSJ articles from the Penn Treebank corpus (Marcus et al., 1993). We trained a fast and robust model² (Heilman and Sagae, 2015) on the treebank, obtaining over 0.84 accuracy on its validation set. Next, we applied the model to segment the documents, and here we firstly conducted sentence splitting as it improved the accuracy of subsequent EDU segmentation. Then, we specified [edu_seg] tags between two EDUs and [sen_seg] tags between two sentences.

¹<https://catalog.ldc.upenn.edu/LDC2002T07>

²<https://github.com/EducationalTestingService/discourse-parsing>

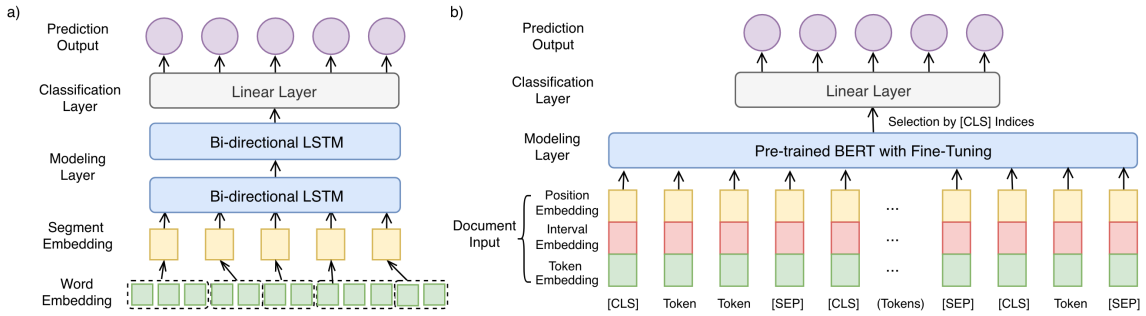


Figure 3: Content selector designs: a) RNN architecture; b) BERT architecture.

As shown in Figure 2a, some EDUs are too semantically fragmented to form an informative segment. In addition, there are inevitable errors in the segmentations, which is not unexpected due to the limited size of the training corpus. In order to balance the segment length and informativeness in addition to mitigating the side-effects from inaccurate EDU segmentation, we therefore defined a set of criteria such as word length, the existence of verbs, and symmetry of quotation marks, to merge short EDUs into longer sub-sentential segments, which are typically at the clause-level. A discourse segment is on average 14 tokens after merging compared to average 7.5 tokens before merging (see Figure 2b).

4 Neural Content Selection

Given a document d containing a number of text spans $[span_1, span_2, \dots, span_n]$, the content selector assigns a score $y_i \in [0, 1]$ to each span i , indicating its probability of being included in the summary. We implemented and compared three neural architectures, which we elaborate below.

4.1 RNN Selector

Recurrent neural network, with its capability of sequential information modeling, is widely applied in extractive summarization.

Here we introduce a multi-layer RNN architecture as the selector, which is simple but competitive as in (Kedzie et al., 2018). As shown in Figure 3a, the input is a sequence of discourse-level segment embeddings, which is calculated by averaging word embeddings. The sentence boundary tags $[sen_seg]$ are converted to a randomly initialized embedding vector. In the modeling layer, a multi-layer Bi-directional LSTM (Schuster and Paliwal, 1997) is used, in which the forward and backward hidden states are concatenated. Then the hidden representation is fed to a linear layer

with a sigmoid function, to predict the probability of extracting each segment.

In our setting, word embeddings were initialized with pre-trained 300-dimension GloVe (Pennington et al., 2014) and fixed during training. Vocabulary size was set to 200k. Out-of-vocabulary words were mapped to a zero embedding. For the modeling layer, it was empirically shown that a two-layer Bi-LSTM worked best. Adam optimizer with $3e-4$ learning rate was used (Kingma and Ba, 2015). Drop-out with $rate = 0.2$ was applied in the modeling and classification layers (Srivastava et al., 2014).

4.2 Transformer Selector

The Transformer (Vaswani et al., 2017) is another effective and efficient neural architecture for language modeling. To compare it with the recurrent encoding scheme, we changed the modeling layer of the design in Section 4.1, by replacing the Bi-directional LSTM with a multi-head attention encoding component. In our setting, we empirically set the layer number of Transformer encoder to 3, and the self-attention head number to 5. The hidden and feed-forward dimension size were set to 400 and 1024 respectively. To better utilize the sequential information, we pre-calculated the position embedding with 100 dimension size and concatenated it with the segment embedding as input. The other hyperparameters of training were set as the same as the RNN selector.

4.3 BERT Selector

Deep contextual representation models with the sophisticated architecture for capturing complex features and unsupervised pre-training on large-scale corpora (e.g. ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018)), have boosted the performance of various NLP tasks. It has been shown that the pre-trained models have implicitly learned linguistic properties such as syntax (Hewitt and

Model	ROUGE-1 F1	ROUGE-2 F1
Lead-3	40.43	17.64
SummaRuNNer*	39.60	16.20
NeuSUM*	41.59	19.01
S-Level Oracle	53.29	32.14
S-Level Bi-LSTM	38.86	17.31
S-Level Transformer	38.57	17.26
S-Level BERT	41.02	19.39
D-Level Oracle	57.74	35.16
D-Level Bi-LSTM	40.36	18.42
D-Level Transformer	40.03	17.83
D-Level BERT	42.78	20.23

Table 1: Experimental results of baselines, oracles and models on Sentence-Level (S-Level) and Discourse-Level (D-Level) segmentation. * denotes results from the papers.

Manning, 2019) and semantic dependency (Conneau et al., 2018).

Since BERT is originally trained as a contextualized language representation model, we adapted and fine-tuned it for discourse-level content selection, as illustrated in Figure 3b. While BERT can be applied to encode sequences separately or jointly, the latter works better for document tasks (Qiao et al., 2019). Therefore, we decided to take the adapted embedding list as our document input. For each segment, we inserted a [CLS] token before and a [SEP] token after it, then converted it to token embeddings with word-piece tokenization (Wu et al., 2016). To distinguish multiple segments, we assigned 0/1 to adjacent segment pairs respectively as interval label. Combined with position embedding, the document input was fed to BERT for contextualized encoding. After that, we collected all the hidden states of [CLS] tokens in the last layer of BERT, which captured the contextual information of segments, then fed them to a linear layer with sigmoid function to get the predicted salient scores.

In our setting, we used the PyTorch version of ‘bert-base-uncased’ BERT³, and fine-tuned all the layers during training. We truncated the lengthy documents to the size of 512 due to the limitation of position index and the significant increase of computational cost by the sliding-window strategy. Adam algorithm (Kingma and Ba, 2015) with warm-up learning was used for optimization. Drop-out rate was set to 0.2 was applied after the modeling layer (Srivastava et al., 2014).

For all models, we obtained the normalized predicted score y_i of each segment i . The loss is calculated as the binary cross entropy of y_i against

³<https://github.com/huggingface/pytorch-transformers>

<p>Reference Summary (Human): Bob barker returned to host " the price is right " on wednesday . Barker , 91 , had retired as host in 2007 .</p> <p>Generated Candidate (S-Level BERT): (cnn) For the first time in eight years , a tv legend returned to doing what he does best . On the april 1 edition of " the price is right " encountered not host drew carey but another familiar face in charge of the proceedings . Instead , there was bob barker , who hosted the tv game show for 35 years before stepping down in 2007 .</p> <p>Generated Candidate (D-Level BERT): On the april 1 edition of " the price is right " encountered not host drew carey , instead , there was bob barker , who hosted the tv game show for 35 years before stepping down in 2007 .</p>
--

Figure 4: Examples of generated summaries. Colored spans contain key information from the gold reference.

ground-truth \hat{y}_i . Each epoch constitutes a full pass through the data with shuffling. During training, the best models were selected with early stopping strategy on the validation set.

5 Experiment & Results

Experiments were conducted on the CNN/Daily Mail dataset (Hermann et al., 2015). We applied discourse-level segmentation in Section 3 on the training, validation, and test set. Since there is no oracle extractive summary set for generating gold labels \hat{y}_i , we constructed them with a greedy algorithm similar to (Kedzie et al., 2018), and obtained the discourse-level oracle summaries by concatenating segments with gold label indices.

Having gotten the prediction outputs, we selected 4 discourse-level segments with the highest scores for each document sample, and then evaluated the candidates against reference summaries with the F1 scores of ROUGE-1 and ROUGE-2 (Lin, 2004). We compared our method with several strong extractive baselines: SummaRuNNer (Nallapati et al., 2017), NeuSUM (Zhou et al., 2018), and Lead-3, a simple but competitive baseline, which takes the first 3 sentences of the document as a summary. Moreover, as control, we split documents into sentences, built a sentence-level oracle set, and trained the selector models in which the most 3 salient sentences were selected.

Results are listed in Table 1, all models with discourse-level segmentation outperform those with sentence-level segmentation, demonstrating the effectiveness of our finer-grained means. Even the vanilla multi-layer Bi-LSTM is competitive when compared to the previous state-of-the-art models, and it slightly outperformed the Transformer architecture. Moreover, the fine-tuned BERT model achieves further improvement, suggesting its contextual modeling which is implicitly

conducted at the sentence-level can be transferred to sub-sentential levels. Additionally, we observed that merging initial EDUs in Section 3 significantly contributed to obtaining better performance, suggesting that preserving semantic congruence is crucial in sub-sentential segmentation.

An example from our results demonstrates that discourse-level extractive summarization retains most of the key information in the reference, and it is more concise than the sentence-level counterpart (see Figure 4). It is able to trim the trivial details that are nonessential to the core meaning of the source text, achieving 19% decrease of the average word length when compared to the sentence-level baseline (from 71 tokens to 57 tokens).

6 Conclusion

In this paper, we proposed using sub-sentential segmentation for single-document extractive summarization. We exploited a discourse-level segmentation scheme and verified its effectiveness by obtaining improvements over sentence-level schemes. We adapted and fine-tuned a deep contextual model for our task and achieved state-of-the-art performance. Incorporating discourse tree structures implicitly or explicitly in the neural network approaches for summarization is an area of interest for future work.

Acknowledgments

The authors would like to thank insightful discussions with Bonnie Webber, Wenqiang Lei, and Ai Ti Aw. This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046).

References

- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. [Building a discourse-tagged corpus in the framework of rhetorical structure theory](#). In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.
- Jianpeng Cheng and Mirella Lapata. 2016. [Neural summarization by extracting sentences and words](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018.

[What you can cram into a single \$\\$&!#*\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. [Bottom-up abstractive summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Michael Heilman and Kenji Sagae. 2015. [Fast rhetorical structure theory discourse parsing](#). *CoRR*, abs/1505.02425.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in neural information processing systems*, pages 1693–1701.

John Hewitt and Christopher D Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Tsutomu Hirao, Masaaki Nishino, Yasuhisa Yoshida, Jun Suzuki, Norihito Yasuda, and Masaaki Nagata. 2015. [Summarizing a document by trimming the discourse tree](#). *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(11):2081–2092.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. [A unified model for extractive and abstractive summarization using inconsistency loss](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141. Association for Computational Linguistics.

Karen Spärck Jones. 2007. [Automatic summarising: The state of the art](#). *Information Processing & Management*, 43(6):1449–1481.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. [Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496,

- Sofia, Bulgaria. Association for Computational Linguistics.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. [A convolutional neural network for modelling sentences](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.
- Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. [Content selection in deep learning models of summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.
- Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. [The role of discourse units in near-extractive summarization](#). In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 147–156. Association for Computational Linguistics.
- William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [Glove: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*.
- Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conferences on Artificial Intelligence.*, pages 2862–2867.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- 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 *Advances in neural information processing systems*, pages 5998–6008.
- 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*.
- Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. [Neural document summarization by jointly learning to score and select sentences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.