SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization

Yang Gao[†], **Wei Zhao**[‡], **Steffen Eger**[‡]
[†] Dept. of Computer Science, Royal Holloway, University of London, UK
[‡] Computer Science Department, Technische Universität Darmstadt, Germany
yang.gao@rhul.ac.uk, {zhao,eger}@aiphes.tu-darmstadt.de

Abstract

We study *unsupervised multi-document summarization evaluation* metrics, which require neither human-written reference summaries nor human annotations (e.g. preferences, ratings, etc.). We propose *SUPERT*, which rates the quality of a summary by measuring its semantic similarity with a *pseudo reference summary*, i.e. selected salient sentences from the source documents, using contextualized embeddings and soft token alignment techniques. Compared to the state-of-the-art unsupervised evaluation metrics, SUPERT correlates better with human ratings by 18-39%. Furthermore, we use SUPERT as rewards to guide a neural-based *reinforcement learning* summarizer, yielding favorable performance compared to the state-of-the-art unsupervised summarizers. All source code is available at https://github.com/yg211/acl20-ref-free-eval.

1 Introduction

Evaluating the quality of machine-generated summaries is a highly laborious and hence expensive task. Most existing evaluation methods require certain forms of human involvement, thus are *supervised*: they either directly let humans rate the generated summaries (e.g. Pyramid (Nenkova and Passonneau, 2004)), elicit human-written reference summaries and measure their overlap with the generated summaries (e.g. using ROGUE (Lin, 2004a) or MoverScore (Zhao et al., 2019)), or collect some human annotations (e.g. preferences over pairs of summaries (Gao et al., 2019a)) to learn a summary evaluation function. Evaluation in *multi-document summarization* is particularly expensive: Lin (2004b) reports that it requires 3,000 hours of human effort to evaluate the summaries from the Document Understanding Conferences (DUC)[1].

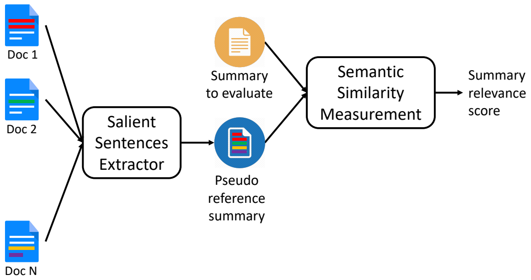

Figure 1: Workflow of SUPERT.

To reduce the expenses for evaluating multi-document summaries, we investigate *unsupervised evaluation* methods, which require neither human annotations nor reference summaries. In particular, we focus on evaluating the *relevance* (Peyrard, 2019) of multi-document summaries, i.e. measuring how much salient information from the source documents is covered by the summaries. There exist a few unsupervised evaluation methods (Louis and Nenkova, 2013; Sun and Nenkova, 2019), but they have low correlation with human relevance ratings at *summary level*: given multiple summaries for the same source documents, these methods can hardly distinguish summaries with high relevance from those with low relevance (see §3).

Contributions. First, to better measure the semantic overlap between source documents and machine-generated summaries, we propose to use state-of-the-art contextualized text encoders, e.g. *BERT* (Devlin et al., 2019) and its variant *Sentence-BERT* (SBERT) (Reimers and Gurevych, 2019), which is optimized for measuring semantic similarity between sentences, to develop unsupervised evaluation methods. We measure the relevance of a summary in two steps: **(i)** identifying the salient information in the input documents, to build a *pseudo reference summary*, and **(ii)** measuring the semantic overlap between the pseudo reference and the

[1] http://duc.nist.gov/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1347–1354
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

summary to be evaluated. The resulting evaluation method is called *SUPERT* (SUmmarization evaluation with Pseudo references and bERT). Fig. 1 illustrates the major steps of SUPERT. We show that compared to state-of-the-art unsupervised metrics, the best SUPERT correlates better with the human ratings by 18-39% (in Kendall's τ).

Second, we use SUPERT as *reward functions* to guide *Reinforcement Learning* (RL) based extractive summarizers. We show it outperforms the state-of-the-art unsupervised summarization methods (in multiple ROUGE metrics).

2 Related Work

Reference-based Evaluation. Popular metrics like ROUGE (Lin, 2004a), BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009) fall into this category. They require (preferably, multiple) human written references and measure the relevance of a summary by comparing its overlapping word sequences with references. More recent work extends ROUGE with WordNet (ShafieiBavani et al., 2018a), word embeddings (Ng and Abrecht, 2015), or use contextualized-embedding-based methods (Zhang et al., 2019; Zhao et al., 2019) to measure the semantic similarity between references and summaries.

Annotation-based Evaluation. Some methods directly ask human annotators to rate summaries following some guidelines, e.g. *Responsiveness*, which measures the overall quality (relevance, fluency and readability) of summaries, and *Pyramid* (Nenkova and Passonneau, 2004), which measures summaries' relevance. Recently, systems have been developed to ease the construction of Pyramid scores, e.g. (Hirao et al., 2018; Yang et al., 2016; Gao et al., 2019b; Shapira et al., 2019), but they still require human-annotated Summary Content Units (SCUs) to produce reliable scores. Besides SCUs, recent work has explored eliciting preferences over summaries (Zopf, 2018; Gao et al., 2018, 2019a) and annotations of important bi-grams (P.V.S and Meyer, 2017) to derive summary ratings.

Some methods collect human ratings on a small number of summaries to train an evaluation function. Peyrard et al. (2017); Peyrard and Gurevych (2018) propose to learn an evaluation function from Pyramid and Responsiveness scores, by using classic supervised learning methods with hand-crafted features. ShafieiBavani et al. (2018b) use the same idea but design corpus based and lexical resource based word embeddings to build the features. Böhm et al. (2019) train a BERT-based evaluation function with 2,500 human ratings for 500 machine-generated summaries from the CNN/DailyMail dataset; their method correlates better with human ratings than ROUGE and BLEU. However, as their method is designed for evaluating single-document summaries, it correlates poorly with the Pyramid scores for multi-document summaries (see §3).

Unsupervised Evaluation. Louis and Nenkova (2013) measure the relevance of a summary using multiple heuristics, for example by computing the Jensen-Shannon (JS) divergence between the word distributions in the summary and in the source documents. Ryang and Abekawa (2012); Rioux et al. (2014) develop evaluation heuristics inspired by the maximal marginal relevance metrics (Goldstein et al., 2000). But these methods have low correlation with human ratings at summary level (see §3). Scialom et al. (2019) propose to generate questions from source documents and evaluate the relevance of summaries by counting how many questions the summaries can answer. However, they do not detail how to generate questions from source documents; also, it remains unclear whether their method works for evaluating multi-document summaries. Sun and Nenkova (2019) propose a single-document summary evaluation method, which measures the cosine similarity of the ELMo embeddings (Peters et al., 2018) of the source document and the summary. In §3, we show that their method performs poorly in evaluating multi-document summaries. SUPERT extends their method by using more advanced contextualized embeddings and more effective text alignment/matching methods (§4), and by introducing pseudo references (§5).

3 Datasets, Baselines and Upper Bounds

Datasets. We use two multi-document summarization datasets from the Text Analysis Conference (TAC)[2] shared tasks: TAC'08 and TAC'09. In line with Louis and Nenkova (2013), we only use the initial summaries (the A part) in these datasets. TAC'08 includes 48 topics and TAC'09 includes 44. Each topic has ten news articles, four reference summaries and 57 (TAC'08) and 55 (TAC'09) machine-generated summaries. Each news article on average has 611 words in 24 sentences. Each summary has at most 100 words and receives a

[2] https://tac.nist.gov/

	TAC'08			TAC'09		
	r	ρ	τ	r	ρ	τ
Baselines (unsupervised evaluation)						
TF-IDF	.364	.330	.236	**.388**	**.395**	**.288**
JS	**.381**	**.333**	**.238**	**.388**	.386	.283
REAPER	.259	.247	.174	.332	.354	.252
C_{ELMo}	.139	.108	.076	.334	.255	.183
Böhm19	.022	-.001	.001	.075	.043	.031
Upper bounds (reference-based evaluation)						
Rouge1	.747	.632	.501	.808	.692	.533
Rouge2	.718	.635	.498	.803	.694	.531
Mover	**.760**	**.672**	**.507**	**.831**	**.701**	**.550**

Table 1: Summary-level correlation between some popular evaluation metrics and human ratings. Unsupervised metrics (upper) measure the similarity between summaries and the source documents, while reference-based metrics (bottom) measure the similarity between summaries and human-written reference summaries.

Pyramid score, which is used as the ground-truth human rating in our experiments.

Baselines & Upper Bounds. For baselines, we consider *TF-IDF*, which computes the cosine similarity of the tf-idf vectors of source and summaries; *JS*, which computes the JS divergence between the words distributions in source documents and summaries; and the *REAPER* heuristics proposed by Rioux et al. (2014). In addition, we use the learned metric from Böhm et al. (2019) (*Böhm19*) and the ELMo-based metric by Sun and Nenkova (2019) (C_{ELMo}, stands for cosine-ELMo; see §2). In all these methods, we remove stop-words and use the stemmed words, as we find these operations improve the performance. For C_{ELMo}, we vectorize the documents/summaries by averaging their sentences' ELMo embeddings. As for upper bounds, we consider three strong reference-based evaluation metrics: ROUGE-1/2 and MoverScore (Zhao et al., 2019); note that references are not available for unsupervised evaluation metrics.

We measure the performance of the baselines and upper bounds by their average summary-level correlation with Pyramid, in terms of Pearson's (r), Spearman's (ρ) and Kendall's (τ) correlation coefficients.[3] Table 1 presents the results. All baseline methods fall far behind the upper bounds. Among baselines, the embedding-based methods (Böhm19 and C_{ELMo}) perform worse than the other lexical-based baselines. This observation suggests that to rate multi-document summaries, using existing single-document summaries evaluation metrics (Böhm19) or computing source-summary embeddings' cosine similarity (C_{ELMo}) is ineffective.

	TAC'08			TAC'09		
	r	ρ	τ	r	ρ	τ
C_{BERT}	.035	.066	.048	.130	.099	.071
$\text{C}_{\text{RoBERTa}}$	.100	.126	.091	.262	.233	.165
C_{ALBERT}	.152	.122	.086	.247	.219	.157
C_{SBERT}	.304	.269	.191	.371	.319	.229
$\text{M}_{\text{RoBERTa}}$	.366	.326	.235	.357	.316	.229
M_{SBERT}	**.466**	**.428**	**.311**	**.436**	**.435**	**.320**

Table 2: Performance of contextual-embedding-based metrics. Soft aligning the embeddings of the source documents and the summaries (the bottom part) yields higher correlation than simply computing the embeddings cosine similarity (the upper part).

4 Measuring Similarity with Contextualized Embeddings

In this section, we explore the use of more advanced contextualized embeddings and more sophisticated embedding alignment/matching methods (rather than cosine similarity) to measure summaries relevance. We first extend C_{ELMo} by considering more contextualized text encoders: BERT, RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) and *SBERT*[4]. We use these encoders to produce embeddings for each sentence in the documents/summaries, and perform average pooling to obtain the vector representations for the documents/summaries. We measure the relevance of a summary by computing the cosine similarity between its embedding and the embedding of the source documents. The upper part in Table 2 presents the results. C_{SBERT} outperforms the other cosine-embedding based metrics by a large margin, but compared to the lexical-based metrics (see Table 1) its performance still falls short.

Zhao et al. (2019) recently show that, to measure the semantic similarity between two documents, instead of computing their document embeddings cosine similarity, minimizing their token embeddings *word mover's distances* (WMDs) (Kusner et al., 2015) yields stronger performance. By minimizing WMDs, tokens from different documents are *soft-aligned*, i.e. a token from one document can be aligned to multiple relevant tokens from the other document. We adopt the same idea to measure the semantic similarity between summaries and

[3] We have also considered the percentage of significantly correlated topics; results can be found in the Github repository.

[4] Model `bert-large-nli-stsb-mean-tokens`.

	TAC'08			TAC'09		
	r	ρ	τ	r	ρ	τ
Random3	.139	.194	.189	.123	.172	.175
Random5	.144	.203	.199	.147	.204	.206
Random10	.163	.228	.229	.201	.279	.284
Random15	.206	.287	.320	.185	.258	.268
Top3	.449	.408	.295	.378	.390	.291
Top5	.477	.437	.316	.413	.421	.314
Top10	**.492**	**.455**	**.332**	.444	.450	.333
Top15	.489	.450	.327	**.454**	**.459**	**.340**

Table 3: Building pseudo references by extracting randomly selected sentences (upper) or the first few sentences (bottom). Results of the random extraction methods are averaged over ten independent runs.

source documents, using RoBERTa and SBERT (denoted by M_{RoBERTa} and M_{SBERT}, respectively). The bottom part in Table 2 presents the results. The WMD-based scores substantially outperform their cosine-embedding counterparts; in particular, M_{SBERT} outperforms all lexical-based baselines in Table 1. This finding suggests that, to rate multi-document summaries, soft word alignment methods should be used on top of contextualized embeddings to achieve good performance.

5 Building Pseudo References

WMD-based metrics yield the highest correlation in both *reference-based* (bottom row in Table 1) and *reference-free* (bottom row in Table 2) settings, but there exists a large gap between their correlation scores. This observation highlights the need for reference summaries. In this section, we explore multiple heuristics to build *pseudo references*.

5.1 Simple heuristics

We first consider two simple strategies to build pseudo references: randomly extracting N sentences or extracting the first N sentences from each source document. Results, presented in Table 3, suggest that extracting the top 10-15 sentences as the pseudo references yields strong performance: it outperforms the lexical-based baselines (upper part in Table 1) by over 16% and M_{SBERT} (Table 2) by over 4%. These findings confirm the *position bias* in news articles (c.f. (Jung et al., 2019)).

5.2 Graph-based heuristics

Graph-based methods have long been used to select salient information from documents, e.g. (Erkan and Radev, 2004; Zheng and Lapata, 2019). These methods build grahs to represent the source documents, in which each vertex represents a sentence and the weight of each edge is decided by the similarity of the corresponding sentence pair. Below, we explore two families of graph-based methods to build pseudo references: *position-agnostic* and *position-aware* graphs, which ignore and consider the sentences' positional information, respectively.

Position-Agnostic Graphs. The first graph we consider is SBERT-based LexRank (*SLR*), which extends the classic LexRank (Erkan and Radev, 2004) method by measuring the similarity of sentences using SBERT embeddings cosine similarity. In addition, we propose an SBERT-based clustering (*SC*) method to build graphs, which first measures the similarity of sentence pairs using SBERT, and then clusters sentences by using the *affinity propagation* (Frey and Dueck, 2007) clustering algorithm; the center of each cluster is selected to build the pseudo reference. We choose affinity propagation because it does not require a preset cluster number (unlike K-Means) and it automatically finds the center point of each cluster.

For each method (SLR or SC), we consider two variants: the *individual-graph* version, which builds a graph for each source document and selects top-K sentences (SLR) or the centers (SC) from each graph; and the *global-graph* version, which builds a graph considering all sentences across all source documents for the same topic, and selects the top-M sentences (SLR) or all the centers (SC) in this large graph. According to our preliminary experiments on 20 randomly sampled topics, we set $K = 10$ and $M = 90$.

Position-Aware Graphs. PacSum is a recently proposed graph-based method to select salient sentences from multiple documents (Zheng and Lapata, 2019). In PacSum, a sentence is more likely to be selected if it has higher average similarity with its succeeding sentences and lower average similarity with its preceding sentences. This strategy allows PacSum to prioritize the selection of early-position and "semantically central" sentences. We further extend PacSum by using SBERT to measure sentences similarity (the resulting method is denoted as *SPS*) and consider both the individual- and global-graph versions of SPS.

Furthermore, we propose a method called *Top+Clique* (*TC*), which selects the top-N sentences and the semantically central non-top-N sentences to build the pseudo references. TC adopts

	TAC'08			TAC'09		
	r	ρ	τ	r	ρ	τ
Position-agnostic graphs						
SLR$_{\text{I}}$	.456	.417	.304	.415	**.423**	**.311**
SLR$_{\text{G}}$	**.461**	**.423**	**.306**	**.419**	**.423**	.310
SC$_{\text{I}}$	.409	.364	.261	.393	.383	.280
SC$_{\text{G}}$	.383	.344	.245	.373	.365	.265
Position-aware graphs						
SPS$_{\text{I}}$	.478	.437	.319	.429	.435	.321
SPS$_{\text{G}}$	.472	.432	.313	.427	.432	.318
TC	**.490**	**.449**	**.329**	**.450**	**.454**	**.336**

Table 4: Building pseudo references by position-agnostic (upper) and position-aware (bottom) graphs.

	TAC'08			TAC'09		
	R$_1$	R$_2$	R$_L$	R$_1$	R$_2$	R$_L$
NTD$_{\text{RP}}$	.348	.087	.276	.360	.090	.187
NTD$_{\text{JS}}$	.353	.090	.281	.368	.095	.192
NTD$_{\text{SP}}$	**.376***	**.102***	**.296***	**.380***	**.103***	**.194**
YLS15	.375*	.096	N/A	.344	.088	N/A

Table 5: Training NTD, a RL-based summarizer, with different rewards (RP: REAPER, SP: SUPERT). NTD performance is averaged over ten runs. R$_{1/2/L}$ stands for ROUGE-1/2/L. *: significant advantage ($p < 0.01$ double-tailed t-tests) over the non-asterisks.

the following steps: **(i)** Label top-N sentences from each document as salient. **(ii)** With the remaining (non-top-N) sentences, build a graph such that only "highly similar" sentences have an edge between them. **(iii)** Obtain the cliques from the graph and select the semantically central sentence (i.e. the sentence with highest average similarity with other sentences in the clique) from each clique as *potentially salient sentences*. **(iv)** For each potentially salient sentence, label it as salient if it is not highly similar to any top-N sentences. Based on preliminary experiments on 20 topics, we let $N = 10$ and the threshold value be 0.75 for "highly similar".

Table 4 presents the graph-based methods' performance. Except for SC$_{\text{G}}$, all other graph-based methods outperform baselines in Table 1. Position-agnostic graph-based methods perform worse not only than the the position-aware ones, but even than the best method in Table 2, which simply uses the full source documents as pseudo references. In addition, we find that the position-aware graph-based sentence extraction methods perform worse than simply extracting top sentences (Table 3). These findings indicate that the position bias remains the most effective heuristic in selecting salient information from news articles; when position information is unavailable (e.g. sentences in source documents are randomly shuffled), it might be better to use all sentences rather than selecting a subset of sentences from the source to build pseudo references.

6 Guiding Reinforcement Learning

We explore the use of different rewards to guide *Neural Temporal Difference* (NTD), a RL-based multi-document summarizer (Gao et al., 2019a). We consider three unsupervised reward functions: two baseline methods REAPER and JS (see §3 and Table 1), and the best version of SUPERT, which selects the top 10 (TAC'08) or 15 (TAC'09) sentences from each source document to build pseudo references and uses SBERT to measure the similarity between summaries and pseudo references.

In addition, we consider a non-RL-based state-of-the-art unsupervised summarizer proposed by Yogatama et al. (2015) (*YLS15*). We use ROUGE to measure the quality of the generated summaries and leave human evaluations for future work. Table 5 presents the results. We find SUPERT is the strongest reward among the considered rewards: it helps NTD perform on par with YSL15 on TAC'08 and perform significantly better on TAC'09.

7 Conclusion

We explored unsupervised multi-document summary evaluation methods, which require neither reference summaries nor human annotations. We find that vectorizing the summary and the top sentences in the source documents using contextualized embeddings, and measuring their semantic overlap with soft token alignment techniques is a simple yet effective method to rate the summary's quality. The resulting method, *SUPERT*, correlates with human ratings substantially better than the state-of-the-art unsupervised metrics.

Furthermore, we use SUPERT as rewards to train a neural-RL-based summarizer, which leads to up to 17% quality improvement (in ROUGE-2) compared to the state-of-the-art unsupervised summarizers. This result not only shows the effectiveness of SUPERT in a downstream task, but also promises a new way to train RL-based summarizers: an infinite number of summary-reward pairs can be created from infintely many documents, and their SUPERT scores can be used as rewards to train RL-based summarizers, fundamentally relieving the *data-hungriness* problem faced by existing RL-based summarization systems.

References

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3108–3118, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Yang Gao, Christian M. Meyer, and Iryna Gurevych. 2018. APRIL: interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4120–4130, Brussels, Belgium.

Yang Gao, Christian M. Meyer, and Iryna Gurevych. 2019a. Preference-based interactive multi-document summarisation. *Information Retrieval Journal*.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019b. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Tsutomu Hirao, Hidetaka Kamigaito, and Masaaki Nagata. 2018. Automatic pyramid evaluation exploiting EDU-based extractive reference summaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4186, Brussels, Belgium. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Subaspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3322–3333, Hong Kong, China. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 957–966.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv e-prints*, page arXiv:1909.11942.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Rouge and its evaluation. In *NTCIR Workshop*.

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *ACL Workshop "Text Summarization Branches Out"*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 145–152.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

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.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark.

Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgements for optimization-based summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 654–660, New Orleans, Louisiana, USA.

Avinesh P.V.S and Christian M. Meyer. 2017. Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1353–1363, Vancouver, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Cody Rioux, Sadid A. Hasan, and Yllias Chali. 2014. Fear the REAPER: A system for automatic multi-document summarization with reinforcement learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 681–690, Doha, Qatar.

Seonggi Ryang and Takeshi Abekawa. 2012. Framework of automatic text summarization using reinforcement learning. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 256–265, Jeju Island, Korea.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3244–3254, Hong Kong, China. Association for Computational Linguistics.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018a. A graph-theoretic summary evaluation for ROUGE. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 762–767, Brussels, Belgium. Association for Computational Linguistics.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018b. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 905–914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 682–687.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1216–1221, Hong Kong, China. Association for Computational Linguistics.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: pyramid evaluation via automated knowledge extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2673–2680.

Dani Yogatama, Fei Liu, and Noah A. Smith. 2015. Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966, Lisbon, Portugal. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv e-prints*, page arXiv:1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore:

Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. 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 1687–1696, New Orleans, Louisiana. Association for Computational Linguistics.