A Graph-theoretic Summary Evaluation for ROUGE

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, Fang Chen
University of New South Wales, Sydney, Australia
Data61 CSIRO, Sydney, Australia
{elahehs,mohammade,wong,fang}@cse.unsw.edu.au

Abstract

ROUGE is one of the first and most widely used evaluation metrics for text summarization. However, its assessment merely relies on surface similarities between peer and model summaries. Consequently, ROUGE is unable to fairly evaluate summaries including lexical variations and paraphrasing. We propose a graph-based approach adopted into ROUGE to evaluate summaries based on both lexical and semantic similarities. Experiment results over TAC AESOP datasets show that exploiting the lexico-semantic similarity of the words used in summaries would significantly help ROUGE correlate better with human judgments.

1 Introduction

Quantifying the quality of summaries is an important and necessary task in the field of automatic text summarization. Among the metrics proposed for this task (Hovy et al., 2006; Tratz and Hovy, 2008; Giannakopoulos et al., 2008), ROUGE (Lin, 2004) is the first and still most widely used one (Graham et al., 2015). This metric measures the concordance of system-generated (peer) summaries and human-generated reference (model) summaries by determining n-grams, word sequences, and word pair matches. ROUGE assumes that a peer summary is of high quality if it shares many words or phrases with a model summary. However, different terminology may be used to refer to the same concepts and hence relying only on lexical overlaps may underrate content quality scores. For clarity, consider the following two sentences:

(i) *They strolled around the city.*
(ii) *They took a walk to explore the town.*

These sentences are semantically similar, but lexically different. If one of them is included in a model summary, while a peer summary contains another one, ROUGE or other surface based evaluation metrics cannot capture their similarity due to the minimal lexical overlap. We aim to help ROUGE with identifying the semantic similarities of linguistic items, and consequently tackling the main problem of its bias towards lexical similarities.

Considering senses instead of words, we use the Personalized PageRank (PPR) algorithm (Haveliwala, 2002) to leverage repetitive random walks on WordNet 3.0 (Fellbaum, 1998) as a semantic network. We disambiguate each word into its intended sense, and obtain the probability distribution of each sense over all senses in the network. Weights in this distribution denote the relevance of the corresponding senses. At each iteration, we measure the semantic similarity by looking at the path taken by the random walker, and weighting the overlaps between a pair of ranked PPR vectors. Our graph-based approach (ROUGE-G) computes semantic similarity scores between n-grams, along with their match counts, to perform both semantic and lexical comparisons of peer and model summaries. The experiment results indicate that ROUGE-G variants significantly outperform their corresponding variants of ROUGE. Beyond enhancing the evaluation prowess of ROUGE, due to its lexico-semantic analysis of summaries, we believe that ROUGE-G has the potential to expand the applicability of ROUGE to abstractive summarization.

2 Background

In the summarization literature, a couple of ROUGE variants (i.e., ROUGE-1, ROUGE-2, ROUGE-SU4) are reported to have a strong correlation with human assessments, and are frequently used to evaluate summaries (Lin and Och, 2004; Owczarzak and Dang, 2011; Over and Yen, 2004). Although ROUGE is a popular evaluation metric, improving the current evaluation metrics is

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 762–767
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

still an open research area. Many of these efforts are analyzed and gathered in the surveys provided by Steinberger and Ježek (2012). Herein, we try to briefly review the most significant ones. Since DUC 2005, the Pyramid metric (Passonneau et al., 2005) was introduced as one of the principal metrics for evaluating summaries in the TAC conference. However, this metric is semi-automated and requires manual identification of summary content units (SCUs). Soon after, Hovy et al. (2006) proposed a metric based on comparison of basic syntactic units (Basic Elements) between peer and model summaries. This metric, BE-HM, was specified as one of the baselines in the TAC AESOP task. Among systems participated in this task from 2009 to 2011, AutoSummENG (DEMOKRITOSGR) (Giannakopoulos et al., 2008) is one of the top systems which compares the graph representations of peer and model summaries. Recently, some evaluation metrics have studied the effectiveness of word semantic similarity to evaluate summaries including terminology variations and paraphrasing (Baroni et al., 2014; ShafieiBavani et al., 2017, 2018). For instance, an automated variant of the Pyramid metric has used distributional semantics to map text content within peer summaries to SCUs (Passonneau et al., 2013). A more recent metric, ROUGE-WE, (Ng and Abrecht, 2015) has also enhanced ROUGE by incorporating the use of a variant of word embeddings, called word2vec (Mikolov et al., 2013).

3 Graph-Theoretic Summary Evaluation

Given a pair of peer and model summaries, we compute PPR vectors at the following levels: (i) *sense level*, to disambiguate each word (having a set of senses); and (ii) *n-gram level*, to measure the semantic similarity. We compare the PPR vectors of each pair of n-grams using the following measures: (i) *Path-based:* considering the path that the random walker takes at each iteration to get to a particular node; (ii) *Rank and Weight:* weighting the overlaps between a pair of ranked PPR vectors.

3.1 Vector Representation

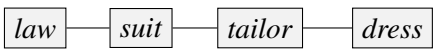
The WordNet graph has edges of various types, with the main types being hypernymy and meronymy to connect nodes containing senses. However, we do not use these types, and consider an edge as an undirected semantic or lexical relation between two synsets. We have utilized the WordNet graph enriched by connecting a sense - irrespective of its part-of-speech (POS) - with all the other senses that appear in its disambiguated gloss (Pilehvar and Navigli, 2015). Dimension of the vector representation is the number of connected nodes in the graph. For better clarity, we consider the adjacency matrix A for our semantic graph, and perform iterative random walks beginning at a set of senses S on WordNet with the probability mass of $p^{(0)}(S)$, which is uniformly distributed across the senses $s_i \in S$, and the mass for all $s_i \notin S$ set to zero. This provides a frequency or multinomial distribution over all senses in WordNet, with a higher probability assigned to senses that are frequently visited. The PPR vector of S is given by:

$$p^{(k)}(S) = dAp^{(k-1)}(S) + (1-d)p^{(0)}(S) \quad (1)$$

At each iteration, the random walker may follow one of the edges with probability d or jump back to any node $s_i \in S$ with probability $(1-d)/|S|$. Following the standard convention, the value of damping factor d is set to 0.85. The number of iterations k is also set to 20, which is sufficient for the distribution to converge.

3.2 Comparing Vectors

Conventional measures for comparing PPR vectors calculate the probability that a random walker meets a particular node after a specific number of iterations, which is potentially problematic (Rothe and Schütze, 2014). For example, consider the following connected nodes:



The PPR vectors of *suit* and *dress* have some weight on *tailor*, which is desirable. However, the PPR vector of *law* will also have a non-zero weight for *tailor*. Consequently, *law* and *dress* are spuriously similar because of the node *tailor*. To prevent this type of false similarity, the random walker needs to take into account the walking path to reach a particular node (Rothe and Schütze, 2014). We formalize this by defining the semantic similarity of two sets of nodes I and J as:

$$Sim_{sem}(I,J) = \sum_{x=0}^{k} c^x \times RW(p^{(x)}(I), p^{(x)}(J)) \quad (2)$$

where damping factor c was optimized on the TAC 2010 (Owczarzak and Dang, 2010) AESOP dataset, and set to 0.7 to ensure that early meetings are more valuable than later meetings. At each iteration x, we compare PPR vectors by ranking their dimensions (senses) based on their values, and weighting the overlaps between them (Equation 3). Hence, we weight the similarity such that differences in the highest ranks (most important senses in a vector) are penalized more than differences in lower ranks. This measure has proven to be superior to cosine similarity, Jensen-Shannon divergence, and Rank-Biased Overlap for comparing vectors (Pilehvar et al., 2013).

$$RW(Y,Z) = \begin{cases} \frac{\sum_{h \in H} (r_h(Y)+r_h(Z))^{-1}}{\sum_{i=1}^{|H|} (2i)^{-1}}, & \text{if } |H| > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where H is the intersection of all senses with non-zero probability in both vectors Y and Z. $r_h(Y)$ denotes the rank of sense h in vector Y, where rank 1 is the highest rank. The denominator is used as a normalization factor that guarantees a maximum value of one. The minimum value is zero and occurs when there is no overlap, i.e., $|H| = 0$.

3.3 Calculating ROUGE-G

We combine lexical and semantic similarities to compute ROUGE-G-N:

$$\text{ROUGE-G-N} = \frac{\sum_{M \in \{ModelSums\}} \sum_{\substack{n\text{-}gram_m \in M, \\ n\text{-}gram_p \in PeerSum}} Sim_{LS}(n\text{-}gram_m, n\text{-}gram_p)}{\sum_{M \in \{ModelSums\}} \sum_{n\text{-}gram_m \in M} Count(n\text{-}gram_m)} \quad (4)$$

where Sim_{LS} is the score of lexico-semantic similarity between a pair of n-grams in model summary ($n\text{-}gram_m$) and peer summary ($n\text{-}gram_p$):

$$\begin{aligned} &Sim_{LS}(n\text{-}gram_m, n\text{-}gram_p) = \\ &\quad \beta \times Count_{match}(n\text{-}gram_m, n\text{-}gram_p) \\ &\quad + (1-\beta) \times Sim_{sem}(n\text{-}gram_m, n\text{-}gram_p) \end{aligned} \quad (5)$$

Scaling factor β was optimized on the TAC 2010 AESOP dataset, and set to 0.5 to reach the best correlation with the manual metrics. $Count_{match}(n\text{-}gram_m, n\text{-}gram_p)$ is the maximum number of the $n\text{-}gram$ co-occurring in a peer summary and a set of model summaries.

3.4 Disambiguation of n-grams

Prior to measuring semantic similarities, each word in n-grams has to be analyzed and disambiguated into its intended sense. However, conventional word sense disambiguations are not applicable due to the lack of contextual information. Hence, we seek the semantic alignment that maximizes the similarity of the senses of the compared words. As an example (Pilehvar et al., 2013), consider two sentences of *"a1. Officers fired."* and *"a2. Several policemen terminated in corruption probe."*, the semantic alignment procedure has been performed as "P_{a1}. $officer_n^3, fire_v^4$", *and* "P_{a2}. $policeman_n^1, terminate_v^4, corruption_n^6, probe_n^1$". t_p^i denotes the i-th sense of a word t in WordNet with POS p. After alignment, among all possible pairings of all senses of $fire_v$ to all senses of all words in *a2*, the sense $fire_v^4$ (employment termination) obtains the maximal similarity value of $Sim_{sem}(fire_v^4, terminate_v^4) = 1$.

3.5 OOV Handling

Out-of-vocabulary (OOV) words are the words that are not defined in the corresponding lexical resource, hence, they will be ignored while generating PPR vectors. The reason is that they do not have an associated node in the WordNet graph for the random walk to be initialized from. To take them into consideration, we add an extra dimension for each OOV term in the resulting PPR vector. Following Pilehvar and Navigli (2015), we set the associated weights of the new dimensions to 0.5 so as to guarantee their placement among the top dimensions in their vectors.

4 Experiments

4.1 Data and Metrics

The only available datasets for the task of Summarization Evaluation are three AESOP datasets[1] provided by TAC 2009, 2010, and 2011. Among them, we optimize scaling factors using the TAC 2010 AESOP dataset, and evaluate ROUGE-G on the TAC 2011 (Owczarzak and Dang, 2011) AESOP dataset for two main reasons: (i) it is the only dataset on which evaluation metrics can be assessed for their ability to measure summary Readability; (ii) To be in line with the most recent work (ROUGE-WE) that has also been evaluated

[1] https://tac.nist.gov/data/

only on this dataset for measuring the Readability scores. This dataset consists of 44 topics, and a set of 10 documents for each topic. There are four human-crafted model summaries for each document set. A summary for each topic is generated by each of the 51 summarizers which participated in the main TAC summarization task. The output of participating automatic metrics is tasked to be compared against human judgments using three manual metrics of *Pyramid*, *Readability*, and *Responsiveness*. Hence, the outputs are scored based on their summary content, linguistic quality, and a combination of both, respectively.

Prior to computing correlation of ROUGE-G variants with manual metrics, ROUGE-G scores have reliably been computed (95% confidence intervals) under ROUGE bootstrap resampling with the default number of sampling point (1000). Correlation of ROUGE-G evaluation scores with the human judgments is then assessed with three metrics of correlation: Pearson r; Spearman ρ; and Kendall τ. We compute scores using the default NIST settings for baselines in the TAC 2011 AESOP task (with stemming and keeping stopwords).

4.2 Results

We evaluate ROUGE-G, against the top metrics (C_S_IIITH3, DemokritosGR1, Catolicasc1) among the 23 metrics participated in TAC AESOP 2011, ROUGE, and the most recent related work (ROUGE-WE) (Table 1). Overall results support our proposal to consider semantics besides surface with ROUGE. Since the large/small differences in competing correlations with human assessment are not an acceptable proof of superiority/inferiority in performance of one metric over another, significance tests should be applied. To better clarify the effectiveness of ROUGE-G, we have used pairwise Williams significance test recommended by Graham et al. (2015) for summarization evaluation. Accordingly, evaluation of a given summarization metric, M_{new}, takes the form of quantifying three correlations: $r(M_{new}, H)$, that exists between the evaluation metric scores for summarization systems and corresponding human assessment scores; $r(M_{base}, H)$, that stands for the correlation of baseline metrics with human judges; and the third correlation, between evaluation metric scores themselves, $r(M_{base}, M_{new})$. It can happen for a pair of competing metrics for which the correlation between metric scores is strong, that a small difference in competing correlations with human assessment is significant, while, for a different pair of metrics with a larger difference in correlation, the difference is not significant (Graham et al., 2015). Using this significance test, the results show that all increases in correlations of ROUGE-G compared to ROUGE and ROUGE-WE variants are statistically significant ($p < 0.05$). We analyze the correlation results reported in Table 1 in the following.

ROUGE-G-2 achieves the best correlation with Pyramid, regarding all correlation metrics. Moreover, every ROUGE-G variant outperforms its corresponding ROUGE and ROUGE-WE variants, regardless of the correlation metric used. However, the only exception is ROUGE-SU4, which correlates slightly better with Pyramid when measuring with Pearson correlation. One possible reason is that Pyramid measures content similarity between peer and model summaries, while the variants of ROUGE-G favor semantics behind the content for measuring similarities. Since some of the semantics attached to the skipped words are lost in the construction of skip-bigrams, ROUGE-SU4 shows a better correlation comparing to ROUGE-G-SU4.

For Responsiveness, ROUGE-G-SU4 achieves the best correlation when measuring with Pearson. We also observe that ROUGE-G-2 obtains the best correlation with Responsiveness while measuring with the Spearman and Kendall rank correlations. The reason is that semantic interpretation of bigrams is easier, and that of contiguous bigrams is much more precise. We also see that every variant of ROUGE-G outperforms its corresponding ROUGE and ROUGE-WE variants.

The readability score is based on grammaticality, structure, and coherence. Although our main goal is not to improve the readability, ROUGE-G-SU4 and ROUGE-G-2 are observed to correlate very well with this metric when measured with the Pearson and Spearman/Kendall rank correlations, respectively. Besides, every variant of ROUGE-G represents the best correlation results comparing to its corresponding variants of ROUGE and ROUGE-WE for all correlation metrics. This is likely due to considering word types and POS tagging while aligning and disambiguating n-grams. POS features are shown by Feng et al. (2010) to be helpful in predicting linguistic quality.

We optimize scaling factor β (Equation 5) on the TAC 2010 AESOP dataset. Figure 1 shows

	Pyramid			Responsiveness			Readability		
Metric	**Pearson**	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**
C_S_IIITH3	0.965	0.903	0.758	0.933	0.781	0.596	0.731	0.358	0.242
DemokritosGR1	0.974	0.897	0.747	0.947	0.845	0.675	0.794	0.497	0.359
Catolicasc1	0.967	0.902	0.735	0.950	0.837	0.666	0.819	0.494	0.366
ROUGE-1	0.966	0.909	0.747	0.935	0.818	0.633	0.790	0.391	0.285
ROUGE-2	0.961	0.894	0.745	0.942	0.790	0.610	0.752	0.398	0.293
ROUGE-SU4	0.981	0.894	0.737	0.955	0.790	0.602	0.784	0.395	0.293
ROUGE-WE-1	0.949	0.914	0.753	0.916	0.819	0.631	0.785	0.431	0.322
ROUGE-WE-2	0.977	0.898	0.744	0.953	0.797	0.615	0.782	0.414	0.304
ROUGE-WE-SU4	0.978	0.881	0.720	0.954	0.787	0.597	0.793	0.407	0.302
ROUGE-G-1	0.971	0.915	0.758	0.944	0.825	0.638	0.791	0.434	0.330
ROUGE-G-2	**0.983**	**0.926**	**0.774**	0.956	**0.869**	**0.713**	0.790	**0.516**	**0.385**
ROUGE-G-SU4	0.979	0.898	0.741	**0.957**	0.814	0.616	**0.823**	0.445	0.334

Table 1: Correlation results with the manual metrics of Pyramid, Responsiveness, and Readability using the correlation metrics of Pearson r, Spearman ρ, and Kendall τ. The best correlations are specified in bold, and the underlined scores show the top correlations in the TAC AESOP 2011.

the correlation results by the variants of ROUGE-G with Pyramid (Pyr) and Responsiveness (Rsp) metrics measured by Pearson. The best results are observed when $\beta = 0.5$. Performance deteriorates when β approaches 1.0 which indicates the ROUGE scores without any touch of semantic similarity. Decreasing β to zero causes the exclusion of lexical match counts, and consequently inappropriateness of the outcomes. This shows the importance of using both lexical and semantic similarities to fairly judge the quality of summaries.

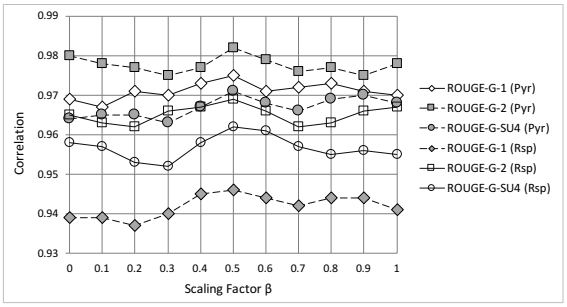


Figure 1: Exploring scaling factor β

It is noteworthy that we have evaluated our approach with the following settings for computing and comparing PPR vectors: (i) *Path-based* with *Rank and Weight* measure (current setting); (ii) *Path-based* with *cosine similarity*; (iii) Excluding *path-based* measure and using *Rank and Weight* measure solely. The results showed that the current setting performs better than the other two.

5 Conclusion

This paper presents ROUGE-G to overcome the limitation of high lexical dependency in ROUGE. Our approach leverages a sense-based representation to calculate PPR vectors for n-grams. The semantic similarity of n-grams are then computed using a formalization of *Path-based* and *Rank and Weight* measures. We finally improve on ROUGE by performing both semantic and lexical analysis of summaries. Experiments over the TAC AESOP datasets demonstrate that ROUGE-G achieves higher correlations with manual judgments in comparison with ROUGE.

In order to demonstrate the effectiveness of ROUGE-G to fairly evaluate abstractive summaries, we need to conduct experiments on a dataset composed of abstractive summaries. However, we evaluated our approach on the TAC 2011 AESOP dataset, which is made of summaries that were generated mostly by extractive systems. Since there is not such dataset at the time of writing this paper, we can continue building on this work by using model summaries, which are abstractive in nature, as a proxy. Thereupon, it is possible to incorporate *jackknifing* procedure in the scoring process in order to see whether our metric can differentiate between peer summaries (naturally extractive) vs. model summaries (naturally abstractive).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. The first author was supported by the "Australian Government Research Training Program Scholarship".

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 238–247.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING 2010)*, pages 276–284. ACL.

George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):1–5.

Yvette Graham et al. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 128–137. ACL.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web (WWW 2002)*, pages 517–526. ACM.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. volume 8. ACL.

Chin-Yew Lin and FJ Och. 2004. Looking for a few good metrics: Rouge and its evaluation. In *Proceedings of the 4th NII Testbeds and Community for Information Access Research: Workshops (NTCIR 2004)*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2013)*, volume 13, pages 746–751. ACL.

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 (EMNLP 2015)*, pages 1925–1930. ACL.

Paul Over and J Yen. 2004. An introduction to duc 2004 intrinsic evaluation of generic new text summarization systems, 2004. *National Institute of Standards and Technology (NIST)*.

Karolina Owczarzak and Hoa Trang Dang. 2010. Overview of the tac 2010 summarization track. In *Proceedings of the 3rd Text Analysis Conference (TAC 2010)*.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*.

Rebecca J Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL 2013)*. ACL.

Rebecca J Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the 2005 Document Understanding Conference (DUC 2005)*.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1341–1351. ACL.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Sascha Rothe and Hinrich Schütze. 2014. Cosimrank: A flexible & efficient graph-theoretic similarity measure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1392–1402.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2017. A semantically motivated approach to compute ROUGE scores. *arXiv preprint arXiv:1710.07441*.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. 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 (COLING 2018)*, pages 905–914. ACL.

Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Stephen Tratz and Eduard Hovy. 2008. Bewte: basic elements with transformations for evaluation. In *Proceedings of the 1st Text Analysis Conference (TAC 2008)*.