Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6656–6662
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Language Agnostic Automatic Summarization Evaluation

Christopher Tauchmann*, Margot Mieskes‡
*Artificial Intelligence and Machine Learning Lab, ‡Information Science,
*Technische Universität Darmstadt, ‡Darmstadt University of Applied Sciences
*tauchmann@cs.tu-darmstadt.de, ‡margot.mieskes@h-da.de

Abstract

So far work on automatic summarization has dealt primarily with English data. Accordingly, evaluation methods were primarily developed with this language in mind. In our work, we present experiments of adapting available evaluation methods such as ROUGE and PYRAMID to non-English data. We base our experiments on various English and non-English homogeneous benchmark data sets as well as a non-English heterogeneous data set. Our results indicate that ROUGE can indeed be adapted to non-English data – both homogeneous and heterogeneous. Using a recent implementation of performing an automatic PYRAMID evaluation, we also show its adaptablilty to non-English data.

Keywords: summarization evaluation, ROUGE, PYRAMID, automatic evaluation

1. Introduction

Research in automatic summarization has long focused on homogeneous, English data, i.e. in the context of the Document Understanding Conferences (DUC) and Text Analysis Conferences (TAC), which resulted in a range of benchmark data sets.[1] In recent years this has expanded to include non-English data (e.g. the MultiLing Shared Task[2] as well as heterogeneous sources (e.g. hMDS and auto-hMDS (Zopf et al., 2016b; Zopf, 2018) or even both (e.g. DBS (Benikova et al., 2016). But evaluation methods are still very strongly focused on the English language. ROUGE (Lin, 2004) for example is based on WordNet and an English stop word list. Using the vast parameter space ROUGE offers, the language specific settings can be switched off or moved offline (Mieskes and Padó, 2019). Similarly, metrics such as the Jensen-Shannon-Divergence or Kullback-Leibler Divergence, which have been used for the evaluation of summarization (Nenkova and Louis, 2008; Zopf et al., 2016b), can be more easily adapted to the target language if semantic resources are available. The PYRAMID method (Nenkova and Passonneau, 2004) was originally developed as a manual evaluation method. Studies show though that if the whole process is automated, the correlation to human judgments drops considerably, therefore keeping the need for high-effort manual annotations (Peyrard and Eckle-Kohler, 2017).

Recently, Gao et al. (2019) presented a method for calculating PYRAMID scores using embeddings called PyrEval, reducing the need for task-specific resources and enabling content-based evaluation of summaries. Considering the easier availability of embeddings for a given language, this would allow for evaluating summaries across a range of languages independent of time-consuming manual annotations.

In our work, we present experiments using embeddings-based PYRAMID evaluations on non-English and heterogeneous data as well as language-independent ROUGE-scores. We aim at answering various research questions with regards to using existing evaluation metrics such as ROUGE and PYRAMID on this data. Our contributions are therefore as follows:

- A new benchmark data set of manual PYRAMID annotations on a German heterogeneous summarization corpus.
- An analysis of the evaluation quality using ROUGE in a language-agnostic way.
- An application of an embeddings-based automatic PYRAMID evaluation method on heterogeneous and/or non-English data.
- Analyzing the evaluation quality in comparison to language-agnostic ROUGE scores and manual PYRAMID annotations, but also in comparison to the language-dependent ROUGE scores.

2. Related Work

There is a vast body of work on automatically and manually evaluating summaries. The most commonly used evaluation method for summarization is ROUGE, which was introduced by Lin and Hovy (2003). As it showed a high correlation with manual evaluation it was quickly adopted for the DUC series and became the de-facto standard evaluation method. ROUGE relies on counting n-grams and calculating Precision, Recall and F-measure by comparing one or several system summaries to one or several manual summaries.

Due to the focus on n-grams it is, however, not able to effectively judge semantically similar summaries. Therefore, Nenkova and Passonneau (2004) developed the PYRAMID method, which was introduced originally as a manual evaluation method. Here, sentences in summaries are split into Summary Content Units for both system and manual summaries and compared based on content. Using a weighting method the final PYRAMID score is calculated. This allows to capture semantically similar parts of a summary.

As this is still fairly time-consuming and similar to ROUGE requires manually written summaries, thus adding to the efforts required for this method, options to evaluate summaries without reference summaries were explored

[1] `https://duc.nist.gov/,https://tac.nist.gov`

[2] http://multiling.iit.demokritos.gr (Giannakopoulos, 2013)

(Nenkova and Louis, 2008). Similar to methods proposed in the AESOP task from 2009 to 2011 (Rankel et al., 2013) none of these methods gained wide-spread use, despite various studies showing problems with ROUGE (see for example (Graham, 2015)).

Ng and Abrecht (2015) combined ROUGE scores with word-sense embeddings to improve evaluation of abstractive summarization but they still require reference summaries. ShafieiBavani et al. (2018) developed an automatic embedding-based method without the need for reference summaries. However, their method offers little information on content semantic, little traceability and is only complementary to existing approaches, as the authors note.

In recent years, methods to partially or fully automate PYRAMID have been developed. Passonneau et al. (2013) used manual pyramids to automatically score summaries. Results on automatic PYRAMID construction were not convincing and/or computationally expensive (Peyrard and Eckle-Kohler, 2017; Yang et al., 2016). Very recently, a more efficient method has been proposed which is based on embeddings (Gao et al., 2019).

All of the methods presented so far have in common that they have primarily been developed and used on English data, while few attempts have been made to evaluate non-English data. In the context of the MultiLing task, a method was presented, but again did not gain wide-spread use (Giannakopoulos and Karkaletsis, 2011).

3. Data

We evaluate both language-agnostic ROUGE and PyrEval on data sets in several languages from two sources. MultiLing (Giannakopoulos et al., 2015) provides multilingual multi-document summarization data sets based on news articles in various languages; our evaluation is performed on English, Spanish and in the experiments on language agnostic ROUGE also on French. An analysis of language-agnostic ROUGE, PyrEval and a discussion of results of a manual PYRAMID annotation is carried out on a small, heterogeneous German summarization evaluation data set based on Benikova et al. (2016).

3.1. DUC

In order to evaluate the effect of removing language specific parameters when using ROUGE, we also use two benchmark data sets from the Document Understanding Conference (DUC), namely DUC 2002 and DUC 2004. The data sets consist of 60 and 50 document collections of 10 documents each. They contain up to four manual summaries per topic cluster as well as automatic summaries submitted at the time. This enables us to quantify the effect of the language specific settings in ROUGE to see whether the difference is significant or not.

3.2. DBS-eval

We extend a heterogeneous data set, DBS, published in Benikova et al. (2016), which contains topically clustered document sets from the educational domain. Topics contain four to 16 documents per cluster; summaries have been created by three or four expert annotators from the field of computational linguistics and are slightly longer than the longest DUC/TAC summaries (approx. 500 words/summary). We provide manual PYRAMID annotations by three expert annotators from the field of computational linguistics on top of DBS and make this heterogeneous evaluation data set in German (henceforth called DBS-eval) available to the research community.[3]

The PYRAMID annotations are compared to ROUGE scores on manual and automatic summaries for this corpus. An analysis of several scores to judge summary quality, such as Jensen-Shannon-Divergence (JS), on both manual and automatic summaries as well as the source documents, gives an overview of the textual quality of DBS-eval. To show the differences to traditional MDS corpora, the experiments on DBS-eval are compared to established benchmark data sets such as those published in the DUC-context.

Table 1 shows a comparison of the mean SCU weight (Pyr, ranging from 1 to $n\_manual\_summaries$) in PYRAMID annotations with several scores. Pyr is compared to a Recall equivalent from peer PYRAMID annotation scores (Pyr auto, ranging from 0 to 1) on automatic summaries, as well as Jensen-Shannon-Divergence (JS, ranging from 0 to 1) and Shannon Entropy scores (ranging from 1 to $\log_2(n\_word\_types)$) on manual summaries. Furthermore, table 1 lists ROUGE-1 and ROUGE-2 Recall scores (ranging from 0 to 1) on manual and automatic summaries for this corpus. The Pearson's correlation of the scores in this table is 0.74^* between PYRAMID and manual peer annotations (Pyr auto), 0.65^* between ROUGE-2 on manual summaries and -0.75^* between PYRAMID and Shannon Entropy. JS correlates with -0.21 and all other scores correlate between 0.5 and 0.6^+.

Table 2 shows JS scores between reference summaries and source documents. As we contrast JS scores with two DUC data sets and also the complete MultiLing data set, differences between DBS-eval and both data sets on news become apparent, which we attribute to the heterogeneity of DBS-eval: DBS-eval shows lower JS.[4] Furthermore, JS scores in DBS-eval vary considerably between individual topics.

Qualitative analysis of PYRAMID evaluation

Information content in PYRAMID annotations can be measured by Summary Content Units (SCUs), where the weights of individual SCUs correspond to the importance of the content they carry (Nenkova and Passonneau, 2004). The maximum number of SCUs corresponds to the number of manual summaries in the PYRAMID annotation.

The mean SCU weight scores (Pyr) in table 1 show that the summaries in DBS-eval contain a lot of SCUs with low weights, which is usual for PYRAMID annotations. For example, topic five has 192 SCUs with weight one that occur in one of three summaries. Only 23 SCUs are contained in two summaries and merely three SCUs are present in all three summaries. We observe that topic five tackles a lot

[3] `https://github.com/ml-research/DBS-eval`

*Correlation is statistically significant below $\alpha = 0.05$.

$^+$Correlation is not statistically significant.

[4] DBS-eval also shows higher Shannon Entropy and higher text dissimilarity – tables are not included but can be added in the camera ready version

Topic	Pyr	R-1 auto	R-2 auto	R-1 man	R-2 man	Pyr auto	JS man	Ent man
1	1.21	0.48	0.20	0.45	0.19	0.32	0.28	6.41
2	1.32	0.51	0.30	0.49	0.22	0.43	0.26	5.82
3	1.2	0.38	0.10	0.46	0.23	0.19	0.29	6.48
4	1.46	0.40	0.15	0.42	0.21	0.35	0.33	6.42
5	1.13	0.35	0.09	0.39	0.12	0.19	0.31	6.79
6	1.27	0.26	0.07	0.36	0.09	0.23	0.24	6.66
7	1.38	0.28	0.05	0.42	0.21	0.32	0.25	5.57
8	1.51	0.63	0.39	0.58	0.36	0.51	0.18	5.38
9	1.46	0.49	0.23	0.55	0.35	0.33	0.27	5.40
10	1.43	0.57	0.22	0.42	0.17	0.33	0.35	6.09
DBS-eval	1.34	0.47	0.2	0.44	0.21	0.32	0.27	5.15

Table 1: Comparison of manual and automatic summaries in DBS-eval per topic (auto created with LSA, LexRank, TexRank, Luhn, Edmundson as presented in (Benikova et al., 2016).

of content that is in itself unrelated to other content in this cluster, which is a pattern also typical for the PYRAMID annotation (see (Passonneau et al., 2006)). For example, in this topic there are many different statements of motivations towards violence as well as psychological treatments, preventive measures, symptoms of stress, role plays, concepts from pedagogy as well as concepts from sociology and psychology and reports of actual crimes. However, there are differences between the individual topics: Topic eight has 55 SCUs with weight one and 27 with weight two, six with weight three and still two with weight four. Half of the information is present in at least two summaries. In contrast, topic eight is focused on providing help for individuals from industry seeking to become teachers and report laws and the procedures to follow in order to apply for candidacy. The scope of topic eight is much smaller and the content is related on a much lower level than in topic five. The observations on mean SCU weight are reflected in the other scores we report and all values show the same pattern: Topics with a low PYRAMID score have a high Entropy value (Pearson's correlation of -0.75**) and frequently also a high JS. The topic with the highest mean SCU weight (topic eight) shows the lowest Entropy within all of the ten topics as well as the lowest JS. We see the highest Pyr auto score as well as the highest scores for ROUGE man and ROUGE auto.

Data	JS mean
DUC02	0.34
DUC04	0.38
MultiLing –EN	0.36
DBS-eval	0.27

Table 2: Jensen-Shannon-Divergence between corpora and reference summaries.

3.3. MultiLing

The data sets for each language in the MultiLing corpus consist of 15 topics extracted from WikiNews with three manual reference summaries and on average 12 system summaries from 240 to 250 words. System summaries come from five systems for Spanish and from seven systems seven for French. The data set also contains document sets for other languages, like Romanian, Czech and Hebrew.

4. Experiments

In the following we present results on our experiments using English and non-English data and automatic evaluation methods based on ROUGE and PyrEval.

4.1. ROUGE

Table 3 shows the results of experimenting with the language specific parameters in ROUGE[5]. We report ROUGE-1 Recall scores. While the "full" parameters keep both stopword filtering and stemming, the language independent setting (langIndep) uses neither, as resources such as WordNet are not easily obtained for most languages. Stopword lists might be easier to get, so we also compare results using only stopword filtering (noStem). Our results indicate that there is indeed no significant difference between using the full feature set as opposed to removing the language dependent components. As stopwords are not filtered in this version, ROUGE scores are higher as they are also counted towards the final score. We also use the English portion of the MultiLing data set for these experiments. We observe that the values change similarly to the DUC data sets, with the stopwords having the largest impact. Therefore, we conclude that having a stopword list in the target language might be beneficial to the results, when using ROUGE as an evaluation metric.

Overall, we conclude that ROUGE can be used for non-English data if manual summaries are available. It is necessary to either do the language dependent steps offline (i.e., outside of ROUGE), if the necessary resources are available, or by removing the language dependent parameters.

Table 4 shows the results for the MultiLing Spanish and French data and for the German DBS-eval data set. While the results for DBS-eval are considerably lower than for French and Spanish they nevertheless give reasonable results, which can be used to judge a summary quality. We observe that the ROUGE results show a higher variance between the various topics, which indicates that some document collections are harder to summarize and/or evaluate than others.

[5] We used standard parameters for these experiments: `n 4 -s -c 95 -r 1000 -f -A -p 0.5 -t 0 -w 1.2 -2 -4 -l 100 -a`. For language independent experiments we dropped the `m` parameter for stemming and the `s` parameter for stopword filtering and adapted the length parameter if necessary.

Data	full	noStem	noStop	langIndep
DUC 2002	0.34	**0.32**	**0.40**	**0.38**
DUC 2004	0.24	0.21	0.33	0.31
MultiLing	0.34	0.31	0.40	0.39

Table 3: Results for ROUGE evaluation with and without language dependent settings on two DUC data sets and the English, Spanish and French portions of the MultiLing data set. Bold figures indicate no statistical difference to the standard parameter settings. Due to the small sample size, statistical significance cannot be determined on the MultiLing data set.

Data	MultiLing Spanish	MultiLing French	DBS-eval
	0.51	0.46	0.36

Table 4: ROUGE with no language dependent settings for non-English data.

4.2. PyrEval

The advantage of PyrEval compared to ROUGE is that it provides semantic information about the summaries. The PyrEval architecture offers several parameters: First, the segmentation after parsing and tagging can be altered.[6] Segments can be an entire sentence or more fine-grained units, such as verb phrases, which can be used as segments for further processing in the set partition algorithm EDUA (Emergent Discovery of Units of Attraction). EDUA creates a content model from the vector representation of the segments. Second, word embeddings from different sources can be used. To illustrate differences in performance, we experiment with two types of embeddings: First, multilingual fastText embeddings which are pretrained on Wikipedia and aligned in a single vector space and consist of a vocabulary of 200,000 words (Conneau et al., 2017), which we refer to as *emb_200k* for the remainder of this study. Second, Bojanowski et al. (2017) provide multilingual fastText embeddings with a larger vocabulary size of 2 million words and vectors for each language in a single vector space, which we refer to as *emb_2m* for the remainder of this study and which are similarly pretrained on Wikipedia. To the best of our knowledge fastText embeddings are so far the only embeddings which are pretrained consistently on one text source for different languages. The Wikipedia corpus provides a clean textual data source with encyclopedic texts of linguistic quality similar to our data sets. Furthermore, Wikipedia is multilingual, and therefore allow pretraining embeddings for various languages. In order to fully employ the lightweight approach of the PyrEval architecture we argue for a setup, which keeps computation time limited and requires little expert knowledge in language. Of course, computation time and performance should be well balanced.

Gao et al. (2019) show that PyrEval performs well on an English benchmark data set (TAC 2010) and also a recent data set of English student summaries for technological topics. Therefore, this work focuses on an evaluation on non-English data. We report quality scores (i.e. Precision) as well as coverage scores (i.e. Recall) and compare them with ROUGE-Precision and ROUGE-Recall on these data sets.

4.2.1. PyrEval on DBS-eval

We measure the performance of PyrEval on three automatic summaries from the original DBS corpus as well as one newer system created by Zopf et al. (2016a) for the manual PYRAMID evaluation with PyrEval's quality scoring function in the way that Gao et al. (2019) propose: We measure Pearson's correlation between the scores from automatic summaries evaluated on the manually created pyramids and those obtained from evaluating them with the automatic pyramids created by PyrEval. PyrEval is evaluated on DBS-eval with the two different embeddings described above. In the manual PYRAMID annotation of DBS-eval, segments frequently are entire sentences. Therefore, the entire sentence from the output of the parser is passed into the EDUA algorithm as one segment. Table 5 shows the quality scores on automatic pyramids and manual pyramids. The scores with *emb_2m* are slightly higher than those with *emb_200k*. Using *emb_200k*, a Pearson's correlation of 0.75** indicates that indeed PyrEval is capable of producing pyramids of similar quality as our German annotators. When we use *emb_2m*, computation time rises by a multiple of 10 per topic on average for the entire PyrEval pipeline but we reach a higher Pearson's correlation of 0.84**.

In general, the scores on the manual pyramids are higher than on automatic pyramids as the average scores in table 5 show. However, the high Pearson's correlation between quality scores on manual and automatic pyramids, especially when we use *emb_2m*, leads us to argue that this could be an issue of coverage. The coverage scores tell us that the automatic summaries achieve a Recall of 0.32 of the content of manual pyramids with *emb_2m*; with *emb_200k* this number drops to 0.28. In the automatic pyramids, automatic summaries achieve a Recall of 0.26 of the content with *emb_2m* and only 0.18 with *emb_200k*. As we see, manual pyramids achieve a better coverage, especially on the vectors with the smaller vocabulary. These coverage scores across the data set, especially on *emb_2m*, reflect the outcome of the experiments on language agnostic ROUGE-1 Recall in section 4.1. Despite some differences in the distribution over the topics, the coverage scores are similar to the Pyr auto scores in table 1. Half of the systems that create automatic summaries are ranked equally in an evaluation with PyrEval and ROUGE-1 Precision with no language specific parameters.[7]

4.2.2. PyrEval on Spanish

We use summaries from the Spanish portion of the MulitLing data from five systems and the evaluation setup with *emb_200k*. Table 6 shows the results for each of the systems. No system scores higher than 0.25. The PyrEval quality scores are rather low compared to those on DBS-eval and so are the ROUGE scores reported by Giannakopoulos et al. (2015) (ROUGE-1 Precision

[6] We use the Stanford CoreNLP parser from Manning et al. (2014)

**Correlation is statistically significant below $\alpha = 0.005$.

[7] Correlation could not be calculated due to the small sample size.

	automatic Pyramids		manual Pyramids		
System	DE-auto-*emb_200k*	DE-auto-*emb_2m*	DE-auto-*emb_200k*	DE-auto-*emb_2m*	R-1 p
LexRank	0.16	0.23	0.45	0.45	0.32
Lsa	0.28	0.28	0.49	0.59	0.36
TexRank	0.15	0.18	0.36	0.38	0.32
MZ	0.23	0.27	0.47	0.51	0.26

Table 5: The average PyrEval quality score per system on German pyramids.

System	ES-auto-*emb_200k*
MMS12	0.10
MMS2	0.25
MMS3	0.25
MMS5	0.20
MMS8	0.11

Table 6: The average PyrEval quality score per system on Spanish automatic pyramids

0.22, ROUGE-1 Recall 0.25 and ROUGE-2 Precision 0.08, ROUGE-2 Recall 0.03 on average over all summaries). This comparison would indicate poor summary quality. However, there were challenges in the ROUGE evaluation[8] and the system summaries were also evaluated with MeMoG, an n-gram graph method, which correlates well with PYRAMID scores and which ranges from zero to one. Some variations in performance can occur which depend on summary quality (Giannakopoulos and Karkaletsis, 2011). The average MeMoG score over all systems is 0.21 and it is similar to the average PyrEval quality score of 0.18.

The Pearson's correlation between the five system summaries (we use one summary per system, even when systems provide multiple summaries) and ROUGE-1 and ROUGE-2 Precision respectively is 0.65^+ and 0.68^+. The Pearson's correlation with MeMoG is 0.47^+.The PyrEval coverage score is 0.16 and slightly lower than on DBS-eval with *emb_200k*. This score corresponds to low ROUGE Recall scores Giannakopoulos et al. (2015) report but is not in line with our language agnostic evaluation in section 4.1.

The findings are not as convincing as those on DBS-eval and we do not have manual pyramids for this data set to evaluate. As the quality of the pyramids that PyrEval produces is not entirely clear at this point, we must take into account that the embedding-word-coverage on *emb_200k* might also be insufficient to cover all information in the documents or it might not be sufficient to take entire sentences as segments. Problems could also be in the quality of the output of the Spanish parser.

5. Conclusion

In this paper we performed experiments to automatically measure summary quality based on English as well as non-English data sets. The results on all evaluation methods indicate that the methods can be used for non-English data – both for homogeneous, as well as heterogeneous data sets. What we observe though is that the performance is not uniform across languages and across document sets. Our analysis of the German data set reveals that ROUGE as well as PYRAMID and the Jensen-Shannon-Divergence show considerable differences between document sets. The correlation between PyrEval scores on automatically and manually constructed pyramids shows that it can be considered a reliable indicator of summary quality.

ROUGE Recall scores with no language settings on the Spanish Mutliling data set are higher than on the German benchmark corpus DBS-eval while the results on PyrEval scores suggest a better quality of German summaries over Spanish summaries. But compared to the English benchmark data, we observe that the differences are in line with using ROUGE on English data with no language specific settings.

The PyrEval evaluation on German shows a high correlation between manual and automatic pyramids and automatic summaries, especially on a larger embedding vocabulary. PyrEval quality scores show comparable results to ROUGE Precision scores in half of the system summaries while a comparison with our ROUGE Recall scores indicates that PyrEval captures a similar amount of content in automatic summaries from four systems. The quality and coverage scores of Spanish MultiLing automatic summaries in automatic pyramids reflect the ROUGE Precision and Recall scores reported by the MultiLing authors, whereas the coverage is considerably lower than our language independent ROUGE Recall score would suggest. As there were reportedly problems with the ROUGE evaluation in MultiLing, we aim to further investigate this outcome with variations of the PyrEval setup.

When comparing the average results of the ROUGE language parameters on the German data to the DUC data, we see that the results are comparable to the standard data sets. This allows the conclusion that the methods yield comparable results on non-English data. Differences in the results need to be examined further but are most likely due to differences in the languages. Results on the automatic PyrEval method indicate that this method is also valid on non-English data, reducing the need for the time-consuming manual PYRAMID annotation.

Lastly, using the Jensen-Shannon-Divergence (JS) allows for evaluating automatic summaries without manual reference summaries, which is in itself also very time-consuming. JS is higher on the heterogeneous DBS-eval data set than on the MutliLing and DUC data sets. Scores on individual topics in DBS-eval vary considerably.

[8] In a forum discussion on the MultiLing task it was noted that no adaptations were made to deal with language specific issues, which might explain the difference to our results.

[+] Correlation is not statistically significant.

5.1. Future Work

The next steps involve the verification of our results on other languages, which are for example available in the MultiLing data. Additionally, so-called Excellent Articles from Wikipedia have been used for automatic summarization and allow us to verify our results on larger data sets and other languages as well. As we only used sentences as segmentation unit so far, using a more fine-grained segmentation method might improve results further. But it has to be taken into account that such fine-grained segmentation might not be applicable for a wide range of languages.

6. Acknowledgements

We would like to thank Yanjun Gao and Rebecca Passonneau for providing the PyrEval code as well as kindly assisting with related questions.

This work has been supported by the research center for Digital Communication and Media Innovation (DKMI) and the Institute for Communication and Media (IKUM) at the University of Applied Sciences Darmstadt. Part of this research further received support from the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

7. Bibliographical References

Benikova, D., Mieskes, M., Meyer, C. M., and Gurevych, I. (2016). Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 1039–1050. Association for Computational Linguistics, December.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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

Giannakopoulos, G. and Karkaletsis, V. (2011). AutoSummENG and MeMoG in evaluating guided summaries. In *Proceedings of the Text Analysis Conference, 2011*.

Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., and Poesio, M. (2015). MultiLing 2015: Multilingual Summarization of Single and Multi-Documents, On-line Fora, and Call-center Conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic, September. Association for Computational Linguistics.

Giannakopoulos, G. (2013). Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August. Association for Computational Linguistics.

Graham, Y. (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*, pages 128–137, Lisbon, Portugal, September. Association for Computational Linguistics.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Association of Computational Linguistics: Text Summarization Workshop*, pages 74–81.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mieskes, M. and Padó, U. (2019). Summarization Evaluation meets Short-Answer Grading. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 79–85, Turku, Finland. LiU Electronic Press.

Nenkova, A. and Louis, A. (2008). Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics 2008*, pages 825—-833.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics 2004*, pages 145–152.

Ng, J.-P. and Abrecht, V. (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, September. Association for Computational Linguistics.

Passonneau, R. J., McKeown, K., Sigelman, S., and Goodkind, A. (2006). Applying the Pyramid Method in the 2006 Document Understanding Conference. pages 1–8.

Passonneau, R. J., Chen, E., Guo, W., and Perin, D. (2013). Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–147, Sofia, Bulgaria, August. Association for Computational Linguistics.

Peyrard, M. and Eckle-Kohler, J. (2017). Supervised learning of automatic pyramid for optimization-based multi-document summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1084–1094, Vancouver, Canada, July. Association for Computational Linguistics.

Rankel, P. A., Conroy, J. M., Dang, H. T., and Nenkova, A. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria, August. Association for Computational Linguistics.

ShafieiBavani, E., Ebrahimi, M., Wong, R., and Chen, F. (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*, pages 905–914, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Yang, Q., Passonneau, R., and De Melo, G. (2016). Peak: Pyramid evaluation via automated knowledge extraction. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 30th AAAI Conference on Artificial Intelligence, AAAI 2016, pages 2673–2679. AAAI press, 1.

Zopf, M., Loza Mencía, E., and Fürnkranz, J. (2016a). Beyond centrality and structural features: Learning information importance for text summarization. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 84–94, Berlin, Germany, August. Association for Computational Linguistics.

Zopf, M., Peyrard, M., and Eckle-Kohler, J. (2016b). The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 1535–1545, Osaka, Japan, December. Association for Computational Linguistics.

Zopf, M. (2018). Auto-hMDS: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).