

Automatic Text Summarization: A State-of-the-Art Review

Oleksandra Klymenko, Daniel Braun^a and Florian Matthes
Technical University of Munich, Department of Informatics, Munich, Germany

Keywords: Text Summarization, Natural Language Generation, Summary Evaluation.

Abstract: Despite the progress that has been achieved in over 50 years of research, automatic text summarization systems are still far from perfect, posing many challenges to the researchers in the field. This paper provides an overview of the most prominent algorithms for automatic text summarization that were proposed in the last years, as well as describes automatic and manual evaluation methods that are currently widely adopted.


1 INTRODUCTION

For more than 5000 years, written language has been the most important medium to document and pass on knowledge. Even more so in our ages of digitization. Due to the decreasing costs of producing, storing and reproducing digital texts, the amount of texts available, e.g. online, is growing rapidly every day. Summarizing these texts becomes necessary in order to help users handle this information overload and better perceive information. Text summarization can be divided into two types: manual summarization and automatic summarization. Manual summarization is the process which includes the creation of summaries manually by human experts. This is an extremely time-consuming, difficult, expensive, and stressful job for humans to perform. Therefore, it became necessary to automate this task, in order to make it faster, cheaper, and easily repeatable. Automatic text summarization is the process of automatically shortening the content of a textual information source in a way that retains its most important information. The goal of summarization systems is to produce a concise and coherent summary which would allow people to understand the content of the input without reading the entire text. A good summary must be fluent and consistent, capture all the important topics, but not contain repetitions of the same information. Automatic text summarization can be practically useful in many different domains. It can be used by companies for generation of reports, by students and scientists for finding most important ideas relevant to their research, by doctors for summarizing patients'

medical information or by journalists for covering the variety of resources, identifying main facts and different viewpoints. For a long time, extractive techniques (cf. Section 2) of text summarization have been the primary focus of research in the field. However, in the past years, there have been less major advances in extractive text summarization. Most of the research in this area proposes either an enhancement of one of the available extractive approaches or an ensemble of several previously known extractive methods. The alternative approach of abstractive summarization (cf. Section 3) has become much more attractive in recent years, especially with the emergence of deep learning technologies. There is a number of scientific works (Nenkova et al., 2011; Lloret and Palomar, 2012; Saggion and Poibeau, 2013) that provide an extensive overview of different automatic summarization methods that were proposed ever since the publication of the first paper on the topic by Luhn in 1958 (Luhn, 1958). In this article, we focus on the most prominent algorithms that were proposed in the last several years and are currently considered to be state-of-the-art, as well as describe the techniques for summarization evaluation which are currently standardly used for assessment of these algorithms.

2 EXTRACTIVE METHODS

Extractive summarization methods produce summaries by concatenating several sentences (text units) from the text being summarized exactly as they occur. The main task of such systems is to determine which sentences are important and should, therefore, be included in the summary. For many years, extrac-

^a  <https://orcid.org/0000-0001-8120-3368>

tive methods have been the main focus of researchers in the text summarization community. Many of the recent approaches address extractive summarization as a sequence labelling task, where each label indicates whether a sentence should be included in the summary or not. Cheng et al. (Cheng and Lapata, 2016) presented a data-driven summarization framework based on neural networks and continuous sentence features. They develop data-driven single-document summarization framework based on a hierarchical document encoder and an attention-based extractor. Such architecture enables development of different classes of summarization models which can extract sentences or words. The models can be trained on large-scale datasets and learn informativeness features based on continuous representations without any access to linguistic annotation. The labels are assigned to each sentence in the document individually based on their semantic correspondence with the gold summary. The authors tested their approach on DailyMail and DUC 2002 datasets and demonstrated results that are comparable to the state of the art. Nallapati et al. (Nallapati et al., 2017) present a Recurrent Neural Network (RNN) based sequence model for extractive single-document summarization which they call SummaRuNNer. The approach is based on the idea of identifying a set of sentences which collectively give the highest ROUGE (evaluation metric discussed in Section 7) with respect to the gold summary. The model allows visualization of its predictions broken up by abstract features such as information content, salience and novelty, thus being very interpretable. Authors also present a novel training mechanism that allows the extractive model to be trained end-to-end using abstractive summaries. The approach was evaluated on three datasets: Daily Mail, joint CNN/DailyMail, originally collected by Hermann et al. (Hermann et al., 2015), and Out-of-Domain DUC 2002 corpus. Results on DailyMail corpus were compared to the ones of Cheng et al. (Cheng and Lapata, 2016) using Rouge recall with summary length restricted to 75 and 275 bytes. The extractively trained SummaRuNNer model showed significant improvement over the comparable model for summary length of 75 bytes while performing comparably for summary length of 275 bytes. The abstractively trained SummaRuNNer model performed comparably for summary length of 75 bytes while underperforming the comparable model for summary length of 275 bytes. On the joint CNN/Daily Mail corpus, SummaRuNNer significantly outperformed the abstractive model of Nallapati et al. (Nallapati et al., 2016), which was the only work at the time, that reported performance on this dataset. On the Out-of-

Domain DUC 2002 corpus, SummaRuNNer was also on par with (Cheng and Lapata, 2016), however, underperformed graph-based TGRAPH (Parveen et al., 2015) and URANK (Wan, 2010) algorithms that were state-of-the-art on this corpus at the time. In (Narayan et al., 2017) authors use an architecture similar to those in previous approaches, however, for sentence ranking, they introduce a global optimization framework, which combines the maximum-likelihood cross-entropy loss with rewards from policy gradient reinforcement learning to directly optimize the final evaluation metric, ROUGE. A sentence gets a high rank for summary selection if it often occurs in high scoring summaries. The model was applied to the CNN/Daily Mail dataset and automatically evaluated using ROUGE, outperforming both of the above extractive systems, which are considered to be state-of-the-art, as well as the most prominent abstractive systems, discussed in the next section. Human evaluations showed that summaries, produced using this approach are also more informative and complete. Yasunaga et al. (Yasunaga et al., 2017) propose a graph-based neural multi-document summarization (MDS) based on the application of Graph Convolutional Networks (GCN) (Kipf and Welling, 2016) on sentence relation graphs. GCN takes in sentence embeddings from Recurrent Neural Networks (RNN) as input node features and through multiple layer-wise propagation generates high-level hidden features for the sentences. Sentence salience is then estimated through a regression on top and the important sentences are extracted in a greedy manner while avoiding redundancy. The model was evaluated on the DUC 2004 multi-document summarization (MDS) task, outperforming traditional graph-based extractive summarizers and the vanilla GRU sequence model with no graph, as well performing competitively to other state-of-the-art MDS approaches. Otherwise, in the past years, there has not been any substantial advancements in extractive text summarization, most publications only propose some improvements to the already long-existing extractive methods. Some researchers assume that extractive summarization methods may have achieved their peak and propose two possible research advancement options: 1) making ensembles of extractive methods and 2) focusing on abstractive techniques. (Mehta, 2016)

3 ABSTRACTIVE METHODS

Simply selecting a subset of sentences from the original text, as done in extractive summarization, leads to various drawbacks such as problems with cohesion

and coherence, caused by inability to combine important information that is spread throughout the document in a short way, loss of meaning due to the usage of out of context pronouns and many others. The goal of automatic text summarization systems, however, is to produce summaries as good as the ones created by humans. When people produce summaries, in order to make them optimal in terms of content and linguistic quality, they tend to edit the text rather than just copy sentences from it as they are. This is why, in order to emulate this process, generation of abstractive summaries is necessary. Abstractive summarization methods produce summaries by rewriting the content in an input, as opposed to extractive methods that simply extract and concatenate important text units from it. Therefore, abstractive summarization requires a deeper semantic and discourse interpretation of the text, as well as a novel text generation process. Carenini and Cheung (Carenini and Cheung, 2008) performed a user study comparing extractive and abstractive summarizers called MEAD* and SEA respectively. While the extractive summarizer simply copied the original sentences from user reviews from the corpus, with the numbers within the summary being footnotes linking to the corresponding review, the abstractive method produced a fluent, coherent text, summarizing the overall feedback from the users. Very recently, there have been several breakthroughs in abstractive text summarization using deep learning. Most of the research relies on Sequence to Sequence Model (Sutskever et al., 2014), which was first introduced as Encoder-Decoder Model by Cho et al. (Cho et al., 2014) and later extended by Bahdanau et al. (Bahdanau et al., 2014) with so-called "attention mechanism". Encoder encodes source sequence into a context vector, while decoder decodes context vector to produce a target sequence. Attention mechanism is used to locate the focus region during decoding. Sequence to Sequence (or seq2seq) model manages text generation pretty well as it was originally developed for machine translation. A group of researchers at Facebook (Rush et al., 2015a) presented a fully data-driven approach to abstractive sentence summarization. Their method utilizes neural attention-based model that generates each word of the summary conditioned on the input sentence. They combine this probabilistic model with a generation algorithm which produces accurate abstractive summaries. The attention-based model provides less linguistic structure comparing to other abstractive summarization approaches, but is easily scalable for training on large amounts of data. Furthermore, the lack of vocabulary constraints in the system makes it possible to train the model on diverse input-output pairs.

In addition to the paper, the source code (Rush et al., 2015b) was also provided to the public. In later work (Chopra et al., 2016) researchers at Facebook extended their model to a Recurrent Neural Network (RNN) architecture. The model includes a more sophisticated encoder which explicitly encodes the position of the input words and uses convolutional network to encode input words. With these modifications, the model showed to significantly outperform the previously proposed system on the Gigaword corpus and perform competitively on the DUC 2004 task. Authors explain the distinctive improvement by the difference between tokenization of DUC 2004 and their training corpus, as well as by the fact that headlines in Gigaword are much shorter than in DUC 2004. Later that year, researchers at IBM (Nallapati et al., 2016) also modeled abstractive text summarization using attentional encoder-decoder recurrent neural networks. To address specific problems in abstractive summarization that are not sufficiently covered by the machine translation based model, they proposed several novel models, yielding further improvement in performance. One of the most interesting models is called "Switching Generator/Pointer". In this model, the decoder is equipped with a "switch" that decides between generating a word based on the context or using a word from the input. In addition, in their work authors proposed a dataset for multi-sentence document summarization and established benchmark numbers on it. The proposed summarization approach showed to significantly outperform the ones of Rush et al. (Rush et al., 2015a) and Chopra et al. (Chopra et al., 2016) and exhibit better abstractive ability. Authors believe that their superior performance in comparison to the above methods was reached by using bidirectional RNN instead of the bag-of-embeddings representation to model the source, as this approach captures richer contextual information of every word. Google has also proposed a sequence-to-sequence with attention model for text summarization, which they called "textsum". The researchers did not provide a paper in support of their work, but openly published the source code (Liu, 2016) at a hosting service. Although these approaches are considered to be the state-of-the-art, they are far from perfect in that they sometimes inaccurately reproduce factual details, are unable to deal with out-of-vocabulary (OOV) words and can only deal with very short documents. Researchers at Facebook use only the first sentence of the source document to train the model, while researchers at Google and IBM use two sentences from the source with a limit of 120 words. However, summarization systems need to be able to deal with much longer documents.

In the works described next, authors tried to address these issues by proposing various modifications to the standard sequence-to-sequence model. In their work, Chen et al. (Chen et al., 2016) explore neural summarization technologies for articles which contain thousands of words. Researchers also base their model on the encoder-decoder framework, however, instead of focusing on attention to get the local context like most of the recent work does, they incorporate coverage mechanism, "distracting" the models to different parts of a document to avoid focusing on only one thing and get the full picture. The authors do not restrict the encoders' architectures to the standard RNN, but use bi-directional gated recurrent units (bi-GRUs) architecture for encoding and decoding. Without engineering any features, they train the models on two large datasets and test them on LCSTS corpus. The proposed approach achieved better performance than the best result reported in (Hu et al., 2015), which the authors used for comparison. See et al. (See et al., 2017) propose a novel architecture that enhances the standard sequence-to-sequence attentional model by using a hybrid pointer-to-pointer generator network and the coverage mechanism. The proposed hybrid network is similar to the ones proposed for short-text summarization by Gu et al. (Gu et al., 2016) and Miao and Blunsom (Miao and Blunsom, 2016). It uses pointing (Vinyals et al., 2015) to copy words from the input text, which provides better accuracy and is better in dealing with OOV words, while retaining the ability to generate words. To ensure coverage of the input document and thus reduce repetitions in the summary, authors propose a version of the coverage vector by adapting the model of Tu et al. (Tu et al., 2016). The model was applied to the CNN/Daily Mail dataset, outperforming the results of (Nallapati et al., 2016) by several ROUGE points.

4 QUERY-FOCUSED

Query-focused summarization became a task in DUC in 2004 in response to researchers' claims that generic summarization is too unconstrained and does not consider special user needs. The aim of such summarization is to generate a summary of a document or multiple documents in the context of a query. Such summarization allows to use more sophisticated, targeted approaches that integrate methods that seek specific types of information with data-driven, generic methods (Nenkova et al., 2011). Below we describe some of the most recent solutions to the problem of query-focused summarization. Wang et al. (Wang et al., 2016) investigate the role of sentence compression

techniques for query-focused multi-document summarization (MDS) and present a respective framework consisting of three steps: Sentence Ranking, Sentence Compression and Post-processing. For sentence ranking, authors experiment with two ranking algorithms - Support Vector Regression (SVR) (Mozer et al., 1997) and LambdaMART (Burges et al., 2007). For sentence compression, which is the main focus of their work, they examine three different approaches to sentence-compression of their own design: rule-based, sequence-based and tree-based. Finally, in the post-processing step, coreference resolution and sentence ordering are performed. All of the proposed models show substantial improvement over pure extraction-based approaches for query-based MDS, with the best-performing system yielding an 11.02 ROUGE-2 score on the DUC 2006 dataset. In (Yousefi-Azar and Hamey, 2017) researchers presented methods for extractive query-oriented single-document summarization using a deep auto-encoder (AE) to compute a feature space from the term-frequency (tf) input. The presented approach is completely unsupervised and does not require queries for any stage of training. Smaller local vocabularies, comparing to other methods, allow to reduce the training and deployment computational costs and thus make the approach more suitable for implementation on mobile devices. The authors perform experiments on two email corpora designed for text summarization and observe results that are much better than those of other unsupervised extractive email summarization techniques and are comparable to the best supervised approaches. In (Litvak and Vanetik, 2017) authors continued their earlier work (Litvak et al., 2015) of applying the Minimum Description Length (MDL) principle for generic summarization by constructing a model where frequent word sets depend on the query. The idea behind the MDL principle is that regularities in data can be used for its compression and the best hypothesis to describe these regularities is the one that can compress the data the most. In the proposed approach, authors select frequent word sets that lead to the best compression of the data and therefore describe the document the best. The extracted summary consists of sentences that provide the best coverage of query-related frequent word sets. The summarization approach was evaluated using ROUGE-1 based on the DUC 2005 and DUC 2006 corpora that were specifically designed for query-based summarization and has shown to perform competitively with the best results. The highly successful Encoder-Decoder Model that was discussed in Section 3 was recently adapted to the query-based summarization by Nema et al. (Nema et al., 2017). The authors

present a model for abstractive query-based summarization based on the encode-attend-decode paradigm with two additions: (1) a query attention model which learns to concentrate on different parts of the query at different points in time, and (2) a diversity based attention model which aims to mitigate the drawback of generating repeated phrases in the summary. The authors also introduced a dataset based on debataepedia and empirically proved that their approach performs significantly better than the plain (vanilla) encode-attend-decode mechanism, gaining 28% in ROUGE-L score, and thus outperforming the previously proposed models.

5 UPDATE SUMMARIZATION

Update summarization is a MDS task of creating a summary under the assumption that the reader is already familiar with the content of older relevant documents. The purpose of an update summary is thus to inform the user about the relevant information on a particular topic. Back in 2003, this task was approached by Allan et al. (Allan et al., 2003) who concluded it to be a hardly feasible challenge. From 2007 to 2011, while update summarization task was part of the DUC challenge, some advances were made in the field, however in the past few years a very limited research has been conducted on the topic. Therefore, in this section we cover some of the related work that has been done in the last 10 years. Gillick et al. (Gillick et al., 2009) improved on their earlier system (Gillick et al., 2008), achieving top performance in 2009 TAC summarization task. Authors adapted their standard system to the update task by taking into account sentence position, more precisely, by assuming that articles on recurrent topics tend to state information at the beginning, before recapping past details. This update showed to significantly improve ROUGE-2 scores. Three years later, Delort and Alfonseca (Delort and Alfonseca, 2012) proposed a unsupervised nonparametric Bayesian approach to model novelty in a document. The model, which is a modification of TopicSum and is called DualSum, is a variant of Latent Dirichlet Allocation (LDA) and learns to distinguish between common information and novel information. The approach was tested on the TAC 2011 dataset and obtained second and third top positions according to different ROUGE scores. Li et al. (Li et al., 2015) adopted the supervised Integer Linear Programming (ILP) framework which has been widely used for generic summarization task in earlier years (Martins and Smith, 2009; Woodsend and Lapata, 2012) for the update summarization task. Authors make two

major improvements: 1) use a set of rich features to measure the importance and novelty of the bigram concepts used in the ILP model and 2) design a sentence reranking component which allows to explicitly model a sentence's importance and novelty. Authors evaluate their methods using several recent TAC datasets, from 2008 to 2011 using ROUGE, showing that both of their additions help to improve the update summarization performance. In one of the most recent works on update summarization, de Chalendar et al. (de Chalendar et al., 2017) extend the framework defined by Gillick and Favre (Gillick and Favre, 2009) by integrating semantic sentence similarity for discarding redundancy in a maximal bigram coverage problem. For evaluation of the idea, authors used DUC 2007, TAC 2008 and TAC 2009 update corpora using different ROUGE metrics and showed that their approach noticeably improves the update summarization performance. In TAC 2011 task, update summarization became a part of the guided summarization task, which was supposed to motivate researchers to work more on abstractive approaches. One of the most prominent works on this task was done by Genest and Lapalme (Genest and Lapalme, 2012), who proposed a fully abstractive approach based on information extraction and natural language generation. For generation of summaries, authors used a rule-based, custom-designed IE module, integrated with Content Selection and Generation.

6 PERSONALIZATION

When summarizing documents, it is important to keep in mind the target audience. For example, when summarizing a scientific article, it makes a difference whether you do it for scientists, practitioners or maybe students who are only in the process of getting acquainted with the topic. Another important point to regard is that it is subjective what information can be considered relevant, since potential readers of produced summaries may differ in their needs, interests and goals. Generic text summarization approaches do not take these important issues into account. Personalization helps to adapt summaries in a way that corresponds to a user's personal preferences and characteristics by gathering additional information about the user. This information can be obtained either by explicit methods, when users enter information themselves, implicit methods, i.e. observing and analyzing user behaviour, or through a combination of both. Some of the existing approaches that were integrated in automatic summarization systems include using interactive user clicks/examinations (Yan et al., 2011)

and annotations (highlights) (Móro et al., 2012) as indicators of user interests. The book of Paris (Paris, 2015) can be referred to for more information on user modelling in text generation.

7 EVALUATION METHODS

With more and more different approaches to automatic summarization emerging, methods to compare and evaluate their performance are necessary. The evaluation process can be either intrinsic or extrinsic. Intrinsic evaluation tests the quality of summarization itself, e.g. its informativeness and coherence, while extrinsic evaluation checks the effectiveness of a summary for another task, for example, answering a query. In this section, we provide an overview of the main evaluation methods that are currently used to assess and report results on automatic summarization. There are two manual methods used at TAC (Text Analysis Conference): Pyramid and Responsiveness. Both these methods are focused on evaluating the informativeness and relevance of the summary content and do not assess its linguistic quality. The Pyramid method (Nenkova et al., 2007) is based on the semantic analysis of multiple human models, which, taken together, according to the authors' assumption, yield a gold-standard for system output. The method weights each Summary Content Unit (SCU) based on the number of human summaries in which it occurred. The Pyramid score, which ranges from 0 to 1 and represents the informativeness of the summary, is equal to the ratio between the total SCUs weights in the created summary and the weight of an optimal summary with the same number of SCUs. The Responsiveness metric is defined for query-focused summarization. For this evaluation, human reviewers are given a query and a summary and are asked to rate on a scale from 1 to 5 how good the summary is in terms of providing the requested information. Manual evaluations require a lot of time and effort and are very expensive and difficult to conduct on a regular basis. Moreover, human summaries are very variable, i.e. different people choose different sentences for extractive summaries and all the more compose different abstractive summaries. This has been proven by several early studies such as (Rath et al., 1961) which have shown not only the low agreement rate between different judges but also that even the same judge may produce a significantly different extract when asked to summarize the same document eight weeks later. Therefore, the results obtained using manual evaluation can be subjective and difficult to reproduce. To address these problems, automatic

methods for evaluation of summaries have been proposed. The most widely used automatic summarization evaluation measure for text summarization which is now standardly used to report results in research papers is called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). ROUGE was inspired by BLEU (Papineni et al., 2002), evaluation measure that was developed for machine translation evaluation. The measure is based on the overlap of units such as n-gram, word sequences, and word pairs in the automatic summary and the optimal manual summaries created by annotators. ROUGE is cheap and fast, and, unlike BLEU, that is oriented at precision, which is overly strict, ROUGE is recall-based, which makes it more preferable for summarization evaluation. One of the main problems of ROUGE is that it relies purely on lexical overlaps, which can significantly underrate summarization score, especially in documents that have a lot of synonyms, terminology variations and paraphrasing. Motivated by this observation, Cohan and Goharian (Cohan and Goharian, 2016) conducted a detailed analysis of ROUGE's effectiveness for evaluation of scientific summaries. In their results, ROUGE showed to be unreliable for evaluation of these types of summaries, with different ROUGE variations producing different correlations with the pyramid scores. The authors propose a metric for summarization evaluation called SERA (Summarization Evaluation by Relevance Analysis), which is based on the content relevance between automatically generated summary and the corresponding manual summaries, written by human annotators. The proposed metric proved to be effective in evaluating summaries of scientific articles, consistently achieving high correlations with manual scores. However, until this day ROUGE is still standardly used evaluation measure for assessment of automatic summarization systems.

8 CONCLUSION

With the development of natural language processing and data collection and analysis opportunities, a significant progress was made in the field of automatic text summarization in the last years. While many researchers still center their work around improving extractive summarization, many shift their focus to abstractive techniques, very recently having several major breakthroughs in the area. However, despite all the research in the field of automatic summarization, current summarizers are still far from perfect and many challenges still remain unsolved. For example, RNN encoder-decoder structures, which are currently con-

sidered to be state-of-the-art, still fail to encode long documents. Many of the current summarizers are still based on identifying frequent word sequences with no semantic processing, which is an especially noticeable problem when it comes to query-focused or guided summarization, which requires "understanding" of the documents. Advancements in this direction could help to eventually produce highly personalized, interactive summaries, tailored to the specific user needs. Furthermore, the field is strongly lacking quality summarization datasets, especially in domains other than science and languages other than English, which slows down and complicates further advancements. Finally, there is also a need in improving methods for evaluation of automatic summarization systems because, as discussed in Section 7, standardly used metric ROUGE is far from perfect in many aspects and with the development of more advanced, focused summaries, the need for more consistent evaluation methods will become essential.

ACKNOWLEDGEMENTS

This work has been sponsored by the German Federal Ministry of Education and Research (BMBF) grant A-SUM 01IS17049.

REFERENCES

- Allan, J., Wade, C., and Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 314–321.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Burges, C. J., Rago, R., and Le, Q. V. (2007). Learning to rank with nonsmooth cost functions. In *Advances in neural information processing systems*, pages 193–200.
- Carenini, G. and Cheung, J. C. K. (2008). Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41.
- Chen, Q., Zhu, X., Ling, Z., Wei, S., and Jiang, H. (2016). Distraction-based neural networks for document summarization. *arXiv preprint arXiv:1610.08462*.
- Cheng, J. and Lapata, M. (2016). Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Chopra, S., Auli, M., and Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 93–98.
- Cohan, A. and Goharian, N. (2016). Revisiting summarization evaluation for scientific articles. *arXiv preprint arXiv:1604.00400*.
- de Chalendar, G., Ferret, O., et al. (2017). Taking into account inter-sentence similarity for update summarization. In *Proceedings of the Eighth International Joint Conference on NLP (Volume 2: Short Papers)*, volume 2, pages 204–209.
- Delort, J.-Y. and Alfonseca, E. (2012). Dualsum: a topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 214–223.
- Genest, P.-E. and Lapalme, G. (2012). Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the ACL: Short Papers-Volume 2*, pages 354–358.
- Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18.
- Gillick, D., Favre, B., and Hakkani-Tür, D. (2008). The icsi summarization system at tac 2008. In *TAC*.
- Gillick, D., Favre, B., Hakkani-Tür, D., Bohnet, B., Liu, Y., and Xie, S. (2009). The icsi/utd summarization system at tac 2009. In *TAC*.
- Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.
- Hu, B., Chen, Q., and Zhu, F. (2015). Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.
- Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Li, C., Liu, Y., and Zhao, L. (2015). Improving update summarization via supervised ilp and sentence reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 1317–1322.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Litvak, M., Last, M., and Vanetik, N. (2015). Krimping texts for better summarization. In *Proceedings of the 2015 EMNLP*, pages 1931–1935.
- Litvak, M. and Vanetik, N. (2017). Query-based summarization using mdl principle. In *Proceedings of the MultiLing 2017 Workshop on Summarization and*

- Summary Evaluation Across Source Types and Genres*, pages 22–31.
- Liu, X. P. (2016). Sequence-to-sequence with attention model for text summarization (textsum).
- Lloret, E. and Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.
- Martins, A. F. and Smith, N. A. (2009). Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 1–9.
- Mehta, P. (2016). From extractive to abstractive summarization: a journey. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 100–106.
- Miao, Y. and Blunsom, P. (2016). Language as a latent variable: Discrete generative models for sentence compression. *arXiv preprint arXiv:1609.07317*.
- Móro, R. et al. (2012). Personalized text summarization based on important terms identification. In *DEXA, 2012 23rd International Workshop on*, pages 131–135.
- Mozer, M., Jordan, M. I., and Petsche, T., editors (1997). *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*. MIT Press.
- Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.
- Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.
- Narayan, S., Pappas, N., Cohen, S. B., and Lapata, M. (2017). Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530*.
- Nema, P., Khapra, M., Laha, A., and Ravindran, B. (2017). Diversity driven attention model for query-based abstractive summarization. *arXiv preprint arXiv:1704.08300*.
- Nenkova, A., McKeown, K., et al. (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.
- Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on ACL*, pages 311–318.
- Paris, C. (2015). *User modelling in text generation*. Bloomsbury Publishing.
- Parveen, D., Ramsel, H.-M., and Strube, M. (2015). Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 EMNLP*, pages 1949–1954.
- Rath, G., Resnick, A., and Savage, T. (1961). The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *Journal of the Association for Information Science and Technology*, 12(2):139–141.
- Rush, A. M., Chopra, S., and Weston, J. (2015a). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.
- Rush, A. M., Chopra, S., and Weston, J. (2015b). Neural attention model for abstractive summarization. <https://github.com/facebookarchive/NAMAS>.
- Saggion, H. and Poibeau, T. (2013). Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 3–21. Springer.
- See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.
- Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.
- Wan, X. (2010). Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1137–1145.
- Wang, L., Raghavan, H., Castelli, V., Florian, R., and Cardie, C. (2016). A sentence compression based framework to query-focused multi-document summarization. *arXiv preprint arXiv:1606.07548*.
- Woodsend, K. and Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint EMNLP and Computational Natural Language Learning*, pages 233–243.
- Yan, R., Nie, J.-Y., and Li, X. (2011). Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the EMNLP*, pages 1342–1351.
- Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., and Radev, D. (2017). Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*.
- Yousefi-Azar, M. and Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68:93–105.