A Repository of Corpora for Summarization

Franck Dernoncourt[1], **Mohammad Ghassemi**[2], **Walter Chang**[1]
[1]Adobe Research, [2]MIT
franck.dernoncourt@adobe.com, ghassemi@mit.edu, wachang@adobe.com

Abstract

Summarization corpora are numerous but fragmented, making it challenging for researchers to efficiently pinpoint corpora most suited to a given summarization task. In this paper, we introduce a repository containing corpora available to train and evaluate automatic summarization systems. We also present an overview of the main corpora with respect to the different summarization tasks, and identify various corpus parameters that researchers may want to consider when choosing a corpus. Lastly, as the recent successes of artificial neural networks for summarization have renewed the interest in creating large-scale corpora for summarization, we survey which corpora are used in neural network research studies. We come to the conclusion that more large-scale corpora for summarization are needed. Furthermore, each corpus is organized differently, which makes it time-consuming for researchers to experiment a new summarization algorithm on many corpora, and as a result studies typically use one or very few corpora. Agreeing on a data standard for summarization corpora would be beneficial to the field.

Keywords: abstractive summarization, extractive summarization, artificial neural networks, corpora

1. Introduction

Automatic summarization has been studied for over half a century (Luhn, 1958). Over the decades, many summarization tasks, systems, metrics, and corpora have been created. Summarization approaches may be categorized in abstractive, extractive, and compressive approaches. When the output of a summarization system is a newly generated text, distinct from the original document, it is referred to as *abstractive*. Systems that compose summaries by combining and restructuring various segments of the original text, are referred to as *extractive*, or *sentential extractive* if sentences from the original text are selected to form the summary. Lastly, systems that compose summaries by pruning tokens from the original text are referred to as *compressive*.[1]

Summarization systems may span single, or multiple-documents, and can produce outputs of varying lengths and structures. When the original text spans over multiple documents, the task is called *multi-document* summarization. The length of the summaries differs across corpora: for example, *sentence-level* summarization aims at summarizing a text into a single sentence, typically abstractively, and *headline generation* aims at summarizing the text into a headline, which tends to be shorter than a sentence.

Summarization is a subjective task (Rath et al., 1961; Lin and Hovy, 2002), requiring human input to assess performance. *Generic* summarization aims at creating a summary that is as reader-independent as possible, i.e. satisfying as many readers as possible. There has been some work on non-generic summarization, such as *query-based* and *topic-based* summarization, which bias the summary toward a query or a topic expressed by the intended reader (Hand, 1997).

Automated evaluation methods have been developed. The most widely used automated evaluation metric for summarization is ROUGE and its variants (Lin and Hovy, 2003; Lin, 2004), followed by METEOR (Banerjee and Lavie, 2005). Other metrics include Basic Elements (Hovy et al., 2005), LSA-based evaluation measures (Steinberger and Ježek, 2012) and SIRA (Cohan and Goharian, 2016). Ideally, one can perform human-based evaluation strategies, such as the pyramid method (Nenkova and Passonneau, 2004).

Given the diversity of summarization approaches, and assessment protocols, it may be challenging for researchers to identify the subset of corpora that are best-suited for a given summarization research task. In this paper, we attempt to solve this problem by presenting an overview of existing corpora, and evaluating their utility for common summarization tasks.

2. Corpora

2.1. Overview

Table 1 presents an overview of the main summarization corpora. The most widely used corpora are the Document Understanding Conference (DUC) and the Text Analysis Conference (TAC) corpora. The DUC corpora were released as part of the summarization shared task hosted at the Document Understanding Conference[2], which took place yearly from 2001 to 2007. Over et al. (2007) provides a detailed overview of the DUC 2001 to 2006 datasets. In 2008, DUC was replaced by the Text Analysis Conference[3], which is organized annually and had a summarization shared task in 2008, 2009, 2010, 2011, and 2014.

Historically, the summarization field has focused on generic extractive summarization. Prior to 2010, work in abstractive summarization had been quite limited (Ganesan et al., 2010). However, over the past few years, artificial neural networks have shown promising results for abstractive summarization. The DUC and TAC corpora, each of them having fewer than 1000 summaries, are too small to train neural networks (Nallapati et al., 2016b; Cheng and

[1]Some researchers use the terms extractive and compressive interchangeably. *Sentential extractive* is unambiguous.

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

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

Dataset	A/E	Lang.	Domain	Multi-doc	Size	Output length	Generic
DUC 2001 (Over and Yen, 2001)	a	en	news	both	60x10	50,100,200,400	y
DUC 2002 (Over and Liggett, 2002)	a,e	en	news	both	60x10	10,50,100,200,400	y
DUC 2003 (Over and Yen, 2003)	a	en	news	both	60x10,30x25	10,100	both
DUC 2004 (Over and Yen, 2004)	a	en,ar	news	both	100x10	10,100	both
DUC 2005 (Dang, 2005)	a	en	news	y	50x32	250	query-focused
DUC 2006 (Dang, 2006)	a	en	news	y	50x25	250	query-focused
DUC 2007 (Dang, 2007)	a	en	news	y	25x10	100	update
TAC 2008 (Dang and Owczarzak, 2008)	a	en	news	y	48x20	100	update,query
TAC 2009 (Dang and Owczarzak, 2009)	a	en	news	y	44x20	100	guided
TAC 2010 (Owczarzak and Dang, 2010)	a	en	news	y	46x20	100	guided
TAC 2011 (Owczarzak and Dang, 2011)	a	en	news	y	44x20	100	guided
ICSI (Janin et al., 2003)	a,e	en	meetings	n	57	390	y
AMI (McCowan et al., 2005)	a,e	en	meetings	n	137	300	y
Opinosis (Ganesan et al., 2010)	a	en	reviews	y	51x100	25	y
Gigaword (David and Cieri, 2003)	a	en	news	n	4,111,240	headline	y
Gigaword 5 (Parker and others, 2011)	a	en	news	n	9,876,086	headline	y
LCSTS (Hu et al., 2015)	a	zh	blogs	n	2,400,591	a few sentences	y
CNN/Daily Mail (Hermann et al., 2015)	a	en	news	n	312,084	50 average	y
MSR Abstractive (Toutanova et al., 2016)	a	en	misc	n	6,000	a few sentences	y
arXiv (Cohan et al., 2018)	a	en	science	n	194,000	220	y
PubMed (Cohan et al., 2018)	a	en	science	n	278,000	216	y

Table 1: Overview of existing datasets for summarization. Abbreviations; a: abstractive; ar: arabic; e: extractive; en: English; multi-doc: multi-document summarization; n: no; y: yes; zh: Chinese. The size is expressed in terms of number of summarized texts. For multi-document summarization corpora, 60x10 means that the corpus contains 60 clusters of documents, each of them is comprised of 10 documents. The output length corresponds to the length of the gold summaries (unless mentioned otherwise, the unit is word). For DUC 2001, 2002, 2003, and 2004, gold abstracts of different lengths are provided (e.g., 50, 100, 200, and 400 words). All datasets are freely available except the Gigaword corpora. Gigaword corpora are also available in Arabic, Chinese, French, German, and Spanish. Aside from Gigaword, any corpus that comprises texts and their titles may be used for title generation.

Lapata, 2016; Nallapati et al., 2017). As a result, recent studies have employed larger datasets, mostly based on Cable News Network (CNN), Daily Mail and Gigaword documents. In Table 2, we present a list of the corpora used in several studies that investigate the use of neural networks for summarization.

2.2. Converting abstractive summaries into extractive

Most corpora for summarization have abstractive summaries as gold-standard targets. In order to circumvent this limitation, several methods have been developed to convert an abstractive summary into an extractive summary. They rely on selecting sentences from the document that maximize a given metric with respect to gold abstractive summaries.

Methods differ with respect to the score, and the sentence selection strategies: Nallapati et al. (2016c) use ROUGE as the score, Cheng and Lapata (2016) use a semantic correspondence metric (Woodsend and Lapata, 2010), Nallapati et al. (2016c) use ROUGE as the score, and Cheng and Lapata (2016) use some semantic correspondence metric (Woodsend and Lapata, 2010). Nallapati et al. (2016c) use a greedy sentence selection approach, Cao et al. (2016a) rely on integer linear optimization for scoring, and Svore et al. (2007) train a neural network.

The choice of abstract-to-extract conversion method is one more parameter making challenging to compare published studies against each other. Note that for the evaluation, one can simply evaluate the predicted extractive summary against a gold abstractive summary with a typical summarization quality metric such as ROUGE, as (Nallapati et al., 2016c) did.

Converting abstractive summaries into extractive is often imperfect though. For example, Jing (2002) analyzed 300 news articles and showed that 19% of human-generated summary sentences contain no matching article sentence, and that only 42% of the summary sentences match the content of a single article sentence (with potentially a few semantic and syntactic modifications between the article sentence and the summary sentence).

2.3. Special types of summarization

There exist many other special types of summarization in addition to the traditional summarization tasks that we have mentioned earlier. These include:

- *Update summarization*: it aims at summarizing what changed between an old text and a more recent text.

Paper	A/E	Corpora
(Cohan et al., 2018)	a	arXiv, PubMed
(Narayan et al., 2017)	e	CNN with image captions
(Paulus et al., 2017)	a	CNN/DM
(See et al., 2017)	a	CNN/DM
(Nallapati et al., 2017)	a,e	CNN/DM, DUC 2002 (t)
(Nallapati et al., 2016c)	e	DM, DUC 2002 (t)
(Cheng and Lapata, 2016)	e	CNN/DM, DUC 2002 (t)
(Ayana et al., 2016)	a	Gw, DUC 2003-4 (t)
(Cao et al., 2016b)	e	DUC 2005, 2006, 2007
(Gu et al., 2016)	a	LCSTS
(Chopra et al., 2016)	a	Gw, DUC 2004
(Nallapati et al., 2016b)	a	Gw, DUC 2003+2004 (t)
(Nallapati et al., 2016a)	a	Gw
(Gulcehre et al., 2016)	a	Gw
(Ranzato et al., 2015)	a	subset of Gw
(Rush et al., 2015)	a	Gw, DUC 2003+2004
(Cao et al., 2015)	e	DUC 2001, 2002, 2004
(Yin and Pei, 2015)	e	DUC 2002 and DUC 2004
(Kågebäck et al., 2014)	e	Opinosis

Table 2: Overview of datasets used in recent studies developing neural network architectures for summarization. Abbreviations; a: abstractive; DM: Daily Mail; e: extractive; Gw: Gigaword (any version); (t): the dataset was used for test only, not training. DUC corpora are typically used for testing only, as they tend to be too small to train neural networks on.

TAC 2008 and 2009 had an update summarization track (Dang and Owczarzak, 2008). The Text Retrieval Conference (TREC) also organized an update summarization shared task yearly from 2013 to 2017, which they sometimes referred to as *temporal summarization* (Aslam et al., 2013; Aslam et al., 2015a; Aslam et al., 2015b) or *real-time summarization* (Lin et al., 2016).

- *Source code summarization*: it aims at either summarizing in a human language what a code snippet performs, or automatically folding blocks of code that are deemed less informative (also referred as the *autofolding problem*). Iyer et al. (2016) compiled a corpus from StackOverflow to summarize source code into English. Fowkes et al. (2017) presented a system to perform autofolding and created a corpus based on the source code of the top six most popular Java projects on GitHub.

- *Overview synthesis*: the task is very similar to multi-document summarization, except that the output is much longer than a typical summary. Zhang and Wan (2017) constructed a corpus based on Wikinews, where each Wikinews is regarded as the gold overview, while the linked news articles are the input of the overview synthesis system.

- *Sentence fusion*: this task is also very similar to multi-document summarization, except that the input is two sentences, and the output is one sentence. It has been shown that generic sentence fusion may lead to a low agreement between humans (Daume III and Marcu, 2004). Sentence fusion may be used to convert an extractive summary into a more abstractive summary (Barzilay and McKeown, 2005).

- *Sentence compression*: the objective is to summarize one single sentence, either abstractively or extractively. Filippova and Altun (2013) constructed the first large corpus for this task, containing 250,000 pairs of sentences. They later created a larger corpus, containing around 2 million pairs, but only 10,000 were publicly released (Filippova et al., 2015).

- *Concept-map-based summarization*: the task is to create a concept map from a text. Falke and Gurevych (2017) created a corpus of 30 educational topics, each containing around 40 source documents and a summarizing concept map that is the consensus of several crowdworkers.

Summarization may also be performed for non-textual input, such as single images (Fan et al., 2008), albums of images (Yu et al., 2017), videos (Evangelopoulos et al., 2008), or voice recording (e.g., meetings or presentations) (Zhang et al., 2007). Different type of inputs may also be combined to perform the summary, which is a task referred to as *multi-modal summarization* (Li et al., 2017).

2.4. Meta-information

When choosing a suitable corpus to train or evaluate a summarization algorithm, many parameters must be taken into account, including:

- Domain of the texts: the majority of corpora concentrate on news aticles. This is a significant shortcoming as supervised models trained on news aticles may have poor performances when applied to another domain. It also limits the evaluation of summarization algorithms to a particular domain.

- Type of gold summaries: abstractive, or extractive. We review in Section 2.2. several methods to convert an abstractive summary into an extractive summary, as most corpora are abstractive.

- Number of gold summaries per texts: typically a corpus contains one gold summary per text in the case of single-document summarization, or one gold summary per group of texts (often referred as *topic* or *cluster*) in the case of multi-document summarization. If each text has more than one gold summary, the corpus may be referred to as *multi-reference* (Toutanova et al., 2016). Note that the task of extractive single-document summarization may be counterintuitively more difficult than extractive multi-document summarization (Nenkova, 2005).

- Language: most existing corpora are in English. The only large-scale, non-English corpus are LCSTS

(Large Scale Chinese Short Text Summarization), which is in Chinese, and the Gigaword corpora, which are available in English, Arabic, Chinese, French, German, and Spanish.

- Length of the text to summarize: typically from a few sentences to a few pages. If the text to summarize is a single sentence, the task is often referred as *sentence compression*, even if the summary is abstractive (Cohn and Lapata, 2013)[4].

- Length of the reference summaries: it typically varies from a headline (headline generation) to several sentences (multi-sentence summary).

- Summarization intent: generic, or non-generic such as query-based or topic-based. Query-based summarization may be viewed as a form of question answering task.

- Presence of side information: some corpora may provide some side information in addition to the text to summarize. For example, Narayan et al. (2017) created a corpus based on CNN news articles that incorporate image captions in addition to the texts of the articles. It is however uncommon.

- Price, license, and access: corpora vary in terms of price, license, and access method. Fortunately, the vast majority of summarization corpora is freely available, with the notable exceptions of the Gigaword corpora, and LCSTS (free for research, potentially non-free for commercial use).

- Number of other studies using it: as a more direct way to assess the popularity of a corpus, one can look at the number of papers that used it. One has to keep in mind that as a result of the evolution of summarization algorithms and research interests, the most used corpora may change over time, as Table 2 shows.

2.5. Repository

The LRE map (Calzolari et al., 2012) contains a list of summarization datasets. However, we found it to have two significant limitations: 1) a few technical issues 2) lack of many summarization-specific metainformation, since it has to support any type of corpus.

In light of the increasing number of summarization corpora, as well as the amount of summarization-specific metainformation, we have created a repository for summarization corpora[5]. The repository aims at providing researchers a synopsis of existing corpora, by displaying metainformation for each corpus. We encourage contributions from anyone, either to improve the metainformation of listed corpora, or adding a new corpus.

[4]If the sentence compression is not abstractive, one can refer to it as *deletion-based* sentence compression (Filippova et al., 2015)

[5]`https://github.com/Franck-Dernoncourt/summarization-corpora`

3. Conclusion

In this paper, we have presented an overview of the main corpora for summarization, and introduced a repository aiming to list corpora for summarization as well as their metainformation. There exist many corpora, but most of them are small and cannot be used to train neural networks. More large-scale corpora for summarization are needed. Furthermore, each corpus has its own data organization; creating a data standard for summarization corpora would make research more efficient.

4. Bibliographical References

Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., and Sakai, T. (2013). Trec 2013 temporal summarization.

Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., and Sakai, T. (2015a). TREC 2014 temporal summarization track overview. Technical report, National Institute of Standards and Technology.

Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., and Sakai, T. (2015b). TREC 2015 temporal summarization track overview. Technical report, National Institute of Standards and Technology.

Ayana, S. S., Liu, Z., and Sun, M. (2016). Neural headline generation with minimum risk training. *arXiv preprint arXiv:1604.01904*.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The lre map. harmonising community descriptions of resources. In *LREC*, pages 1084–1089.

Cao, Z., Wei, F., Dong, L., Li, S., and Zhou, M. (2015). Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.

Cao, Z., Chen, C., Li, W., Li, S., Wei, F., and Zhou, M. (2016a). Tgsum: Build tweet guided multi-document summarization dataset. In *AAAI*, pages 2906–2912.

Cao, Z., Li, W., Li, S., Wei, F., and Li, Y. (2016b). Attsum: Joint learning of focusing and summarization with neural attention. *arXiv preprint arXiv:1604.00125*.

Cheng, J. and Lapata, M. (2016). Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

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 Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Cohan, A. and Goharian, N. (2016). Revisiting summarization evaluation for scientific articles. *Language Resources and Evaluation Conference (LREC)*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. Submitted to NAACL.

Cohn, T. and Lapata, M. (2013). An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):41.

Dang, H. T. and Owczarzak, K. (2008). Overview of the tac 2008 update summarization task. In *Proc. of the First Text Analysis Conference*.

Daume III, H. and Marcu, D. (2004). Generic sentence fusion is an ill-defined summarization task. In *Proceedings of the ACL Text Summarization Branches Out Workshop*, pages 96–103.

Evangelopoulos, G., Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., and Avrithis, Y. (2008). Movie summarization based on audiovisual saliency detection. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 2528–2531. IEEE.

Falke, T. and Gurevych, I. (2017). Bringing structure into summaries: Crowdsourcing a benchmark corpus of concept maps. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2969–2979. Association for Computational Linguistics.

Fan, J., Gao, Y., Luo, H., Keim, D. A., and Li, Z. (2008). A novel approach to enable semantic and visual image summarization for exploratory image search. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 358–365. ACM.

Filippova, K. and Altun, Y. (2013). Overcoming the lack of parallel data in sentence compression. In *EMNLP*, pages 1481–1491.

Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with lstms. In *EMNLP*, pages 360–368.

Fowkes, J., Chanthirasegaran, P., Ranca, R., Allamanis, M., Lapata, M., and Sutton, C. (2017). Autofolding for source code summarization. *IEEE Transactions on Software Engineering*.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics.

Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y. (2016). Pointing the unknown words. *arXiv preprint arXiv:1603.08148*.

Hand, T. F. (1997). A proposal for task-based evaluation of text summarization systems. *Intelligent Scalable Text Summarization*.

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.

Hovy, E., Lin, C.-Y., and Zhou, L. (2005). Evaluating duc 2005 using basic elements. In *Proceedings of DUC*, volume 2005.

Hu, B., Chen, Q., and Zhu, F. (2015). Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.

Iyer, S., Konstas, I., Cheung, A., and Zettlemoyer, L. (2016). Summarizing source code using a neural attention model. In *ACL (1)*.

Jing, H. (2002). Using hidden markov modeling to decompose human-written summaries. *Computational linguistics*, 28(4):527–543.

Kågebäck, M., Mogren, O., Tahmasebi, N., and Dubhashi, D. (2014). Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*, pages 31–39.

Li, H., Zhu, J., Ma, C., Zhang, J., and Zong, C. (2017). Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1113. Association for Computational Linguistics.

Lin, C.-Y. and Hovy, E. (2002). Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pages 45–51. 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, J., Roegiest, A., Tan, L., McCreadie, R., Voorhees, E., and Diaz, F. (2016). Overview of the TREC 2016 real-time summarization track. In *Proceedings of the 25th Text REtrieval Conference, TREC*, volume 16.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the Association for Computational Linguistic workshop*, volume 8.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Nallapati, R., Xiang, B., and Zhou, B. (2016a). Sequence-to-sequence rnns for text summarization. *arXiv preprint arXiv:1602.06023*.

Nallapati, R., Zhou, B., dos Santos, C., glar Gulçehre, Ç., and Xiang, B. (2016b). Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Nallapati, R., Zhou, B., and Ma, M. (2016c). Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244*.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *AAAI*.

Narayan, S., Papasarantopoulos, N., Lapata, M., and Cohen, S. B. (2017). Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530*.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-NAACL 2004*.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, volume 5, pages 1436–1441.

Over, P., Dang, H., and Harman, D. (2007). DUC in context. *Information Processing & Management*, 43(6):1506–1520.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

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. (2015). A neural attention model for abstractive sentence summarization. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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

Svore, K. M., Vanderwende, L., and Burges, C. J. (2007). Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL*, pages 448–457.

Toutanova, K., Brockett, C., Tran, K. M., and Amershi, S. (2016). A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *EMNLP*, November.

Woodsend, K. and Lapata, M. (2010). Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574. Association for Computational Linguistics.

Yin, W. and Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In *IJCAI*, pages 1383–1389.

Yu, L., Bansal, M., and Berg, T. (2017). Hierarchically-attentive rnn for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 977–982. Association for Computational Linguistics.

Zhang, J. and Wan, X. (2017). Towards automatic construction of news overview articles by news synthesis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2101–2106. Association for Computational Linguistics.

Zhang, J. J., Chan, H. Y., and Fung, P. (2007). Improving lecture speech summarization using rhetorical information. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 195–200. IEEE.

5. Language Resource References

Dang, H. T. and Owczarzak, K. (2008). Overview of the tac 2008 opinion question answering and summarization tasks. In *Proc. of the First Text Analysis Conference*, volume 2.

Dang, H. and Owczarzak, K. (2009). Overview of the tac 2009 summarization track (draft). In *In Proceedings of the Second Text Analysis Conference (TAC2009)*.

Dang, H. T. (2005). Overview of DUC 2005.

Dang, H. T. (2006). Overview of DUC 2006.

Dang, H. T. (2007). Overview of DUC 2007.

David, G. and Cieri, C. (2003). English gigaword LDC2003t05. *Linguistic Data Consortium*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Over, P. and Liggett, W. (2002). Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems. *ACL 2002, Workshop on Text Summarization*.

Over, P. and Yen, J. (2001). Introduction to DUC-2001: an intrinsic evaluation of generic news text summarization systems.

Over, P. and Yen, J. (2003). Introduction to DUC-2003: an intrinsic evaluation of generic news text summarization systems.

Over, P. and Yen, J. (2004). Introduction to DUC-2004: an intrinsic evaluation of generic news text summarization systems.

Owczarzak, K. and Dang, H. T. (2010). Overview of the tac 2010 summarization track. In *Proceedings of the Third Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.*

Owczarzak, K. and Dang, H. T. (2011). Overview of the tac 2011 summarization track: Guided task and aesop

task. In *Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November*.

Parker, R. et al. (2011). English gigaword fifth edition LDC2011T07. *Linguistic Data Consortium*.