How Domain Terminology Affects Meeting Summarization Performance

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, Fei Liu

Computer Science Department, University of Central Florida
Orlando, FL 32816, USA

{jjkoay,alexroustai,xd.zangyiwu,aleckerrigan}@knights.ucf.edu
feiliu@cs.ucf.edu

Abstract

Meetings are essential to modern organizations. Numerous meetings are held and recorded daily, more than can ever be comprehended. A meeting summarization system that identifies salient utterances from the transcripts to automatically generate meeting minutes can help. It empowers users to rapidly search and sift through large meeting collections. To date, the impact of domain terminology on the performance of meeting summarization remains understudied, despite that meetings are rich with domain knowledge. In this paper, we create gold-standard annotations for domain terminology on a sizable meeting corpus; they are known as jargon terms. We then analyze the performance of a meeting summarization system with and without jargon terms. Our findings reveal that domain terminology can have a substantial impact on summarization performance. We publicly release all domain terminology to advance research in meeting summarization.[1]

1 Introduction

A vast number of meetings are being held and recorded everyday, far more than can ever be comprehended. With this explosion of meetings comes a pressing need to develop summarization techniques to assist in browsing meeting archives (Carletta et al., 2006; Ailomaa et al., 2006). A meeting summarization system takes a meeting recording and its transcript as input and produces a concise text summary as output, which preserves the most important content of the meeting discussion (Murray and Carenini, 2008; Liu and Liu, 2009; Shang et al., 2018; Li et al., 2019). The techniques hold great potential to make large archives of meetings substantially more efficient to browse, search and facilitate information sharing.

We envision an automated summarizer that is capable of generating meeting minutes by identifying salient utterances from transcribed meeting recordings. Neural text summarization has seen significant progress (See et al., 2017; Tan et al., 2017; Chen and Bansal, 2018; Narayan et al., 2018; Lebanoff et al., 2018; West et al., 2019; Liu and Lapata, 2019; Laban et al., 2020), but most prior work focused on written texts. In contrast, recent years have seen a growing interest in summarizing spoken texts (Tardy et al., 2020). Particularly, the characteristics of meetings, domain terminology and limited annotated data pose novel challenges to neural summarization models. We favor extractive over abstractive models as the latter are prone to hallucinate content that is unfaithful to the input (Kryscinski et al., 2019).

In this paper, we investigate how domain terminology impacts meeting summarization performance, especially in the context of neural extractive summarization. Jargon is the specialized terminology associated with a particular domain (Meyers et al., 2014). It is employed in a communicative context and may not be well understood outside that context. Because meetings are usually held among professionals, jargon is ubiquitous in meeting discussions. In Table 1, we provide an example of jargon terms identified by human experts. Without a thorough study of technical jargon in the meeting domain, it is unclear how best to optimize a meeting summarizer to incorporate domain knowledge.

We present an assessment of the meeting summarization performance by comparing models trained with and without jargon. A collection of jargon terms are meticulously compiled by our expert annotators from

[1] https://github.com/ucfnlp/meeting-domain-terminology

Proceedings of the 28th International Conference on Computational Linguistics, pages 5689–5695
Barcelona, Spain (Online), December 8-13, 2020

Start	End	Spoken Utterance
247.255	252.672	with Andreas' help um Andreas put together a sort of no frills recognizer which is uh
252.672	258.837	gender-dependent but like no adaptation, **no cross-word models, no trigrams - a bigram recognizer**
258.837	262.221	and that's trained on **Switchboard** which is telephone conversations.
263.983	267.154	and thanks to Don's help wh- who - Don took
267.154	270.431	the first meeting that Jane had transcribed
270.431	277.520	and um you know separated - used the individual channels we segmented it in- into the segments that Jane had used
277.520	279.952	and uh Don sampled that so -
281.374	289.611	um and then we ran up to I guess the first twenty minutes, up to **synch time** of one two zero zero so is that - that's twenty minutes or so?
289.611	296.601	Um yeah because I guess there's some, and Don can talk to Jane about this, there's some bug in **the actual synch time file** that

Table 1: A snippet of a human transcript that contains spoken utterances and their start/end times. Domain terminology is in bold.

a meeting corpus containing multi-party conversations on the topic of speech and signal processing (Janin et al., 2003). Such jargon terms are distinct from speech recognition errors; the latter substitutes one word for another similar-sounding word during automatic transcription. The users can eliminate transcription errors using a modern interactive transcript editor. However, there remains a pressing need to understand how domain terminology affects the meeting summarization performance.

Our contributions are twofold. First, we create gold-standard annotations for domain terminology on a large meeting corpus; they are known as jargon terms. Prior work has not explored such domain-specific thesauri and thus there is limited knowledge of the target domain. Second, we analyze the performance of a meeting summarization system with and without jargon. Due to the nature of sound, such a summarizer is highly desirable to aid users in navigating through meeting recordings. Our findings suggest that domain terminology has a substantial impact on summarization performance, which should not be overlooked.

2 Data and Annotation

We extend the ICSI meeting corpus (Janin et al., 2003) for this study, which contains 75 meetings recorded at the International Computer Science Institute, Berkeley.[2] The meetings are primarily between speech group members of ICSI. An average meeting lasts an hour and has up to 10 participants. Each participant wore a close-talking microphone and they sat around a meeting table equipped with far-field microphones. The corpus is one of the larger resources in this area (Renals et al., 2012). It contains rich annotations including human transcripts, segmentation of utterances and further annotations of extractive summaries[3], making the corpus suitable for summarization. We have chosen ICSI over the AMI corpus (Carletta et al., 2006); both are natural conversations, but the scenarios in AMI meetings are artificial.

Annotating *domain terminology* is non-trivial as there lacks a universal definition. Instead, we solicit annotations from undergraduate students majoring in computer science and designate words and expressions that are beyond the scope of their knowledge as domain terminology. Interestingly, modern deep neural models often acquire such generic knowledge through unsupervised pretraining (Lewis et al., 2020). The annotators are instructed to identify words and expressions from human transcripts; they are called jargon terms and usually have particular meaning in the speech and language processing field.

The student annotators are able to annotate all of the 75 meetings for jargon terms. Meeting transcripts are substantially longer than typical news articles. A transcript contains 1,731 utterances on average and 7 words per utterance. Each meeting is annotated by one student due to the sheer size of the transcripts. However, one of the meetings has been annotated by all of the four annotators. Their average pairwise inter-annotator agreement is 0.69, indicating a moderate to high agreement between the annotators. We find that an average meeting contains 92 jargon expressions and each expression contains about 3 words. Jargon terms are observed in 5.2% of the utterances; when short utterances containing less than 5 words

[2] http://www1.icsi.berkeley.edu/Speech/mr/mtgrcdr.html

[3] http://groups.inf.ed.ac.uk/ami/icsi/

		ROUGE-1			ROUGE-2			ROUGE-SU4		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
HUMAN	(Xie and Liu, 2010)	–	–	**69.1**	–	–	33.3	–	–	–
	Ours (w/o Jargon)	59.5	63.8	59.9	32.9	34.4	32.8	33.5	35.3	33.6
	Ours (w/ Jargon)	57.0	70.6	60.7	34.9	43.0	**37.1**	34.9	43.2	**37.2**
ASR	(Shang et al., 2018)	27.6	36.3	31.0	4.4	5.6	4.8	9.9	13.5	11.3
	Ours (w/o Jargon)	41.7	55.1	**46.8**	15.1	20.5	17.2	18.7	25.3	21.3
	Ours (w/ Jargon)	39.7	57.5	46.6	15.1	21.9	**17.7**	18.2	26.5	**21.4**

Table 3: Results of our summarizer on the ICSI test set. We report ROUGE scores for our summarizer, with and without using jargon, and contrast it with strong baseline systems (Xie and Liu, 2010; Shang et al., 2018). Experimental results on human transcripts and speech recognition outputs (ASR) suggest that our model performs on par with prior state of the art.

are removed from consideration, the percentage is fairly significant (11.6%).

Our collection of domain terminology will be a valuable resource to investigate a variety of research questions regarding domain adaptation. Importantly, if a summarizer performs better when jargon terms are excluded, it indicates domain terminology may have only limited impact on determining utterance salience, or the summarizer has been ineffective in using domain knowledge. Conversely, if the summarizer performs less well, domain terminology is considered essential and it is important for speech recognizers to correctly transcribe these terms to avoid any loss in summarization performance. In what follows, we describe our meeting summarizer and examine how domain terms are processed by a modern tokenizer.

3 Meeting Summarization

The very first step that one must take to build a meeting summarizer is tokenization, which transforms an input utterance to a sequence of *sub-word units*. WordPiece (Schuster and Nakajima, 2012) and BPE (Sennrich et al., 2016) are two modern methods for tokenization. We use WordPiece that has a total vocabulary of 30,522 sub-words. The method builds a vocabulary of the desired size by iteratively combining word parts into a sub-word if doing so increases the language model likelihoods. Given the vocabulary and any input word, it uses a greedy longest-match-first algorithm to tokenize it into sub-word units; the longest sub-word will be matched first. We provide example tokenization outputs in Table 2.

Jargon Term	Tokenization
SmartKom system	**smart-ko-m** system
discourse annotations	discourse **ann-ota-tions**
situational context factors	situation-al context factors
modifiers, auxiliaries	mod-ifiers , aux-ilia-ries
JavaBayes belief-net	**java-bay-es** belief - net
a real wizard system	a real wizard system
the L_D_C	the **l _ d _ c**
the near field mikes	the near field **mike-s**

Utterance with Jargon
she wanted to display the **stylized F_ zeroes,** I think they're called?

Utterance without Jargon
she wanted to display the **[MASK]** I think they're called?

Table 2: An example showing how jargon terms are processed by a modern tokenizer, WordPiece. E.g., *smart-ko-m* means the jargon *SmartKom* was split into three tokens. Moreover, our method allows jargon to be masked-out of the utterances for summarization.

We show that most domain terminology can be properly processed by the WordPiece tokenizer. There are two immediate issues that require attention. First, it has considerable difficulties processing infrequent entities and terms, e.g., *smart-ko-m*, *java-bay-es* and *ann-ota-tions* are not well tokenized. Moreover, entities such as "LDC" need to be spelled out, the tokenizer transforms it into three individual letters, thus losing the original meaning.

Our meeting summarizer takes as input an utterance and outputs a binary label indicating if the utterance should be included in the summary. Due to data scarcity, we refrain from using sequential prediction or a more sophisticated approach that may overfit, but focus primarily on demonstrating the impact of domain terminology on model performance. Our summarizer is based on BERT-LARGE that contains 24 layers of Transformer blocks, 16 attention heads and 1024-dimensional hidden vectors (Devlin et al., 2019). The top-layer hidden vector of the [CLS] token is used as the representation of the input utterance. We apply a linear and a softmax layer to predict a binary label. Importantly, jargon terms can be masked-out of the input utterances by replacing each term with [MASK] token prior to training (Table 2). The method thus

	Without Jargon						With Jargon					
Summ	Classifier			Summarizer			Classifier			Summarizer		
Length	P(%)	R(%)	F(%)	R-1	R-2	BERTScore	P(%)	R(%)	F(%)	R-1	R-2	BERTScore
5%	37.4	10.0	15.8	36.4	18.5	54.0	39.3	10.5	16.6	37.2	19.1	53.6
10%	34.3	18.3	23.8	51.9	26.2	57.6	37.1	19.8	25.8	53.8	29.7	57.3
15%	32.2	25.7	28.6	58.0	29.7	59.8	37.8	30.2	33.5	60.9	35.5	60.3
20%	32.0	34.1	33.0	60.3	33.5	**62.0**	35.1	37.4	36.2	62.0	37.1	**62.1**
Gold	33.5	33.5	**33.5**	**61.7**	**33.5**	61.4	37.0	37.0	**37.0**	**63.7**	**38.6**	61.3

Table 4: Results of our meeting summarizer, using jargon or not, while varying the length of output summaries.

employs a single architecture to assess model performance with and without jargon.

We train the summarizer on 38,657 utterances from 54 meetings; each meeting was annotated by a single annotator. Utterances containing less than 5 words are removed from consideration. The summarizer is evaluated on the standard test set containing 6 meetings; each of these meetings have been annotated by three annotators (Carenini et al., 2011). Our experiments are performed on human transcripts and ASR outputs, respectively, the latter are acquired from the SRI speech recognizer. In the following, we discuss our findings in terms of how domain terminology affects summarization.

4 Results and Analysis

Our experimental results are presented in Table 3. We evaluate against two strong baseline systems. Xie and Liu (2010) describe an extractive meeting summarizer utilizing maximum marginal relevance and speech-specific features. Shang et al. (2018) introduce a graph framework to group utterances into clusters, perform multi-sentence compression then selection under a budget constraint. Our experiments show that, despite its simplicity, our meeting summarizer can outperform or perform on par with prior state of the art, showing a remarkable advancement of pretrained deep models in the meeting domain.

We observe that summarizing with jargon terms yields substantially better performance (an absolute gain of +4.3% R-2 F-score) on human transcripts, comparing to the alternative that masks jargon out of input utterances. The performance gap has narrowed on ASR transcripts, as domain terminology contains infrequent entities and terms, which are subject to transcription errors. Our findings suggest that domain terminology plays a significant role in determining utterance salience. Its impact on summarization and other downstream meeting applications should not be underestimated.

In Table 4, we assess the model performance on human transcripts, using jargon or not during training, and generate output summaries of varying length. We rank the utterances by their confidence scores and select a portion of them. *Gold* uses the length of ground-truth summaries. We show the precision, recall and F-scores of our classifier, ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) for summaries.[4] We find that across all lengths and evaluation metrics, summarizing with jargon can lead to a performance boost for meeting summarization. While this work has primarily experimented with the ICSI corpus, the results are sufficiently substantial that we expect them to hold over similar meeting corpora.

5 Related Work

Generating meeting summaries is a challenging problem with a great application potential. A significant number of techniques have been attempted in the past, including extraction of utterances and keyphrases from transcripts (Galley, 2006; Murray and Carenini, 2008; Liu et al., 2009; Gillick et al., 2009) and taking advantage of prosodic and speaker-related features (Maskey and Hirschberg, 2005; Zhu et al., 2009; Chen and Metze, 2012). As spoken utterances are verbose with low information density, some methods further compress and merge utterances (Liu and Liu, 2013; Wang and Cardie, 2013; Mehdad et al., 2013). Despite these valuable contributions, a closer investigation remains necessary to develop an understanding of how domain terminology affects meeting summarization performance.

Recent years have seen a renewed interest in summarizing meeting transcripts (Shang et al., 2018; Zhu et al., 2020; Tardy et al., 2020) and other types of online and transcribed conversations (Goo and Chen, 2018;

[4] The hash code for BERTScore is xlnet-base-cased_L12_no-idf_version=0.3.4(hug_trans=2.5.1)-rescaled

Yuan and Yu, 2020; Gliwa et al., 2019). In particular, Tardy et al. (2020) create a corpus containing 22 public meetings including their automatic transcriptions from audio recordings and meeting reports written by a professional. Li et al. (2019) develop a multi-modal hierarchical attention mechanism for abstractive summarization, where attention is applied to topics, utterances and words to narrow the focus to salient content; their experiments were performed the AMI corpus, thus results are not directly comparable. Our work excludes prosodic and speaker-related features to focus solely on domain terminology. It provides a new baseline for future research toward building effective meeting summarizers.

6 Conclusion

We seek to better understand how domain terminology impacts meeting summarization performance in the context of neural extractive summarization. We solicit quality annotations from expert annotators to compile a list of jargon terms from a sizable meeting corpus, which is a valuable resource to investigate a variety of research questions regarding domain adaptation. Our extensive experiments show that domain terminology has a substantial impact on summarization performance that should not be neglected. Future work may address the questions of how to obtain domain terminology in a semi-automatic way and inject domain knowledge into a meeting summarization system.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful feedback. This research was supported in part by the NSF grant IIS-1909603. Xiaojin Dai was supported by NSF DUE-1643835.

References

Marita Ailomaa, Miroslav Melichar, Agnes Lisowska, Martin Rajman, and Susan Armstrong. 2006. Archivus: A multimodal system for multimedia meeting browsing and retrieval. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 49–52, Sydney, Australia, July. Association for Computational Linguistics.

Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2011. Methods for mining and summarizing text conversations. *Morgan and Claypool Publishers*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July. Association for Computational Linguistics.

Yun-Nung Chen and Florian Metze. 2012. Two-layer mutually reinforced random walk for improved multi-party meeting summarization. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, pages 461–466. IEEE.

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

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 364–372, Sydney, Australia, July. Association for Computational Linguistics.

Daniel Gillick, Korbinian Riedhammer, Benoît Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop*.

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

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China, November. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online, July. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy, July. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy, July. Association for Computational Linguistics.

Fei Liu and Yang Liu. 2009. From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264, Suntec, Singapore, August. Association for Computational Linguistics.

Fei Liu and Yang Liu. 2013. Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression. *IEEE Transactions on Audio, Speech and Language Processing*, 21(7):1469–1480.

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 620–628, Boulder, Colorado, June. Association for Computational Linguistics.

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, pages 621–624. ISCA.

Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond T. Ng. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146, Sofia, Bulgaria, August. Association for Computational Linguistics.

Adam Meyers, Zachary Glass, Angus Grieve-Smith, Yifan He, Shasha Liao, and Ralph Grishman. 2014. Jargon-term extraction by chunking. In *Proceedings of the COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language*, pages 11–20, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 773–782, Honolulu, Hawaii, October. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis. 2012. Multimodal signal processing: Human interactions in meetings. *Cambridge University Press.*

M. Schuster and K. Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia, July. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada, July. Association for Computational Linguistics.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then summarize: Automatic alignment methods for summarization corpus creation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6718–6724, Marseille, France, May. European Language Resources Association.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria, August. Association for Computational Linguistics.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China, November. Association for Computational Linguistics.

Shasha Xie and Yang Liu. 2010. Using confusion networks for speech summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–54, Los Angeles, California, June. Association for Computational Linguistics.

Lin Yuan and Zhou Yu. 2020. Abstractive dialog summarization with semantic scaffolds. *arXiv:1910.00825*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaodan Zhu, Gerald Penn, and Frank Rudzicz. 2009. Summarizing multiple spoken documents: finding evidence from untranscribed audio. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 549–557, Suntec, Singapore, August. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: 2020 Conference on Empirical Methods in Natural Language Processing*.