Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference

Tobias Falke[1*], **Leonardo F. R. Ribeiro**[2], **Prasetya Ajie Utama**[2],
Ido Dagan[3] **and Iryna Gurevych**[2]

[1]Amazon, [2]Research Training Group AIPHES and UKP Lab, Technische Universität
Darmstadt, Germany, [3]Bar-Ilan University, Ramat-Gan, Israel
falket@amazon.com, {ribeiro,utama}@aiphes.tu-darmstadt.de,
dagan@cs.biu.ac.il, gurevych@ukp.informatik.tu-darmstadt.de

Abstract

While recent progress on abstractive summarization has led to remarkably fluent summaries, factual errors in generated summaries still severely limit their use in practice. In this paper, we evaluate summaries produced by state-of-the-art models via crowdsourcing and show that such errors occur frequently, in particular with more abstractive models. We study whether textual entailment predictions can be used to detect such errors and if they can be reduced by reranking alternative predicted summaries. That leads to an interesting downstream application for entailment models. In our experiments, we find that out-of-the-box entailment models trained on NLI datasets do not yet offer the desired performance for the downstream task and we therefore release our annotations as additional test data for future extrinsic evaluations of NLI.

1 Introduction

The general success of deep learning techniques and the availability of large-scale single-document summarization datasets, such as the CNN-DailyMail (CNN-DM) corpus (Hermann et al., 2015), have recently led to a renewed interest in abstractive summarization. Following the pioneering works of Rush et al. (2015), Chopra et al. (2016) and Nallapati et al. (2016), many models have been developed in recent years that can all generate summaries by freely choosing words from a large vocabulary rather than reusing full sentences from the input document.

While neural models have been very successful at producing fluent text with this approach, a downside is that there is less guarantee than in extractive approaches that the content of the summary is *factually correct*. Such models regularly introduce errors as illustrated in Figure 1, where the summary sentence is clearly not supported by the document. For sentence summarization, Cao et al. (2018) found up to 30% of summaries to be incorrect. That greatly reduces their usefulness, as a user cannot trust the content of the summary.

Source Sentence: prince george could be days away of becoming an older brother as the duchess is due to give birth to her second child mid-to-late april.

Summary Sentence: *prince george* is due to give birth to her second child mid-to-late april.

Figure 1: Example of an incorrect summary sentence produced by PGC (see Section 4) on CNN-DM.

In this paper, we follow the idea that all information in a summary should be *entailed* by the source document. We study the use of natural language inference (NLI) (Bowman et al., 2015), also known as textual entailment (Dagan et al., 2006), to detect factual errors. In particular, we test whether entailment predictions of NLI models can be used to rerank generated summaries such that more correct ones are preferred. Such a reranking approach can be easily combined with any recent summarization model and allows us to clearly quantify the impact of using NLI.

Our contributions and the organization of this paper are the following: First, we describe how the correctness of a generated summary can be verified efficiently via crowdsourcing. Second, we report correctness estimates for summaries generated by three recent abstractive summarization

[*]The work was done while the first author was also affiliated to the research training group AIPHES at TU Darmstadt.

systems, showing that even recent state-of-the-art models have errors in 25% of their summaries. Finally, we compare different NLI models regarding their ability to rank more correct summaries above incorrect alternatives. Here, our main finding is that models trained on NLI datasets transfer poorly to our downstream task, limiting the effectiveness of reranking. To improve NLI models for this setup, we release our collected annotations to be used as additional test data in future work.[1]

2 Related Work

Previous work already proposed the use of explicit proposition structures (Cao et al., 2018) and multi-task learning with NLI (Li et al., 2018; Pasunuru et al., 2017) to successfully improve the correctness of abstractive sentence summaries. In this work, we instead focus on the more challenging single-document summarization, where longer summaries allow for more errors. Very recently, Fan et al. (2018) showed that with ideas similar to Cao et al. (2018)'s work, the correctness of document summaries can also be improved.

Moreover, Guo et al. (2018) and Pasunuru and Bansal (2018) proposed to use NLI-based loss functions or multi-task learning with NLI for document summarization. But unfortunately, their experiments do not evaluate whether the techniques improve summarization correctness. We are the first to use NLI in a reranking setup, which is beneficial for this study as it allows to us to clearly isolate the net impact of the NLI component.

3 Evaluating Summary Correctness

Similar to previous work by Cao et al. (2018) and Li et al. (2018), we argue that the correctness of a generated summary can only be reliably evaluated by manual inspection. But in contrast to previous studies, we rely on *crowdsourcing* to make the evaluation more efficient.

In our crowdsourcing interface, we show a summary sentence by sentence on the left and the full source document on the right. For every summary sentence, a worker assigns the label *correct*, if the information is entailed by the document, *incorrect*, if it contradicts the document or contains information not present[2], or *unclear*, if the worker cannot decide. In particular, as we cannot assume that crowdworkers are familiar with the term entailment, we ask them whether a summary sentence is "correct given the information in the article". As many generated sentences are largely extractive, our interface also highlights the sentence in the source document with the highest word overlap, helping the worker to find the relevant information faster. We pay workers \$0.20 per task (labeling all sentences of one summary).

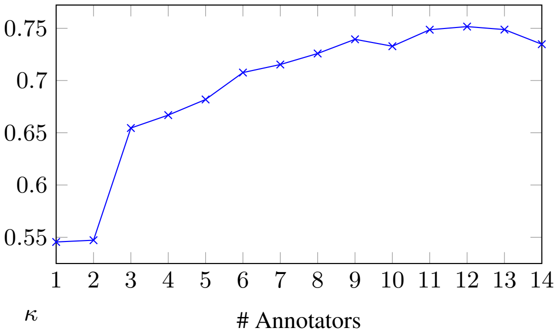

Figure 2: Agreement between crowdsourced and expert annotations at increasing numbers of workers.

Given the correctness labels for every sentence, we first merge the labels collected from different annotators. A summary then receives the label *incorrect* if at least one of its sentences has been labeled as such, otherwise, it is labeled as *correct*.

A challenge of crowdsourcing is that workers are untrained and some might produce low quality annotations (Sabou et al., 2014). For our task, an additional challenge is that some errors are rather subtle, while on the other hand the majority of summary sentences are correct, which requires workers to carry out the task very carefully to catch these rare cases.

We use MACE (Hovy et al., 2013), a Bayesian model that incorporates the reliability of individual workers, to merge sentence-level labels. We also performed an experiment to determine the necessary number of workers to obtain reliable labels. Two annotators from our lab labeled 50 generated summaries (140 sentences) manually and then merged their labels to obtain a gold standard. For the same data, we collected 14 labels per sentence from crowdworkers. Figure 2 shows the agreement, measured as Cohen's κ, between the MACE-merged labels of different subsets of the crowdsourced labels and the gold standard. We find that the agreement is substantial with at least 3 workers and that it plateaus at 9, with κ at 0.74.

[1] The data is available at `https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2002`.

[2] In NLI terms, information not present in the document would be *neutral* w.r.t the document, but in a summary it is still undesired, as all its content should be entailed.

Model	Incorrect	ROUGE-1	ROUGE-2	ROUGE-L	Length
PGC (See et al., 2017)	8%	39.49%	17.24%	36.35%	59.7
FAS (Chen and Bansal, 2018)	26%	40.88%	17.80%	38.53%	72.1
BUS (Gehrmann et al., 2018)	25%	41.52%	18.76%	38.60%	54.4

Table 1: Fraction of incorrect summaries produced by recent summarization systems on the CNN-DM test set, evaluated on a subset of 100 summaries. ROUGE scores (on full test set) and average summary length for reference.

Source:	[...] jim jepps used a blog called the daily maybe to defend "rape fantasies", describe paedophiles as "complex human beings" and question why teachers who have relationships with pupils are put on the sex offenders register. [...]
PGC:	*green party leader natalie bennett* used a blog called the daily maybe to defend [...]
Source:	(cnn) if newly revised nypd training materials are approved by a federal judge, new cadets could be taking courses reminding them "not to engage in racial profiling." [...]
FAS:	new: new nypd training materials are approved by a federal judge. [...] [*if* missing]
Source:	england's first-choice right-back at the world cup looks set to leave liverpool after six years this summer. [...]
BUS:	*england's premier league clubs* set to leave liverpool after six years this summer. [...]

Figure 3: Examples of incorrect sentences produced by different summarization models on the CNN-DM test set.

4 Correctness of State-of-the-Art Models

Using the crowd-based evaluation, we assessed the correctness of summaries for a randomly sampled subset of 100 summaries from the CNN-DM test set. We included three summarization models:

PGC The pointer-generator model with coverage as introduced by See et al. (2017).

FAS The hybrid extractive-abstractive system proposed by Chen and Bansal (2018) including their redundancy-based reranking.

BUS The bottom-up summarization system recently proposed by Gehrmann et al. (2018).

To the best of our knowledge, BUS is the state-of-the-art abstractive model on the non-anonymized version of CNN-DM as of writing this, while FAS is only slightly behind. We use the original generated summaries provided by the authors and crowdsource correctness labels using 9 workers.

Table 1 shows the evaluation results[3]. In line with the findings for sentence summarization (Cao et al., 2018; Li et al., 2018), we observe that factual errors are also a frequent problem for document summarization. Interestingly, the fraction of incorrect summaries is substantially higher for FAS and BUS compared to PGC. The length of the generated summaries appears to be unrelated to the number of errors. Instead, the higher abstractiveness of summaries produced by FAS and BUS, as analyzed in their respective papers, seems to also increase the chance of introducing errors. In addition, we also observe that among the three systems correctness and ROUGE scores do not correlate, emphasizing one more time that a ROUGE-based evaluation alone is far too limited to account for the full scope of the summarization task.

Figure 3 shows an incorrect summary sentence for each model. Common mistakes are using wrong subjects or objects in a proposition (examples 1 and 3), confusing numbers, reporting hypothetical facts as factual (example 2) or attributing quotes to the wrong person. Especially BUS and FAS often combine a subject and an object from different parts of a complex sentence such that a new, not-entailed proposition is formed, as demonstrated by the example in Figure 1.

5 Reranking based on NLI Predictions

Having seen that incorrect facts are an issue in state-of-the-art summarization models, we now turn to leveraging NLI to address this issue.

5.1 Reranking Approach

Our reranking approach follows the idea that everything in a summary should be entailed by the source document. Given a document D and sum-

[3] The ROUGE scores have been recomputed by us on the used data and match the reported scores very closely.

marization system $\mathcal{S}$, we assume that $\mathcal{S}$ can produce a list of k alternative summaries $S_0, ..., S_k$ of D. As most models typically search for the best summary sequence with beam search, k alternative summaries can be easily obtained by keeping all hypotheses from a beam search with size k.

Let $\mathcal{N}$ be an NLI model that predicts the probability $\mathcal{N}(p, h)$ that sentence h is entailed by sentence p. We score each summary alternative S_i, consisting of sentences $s_{i0}, ..., s_{in}$, heuristically based on its entailment probability given the document D, with sentences $d \in D$, as follows:

$$\sigma(S_i) = \frac{1}{n} \sum_{j=1}^{n} \max_{d \in D} \mathcal{N}(d, s_{ij})$$

We max over the sentences of the source document, as it is sufficient for a summary sentence to be entailed by one source sentence, but average over the summary sentences, as all of them should be entailed. Out of the k summary alternatives, the one with the highest score $\sigma(S_i)$ is the new predicted summary after reranking.

5.2 Experiments

We perform two experiments using NLI models for summary-level and sentence-level reranking.

NLI Models In our experiments, we test five NLI models. We use Parikh et al. (2016)'s decomposable attention model (DA) and Chen et al. (2017)'s enhanced sequential inference model (ESIM) as reimplemented and augmented with ELMO embeddings (Peters et al., 2018) by AllenNLP.[4] Further, we also include our own implementations of InferSent (Conneau et al., 2017) and shortcut-stacked encoders (SSE) (Nie and Bansal). And finally, we include a version of BERT-base (Devlin et al., 2019) fine-tuned on MultiNLI (Williams et al., 2018). DA and ESIM have been trained on SNLI 1.0 (Bowman et al., 2015), achieving 86.4% and 88.5% accuracy; InferSent and SSE were trained on MultiNLI, achieving 70.3% and 73.7% mismatched dev set accuracy. The fine-tuned BERT model has 83.6% mismatched accuracy on MultiNLI.

Summary Reranking To avoid the repeated effort of post-hoc correctness evaluations, we first created an annotated dataset from the *validation* part of CNN-DM. For 200 documents, we sampled 5 hypotheses out of a beam with size 100 and crowdsourced correctness labels for the resulting 1000 summaries. Since the availability of at least one correct summary hypothesis is a prerequisite of the reranking approach, we rely on FAS which uses a variant of beam search yielding more diverse hypotheses (Li et al., 2016). We use the code and pretrained model provided by the authors.

Split	NLI Model	Incor.	Δ	↑	↓
Val	*Original*	42.1%			
	Random	50.7%	+8.6	16	26
	DA	51.4%	+9.3	13	23
	SSE	45.8%	+3.7	18	22
	ESIM	39.3%	-2.8	23	20
	InferSent	38.3%	-3.8	24	20
	BERT	28.0%	-14.1	25	10
Test	*Original*	26.0%			
	ESIM	29.0%	+3.0	11	14

Table 2: Fraction of incorrect summaries at first position after reranking with different NLI models. ↑ and ↓ show the absolute number of improved (incorrect replaced by correct) and worsened (vice versa) instances.

For 107 out of the 200 documents, an incorrect and correct summary is among the 5 alternatives. Table 2 shows that in this sample from the validation data, the fraction of incorrect summaries at first position, when the 5 alternatives are ranked as during beam search, is at 42.1%.

Using entailment probabilities of ESIM and InferSent, we can slightly improve upon that and reduce incorrect summaries. However, with DA and SSE, more incorrect summaries end up in the first position. Note that these results are not in line with the model's NLI accuracies, underlining that performance on NLI does not directly transfer to our task. Only for BERT, which outperforms the other models on NLI by a large margin, we also see substantially better reranking performance. But even for this powerful model, more than half of the errors still remain in the summaries.[5] Interestingly, we also find that for ESIM and InferSent, reranking hurts in many cases, leaving just a few cases of net improvement.

Given the validation results, we then applied reranking to the CNN-DM *test* data followed by a post-hoc correctness evaluation as in Section 4. We used the ESIM model and reranked all 100

[4] https://allennlp.org/

[5] Note that the construction of the validation dataset ensures that the fraction of incorrect summaries can be reduced to 0% by reranking. For the test data, the lower bound is not known (as not all 100 hypotheses have been annotated).

Source:	the home which was built for former australian prime minister malcolm fraser and his wife tamie has been opened for inspection just a day after his sudden passing.	IS	DA	SSE	ESIM	BERT
Correct:	the home was built for former prime minister malcolm fraser and his wife tamie.	34%	86%	54%	94%	99%
Incorre.:	the home was built for inspection, just a day after his sudden passing.	99%	96%	99%	96%	96%

Figure 4: Two alternative sentences from generated summaries, one correct and one incorrect, for the given source sentence. All tested NLI models predict very high entailment probabilities for the incorrect sentence, with only BERT estimating a slightly higher probability for the correct alternative.

NLI Model	**Incorrect**	Δ
Random	50.0%	
DA	42.6%	-7.4
InferSent	41.3%	-8.7
SSE	37.3%	-12.7
BERT	35.9%	-14.1
ESIM	32.4%	-17.6
Human	16.1%	-33.9

Table 3: Fraction of incorrectly ordered sentence pairs using different NLI models' entailment predictions and crowdsourced human performance on the dataset.

beam hypotheses generated by FAS.[6] In contrast to the validation sample, the fraction of incorrect summaries increases from 26% to 29% (Table 2), demonstrating that the slight improvement on the validation data does not transfer to the test set.

Sentence Ranking To better understand the effect of NLI models, we carried out a second experiment that factors out some complexities of reranking. From the sampled and annotated validation data, we derived 373 triples of a source sentences d and two summary sentences, one correct (s^+) and one incorrect (s^-), covering the same content. We test how often the NLI models prefer the wrong sentence, i.e. $\mathcal{N}(d, s^-) \geq \mathcal{N}(d, s^+)$.

Table 3 shows the results. Here, ESIM performs best, followed by BERT. InferSent, while being slightly better than ESIM before, performs worse in this setup, demonstrating that the raw NLI performance does not directly correspond to the reranking performance. In general, we see that all five models leave a large gap to human performance, which we determined via crowdsourcing.

Discussion Looking at the data, we found many examples for which the NLI predictions are not as expected (as shown in Figure 4), although the incorrect sentence can be easily spotted by humans. One reason for this could be the domain shift from SNLI and MultiNLI to the newswire text of CNN-DM, suggesting that data from more diverse genres is needed. Another known issue is that NLI models tend to rely on simplifying heuristics such as lexical overlap (McCoy et al., 2019), explaining the high entailment probability that even BERT predicts for the incorrect sentence in Figure 4. These results and examples illustrate that current NLI models are not yet robust enough for our downstream task. On the other hand, the state-of-the-art performance on common NLI datasets is already very close to human performance (Nikita and Bowman, 2019), suggesting that new datasets, such as the one presented here, are necessary to expose the models' remaining limitations.

6 Conclusions

We addressed the issue of factual errors in abstractive summaries, a severe problem that we demonstrated to be common even with state-of-the-art models. While entailment predictions should help with this issue, out-of-the-box NLI models do not perform well on the task. Our proposed task and collected data can therefore be a valuable resource for future extrinsic evaluations of NLI models.

Acknowledgements

This work has been supported by the German Research Foundation through the research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1) and the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1).

[6] When performing this manual evaluation, we unfortunately did not have the fine-tuned BERT model available.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. In *Proceedings of the Thiry-Second AAAI Conference on Artificial Intelligence*, pages 4784–4791, New Orleans, LA, USA.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668, Vancouver, Canada.

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*, pages 675–686, Melbourne, Australia.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 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, San Diego, CA, USA.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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*, pages 4171—4186, Minneapolis, MN, USA.

Lisa Fan, Dong Yu, and Lu Wang. 2018. Robust Neural Abstractive Summarization Systems and Evaluation against Adversarial Information. In *NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop*, Montreal, Canada.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 687–697, Melbourne, Australia.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1693–1701, Montreal, Canada.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, GA, USA.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, NM, USA.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A Simple, Fast Diverse Decoding Algorithm for Neural Generation. ArXiv preprint 1611.08562.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany.

Yixin Nie and Mohit Bansal. Shortcut-Stacked Sentence Encoders for Multi-Domain Inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark.

Nangia Nikita and Samual R. Bowman. 2019. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In *Proceedings* *of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-Reward Reinforced Summarization with Saliency and Entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–653, New Orleans, LA, USA.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2017. Towards Improving Abstractive Summarization via Entailment Generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 27–32, Copenhagen, Denmark.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA, USA.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

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*, pages 1073–1083, Vancouver, Canada.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122, New Orleans, Louisiana.