On the Summarization of Consumer Health Questions

Asma Ben Abacha
asma.benabacha@nih.gov

Dina Demner-Fushman
ddemner@mail.nih.gov

LHNCBC, U.S. National Library of Medicine, Bethesda, MD

Abstract

Question understanding is one of the main challenges in question answering. In real world applications, users often submit natural language questions that are longer than needed and include peripheral information that increases the complexity of the question, leading to substantially more false positives in answer retrieval. In this paper, we study neural abstractive models for medical question summarization. We introduce the MeQSum corpus of 1,000 summarized consumer health questions. We explore data augmentation methods and evaluate state-of-the-art neural abstractive models on this new task. In particular, we show that semantic augmentation from question datasets improves the overall performance, and that pointer-generator networks outperform sequence-to-sequence attentional models on this task, with a ROUGE-1 score of 44.16%. We also present a detailed error analysis and discuss directions for improvement that are specific to question summarization.

1 Introduction

Teaching machines how to automatically understand natural language questions to retrieve relevant answers is still a challenging task. Different factors increase the complexity of the task such as the question length (cf. Figure 1), the lexical heterogeneity when describing the same information need, and the lack of domain-specific training datasets. Improving Question Answering (QA) has been the focus of multiple research efforts in recent years. Several efforts proposed interactive and non-interactive query relaxation techniques to translate the input questions into structured queries covering specific elements of the questions (Yahya et al., 2013; Mottin et al., 2014; Ben Abacha and Zweigenbaum, 2015; Meng et al., 2017). Other efforts focused on (i) identifying question similarity (Nakov et al., 2016, 2017) and question entailment (Ben Abacha and Demner-Fushman, 2019b) in order to retrieve similar or entailed questions that have associated answers, or (ii) paraphrasing the questions and submitting the simplified versions to QA systems (Bordes et al., 2014; Dong et al., 2017).

Question: **polymicrogyria**. My 16 month old son has this. Does not sit up our crawl yet but still trying and is improving in grabbing things etc etc. Have read about other cases that seem 10000 time worse. It's it possible for this post of his brain to grown to normal and he grow out of it?

➡ **Summary:** What is the **prognosis** for **polymicrogyria**?

Question: I have had many gout attacks since i have been 30 years old now 70. I take **allopurinol** and blood pressure meds. Before i took **allopurinol** i never had high blood pressure. Also i have developed basal cell skin cancer i have heard **allopurinol** will cause that. Reduces acid in your system?

➡ **Summary:** What are the **side effects** of **allopurinol**?

Figure 1: Consumer health questions and associated summaries from the gold standard. The entities in Red are the foci (main entities). The words in Blue and underlined are the triggers of the question types.

Question simplification or summarization was less studied than the summarization of news articles that has been the focus of neural abstractive methods in recent years (Rush et al., 2015; Nallapati et al., 2016; Chopra et al., 2016; See et al., 2017). In this paper, we tackle the task of consumer health question summarization. Consumer health questions are a natural candidate for this task as patients and their families tend to provide numerous peripheral details such as the patient history (Roberts and Demner-Fushman, 2016), that are not always needed to find correct answers. Recent experiments also showed the

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2228–2234
Florence, Italy, July 28 - August 2, 2019.

key role of question summarization in improving the performance of QA systems (Ben Abacha and Demner-Fushman, 2019a).

We present three main contributions: (i) we define *Question Summarization* as generating a condensed question expressing the minimum information required to find correct answers to the original question, and we create a new corpus[1] of 1K consumer health questions and their summaries based on this definition (cf. Figure 1); (ii) we explore data augmentation techniques, including semantic selection from open-domain datasets, and study the behavior of state-of-the-art neural abstractive models on the original and augmented datasets; (iii) we present a detailed error analysis and discuss potential areas of improvements for consumer health question summarization.

We present related work in the following section. The abstractive models and data creation and augmentation methods are presented in section 3. We present the evaluation in section 4 and discuss the results and error analysis in section 5.

2 Related Work

With the recent developments in neural machine translation and generative models (Bahdanau et al., 2014), text summarization has been focusing on abstractive models for sentence or headline generation and article summarization (Rush et al., 2015; Nallapati et al., 2016; Gehrmann et al., 2018). In particular, Rush et al. (2015) proposed an approach for the abstractive summarization of sentences combining a neural language model with a contextual encoder (Bahdanau et al., 2014). For text summarization, Nallapati et al. (2016) proposed a recurrent and attentional encoder-decoder network that takes into account out-of-vocabulary words with a pointer mechanism. This copy mechanism can combine the advantages of both extractive and abstractive summarization (Gu et al., 2016). See et al. (2017) used a hybrid pointer-generator network combining a sequence-to-sequence (seq2seq) attentional model with a similar pointer network (Vinyals and Le, 2015) and a coverage mechanism (Tu et al., 2016). They achieved the best performance of 39.53% ROUGE-1 on the CNN/DailyMail dataset of 312k news articles. Abstractive summarization models have mainly been trained and evaluated on news articles due to the availability of large scale news datasets. Fewer efforts tackled other subtasks with different inputs, such as summarization of opinions, conversations or emails (Duboué, 2012; Li et al., 2016; Angelidis and Lapata, 2018).

In this paper we focus on the summarization of consumer health questions. To the best of our knowledge, only Ishigaki et al. (2017) studied the summarization of lengthy questions in the open domain. They created a dataset from a community question answering website by using the question-title pairs as question-summary pairs, and compared extractive and abstractive summarization models. Their results showed that an abstractive model based on an encoder-decoder and a copying mechanism achieves the best performance of 42.2% ROUGE-2.

3 Methods

We define the question summarization task as generating a condensed question expressing the minimum information required to find correct answers to the original question.

3.1 Summarization Models

We study two encoder-decoder-attention architectures that achieved state-of-the-art results on open domain summarization datasets.

Sequence-to-sequence attentional model. This model is adopted from Nallapati et al. (2016). The encoder consists of a bidirectional LSTM layer fed with input word embeddings trained from scratch for the summarization task. The decoder also consists of a bidirectional LSTM layer. An attentional distribution (Bahdanau et al., 2014) is computed from the encoder's LSTM to build a context vector that is combined with the decoder embeddings to predict the word that is most likely to come next in the sequence.

Pointer-generator network. This model is adopted from See et al. (2017). It extends the sequence-to-sequence attentional model with pointer network (Vinyals and Le, 2015) that has a flexible copying mechanism allowing to either generate the next word or point to a location in the source text. The decision on whether to generate the new word or to point back to a source location is made by using a probability function as a soft switch. This probability is computed from dense connections to the decoder's input and hidden state and the context vector. This design is particularly suited to deal with words outside of

[1] github.com/abachaa/MeQSum

Method	Type	Examples
#1 MeQSum	**Consumer Health Question**	I suffered a massive stroke on [DATE] with paralysis on my left side of my body, I'm home and conduct searches on the internet to find help with recovery, and always this product called neuroaid appears claiming to restore function. to my knowledge it isn't approved by the FDA, but it sounds so promising. do you know anything about it and id there anything approved by our FDA, that does help?
	Summary	What are treatments for stroke paralysis, including neuroaid?
#2 Augmentation with Clinical Data	**Clinical Question**	55-year-old woman. This lady has epigastric pain and gallbladder symptoms. How do you assess her gallbladder function when you don't see stones on the ultrasound? Can a nonfunctioning gallbladder cause symptoms or do you only get symptoms if you have stones?
	Summary	Can a nonfunctioning gallbladder cause symptoms or do you only get symptoms if you have stones?
#3 Augmentation with Semantic Selection	**Medical Question**	Is it healthy to ingest 500 mg of vitamin c a day? Should I be taking more or less?
	Summary	How much vitamin C should I take a day?

Table 1: Examples of question-summary pairs from the created datasets.

the target vocabulary in production or test environments. We also test the coverage variant of this model which includes an additional loss term taking into account the diversity of the words that were targeted by the attention layer for a given text Tu et al. (2016). This variant is intended to deal with repetitive word generation issue in sequence to sequence models.

3.2 Data Creation

We manually constructed a gold standard corpus, MeQSum, of 1,000 consumer health questions and their associated summaries. We selected the questions from a collection distributed by the U.S. National Library of Medicine (Kilicoglu et al., 2018). Three medical experts performed the manual summarization of the 1K questions using the following guidelines: (i) the summary must allow retrieving correct and complete answers to the original question and (ii) the summary cannot be shortened further without failing to comply with the first condition. All the summaries were then double validated by a medical doctor who also gave the following scores: 1 (perfect summary), 0.5 (acceptable), and 0 (incorrect, and replaced the summary in this case). Based on these scores, the inter-annotator agreement (IAA) was 96.9%. In method **#1**, we used 500 pairs for training and 500 pairs for the evaluation of the summarization models.

We augmented the training set incrementally with two different methods. In the first augmentation method (**#2**) we added a set of 4,655 pairs of clinical questions asked by family doctors and their short versions (Ely et al., 2000). The second (augmented) training set has a total of 5,155 question-summary pairs.

Our third method (**#3**) relies on the semantic selection of relevant question pairs from the Quora open-domain dataset (Shankar Iyer and Csernai, 2017). The source Quora dataset consists of 149,262 pairs of duplicate questions. We selected a first set of candidate pairs where a question A had at least 2 sentences and its duplicate question B had only one sentence. Sentence segmentation was performed using the Stanford parser. This first selection led to a subset of 11,949 pairs. From this subset, we targeted three main medical categories: *Diseases*, *Treatments*, and *Tests*. We extracted the question pairs that have at least one medical entity from these categories. We used MetaMapLite (Demner-Fushman et al., 2017) to extract these entities by targeting a list of 35 UMLS (Lindberg et al., 1993) semantic types[2]. The final Quora subset constructed by this method contains 2,859 medical pairs. The third (augmented) training set includes the data from the three methods (8,014 training pairs). Table 1 presents example question-summary pairs from each dataset.

4 Experiments and Results

In the pointer generator and the seq2seq models, we use hidden state vectors of 256 dimensions and word embedding vectors of 128 dimensions trained from scratch. We set the size of the source and target vocabularies to 50K and the minimum length of the question summaries to 4 tokens. When applied, the coverage mechanism was started from the first iteration. We use the Adagrad

[2] *acab, anab ,comd, cgab, dsyn, inpo, mobd, neop, patf, sosy, bact, virs, lbpr, diap, lbtr, irda, nsba, vita, strd, phsu, antb, clnd, horm, carb, lipd, topp, aapp, nnon, elii, hops, orch, imft, bacs, inch, opco*

optimizer with a learning rate of 0.15 to train the network. At decode-time, we used beam search of size 4 to generate the question summary.

Method	Training Set	R-1	R-2	R-L
Seq2seq Attentional Model	#1	24.80	13.84	24.27
	#2	28.97	18.34	28.74
	#3	27.62	15.70	27.11
Pointer Generator (PG)	#1	35.80	20.19	34.79
	#2	42.77	25.00	40.97
	#3	44.16	27.64	42.78
PG+Coverage	#1	39.57	23.05	38.45
	#2	40.00	24.13	38.56
	#3	41.76	24.80	40.50

Table 2: Results of the question summarization models on the gold standard dataset.

Results are reported using the ROUGE-1, ROUGE-2, and ROUGE-L measures and presented in Table 2. The pointer generator achieves a ROUGE-1 score of 44.16% when trained on the full training dataset of 8k pairs (Method #3). The coverage mechanism improved the results of the first training set, with a limited number of training pairs (500), but decreased performance on the other training sets. This is maybe explained by the fact that the systems did not generate frequent repetitions when using the second and third training sets, which suggests that the data augmentation methods provided enough coverage and better training for the generation of relevant summaries from the test data. Figure 2 presents an example of a generated summary.

5 Discussion

The best performance of 44.16% is comparable to the state-of-the-art results in open-domain text summarization. Interestingly this performance was achieved using a relatively small set of 8K training pairs (2.5% of the size of the CNN-DailyMail dataset). Although this observation can be partially explained by the shorter average length of question summaries when compared to news summaries, a ROUGE-1 score of 44.16% suggests that the trained model reached a relatively efficient local optimum with a useful level of abstraction for consumer health question summarization. This result is especially promising, considering (i) the low-frequency nature of most medical entities and (ii) the fact that the model did not rely on external sources of medical knowledge.

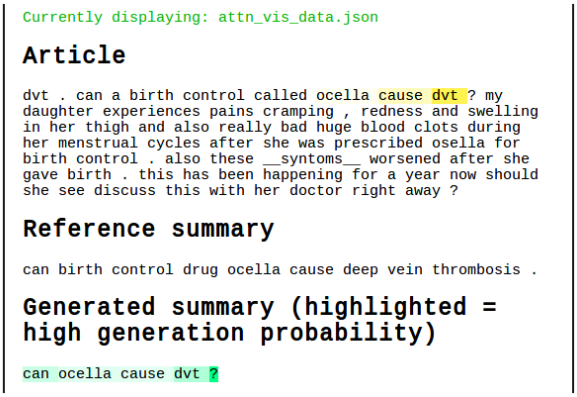

Figure 2: A summary generated by PG+M#2 method.

ROUGE (Lin and Hovy, 2003) is based on n-gram co-occurrences and despite its wide use in summary evaluation, it has some limitations. Metrics specific to question answering, such as POURPRE for the evaluation of answers to definition questions (Lin and Demner-Fushman, 2005), share some of the same limitations and do not capture fluency or semantic correctness of the summary. To study the correlation between ROUGE and human judgment in question summarization, we manually evaluated a subset of 10% of the generated summaries. We randomly selected 50 summaries produced by each PG method (M#1, M#2, and M#3) from the test set. To judge the correctness of the generated summaries, we used three scores: 0 (incorrect summary), 1 (acceptable summary), and 2 (perfect). Table 4 presents the results of the manual evaluation of the summaries. Table 3 presents examples of the generated summaries by each evaluated method. A fair amount of the manually evaluated summaries were extractive, but many were correctly generated, as can be seen in the examples.

We manually evaluated the three PG methods that achieved the best performance. These methods do not include coverage which aimed to deal with repetitive word generation issue. From our observations, few generated summaries had the repetition issue (e.g. "*where can i find information on genetic genetic genetic genetic genetic ...*"). All repetitions were generated by the M#1 method having the smallest training set (500 pairs), which means that having more training instances (5K for M#2 and 8K for M#3) alleviated the repetition problem in question summarization.

For a more in-depth analysis, we studied the

Question #1	Kidney failure 3rd stage What foods do I eat? and if I drink lots of water will that help? Is there a book that I can get to understand this disease?
Reference	where can i find information on stage three kidney failure and what are the nutritional guidelines for it?
M1	what are the treatments for failure?
M2	kidney failure 3rd stage what foods do i eat?
M3	what are the treatments for kidney failure?
Question #2	pseudogout @ http://www.nlm.nih.gov/medlineplus/ency/article/000421.htm I see the statement "There is no known way to prevent this disorder. However, treating other problems that may cause pseudogout may make the condition less severe" which I would like to have explained, especially what those other problems are &how they may be treated. I'm especially interested in whether supplemental calcium may not be good to take.
Reference, M1 & M2	what are the treatments for pseudogout?
M3	what are the treatments for pseudogout http://www.nlm.nih.gov/medlineplus/ency/article/000421.htm?

Table 3: Examples of summaries generated by the three PG methods vs. manually created reference summaries.

Score	**PG+M#1**	**PG+M#2**	**PG+M#3**
Manual	13%	46%	37%
ROUGE-1	*35.80%*	*42.77%*	*44.16%*

Table 4: Manual Evaluation of the PG methods' summaries on 10% of the test set. The manual score is the normalized average score over all summaries.

manually generated summaries of the PG+M#3 method on a random 10% subset of the test data. We identified 4 main types of errors that should be tackled in future efforts: **T1 (Question Focus[3]):** The question focus is missing or not correctly identified (e.g. "*What are the treatments?*"). **T2 (Question Type):** The question type is not the same (e.g. "*what are the treatments for williams syndrome?*" instead of "*where can I get genetic testing for william's syndrome?*"). **T3 (Semantic inconsistency):** The question type does not apply to the focus category: e.g., "*what are the treatments for nulytely?*", where nulytely is a drug name). **T4 (Summarization)**: The summary is either not minimal, or not complete: e.g., the original question contains several sub-questions, but the summary contains only one of them. The examples above are from the results of the method PG+M#3. Table 5 presents the distribution of error types, taking into account multiple error types per summary when they occur. 76% of the errors are related to the question focus and the question type. Interestingly, only 7% of the summaries are semantically inconsistent. These findings suggest that training the networks to take into account the question focus and type is a promising direction for improvement. Such approach could be achieved either through multitask training or through additional input features, and will be investigated further in our future work.

Method	**T1**	**T2**	**T3**	**T4**
PG+M#3	38%	31%	7%	24%

Table 5: Distribution of error types.

6 Conclusion

We studied consumer health question summarization and introduced the MeQSum corpus of 1K consumer health questions and their summaries, which we make available in the scope of this paper[4]. We also explored data augmentation methods and studied the behavior of abstractive models on this task. In future work, we intend to examine multitask approaches combining question summarization and question understanding.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We would like to thank Sonya E. Shooshan and Mark Sharp for their help with the manual summarization.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *CoRR*, abs/1808.08858.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[3] Main entity in the question.

[4] github.com/abachaa/MeQSum

Asma Ben Abacha and Dina Demner-Fushman. 2019a. On the role of question summarization and information source restriction in consumer health question answering. In *Proceedings of the AMIA 2019 Informatics Summit, San Francisco, CA, USA, 2019*.

Asma Ben Abacha and Dina Demner-Fushman. 2019b. A question-entailment approach to question answering. *CoRR*, abs/1901.08079.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Inf. Process. Manage.*, 51(5):570–594.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 615–620.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, June 12-17, 2016*, pages 93–98.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *JAMIA*, 24(4):841–844.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 875–886.

Pablo Ariel Duboué. 2012. Extractive email thread summarization: Can we do better than he said she said? In *INLG 2012 - Proceedings of the Seventh International Natural Language Generation Conference, 30 May 2012 - 1 June 2012, Starved Rock State Park, Utica, IL, USA*, pages 85–89.

John W. Ely, Jerome A. Osheroff, Paul N. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, and P. Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429–432.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4098–4109.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Tatsuya Ishigaki, Hiroya Takamura, and Manabu Okumura. 2017. Summarizing lengthy questions. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 792–800.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1):34:1–34:28.

Qiudan Li, Zhipeng Jin, Can Wang, and Daniel Dajun Zeng. 2016. Mining opinion summarizations using convolutional neural networks in chinese microblogging systems. *Knowl.-Based Syst.*, 107:289–300.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*.

Jimmy J. Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 931–938.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32:281–291.

Xiangfu Meng, Xiaoyan Zhang, Yanhuan Tang, and Chongchun Bi. 2017. Adaptive query relaxation and top-k result ranking over autonomous web databases. *Knowl. Inf. Syst.*, 51(2):395–433.

Davide Mottin, Alice Marascu, Senjuti Basu Roy, Gautam Das, Themis Palpanas, and Yannis Velegrakis. 2014. IQR: an interactive query relaxation system for the empty-answer problem. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1095–1098.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, San Diego, CA, USA, June 16-17, 2016*, pages 525–545.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. In *International Conference on Learning Representations, Workshop track*.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *JAMIA*, 23(4):802–811.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Nikhil Dandekar Shankar Iyer and Kornl Csernai. 2017. First quora dataset release: Question pairs.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. 2013. Robust question answering over the web of linked data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1107–1116.