INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

FinChat: Corpus and evaluation setup for Finnish chat conversations on everyday topics

Katri Leino[1], *Juho Leinonen*[1], *Mittul Singh*[1], *Sami Virpioja*[2], *Mikko Kurimo*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Department of Digital Humanities, University of Helsinki, Finland

katri.k.leino@aalto.fi, juho.leinonen@aalto.fi, mittul.singh@aalto.fi, sami.virpioja@helsinki.fi, mikko.kurimo@aalto.fi

Abstract

Creating open-domain chatbots requires large amounts of conversational data and related benchmark tasks to evaluate them. Standardized evaluation tasks are crucial for creating automatic evaluation metrics for model development; otherwise, comparing the models would require resource-expensive human evaluation. While chatbot challenges have recently managed to provide a plethora of such resources for English, resources in other languages are not yet available. In this work, we provide a starting point for Finnish open-domain chatbot research. We describe our collection efforts to create the Finnish chat conversation corpus FinChat, which is made available publicly. FinChat includes unscripted conversations on seven topics from people of different ages. Using this corpus, we also construct a retrieval-based evaluation task for Finnish chatbot development. We observe that off-the-shelf chatbot models trained on conversational corpora do not perform better than chance at choosing the right answer based on automatic metrics, while humans can do the same task almost perfectly. Similarly, in a human evaluation, responses to questions from the evaluation set generated by the chatbots are predominantly marked as incoherent. Thus, FinChat provides a challenging evaluation set, meant to encourage chatbot development in Finnish.

Index Terms: Finnish corpora, chatbot evaluation, open-domain chatbots, conversational language modeling

1. Introduction

Recently, open-domain conversational agents or chatbots, capable of casual conversation, have received much attention from NLP researchers. This trend has been supported by regularly organized chatbot challenges like Dialogue State Tracking Challenges[1], NeurIPS' Conversational Intelligence Challenge[2] and Amazon Alexa prize[3]. Nevertheless, training good open-domain chatbots is challenging. The chatbot training requires large amounts of conversational data and an evaluation setup to develop them. It is often easier to extract conversational data from *online* sources for training than constructing standardized evaluations. Yet, the latter is essential for model development because the alternative is to employ expensive human evaluation for comparing different conversational agents.

Chatbot challenges have overcome this issue by providing standardized evaluation setups to compare models [1, 2, 3]. The growth of resources, however, has been restricted to English. For Finnish, like many other languages, there are no chatbot evaluation setups available. Meanwhile, using machine translated corpora from well-resourced languages is dependent on the translation quality which for Finnish is a concern at the moment. In this work, our focus is to bridge this gap for Finnish and to bootstrap open-domain chatbot research in the language. We provide the FinChat corpus and evaluation setup to support this aim.

The FinChat corpus consists of Finnish chat conversations on everyday topics collected from voluntary participants. Our goal is to collect conversations that are natural and engaging, which are two important qualities of a human conversation [4]. To ensure naturalness, we do not restrict our participants to a script, specify the language style (formal or informal) or restrict them to specific discussion points. To ensure engaging conversations, we provide participants with seven broad and diverse topics to guide their conversation. Later, the participants self-evaluate each conversation to be engaging or not.

The FinChat evaluation setup includes a retrieval task to help automatically compare chatbot models. The task provides chatbot with a sentence from a conversation and asks to predict the answer as the continuation from the given list. The task is easy for humans, who achieve 95.1% accuracy, whereas off-the-shelf chatbot models trained on large Finnish conversational datasets perform much worse, barely achieving the accuracy of a random choice, 10%. We also perform a human evaluation where responses generated by the chatbots to the questions from the FinChat evaluation set are marked for grammatical correctness, intelligibility and coherence. The best generated responses score high, close to original human responses, on grammar and intelligibility but much worse on coherence. Thus, FinChat poses a challenging task for chatbot modelling. To support further research, we publicly release the FinChat corpus, our evaluation setup, and training recipes to recreate the considered chatbots at `https://github.com/aalto-speech/FinChat`.

2. Related Work

Conversational chat corpora in English are mostly knowledge-grounded [5, 6, 7] or partly scripted [8]. Conversations are knowledge-grounded by providing participants with a highly specific topic and background reading beforehand. This data generation style results in topical and more coherent conversations. Such data is useful to create chatbots suitable for information retrieval. For more casual conversations, there are only a few corpora such as PersonaChat [8] that have scripted small-talk conversations. Alternatively, real conversations from social media corpora can be extracted. However, they require significant filtering effort to ensure the quality and appropriateness of the content. Our approach for conversational chat

[1]`https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge`
[2]`http://convai.io/`
[3]`https://developer.amazon.com/alexaprize`

http://dx.doi.org/10.21437/Interspeech.2020-2511

corpus is to collect diverse casual conversations by not restricting the content and only providing a broad topic as guidance. We also promote diversity by having participants of different ages and giving them the freedom to converse with the conversational language they use in real life. Additionally, we aim for longer conversation by providing participants more time. In most of the above data sets, the length of the conversation are often short with only 5-8 turns except for Topical-Chat [5] that has 20 turn conversations. In comparison, the FinChat corpus has conversation length of 14 turns on the average.

Crowd-sourcing is a popular option to gather chat conversations because it gives easy access to a large amount of participants, and the content of the conversation can be controlled for quality. Unfortunately, all languages are not well represented in the crowd-sourcing platforms, and therefore, gathering substantial amounts of data for them is challenging. As this concerns Finnish as well, we created a collection setup and recruited volunteers to have a casual chat with each other.

For chatbot evaluation, using perplexity, cross-entropy, and translation metrics such as BLEU [9], METEOR [10], ROUGE [11], are straightforward to calculate, but show little correlation with human evaluation [12, 13]. Besides, PersonaChat corpus introduced hits@1/N, which is the success rate of predicting the correct reply out of $N - 1$ random sentences. N-choose-k [14] could be seen as an extension of hits metric, where it is enough for the correct response to appear in top-k. In our work, we employ these different metrics to evaluate chatbots on the evaluation task.

In conversation modeling, most recent advances employ Transformer models [15, 16]. However, many present approaches still use RNN-based encoder-decoder architecture [17, 18]. In our work, we test off-the-shelf systems from both these two approaches on the FinChat evaluation task.

3. FinChat dataset

FinChat corpus provides a set of natural, diverse and engaging conversations on seven topics collected from voluntary participants. To promote natural conversations, we did not provide any scripts to the participants except for the one-word topics. This way, they could steer the content of the conversation based on their knowledge of that topic. In our study, the participants were of different ages and belonged to different academic backgrounds. They also have minimal restrictions on language and conversation style to allow collection of diverse conversation styles. Each conversation is self-evaluated by the participants for engagingness. In this section, we first describe the details on the setup of the collection effort and instructions given to the participants, and then provide essential statistics of the dataset.

3.1. Collection Setup

For Finnish, crowd-sourcing platforms could not be used due to the lack of native speakers on them. We also tried machine translating PersonaChat, but the results were of poor quality and even incoherent at times. Instead, we setup a chat server[4] and invited voluntary participants for the collection effort. The participants were Finnish natives in three age-based user groups: high school students (16-19 years), university students (20-25 years) and university staff (25 years or above).

The data was collected in multiple sessions, where each session had participants from the same group. In each session, participants were paired randomly and used fake names to maintain anonymity. At the beginning of the session, participants were given a topic to discuss. They were also instructed to (1) not reveal any personal information, (2) ask one question at the time and wait for their partner's reply, (3) use conversational language, and (4) not use any abusive language. After chatting 10-15 minutes, conversation partners were switched and a new topic was given. In a session, each participant had two or three conversations. After each conversation, participants self-evaluated their conversation with a questionnaire. The specific questions are reported in Figure 1. In the case of violating the instructions related to personal information, the data was anonymized following GDPR.

[4] https://github.com/sdelements/lets-chat

Table 1: *FinChat data statistics: the number of conversations (Conv), messages (Mes), and words and the rate of interesting conversations for each topic and group. The groups are university staff, university students and high school students (HS)*

Topics	Conv	Mes	Words	Interesting
Sports	24	1,054	5,703	77 %
Literature	15	655	3,179	61 %
TV	15	900	4,132	71 %
Traveling	12	463	4,418	83 %
Food	9	240	2,140	78 %
Movies	7	209	1,546	57 %
Music	4	149	1,263	100 %
Univ. staff	41	1,526	12,700	77 %
Univ. students	10	239	2,769	85 %
HS students	34	1,863	6,733	66 %
All	86	3,630	22,210	74 %

Table 2: *Group statistics: The number of conversations (Conv), the average word length, the average number of messages in each conversation (Mes / Conv), the number of words in each message (Words / Mes), and the rate of interesting conversations in each age group.*

Groups	Word length	Mes / Conv	Words / Mes
Univ. staff	6.0	37.2	8.3
Univ. students	5.5	23.9	11.6
HS students	5.0	54.8	3.6

3.2. Statistics

FinChat corpus contains conversations with message timestamps, sender's id, and metadata information. The metadata includes information on participant id, age group, topic, and questionnaire results. Table 1 shows the conversation statistics for each topic and for each user group. The number of conversations on each topic varies because of different number of topics and participants per session. The corpus has 86 conversations with 3,630 messages, 22,210 words with the average word length of 5.6, and on the average 14 turns per each conversation.

User group statistics are reported in the Table 2. The majority of the participants were university staff and high school students. It is possible to see interesting differences between these two groups. High schools student sent more messages than other groups. However, their messages were a lot shorter and they used smaller words than other participants. They were also more unsatisfied with the conversations: only 66% reported that their conversation was interesting. In contrast, university staff and students rated 77% and 85% of their conversations in-

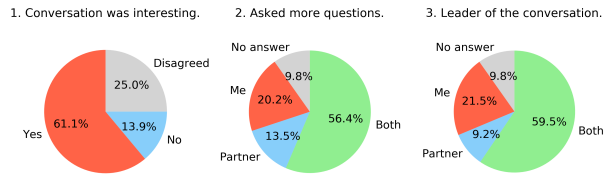

Figure 1: *Questionnaire results.*

teresting. Based on the informal feedback, high school students struggled more on keeping the conversation going, which could partly explain the results. Adults, on the other hand, seemed to enjoy their time talking with others. High school students also tended to go off-topic more often than other groups. In the whole corpus, 21.5% of the conversations contained other than the given topic.

As we wanted to collect engaging conversations, the participants were asked to evaluate their conversations by filling a questionnaire. The questions and results of the questionnaire are summarized in Figure 1. 86.5% of the conversations were rated as enjoyable by either one or both the participants, and thus, we were able to collect engaging conversations. We also asked the participants to answer who asked more questions and who led the conversation. 68% of the participants answered in the same way for both questions. In 84% of the cases, participants agreed who was asking more questions. However, conversation attendees did not often agree on who led the conversation, as only in 44% of the cases they agreed on this aspect.

Table 3: *Evaluation data statistics: the number of conversations (Conv), messages (Mes), and words for each topic.*

Topics	Conv	Mes	Words
Sports	4	200	966
Literature	2	74	300
TV	2	77	573
Traveling	2	55	620
Food	6	140	1,311
Music	3	103	935

3.3. Evaluation setup

Before extracting the evaluation set, we fuse a user's consecutive messages. Then, we select adjacent sentences, which now belong to two different users, that both have more than ten characters. From this set, we select a hundred sentence pairs for every user group. Human evaluators inspected these sentence pairs discarding pairs which are not discernible among false alternatives. The filtered sentence pairs, a total of 226, are utilized for evaluation, as described in Section 4.2. The corresponding conversations form the evaluation set, details of which are presented in Table 3.

4. Chatbot evaluation

In this section, we present the chatbot models used in our study and automatic evaluation metrics to evaluate them. We also describe the evaluation setup to compare these models and conduct a human evaluation to understand their limitations.

4.1. Models

We utilize two popular architectures to train our chatbot models: the encoder-decoder (ED) based model [19] and the Transformer based model [8, 15, 20]. In the encoder-decoder model, both the encoder and decoder are 2-layered bidirectional gated recurrent units (GRU) [21] with 500 neurons, with a dropout of 0.2. The decoder applies global attention [19] with dot alignment function and softmax output layer. Masked negative log-likelihood loss is used with Adam [22] optimizer. The learning rate is 0.0001, and the gradient clipping value is 50. Our Transformer model uses a language modeling task similar to Hugging Face's ConvAI2 submission [8, 20]. The encoder part of the model has four layers with 400 neurons, four attention heads and a 0.2 dropout. To predict the actual word, a linear layer with log softmax is applied. The loss function is the negative log-likelihood and the optimizer is Adam with a learning rate of 0.00001 and gradient clipping of 0.5. For chatbot training, we modify the recipes from [23, 24] to work with subword units generated using Morfessor [25, 26], which are essential for modeling an agglutinative language like Finnish.

Table 4: *Mean scores with a standard deviation of human evaluation of question-answer pairs generated by Encoder-Decoder (ED) and Transformer models trained with OpenSubtitles (OS) and Suomi24 (S24) data sets.*

	Human evaluation score		
Model	Intelligible	Coherence	Grammar
ED OS	4.51 ± 1.19	1.83 ± 0.39	4.35 ± 1.10
ED S24	4.10 ± 0.87	1.67 ± 0.19	3.95 ± 0.97
Transformer OS	2.38 ± 0.92	1.28 ± 0.43	2.57 ± 0.66
Transformer S24	1.95 ± 0.07	1.33 ± 0.20	2.03 ± 0.34
Human	4.97 ± 0.70	4.85 ± 0.75	4.47 ± 0.86

We train each of these models on two different Finnish corpora that can be considered out-of-domain with respect to FinChat, but still include conversational language: Suomi24 (S24) corpus [27], based on extracted messages from a popular Finnish conversation forum, and OpenSubtitles[5] (OS) corpus [28]. We form a one million sentence subset from the original corpus.

4.2. Experiments

We setup a prediction task where the first sentence of an evaluation sentence pair is fed to the model. The output can then be evaluated using character-based cross-entropy averaged over all the next sentences (CE), character n-gram F-score (chrF) [29], hits@1/N, and N-choose-k [14] with $N = 10$ and $k = 5$ (10C5). We calculate the CE for the next sentence in the pair and average it across all pairs. The chrF score [29] compares the model-generated next sentence with the correct sentence on a character n-gram basis. For hits@1/10 and 10-choose-5, we use the pair's first sentence as the question and create a possible answer set by mixing the pair's correct next sentence with randomly chosen nine other sentences from the evaluation. 24.7% generated questions did not have a clear, correct answer and were removed manually. Given the question, the chatbot chooses from the answer list. For ranking and predicting the sentences in the list, we use their cross-entropy value assigned by chatbot given the question.

Humans tested the question and answer set with 95.1% accuracy. According to feedback from human evaluators, some considered the task challenging regardless of the high accuracy, and many had to think of the context and use style cues to deduct the correct sentence. From both Suomi24 and OpenSubtitles, we separated a development set from a held-out set and correspondingly generated one thousand question-answer pairs as

[5] http://www.opensubtitles.org/

Table 5: *Results of automatic metrics for Encoder-Decoder (ED) and Transformer models trained with OpenSubtitles (OS) and Suomi24 (S24) data sets. For each training set, the models are evaluated on the corresponding development set and the FinChat evaluation set.*

		Development set (S24 or OS)				FinChat evaluation set			
Model	Train data	CE	chrF	hits@1/10	10C5	CE	chrF	hits@1/10	10C5
Transformer	S24	0.847	0.132	0.126	0.532	1.143	0.104	0.0619	0.469
	OS	0.701	0.132	0.100	0.499	1.36	0.0889	0.0664	0.527
Encoder-Decoder	S24	1.07	0.0943	0.141	0.607	1.30	0.0787	0.0973	0.540
	OS	0.993	0.0813	0.103	0.518	1.53	0.0554	0.0841	0.496

FinChat.

Ten human evaluators also evaluated the chatbot models. They were shown ten questions and were asked to score the model-generated answers for each question on three metrics: 1) *intelligible*: the answer is an understandable sentence in some context, 2) *coherence*: sentence answers the question and 3) *grammar*: the sentence is the grammatically correct form. The standard scale from 1 (very poor) to 5 (very good) was used. Original answers were rated in the same manner.

4.3. Results

According to human evaluation scores in Table 4, encoder-decoder models surpass transformers in every metric, with the model trained on OpenSubtitles data being marginally better than the one trained on Suomi24. The evaluation also suggests that encoder-decoder models can generate intelligible and grammatically correct sentences, but they do not predict coherently based on the previous message. The ED model trained with OpenSubtitles received the best scores among all the models. On the other hand, transformer models perform poorly in every human evaluation metric: they often produced unintelligible answers and nonsense words.

Despite the problems in text generation, the Transformer models are competitive with encoder-decoder models in terms of the automatic evaluation metrics based on development sets. These results are shown in Table 5. Cross-entropy results suggest that the Transformers had learned the domain of the training data better. They are also better in cross-entropy and chrF for the FinChat data. In contrast, the metric that shows encoder-decoders as better is hits@1/10. With Transformers, the correct reply is less probable than by chance, suggesting they have learned a wrong model of a conversation. While encoder-decoders do not produce coherent reactions to previous messages either, the probability distribution they have modeled might be more accurate. In 10-choose-5, there might be so many wrong sentences that the correct one easily ends up at the top half, but with hits@1/10, some learning needs to take place.

5. Discussion

While this paper had success with hits@1/N evaluation metric, the problem of automatically evaluating chatbots is far from solved, and we will continue to develop better automatic metrics for chatbot evaluation. The hits@1/N results showed a clear difference between encoder-decoders and Transformers, which suggests that curated metrics are valuable. Generating the evaluation set automatically, and then using automatic metrics with it, does not seem feasible at the moment.

FinChat is a challenging data set because of the free nature of the conversations. The messages are not strictly organized as question-answer pairs, as it is common in the more restricted and scripted chats. The messages do not always answer to the previous message but may refer to statements in the conversation history.

Recruiting volunteers to generate chat conversation instead of funded crowd-sourcing is difficult and time-consuming. However, we managed to collect a corpus with size adequate to be used as an evaluation set. Unfortunately, a lot larger data set would be needed for training chatbot models. In the future, we will continue expanding the data set in order to provide also training material for the models. We will also aim to balance topics and possibly introduce new topics. In addition, we are interested in including older participants to have a more versatile data set, as the way people discuss over chat differs a lot based on their age and background. We also aim to include new metrics that would correlate better with human evaluation and measure longer conversation history. New metrics will be necessary, especially when more advanced models are developed. Furthermore, additional work needs to be put into modeling. Using a much larger pre-trained Transformer model and fine-tune it with the chat corpus is the obvious next step. The current models, both Transformers and encoder-decoders, might also benefit from more thorough hyper-parameter tuning. In addition, more advanced decoding methods have recently shown promising results for increasing coherence and engagingness [30, 31]. Finally, since FinChat has topical information, fusing that to the models is an exciting avenue.

6. Conclusion

Other languages aside from English do not have an established evaluation setup for open-domain chatbot research. In this paper, we presented Finnish chat conversation corpus, FinChat, and showed that it could be used to evaluate open-domain chatbots. In our experiments, off-the-shelf chatbots based on encoder-decoder and transformer models performed much worse than humans on the FinChat evaluation task. Thus, FinChat posed a challenging problem. We hope these resources will encourage further research on Finnish chatbots and inspire similar efforts in other languages.

7. Acknowledgements

We would like to thank all volunteers who participated in FinChat data collection and human evaluation studies. This work was supported by the Emil Aaltonen Foundation, Kone Foundation and the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

8. References

[1] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston, "Parlai: A dialog research software platform," *arXiv preprint arXiv:1705.06476*, 2017.

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[3] J. Sedoc, D. Ippolito, A. Kirubarajan, J. Thirani, L. Ungar, and C. Callison-Burch, "Chateval: A tool for chatbot evaluation," in *NAACL 2019 (Demonstrations)*, 2019, pp. 60–65.

[4] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," in *NAACL-HLT 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1702–1723. [Online]. Available: https://www.aclweb.org/anthology/N19-1170

[5] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, and A. A. AI, "Topical-chat: Towards knowledge-grounded open-domain conversations," *Proc. Interspeech 2019*, pp. 1891–1895, 2019.

[6] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, "Towards exploiting background knowledge for building conversation systems," *arXiv preprint arXiv:1809.08205*, 2018.

[7] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[8] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe *et al.*, "The second conversational intelligence challenge (convai2)," in *The NeurIPS'18 Competition*. Springer, 2020, pp. 187–208.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL 2002*. Association for Computational Linguistics, 2002, pp. 311–318.

[10] S. Banerjee and A. Lavie, "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*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[12] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *EMNLP 2016, month = nov, year =*.

[13] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic Turing test: Learning to evaluate dialogue responses," in *ACL 2017*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1116–1126. [Online]. Available: https://www.aclweb.org/anthology/P17-1103

[14] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, and R. Kurzweil, "Generating high-quality and informative conversation responses with sequence-to-sequence models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2210–2219. [Online]. Available: https://www.aclweb.org/anthology/D17-1235

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[16] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," 2019.

[17] F.-G. Su, A. R. Hsu, Y.-L. Tuan, and H.-Y. Lee, "Personalized Dialogue Response Generation Learned from Monologues," in *Proc. Interspeech 2019*, 2019, pp. 4160–4164. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1696

[18] A. Baheti, A. Ritter, J. Li, and B. Dolan, "Generating more interesting responses in neural conversation models with distributional constraints," in *EMNLP 2018*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3970–3980. [Online]. Available: https://www.aclweb.org/anthology/D18-1431

[19] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP 2015*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1412–1421. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1166

[20] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *CoRR*, vol. abs/1901.08149, 2019. [Online]. Available: http://arxiv.org/abs/1901.08149

[21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP 2014*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] PyTorch, "Word-level language modeling rnn," https://github.com/pytorch/examples/tree/master/word_language_model, 2020.

[24] M. Inkawhich, "Chatbot tutorial," https://github.com/pytorch/tutorials/blob/master/beginner_source/chatbot_tutorial.py, 2019.

[25] M. Creutz, K. Lagus, and S. Virpioja, "Unsupervised morphology induction using morfessor," in *Finite-State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, ser. Lecture Notes in Computer Science, A. Yli-Jyrä, L. Karttunen, and J. Karhumäki, Eds., vol. 4002. Springer, 2005, pp. 300–301.

[26] P. Smit, S. Virpioja, S. Grönroos, and M. Kurimo, "Morfessor 2.0: Toolkit for statistical morphological segmentation," in *EACL 2014*, G. Bouma and Y. Parmentier, Eds. The Association for Computer Linguistics, 2014, pp. 21–24.

[27] Aller Media Oy, "Suomi24 -korpus 2001-2017, VRT-versio 1.1," 2020. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2020021801

[28] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://www.aclweb.org/anthology/L16-1147

[29] M. Popović, "chrF: character n-gram f-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://www.aclweb.org/anthology/W15-3049

[30] I. Kulikov, A. Miller, K. Cho, and J. Weston, "Importance of search and evaluation strategies in neural dialogue modeling," in *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, Oct.–Nov. 2019, pp. 76–87. [Online]. Available: https://www.aclweb.org/anthology/W19-8609

[31] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.