Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2236–2244
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Creating a Data Set of Abstractive Summaries of Turn-labeled Spoken Human-Computer Conversations

Virginia Meijer and Iris Hendrickx
Centre for Language Studies, Centre for Language and Speech Technology
Radboud University, Erasmusplein 1, 6500 HD, Nijmegen, The Netherlands
virginiapatricia96@gmail.com, iris.hendrickx@ru.nl

Abstract

Digital recorded written and spoken dialogues are becoming increasingly available as an effect of the technological advances such as online messenger services and the use of chatbots. Summaries are a natural way of presenting the important information gathered from dialogues. We present a unique data set that consists of Dutch spoken human-computer conversations, an annotation layer of turn labels, and conversational abstractive summaries of user answers. The data set is publicly available for research purposes. We used this data set to train and evaluate two transformer abstractive summarization models for Dutch dialogues. We briefly discuss the performance and errors of these summarization models and conclude that these models are not yet sufficient to be used in practice.

Keywords: automatic summarization, human-computer interaction, corpus creation

1. Introduction

The aim of summarizing a piece of text is to condense the important information into a shorter and fluently written text. Humans often reuse text from the original document. Through trimming, rephrasing, and reordering they compress the text into a summary containing the most important information (Jing and McKeown, 1999). In abstractive summarization, the condensed summary contains text that does not stem from the original document while extractive summaries consist of pieces of the original text. Gliwa et al. (2019) note that for certain text genres such as news articles and scientific publications a large number of high-quality data sets exist that include manually written summaries. However, for the emerging field of dialogue summarization high-quality data is less available. At the same time the authors also noted that the availability of raw digitally recorded dialogue data is increasing due to the growing popularity of messenger services and chatbots.

In this paper we focus on summarizing spoken human-computer dialogues. Ultimately we would like to create an automatic summarization system that can be integrated into a conversational agent. We are interested in building spoken conversational agents who talk with users about their daily life and well-being. The agent aims to create a personal profile that can provide insights in someone's health and happiness (van Waterschoot et al., 2020).

We believe that summaries are a natural way of informing the user about what information was stored by the agent during the conversation and to provide a mechanism for users to signal errors in the stored information. At the end of the conversation the agent summarizes the answers of the user to the user for both transparency of the data collection process and for user feedback (Hendrickx et al., 2021).

We created a data set of summaries as a first step towards the end goal of integrating summaries for transparency and as a feedback mechanism in conversational agents. This data set is unique as it consists of Dutch spoken human-computer conversations and is available for research. The conversations are manually annotated with task-specific turn labels to distinguish different sections in the dialogues, because the summaries focus only on the informative answers from the user. The summaries are formulated in a second person perspective, such that the summary can be read back naturally by the agent to the user.

We used this data set as training and evaluation material for an automatic abstractive summarization Transformer system of Dutch spoken human-computer conversations that we developed based on the approaches of Raffel et al. (2020) and Rothe et al. (2020). We refer to the work of Meijer (2021) for more details on the automatic summarization experiments as this paper first discusses other related data sets in Section 2. We describe dialogue data in Section 3 and the turn labeling and summarization process in Section 4. We provide an overview of the data set characteristics and some examples in 5. We briefly report on the results of the abstractive summarization experiments in Section 6 and conclude in Section 7.

2. Related work

As the state of the art in developing automatic abstractive summarization systems is progressing (Gupta and Gupta, 2019; Lin and Ng, 2019), we also observe an increased interest in dialogue summarization (Liu et al., 2019b; Liu et al., 2019a; Zhao et al., 2020). We briefly discuss three publicly available corpora that are used in abstractive spoken dialogue summarization research.

The ISCI meeting corpus provides 72 hours of recorded audio from multi-party meetings in English (Janin et al., 2003). In subsequent work, the authors added abstractive reference summaries, providing a foundation for future research in dialogue and multi-party summarization (Murray et al., 2006).

The AMI meeting corpus contains 100 hours of English spoken audio and video recordings (Mccowan et al., 2005). The corpus has been used for abstractive summarization by utilizing annotated topic descriptions as reference summaries in a recent study (Goo and Chen, 2018). However, the task moved from summarization to classification because several dialogues have the same topic.

The Spotify Podcast data set is a recently published corpus of Spotify owned-and-operated English spoken podcasts (Clifton et al., 2020). The corpus contains audio and automatically generated transcriptions of the podcasts. For summarization, the authors created two variants of the corpus that include abstractive reference summaries. Evaluation of these summaries provide a benchmark for future studies.

Shared characteristics between these corpora and our data set are the spoken conversation nature, the multiple participants in the conversations, and the inclusion of abstractive rather than extractive summaries. Our data set differs from these corpora as it consists of data other than English, contains human-computer conversations, uses turn labels to focus which parts are summarized, and focuses on user-answer summarization.

3. Data

We build our summarization data set on the basis of two already publicly available data sets of Dutch recorded spoken human-computer conversations: the JASMIN-CGN data set (Cucchiarini et al., 2006) and the BLISS data set (van Waterschoot et al., 2020).

All conversations in the different data sets are agent-initiative dialogues. Through its prompts, the conversational agent tries to elicit information about the user's daily life.

3.1. JASMIN-CGN

The JASMIN-CGN data set contains about 90 hours of two types of Dutch spoken material: read speech and mock human-machine dialogues (Cucchiarini et al., 2006). For our study, only 489 human-machine dialogues are considered. These dialogues were collected using a Wizard of Oz experiment. A person simulates the behavior of an artificial agent and converses with a human user. Importantly, the user is not aware that the artificial agent opposite them is also a person. Next to the audio material, the data set contains verbatim transcriptions for each dialogue.

3.2. BLISS

The data set collected by the BLISS project consists of 55 sessions between the BLISS conversational agent and a human user (van Waterschoot et al., 2020). The project aims to gather information about the health and well-being of the user through conversation. The dialogues relate to the user's daily activities and motivations for these activities. The dialogues average at a length of 2 minutes and 34 seconds. We also use ten recently recorded sessions where an agent and user discuss the effects of the COVID-19 pandemic on their health and well-being (average session length 12 minutes). We also use seven dialogues discussing childhood activities (average session length 4 minutes and 55 seconds).

4. Methodology

Our goal is to extend the above described data sets with hand-written abstractive summaries. The added summaries are vital for the evaluation of the automatic summarization transformer implementation that we developed. They serve as a 'golden standard' to compare the models' output summaries to. We use manual dialogue transcripts of the different data sets, because automated transcripts would negatively impact the quality of the summaries.

4.1. Data pre-processing

The JASMIN-CGN dialogues are split into two parts following a rule-based method. One part of the dialogue discusses what a trip planned by the user would look like. The second part discusses daily activities of the user and the motivations for these activities. The second part of the dialogues are used for summarization. The content of the BLISS dialogues are not adjusted. The BLISS data set only contained automatically created transcriptions and were manually transcribed to match the manual JASMIN-CGN transcriptions.

4.2. Turn labeling

The key challenge when manually writing a summary is to determine which information is important and should be included in the summary. We use turn labeling to focus on only these informative dialogue parts.

Within the dialogues, a turn consists of several utterances that are spoken directly after each other. Most turns of the dialogues are part of a question-and-answer (QA) pair. In agent-initiative dialogues, the questions mostly come from the agent's side and elicit answers from the user's side. Turns not belonging to a QA pair are single turns. These might still provide important information even if they are not elicited directly after a question.

Each dialogue is annotated manually by splitting the dialogue into QA pairs and single turns, and labeling these content units. As suggested by Lemon and colleagues (Lemon et al., 2003) we distinguish between informative content-level and interaction level communication processes and we divided our turn labels into these two coarse categories. The first category contains informative labels, in our case study the content relates to information the user shares about their own life. We broadly followed the type of turn labels that were used for the BLISS data (van Waterschoot et al., 2020). Note that we use domain and project-specific informative turn labels related to the daily activities and motivations of the user.

The second category contains other, generic labels pertaining to interaction communication such as greetings and conversation flow problems. The turn labels provide information along three dimensions, which we detail in Section 4.4.

4.3. Summarizing

The annotations are used to filter the turns that do not belong to health and well-being information. With the remaining content units a summary of the user's daily activities and motivations is written. We also do not include any of the typical speech characteristics like unclear words or laughs in the summaries.

From the 557 dialogues in the combined data sets, 459 dialogues are annotated and summarized by a single annotator (and are used as training material in our future automatic summarization experiments). The remaining 98 dialogues are annotated and summarized by two annotators (and serve as held-out test material). Using the test set, we computed inter-annotator agreement as precision on turn labeling and obtained a precision of 0.73 between both annotators. We computed unigram overlap, a ROUGE-1 F1 score of 54.2, between the two summaries to get an indication of annotator agreement. We observe that annotator one consistently wrote shorter summaries including less information than annotator two.

The goal of these summaries is that they can be repeated back to the user. Through the summary the user is informed about the data that is collected and the user will be able to verify the collected data. To achieve this, the summaries are written in informal Dutch from the second singular person perspective. This perspective is used by the agent in the original dialogues and will preserve the flow of the conversation. Additionally, to provide a summary that the user can recognize and verify, the summaries follow the user's own word choice.

4.4. Turn label dimensions

The turn labels provide information along three dimensions:

- Specificity: The category of the label denotes the specificity of the turn. Informative content labeled turns indicate the domain-specific information about the user's daily activities and motivations for these activities. Interaction communication labeled turns occur more generally in dialogues from any domain. For example, greetings and misunderstandings between speakers are not domain-specific information.
- Number of participants: This dimension notes whether a single participant is speaking (single turn) or whether there is an interaction between the human user and the artificial agent (QA pair). The participants cannot have a QA pair with themselves as a QA pair denotes an interaction. Single turns can be spoken by the user or the agent.
- Repetition: This dimension denotes if the information provided in a turn has been shared in an earlier part of the conversation. Information can be repeated because of a prompt by the agent such as "can you rephrase that?". Information can also be repeated because the user returns to a subject later in the conversation.

All turn labels convey information along these three dimensions, but some combinations do not occur in the text. We detail all turn labels in Appendix A.

5. Results & discussion

In total, the data set created in this project contains 68 dialogues from the BLISS data set and 489 split dialogues from the JASMIN-CGN data set. We have a total of 685 summaries, as for 98 dialogues we have a second annotation and summary. This Dialogue Summaries Dataset is publicly available for research purposes here: `http://hdl.handle.net/10032/tm-a2-r5`.

Tables 1 and 2 show the characteristics of the annotated BLISS human-machine conversations and JASMIN-CGN Wizard of Oz human-machine conversations. In both corpora the user speaks slightly less than the agent. The BLISS summaries have a higher number of tokens (words) on average. This is because the BLISS COVID-19 dialogues are much longer than the other BLISS dialogues and have longer summaries. However, the compression rate (the size of the summary in relation to the original text) is similar between the two corpora, which indicates that the amount of relevant information for the summary elicited from the user is similar in both corpora.

Data characteristics	**Size**
# of dialogues	68
# of turn labels	1243
# of tokens in user turns	9521
% tokens spoken by user	40%
# summaries	136
# av. tokens in summaries	90
compression rate	24%

Table 1: BLISS statistics

We show an example of the summarized form of a dialogue taken from the BLISS data in Example 1. As you can see the summary is written from a second person perspective. Note that although the summary is focused on conveying the user answers, the agent prompts are also present in the summary.

Data characteristics	Size
# of dialogues	489
# of turn labels	7866
# of tokens in user turns	57007
% tokens spoken by user	43%
# summaries	549
# av. tokens in summaries	58
compression rate	21%

Table 2: JASMIN-CGN statistics

(1) Uit eten gaan is een activiteit die je dit weekend graag zou willen doen. Je vindt het leuk om uit eten te gaan omdat je van lekker eten en gezelligheid houdt. Iets anders wat je dit weekend graag zou willen doen is met je gezin wandelen in het bos. Als je niet zou kunnen wandelen in het bos, zou je de natuur het meeste missen. Afgelopen week was je in Ibiza en vond je aan het strand liggen het allerleukste. ('Going out for dinner is an activity that you would like to do this weekend. You like going out for dinner because you love good food and good company. Something else that you would like to do this weekend is taking a walk in the forest with your family. If you could not walk in the forest you would miss nature the most. You were in Ibiza last week and you enjoyed laying on the beach the most.')

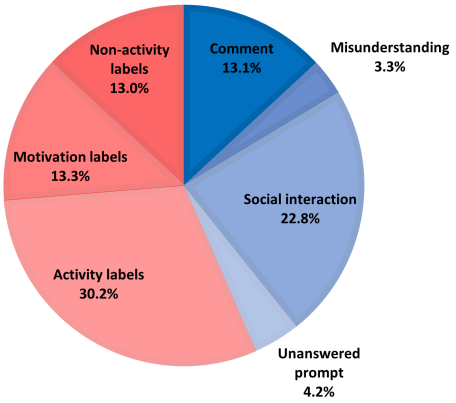

Figure 1: BLISS turn label distribution. Blue-tinted slices represent interaction communication labels and red-tinted slices indicate informative content labels.

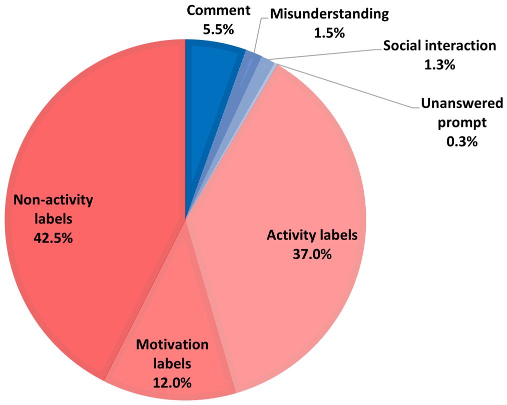

Figure 2: JASMIN-CGN turn label distribution. Blue-tinted slices represent interaction communication labels and red-tinted slices indicate informative content labels.

The BLISS dialogues (Figure 1) contain more comments and misunderstandings than the JASMIN-CGN dialogues (Figure 2). The BLISS dialogues also contain more social interactions. We only used the second part of the JASMIN-CGN dialogues in this data set that did not include the start of the conversation, and therefore fewer greetings are present in these dialogues. Furthermore, most turns in both corpora are labeled with a domain-specific informative content label. Figure 2 shows that especially in the JASMIN-CGN dialogues most turn labels are informative and provide information about the user not relating to their daily activities or motivations. An example of this is that in the JASMIN-CGN dialogues, users are asked what their favorite food is. Users often share this information but do not relate the information to an activity such as going out for dinner.

Looking at the information that is considered relevant for the summaries in this project, Figures 1 and 2 show that in both BLISS and JASMIN-CGN dialogues a high percentage of turns are labeled as activity or motivation. The overlap in these high percentages show that the content of the dialogues is indeed comparable. Together with the overlapping statistics, the similarities between the corpora show that combining the two was a fruitful design choice.

During annotation of the dialogues, we encountered several situations that did not fit the annotation scheme. These difficulties arose because the intent of the user was not always clear. This especially occurred when the users answered a prompt with "yes". It can be used as an affirmative response or as an indication that the user is thinking. Such ambiguities complicated the annotation process of the dialogues.

Additionally, the specificity of the informative labels narrows the scope of the label types. This data set centers on health and well-being dialogues, using very domain-specific informative labels. In other domains, these labels might not cover all the content that is discussed in a dialogue. In contrast, the interaction communication labels can occur in dialogues of any topic

due to their generality.
Finally, for model development, we have divided the dialogues and summaries into a training set and a test set. The training set comprises 459 extended JASMIN-CGN dialogues that all contain annotations and summaries by a single author. The test set consists of 30 extended JASMIN-CGN dialogues and 68 extended BLISS dialogues. These dialogues contain annotations and summaries by two authors. We feel that a second summary supports more comprehensive evaluation analysis of automatically generated summaries.

6. Automatic Summarization

We used the corpus for training and evaluation of an automatic summarization system. We refer to Meijer (2021) for all details on these experiments as here we only briefly describe the main outcomes. We created two transformer abstractive summarizers for Dutch dialogues. We implemented a T5 transformer using the multilingual pre-trained mT5 model (Xue et al., 2020) and we leveraged the Dutch pre-trained model RobBERT V2 by Delobelle et al. (2020) to implement a RoBERTaShare transformer (here refered to as RobBERTShare).
For evaluation we used the 3-longest user utterances as baseline summaries following the approach of Zhao et al. (2020). We fine-tuned the pre-trained transformers on the training set and evaluated the generated summaries against the manual summaries in the test set computing ROUGE scores. Note that we have two different summaries for each dialogue in the test set. We present the results in Table 3. The overall ROUGE scores are generally higher for the summaries of the first author. A possible contributing factor to this is that the the first author also wrote all reference summaries in the training set. The models could have picked up author-specific stylistic features, resulting in a higher overlap between summaries and in a higher ROUGE score.
As ROUGE is not an optimal evaluation metric for abstractive summarization, we also performed a manual error analysis on 30 dialogue summaries to investigate what type of errors the transformer summarizers are making. We observed that the most typical errors are fabrication errors, leading to summaries containing information that was not present in the actual original dialogues. This type of information fabrication is one of the known problems of this type of abstractive automatic summarizers. Additionally, summaries generated by the mT5 transformer are often incoherent as they contain short sentences that cut off after a conjunction. In contrast, summaries generated by the RobBERTShare transformer are generally closer to the reference summaries in length and fluency. However, the fabrication errors render the summaries unusable as a means to inform the user.
We conclude that the quality of the automatically generated summaries is currently insufficient to be readily integrated in a conversational agent as we had intended.

7. Conclusion & future work

We described the creation of a data set consisting of 557 Dutch human-computer conversations manually annotated with turn labels and abstract summaries of the user's answers.
We have built an automatic transformer summarizer using the summarization data set for tuning and testing of the models (Meijer, 2021). Ultimately we aim to create summaries of the key information disclosed by the users in a human-machine interaction setting. This summary is written from a second person perspective and may serve as a verification point for the user. Providing a summary of the collected data to the user will inform the user what data is collected and simultaneously provide the user with the opportunity to correct erroneous data.
However, this data set can also be useful for other research purposes such as developing methods for automatic dialogue turn splitting and turn labeling. This data set can also be studied from a discourse oriented perspective, for example to look at dialogue flow or human-computer misunderstandings. The dialogues and annotations can also be used to research whether the use of a conversational agent in the BLISS project and a Wizard of Oz setup in the JASMIN-CGN project contributes to differences in the content of the dialogues.

8. Acknowledgments

This work is part of the research programme Data2-Person with project no. 628.011.029, which is (partially) financed by the Dutch Research Council (NWO). We thank Zjoske Koks for helping out with the annotation and we would like to thank the reviewers for their valuable comments.

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G. J. F., Karlgren, J., Carterette, B., and Jones, R. (2020). 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, December. International Committee on Computational Linguistics.

Cucchiarini, C., Van Hamme, H., Van Herwijnen, O., and Smits, F. (2006). JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 135–138, Genoa, Italy. European Language Resources Association (ELRA).

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265,

Approach	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Summaries first author			
Longest-3-user	28.24	**11.21**	24.54
mT5	21.33	7.88	22.56
RobBERTShare	**32.60**	9.21	**28.42**
Summaries second author			
Longest-3-user	**29.57**	**12.00**	**26.27**
mT5	16.40	4.28	17.79
RobBERTShare	24.95	3.67	22.34

Table 3: Evaluation of the generated summaries with ROUGE metrics. The reported ROUGE F1 scores are averaged over the 98 test set dialogues. Best performing model scores are reported in bold text.

Online, November. Association for Computational Linguistics.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (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. Association for Computational Linguistics.

Goo, C. W. and Chen, Y. N. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Gupta, S. and Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Hendrickx, I., van Waterschoot, J., Khan, A., ten Bosch, L., Cucchiarini, C., and Strik, H. (2021). Take Back Control: User Privacy and Transparency Concerns in Personalized Conversational Agents. In *Proceedings of the ACM IUI 2021 Workshops*, page 6, College Station, Texas, USA, April. CEUR-WS.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, pages I–364–I–367, Hong Kong. IEEE.

Jing, H. and McKeown, K. R. (1999). The decomposition of human-written summary sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136, New York, USA.

Lemon, O., Cavedon, L., and Kelly, B. (2003). Managing dialogue interaction: A multi-layered approach. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 168–177.

Lin, H. and Ng, V. (2019). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9815–9822.

Liu, C., Wang, P., Xu, J., Li, Z., and Ye, J. (2019a). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965. Association for Computing Machinery.

Liu, Z., Ng, A., Lee, S., Aw, A. T., and Chen, N. F. (2019b). Topic-Aware Pointer-Generator Networks for Summarizing Spoken Conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Mccowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI Meeting Corpus. In L. P. J. J. Noldus, et al., editors, *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology.

Meijer, V. (2021). Automatically summarizing dutch human-machine dialogues using transfer learning approaches. Master's thesis, Radboud University, Nijmegen,The Netherlands.

Murray, G., Renals, S., Carletta, J., and Moore, J. (2006). Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 367–374, New York City, USA, jun.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Peter, W. L., and Liu, J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

van Waterschoot, J., Hendrickx, I., Khan, A., Klabbers, E., de Korte, M., Strik, H., Cucchiarini, C., and Theune, M. (2020). BLISS: An agent for collecting spoken dialogue data about health and well-being. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 449–458. European

Language Resources Association.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer.

Zhao, L., Xu, W., and Guo, J. (2020). Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449. International Committee on Computational Linguistics.

A. Turn Labels

The labels are divided into two broad categories: informative content labels and interaction communication labels. The informative content labels include:

- Activity information interaction: In the QA pair, the user divulges information about their daily activities as a response to the agent's prompt.
- Repeat activity information interaction: In the QA pair, the user repeats information about their daily activities as a response to the agent's prompt. The information has already been disclosed earlier in the conversation.
- Motivation information interaction: In the QA pair, the user divulges information about the motivation for their daily activities as a response to the agent's prompt.
- Repeat motivation information interaction: In the QA pair, the user repeats information about the motivation their daily activities as a response to the agent's prompt. The information has already been disclosed earlier in the conversation.
- Non-activity information interaction: In the QA pair, the user divulges information about himself that does not relate to activities or motivations as a response to the agent's prompt.
- Activity comment: The single turn contains a comment about an activity of the user without a prompt from the agent.
- Motivation comment: The single turn contains a comment about the motivation for an activity of the user but does not follow a prompt of the agent.

The following labels fall into the second category of interaction communication labels:

- Social interaction: In the QA pair a social interaction occurs between the user and the agent in response to the agent's prompt. Examples include greetings and farewells.
- Misunderstanding: In the QA pair, the user signals a misunderstanding of the agent's prompt such as asking the agent to repeat or reformulate the prompt.
- Comment: The single turn does not provide information about the user's motivations or activities but does contain meaning. Examples include a comment on the dialogue system or a lack of response from the user to the agent's prompt.
- Filler: The single turn solely contains utterances that are transcribed but do not contain the expected elicited intent such as "Uhm. Well uh uh uh. Mm.".

Dutch dialogue	**English dialogue**
SOCIAL INTERACTION Agent: Hallo! Heb je wel eens tegen een computer gepraat die terug praat? User: Uh. Uh. Nee	SOCIAL INTERACTION Agent: Hi! Have you ever spoken with a computer that talks back? User: Uh. Uh. No
SOCIAL INTERACTION Agent: Oh, dat is toevallig, ik heb ook nooit eerder met mensen gepraat. Hoe vind je het om zo met mij te praten? User: Bijzonder.	SOCIAL INTERACTION Agent: Oh, that is a coincidence, I have never spoken with people before. How do you like talking with me like this? User: Special.
ACTIVITY INFORMATION INTERACTION Agent: Ik ga er van blozen! Nu ben ik wel nieuwsgierig naar jou! Als je nu één ding mocht kiezen wat je komend weekend zou willen doen. Wat zou dat zijn? User: Uit eten.	ACTIVITY INFORMATION INTERACTION Agent: You are making me blush! Now I am curious about you! If you could now choose one thing you would like to do this weekend. What would that be? User: Going out for dinner.
MOTIVATION INFORMATION INTERACTION Agent: Klinkt goed. Waarom vind je dat leuk? User: Ik hou van lekker eten en gezelligheid.	MOTIVATION INFORMATION INTERACTION Agent: Sounds good. Why do you like that? User: I like good food and good company.
ACTIVITY INFORMATION INTERACTION Agent: En als je nog een tweede activiteit zou mogen kiezen? Wat zou dat zijn? User: Met mijn gezin uh wandelen in het bos.	ACTIVITY INFORMATION INTERACTION Agent: And if you could choose a second activity? What would that be? User: Taking a walk with my family in the woods.
MOTIVATION INFORMATION INTERACTION Agent: Wat zou je het meeste missen als je wandelen niet zou doen? User: Uhm de natuur.	MOTIVATION INFORMATION INTERACTION Agent: What would you miss most if you could not take a walk? User: Uhm nature.
ACTIVITY INFORMATION INTERACTION Agent: Als je naar de afgelopen week kijkt. Welke activiteit vond je het allerleukst? User: Ik ben uh net terug uit uh ibiza dus ik vond aan het strand liggen het uh allerleukste afgelopen week.	ACTIVITY INFORMATION INTERACTION Agent: If you look at the past week. Which activity did you like the most? User: I just uh got back from uh Ibiza so I enjoyed lying on the beach the uh most last week.
MISUNDERSTANDING Agent: Tof! Wat voor gevoel geeft voelen jou? User: Wat voor gevoel geeft voelen jou ik snap het niet helemaal kun je te herhalen of anders formuleren. Ok.	MISUNDERSTANDING Agent: Great! How does feeling make you feel? User: How does feeling make you feel I do not really understand can you repeat it or reformulate it. Ok.
SOCIAL INTERACTION Agent: Dank je! Dat was het voor nu. Denk je nog steeds hetzelfde over mij? User: Ja ik vind het nog steeds bijzonder.	SOCIAL INTERACTION Agent: Thank you! That is it for now. Do you still think the same about me? User: Yes I still think it is special.
SOCIAL INTERACTION Agent: Leuk dat je hebt meegedaan! User: Ja dank je wel.	SOCIAL INTERACTION Agent: Thanks for participating! User: Yes thank you.

Table 4: An example of an annotated dialogue and English translation.