Intentional Chatting: Mapping Automatic Evaluation Metrics to High-Level Conversational Intentions in Artificial Agents*

Marije Brandsma and **Bas Diender** and **Mekselina Doğanç**
and **Caroline Hallmann** and **Alice Ye**
Vrije Universiteit Amsterdam

1 Introduction

Conversations are diverse and many factors may influence them, such as separate (sub-)goals, the relationship that exists between the different participants, and their behavior. In a time where conversational AI is increasingly implemented, the need for automatic evaluation arises. However, it is hard to evaluate these conversations automatically. They might not understand the context or cannot interpret creativity in responding. As a result, most evaluations are done by humans, which is very time-consuming.

In this paper, we explore the capabilities and challenges of automatic dialogue evaluation metrics. Two datasets were compared: DialyDialog and ConvAI2. For these datasets, the conversational flow and the agent's memory were analyzed, based on the intention of the conversation. These intentions are identified with the speech acts of Searle (1976) in mind. This way, we aim to give insight to the possibility of automatically mapping evaluation metrics to high-level conversational intentions.

We define the intentions of utterances and conversations through the speech acts as proposed by Searle (1976), a linguistic theory broadly accepted and used to classify illocutionary speech acts. This theory describes five different intentions in a conversation. These speech acts can identify the purpose of a conversation or conversational agent.

First, the *assertive* (or representive) speech act describes an utterance with the attention of the speaker to represent their knowledge about the state of the world. This can contain factual information ("Amsterdam is the main capital of the Netherlands") or information about agents in them ("Tom bought a new shirt"). Secondly, when a speaker commits themselves to an action ("I will wash the dishes"), this is considered as a *commissive* speech act. The *directive* speech act is closely related, but it refers to the speaker attempting to make the addressee do something ("Would you wash the dishes?). *Declarations* refer to an act of the speaker by solely pronouncing the utterance ("I now declare you man and wife" or "I nominate Lisa as our new class president"). Finally, an *expressive* speech act expresses the speaker's attitude about objects and facts about the world ("I don't like Tom's new shirt", "Apeldoorn should be the main capital of the Netherlands").

These speech acts describe the different intentions with which an (artificial) conversational agent can communicate. As an extension, we assume that the speech act that is present the most in a conversation can be considered the intention of this conversation. For example, if the intention is to share information, it is expected that the assertive speech act is most present.

In chapter 2, the method that was used for this paper will be explained. This includes the datasets that were used, the dialog act classification model, pre-processing steps, the automatic evaluation method, and the analysis of the results. Chapter 3 will present these results, talking about the conversation flow and agent's memory of the two different datasets. Finally, chapter 4 and 5 will explain the results, make observations based on them, and identify challenges that still lie ahead of us.

2 Method

2.1 Datasets

2.1.1 DailyDialogue (DD)

The dataset DialyDialogue (Li et al., 2017) consists of dialogues for exchanging information and

[0] https://github.com/cltl/ma-communicative-robots/tree/2022/projects/intentional_chatting/

enhancing social bonding on a daily based chit-chat. The dialogues are human-written with always two people participating in bi-turn flows with Questions-Inform and Directives-Commissives, which are also the 4 intention classes in the dataset. As with the dataset EmoryNLP, DD is also rich in emotion but the annotations, in this case, are emotions of the Big Six Theory: anger, disgust, fear, happiness, sadness, surprise and one for "other" emotions. The largest three categories the dialogues develop around are relationships (33.33%), ordinary Life (28.26%), and work (14.49%). There are a little more than 13k multi-turn dialogues in this dataset, with approximately 8 turns on average.

2.1.2 ConvAI2

The ConvAI2 dataset (Dinan et al., 2020) was used during the Conversational Intelligence Challenge conducted under the scope of NeurIPS and is based on the `Persona-Chat`: "The speaker pairs each have assigned profiles coming from a set of 1155 possible personas (at training time), each consisting of at least 5 profile sentences, setting aside 100 never seen before personas for validation. Person 1 is given their own persona at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation". The competitors additionally crowdsourced a test set for automatic evaluation consisting of 100 new personas and over 1,000 dialogues. Hence, the complete dataset has almost 20k dialogues.

2.2 Models

2.2.1 `MIDAS`

`MIDAS` is a dialog act prediction model, originally created to assist machines in conversations with human partners (Yu and Yu, 2019). `MIDAS` identifies a hierarchical set of 23 dialog acts, that can be found in table 1. With a reported F1-score of 0.79, this model performs well in identifying dialog acts in conversations. These classes have been used to identify which of the speech acts of Searle (1976) are relevant for a conversational turn. For an overview of the dialog acts that `MIDAS` uses, see (Yu and Yu, 2019).

These classes have been mapped to speech acts with conversational intentions in mind. This means that, for example, a factual question is assumed to be assertive. Even though a question is not considered to be an assertive speech act, it does aim toward an assertive intention of the conversation as a whole. Besides the five speech acts identified by Searle (1976), the category 'other' has been added to account for instances where a dialog act cannot be fitted in any of the other speech acts, as is the case for the class "nonsense" (language that is uninterpretable) and "hold" and "abandon", that serve a purely technical purpose. The mapping can be found in table 1 as well. As can be observed, a distinction between assertive and expressive speech acts can not always be made based on the `MIDAS` classifications.

As can be observed in table 1, the commissive speech act has not been mapped to any dialog acts. There was no classification label that could be considered to be commissive, something that is rare to occur in an online conversation with a chatbot, as they do not tend to give commands to the other party. Human conversers also do not make promises to chatbots. Because of this, the commissive speech acts are not included in the mapping. However, in the analysis the commissive speech act, even though not present, is named. This decision was made to keep the theory Searle (1976) complete and not split it up.

2.3 Pre-processing

The first step was to convert the text in the datasets into knowledge graphs and to assign a speech act to each sentence in the datasets. To this end, each dataset was first converted to an intermediate structure where each row contained an utterance, a speaker for that utterance, and some identifier to link it to a dialog, the latter two were generated arbitrarily if they were not already present in the source data. Instances, where a dialog had more than two speakers, were omitted. In the next step, utterances were annotated for one of twenty-three speech acts using the MIDAS dialog act prediction model (Yu and Yu, 2019), and if possible split into subject-predicate-object triples using the spacy-based analyzer from the triple extraction library by CLTL[1]. These triples formed the basis for the knowledge graphs, with the subjects and objects being nodes and the predicates the edges linking them.

With each sentence in the dataset having been converted into a data structure that can easily be turned into graphs, the final pre-processing step looked at the data on a dialog-by-dialog basis. Then, starting at the first turn in the dialog, a graph

[1] https://github.com/leolani/cltl-knowledgeextraction

Table 1: Mapping of Yu and Yu (2019) dialog acts to Searle (1976) speech acts.

MIDAS dialog acts	Searle's speech acts
factual open question	assertive
positive answer	assertive, expressive
command	directive
opinion	expressive
statement	assertive, expressive
back channeling	other
yes/no question	assertive, expressive
appreciation	expressive
other answers	assertive, expressive
thanking	expressive
opinion open question	expressive
hold	other
closing	declarative
comment	assertive, expressive
negative answer	assertive, expressive
complaint	expressive
abandon	other
dev-command	directive
apology	expressive
nonsense	other
other	other
opening	declarative
respond to apology	expressive

was initialized, if a triple was extracted from the utterance at that turn of the dialog, it was added to the graph. The graph was then saved to a file and passed on to the next turn in the dialog, where this process was repeated. The end result was a set of increasingly informative graphs for every conversation in the dataset.

2.4 Automatic evaluations

Episodic knowledge graph evaluation (eKG) (Báez Santamaría et al., 2022) is a reference-free and explainable method for evaluating information and knowledge conveyed during dialogues via episodic knowledge graphs. Therefore, the approach is suitable to evaluate multiple dialogues generated by different interlocutors. Provided the datasets used in the current report are from different participants and the aim is to characterize conversations with multiple variables in the datasets, this evaluation metric meets the requirements. Based on the properties and attributes from knowledge graphs formatted in OWL, eKG is mapped to generate three groups of metrics with submetrics. Group A focuses on mathematical aspects such as measuring the average degree, the number of components, centrality entropy, and sparseness. Group B is about the semantic parts for example average population of the extracted triples. Group C indicates the integrity of the conversation, e.g. ratio claim to triples and ratio perspective to claims.

To implement the evaluation, we use the CLTL library for dialogue evaluation.[2] The tool uses *rdflib* to map through eKGs and generates all submetrics in the Appendix (Table 2) in CSV files and creates visual plots for the submetrics mentioned above in the paragraph.

2.5 Analysis

The analysis consisted of two main parts. One, conversation flow, and two, the agent's memory. To evaluate the conversational flow of DD and ConvAI2, sparseness and average degree scores were calculated. For the agent's memory, MIDAS dialog acts were mapped onto Searle's speech acts and evaluated respectively.

2.5.1 Sparseness

For DD and ConvAI2, corresponding eKGs were produced to measure the level of sparseness in the

[2] https://github.com/selbaez/evaluating-conversations-as-ekg

given conversations. Sparseness is defined as the ratio of the number of edges in the graph to the maximum number of edges the graph could have, both in terms of two nodes being connected or not, and in terms of different types of edges. Sparseness can either go up by adding new nodes to the graph or by adding different types of edges to the graph. As such, a dialog with a sparse graph is one where new information is added without it necessarily being linked to the information that was already discussed. Thus, it measures how semantically connected a given dialog is (Báez Santamaría et al., 2022).

2.5.2 Average degree

The flow of a conversation is calculated through the average node degree. It counts the incoming and outgoing edges per node in a knowledge graph (Báez Santamaría et al., 2022). In theory, Human-Human conversation is most fluent, which is shown by a steep downward slope on the knowledge graph. Machine-machine interaction is the least fluent, having a steep upward slope (Báez Santamaría et al., 2022).

2.5.3 Intention classifier

For the agent's memory, 23 MIDAS speech acts were evaluated from the DD and ConvAI2 datasets. First, MIDAS dialog acts were mapped to 6 Searle's speech acts (see table.1). During the mapping process, no match was identified for the commissive Searle speech act. Thus, generating an overall score of 0 during the evaluation of both datasets. After mapping all the speech acts in all turns, the average intention score per conversation was calculated. The most occurring Searle speech acts were stored and calculated to get the dataset distribution. This procedure was applied for both datasets, DD and ConvAI2. Once the dataset distribution scores of intentions were given, two pie charts were created for the descriptive statistical analysis procedure.

3 Results

After conducting the evaluation and analysis for DD and ConvAI2, eKGs and statistical figures were produced to show the trends of conversation flow and evaluate the intentions of an agent's memory.

3.1 Conversation flow

Two types of eKGs were produced to evaluate the conversation flow in DD and ConvAI2. First, the sparseness eKG and second, the average degree score eKG.

3.1.1 Sparseness

The change in graph sparseness per turn in the dialog for DD and ConvAI2 are shown in figure 1, with each line representing one dialog. As stated, Báez Santamaría, Vossen, and Baier (2022) found sparseness in eKGs to be correlated with the overall human evaluation of a dialog, with a downward trend of sparseness per turn being linked to a more successful dialog, as it indicates an increased semantic interconnectivity of the concepts in the dialog. In both datasets, the sparseness per turn follows a similar trend in all dialogs in both datasets: A conversation starts out with a sparseness of zero, since the knowledge graph does not contain any nodes yet, and consequently, cannot contain any edges. Then, once the triple extractor has found the first triple, the sparseness goes up to around 0.02. The sparseness then drops at a similar rate for all conversations. Some outliers include flat lines at a sparseness of 0 or 0.02, indicating that the triple extractor failed to extract (more than one) triple, and as such did not update the knowledge graph in the course of the dialog. These results indicate that the dialogs in both datasets were generally successful with little difference between the two datasets.

3.1.2 Flow

Figure 2 shows the average degree of the eKGs produced from the DD and ConvAI2 data. The average degree represents the number of edges per node in the DD and ConvAI2 eKGs. The average degree curves from the DD knowledge graph (left) do not follow a linear trend and show different degrees of fluency.

Out of the 100 conversations in DD, only six follow a steep downward slope represented by the bottom lines starting at 2.5 average degrees. The eKG shows outliers as nonfluctuating downward curves, which can not be accounted for flow. The majority of curves show a steep increase in the average degree as the conversation proceeds. This trend may be due to the repetitive nature of the chatbot. After several turns, it starts to repeat itself which then leads to a drastic increase in the average degree. This is evidently shown by the blue increasing to an average degree score 3.3 at turn 18.5.

In comparison, the ConvAI2 average node degree eKG showed less fluency than in the DD eKG.

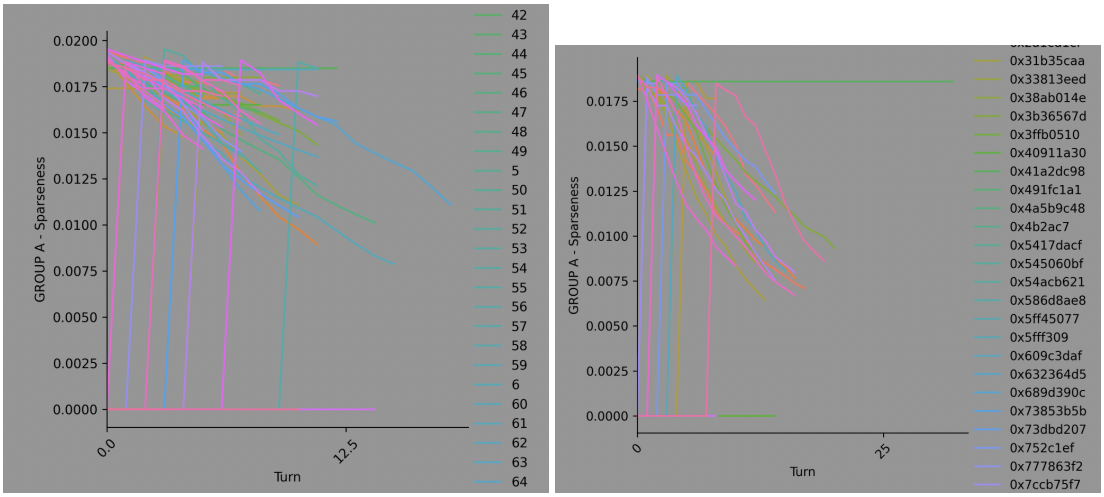

Figure 1: Sparseness per turn for DailyDialogue (left) and ConvAI2 (right)

When looking at the slopes, the majority follow a steep upward trend, representing low fluency. Furthermore, the average degree scores are higher in ConvAI2 than in DD. These scores may also be explained due to the repetitive nature of the chatbot. Approximately 4 ConvAI2 conversations follow a fluent trend at an average degree score of 3.6. As the number of turns increases, the green, pink, purple and orange slopes start to become steeper and follow a downward trend, indicating that the conversation becomes more fluent over the number of turns.

3.2 Agent's memory

After mapping the identified speech acts from DD and ConvAI2 to Searle's intentional speech act paradigm, the results were used to generate the corresponding pie charts in figure 3.

Comparing the results of the two datasets, a variation between intention scores can be identified. Overall, ConvAI2 had a higher distribution of intentions than DD. Declarative speech acts occur more frequently in ConvAI2 with a score of 17 % compared to DD with a score of 1 %. Directive speech acts only account for 2 % in ConvAI2 whereas for 10 % in DD. The category 'other' scored higher for ConvAI2 with 18 % compared to 5 % in DD.

Despite these differences, positive correlations are found for two intentions: 1) Expressive and 2) Assertive. Both, DD and ConvAI2 scored highest on expressive and assertive speech acts. DD had 48 % expressive speech acts and ConvAI2 had 39 %. A total share of 36 % were assertive speech acts in DD and 24 % in ConvAI. Furthermore, both datasets scored 0 on the commissive intention as it was not classifiable. This shows that only five intentions were fulfilled throughout the interactions.

4 Discussion

Overall, the average degree scores of DD and ConvAI2 show low levels of fluency. The increase of average node degrees over turns shown through the steep upward slopes in both eKGs may be due to the repetitive nature of the chatbot. After several turns, it starts to repeat itself which then leads to a drastic increase in the average degree.

The results of the sparseness analysis showed that the trend of sparseness per turn was similar in both the DailyDialog and ConvAI2 datasets. This suggests that the fluency of a dialog, as indicated by the knowledge graph sparseness per turn, may not be significantly affected by the specific setup of each of the datasets. The initial increase in sparseness followed by a drop may be a general pattern observed in dialogs, as the knowledge graph suddenly goes from empty to not empty, and then gradually becomes more interconnected as the conversation progresses.

After evaluating the intention scores of DD and ConvAI2, some intentional speech acts occurred more frequently than others. Both datasets did not classify any commissive speech acts, which may be explained by the nature of human-robot interaction. Chatbots are not likely to give commissive speech acts such as promises, oaths, pledges, threats, or vows, but rather adapt more interactive intentional speech acts such as ones of assertive and expressive nature. The results show that the most frequently identified intentions were assertive and expressive in DD and ConvAI2. These find-

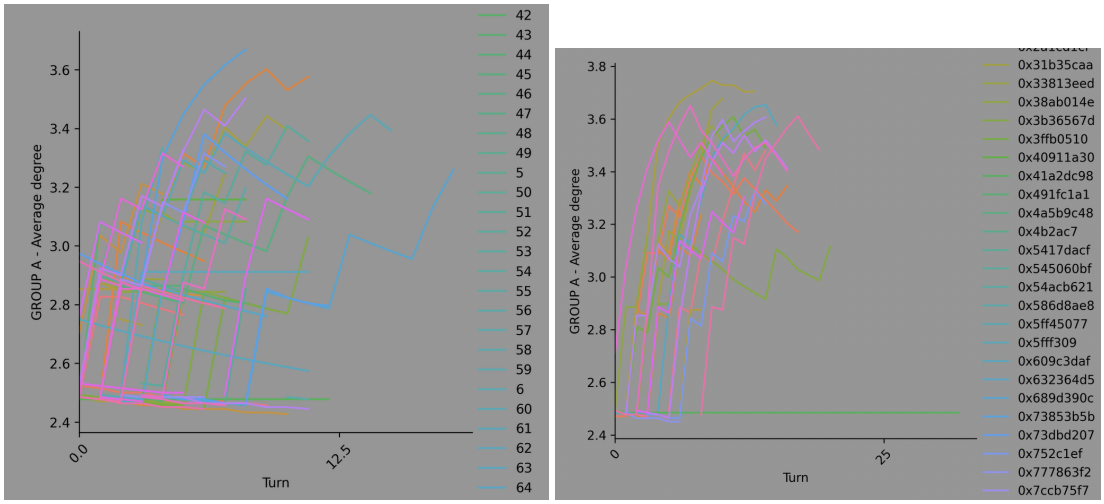

Figure 2: Average degree per turn in the DailyDialogue (top) and ConvAI2 (bottom) dialogs.

Figure 3: DailyDialogue (top) and ConvAI2 (bottom) Searle Intentions Distribution

ings show a positive correlation between the two datasets and may be explained by the dynamic nature of dialogue. Furthermore, the 5 intentions used in ConvAI2 were more normally distributed than in DD. The difference in intentions is more severe in DD than ConvAI2, as the percentage of expressive and assertive intentions is highest in DD.

Originally, 8 datasets were researched for implementation purposes. However, due to various limitations, only three datasets were loaded, of which only two were evaluated. The datasets that were not evaluated can be found in appendix A. In the future, it is important for the six remaining datasets to be evaluated to obtain insight in a wider variety of dialogs – including human-machine and machine-machine dialogs – in order to achieve results with higher reliability and validity.

After evaluating the intentions of DD and ConvAI2, the high scores of the assertive and expressive speech acts may be due to the difficulty of distinguishing between these two speech acts. When mapping the MIDAS dialog acts to Searle's speech acts, it is difficult to assign the correct speech act accordingly. Thus, in the future, the dialog acts should be mapped onto different types of speech act paradigms. By mapping, the dialog acts as multiple intention classifiers, scores can be cross-validated.

5 Conclusion

In this report, we conducted an analysis on intentional chatting with a chatbot. We made use of the automatic evaluation matrices and selected a speech act classifier to categorise intentions for two types of datasets. At the start of the project, it was required to conduct research on NLP datasets that are used for classifying speech acts. We found 8 datasets and chose DailyDialogue(DD) and ConvAI for our exploratory analysis. Next, we also looked at speech act classifier and instead of using the 4 intentions defined in DD, we decided to adapt the Searle's speech act classifier for our report. We used libraries from CLTL to extract triples to generate knowledge graphs and to measure the graph of each dialogue to make evaluation metrics. The results showed that both ConvAI and DD succeeded in carrying out most of the conversations according to the tendency of reduction in sparseness, but

neither of them have high fluency. Overall, ConvAI2 has a higher distribution of intentions than DD, and the most frequently identified intentions were assertive and expressive in DD and ConvAI2.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

Selene Báez Santamaría, Piek Vossen, and Thomas Baier. 2022. Evaluating agent interactions through episodic knowledge graphs. *arXiv preprint arXiv:2209.11746*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ–A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *arXiv preprint arXiv:1810.00278*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The Second Conversational Intelligence Challenge (ConvAI2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-Turn Dialogue Dataset. *arXiv preprint arXiv:1710.03957*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. *arXiv preprint arXiv:2006.12719*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset. *arXiv preprint arXiv:1811.00207*.

John R Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1).

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a Conversational Agent Overnight with Dialogue Self-Play. *arXiv preprint arXiv:1801.04871*.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *Workshops at the 32 AAAI Conference on Artificial Intelligence*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. *arXiv preprint arXiv:2109.06427*.

A Unused datasets

A.1 Fine-grained Evaluation of Dialog (FED)

The FED dataset (Mehri and Eskenazi, 2020) consists of both human-machine and human-human dialogues which are annotated on turn-level and dialogue-level. The machines used for this purpose are the bots Meena (Adiwardana et al., 2020) and Mitsuku. The collected conversations are open-domain chit chat conversations for means of measuring dialogue quality of rather casual conversations. Hence, the annotations include 18 fine-grained dialogue qualities, of which 8 are at turn-level and 10 at dialogue-level. One example for a dialogue quality at turn-level is a classification whether the response is engaging, whereas at dialog-level an example is whether there is diversity in the system responses. In total, the dataset consists of 124 conversations and 4712 data points.

A.2 Multi-Domain Wizard-of-Oz (MultiWOZ)

MultiWOZ is a dataset that was collected through the Wizard-of-Oz framework (WOZ) and encompasses multiple domains and topics over around 10k dialogues (Budzianowski et al., 2018). The dialogues are highly natural conversations between a tourist and a clerk from an information center in a touristic city. The linguistic variation in this dataset is rich and complex and around 70% of dialogues have more than 10 turns.

A.3 EmoryNLP

The EmoryNLP dataset (Zahiri and Choi, 2018) was collected for the specific task of Emotion Detection and consists of dialogues from the Friends TV show characters. It is comprised of 97 episodes which make around 12k utterances, hence the annotations are made in turn-level. The emotions used for annotation are the six primary emotions in the Willcox's feeling wheel: sad, mad, scared, powerful, peaceful, joyful—and a default emotion of neutral which results in 7 possible annotations. This dataset is already used in a previous work for evaluating with the episodic knowledge graphs. Nevertheless, since it is a valuable dataset for rich emotion, it will be used in this project as well.

A.4 Machines Talking to Machines (M2M)

For a variation in the way how the dialogues were constructed, we want to also use the M2M data (Shah et al., 2018), which consists of computer generated conversations between a user and an assistant and is query-based. The conversations have a human component to them since the generated conversations were paraphrased via crowdsourcing. More specifically, the paper describes this procedure as: "a simulated user bot and a domain-agnostic system bot converse to exhaustively generate dialogue 'outlines', i.e., sequences of template utterances and their semantic parses which are then connected with context by crowd workers while preserving meaning". The contribution of this dataset is the goal to create dialogue agents with goal-oriented dialogues applicable in different domains. In total, 3000 dialogues were collected through the procedure described above.

A.5 Commonsense-Dialogues

The Common Sense Dialogue dataset (Zhou et al., 2021) is made for response generation with plausible common sense inferences. The dataset consists partly of the previously presented DailyDialog as well as EmpatheticDialog and MuTual datasets. On top of that, it includes additional crowdsourced data with 4 to 6 turns between two friends about an event described in the given context. The authors published 11k dialogues which are filtered through ConceptNet.

A.6 EmpatheticDialogues

This dataset (Rashkin et al., 2018) consists of around 25k conversations about a situation description, crowdsourced from 810 different participants in the form of speaker and listener. The goal of this dataset is for conversational agents to generate empathetic responses.

B Criteria for dialogue evaluation

Group	Criterion
Group A	total nodes
	total edges
	average degree
	average degree centrality
	average closeness
	average degree connectivity
	average assortativity
	average node connectivity
	number of components
	number of strong component
	centrality entropy
	closeness entropy
	sparseness
Group B	total classes
	total properties
	average population
	attribute richness
	relationship richness
	total axioms
Group C	total triples
	total world instances
	total claims
	total perspectives
	total mentions
	total conflicts
	total sources
	total interactions
	total utterances
	ratio claim to triples
	ratio perspectives to triples
	ratio conflicts to triples
	ratio perspectives to claims
	ratio mentions to claims
	ratio conflicts to claims
	average perspectives per claim
	average mentions per claim
	average turns per interaction
	average claims per source
	average perspectives per source

Table 2: List of criteria for dialogue evaluation