GiCCS: A German in-Context Conversational Similarity Benchmark

Shima Asaadi* and **Zahra Kolagar*** and **Alina Liebel**
Fraunhofer IIS
{shima.asaadi,zahra.kolagar,alina.liebel}@iis.fraunhofer.de

Alessandra Zarcone
Hochschule Augsburg
alessandra.zarcone@hs-augsburg.de

Abstract

The Semantic textual similarity (STS) task is commonly used to evaluate the semantic representations that language models (LMs) learn from texts, under the assumption that good-quality representations will yield accurate similarity estimates. When it comes to estimating the similarity of two utterances in a dialogue, however, the conversational context plays a particularly important role. We argue for the need of benchmarks specifically created using conversational data in order to evaluate conversational LMs in the STS task. We introduce GiCCS, a first conversational STS evaluation benchmark for German. We collected the similarity annotations for GiCCS using best-worst scaling and presenting the target items in context, in order to obtain highly-reliable context-dependent similarity scores. We present benchmarking experiments for evaluating LMs on capturing the similarity of utterances. Results suggest that pretraining LMs on conversational data and providing conversational context can be useful for capturing similarity of utterances in dialogues. GiCCS is publicly available to encourage benchmarking of conversational LMs.

1 Introduction

The Semantic Textual Similarity (STS) framework is typically used for the extrinsic evaluation of NLP models (Agirre et al., 2012), and in particular for language models (LMs): if a model can successfully estimate the similarity between sentences, it's a good sign that it has learned good-quality semantically meaningful representations. In natural language generation, STS can provide a useful alternative to word overlap measures to analyse system output similarity (Dušek et al., 2020; Novikova et al., 2016). STS has been used both for training (Reimers and Gurevych, 2019; Vulic et al., 2021) and for evaluating LMs (Yang et al., 2018).

To the best of our knowledge, the majority of STS benchmarks have been created from written language resources using non-conversational data, and are mainly in English. Conversational data, however, has some peculiarities which make it potentially challenging for the STS task: (1) questions and requests are frequent, (2) whether two sentences are semantically similar may depend on pragmatic factors triggered by the conversational context. For example, it may be challenging for human annotators to assess that *Could you turn it up a bit?* and *I'd like the AC to be colder* advance the conversation in a similar way, without (1) sufficient conversational context and (2) an operative definition of what it means for a question and a declarative sentence to be semantically similar. Moreover, STS datasets are typically annotated on a rating scale, which may lead to inconsistencies in annotation, scale region bias, and fixed granularity issues (Kiritchenko and Mohammad, 2017).

In this paper, we introduce GiCCS, the first German in-Context Conversational benchmark for evaluating LMs on the STS task. GiCCS is a multi-domain dataset containing 300 items, each consisting of a domain name, a multi-turn German dialogue context (i.e., dialogue history) where the last utterance is paired with a target utterance, and a semantic similarity score between the paired utterances. GiCCS contains data from German dialogue resources as opposed to machine-translated data (e.g. the GLUE benchmark for STS; Wang et al. 2018). The similarity scores were crowdsourced using the Best-Worst Scaling (BWS) annotation technique (Louviere et al., 2015), as it was shown to address the limitations of the rating scales technique (Kiritchenko and Mohammad, 2017; Asaadi et al., 2019). The dialogue history (the previous 3

*These authors contributed equally to this work.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 351 - 362

or 5 turns) was presented to the crowd-workers during crowdsourcing for better similarity judgements based on the conversational context.

Adopting BWS in order to overcome the limitations of rating scales led to a high inter-annotator agreement with an overall Krippendorff's α of 0.74. Furthermore, we present a reliability study of the similarity scores obtained from the BWS annotations and based on the conversational context.

Lastly, we present benchmarking experiments to evaluate different LMs on the STS task using GiCCS. Experiments show that pre-training LMs on conversational data is beneficial for capturing conversational representations in downstream tasks and conversational applications, such as dialogue systems. GiCCS has a wide range of further applications, such as the evaluation of dialogue generation models, answer selection and ranking systems, and question answering based on dialogue history.

2 Background and Related Work

2.1 Semantic Textual Similarity

Agirre et al. (2012) introduced a large-scale STS benchmark, consisting of pairs of sentences and similarity scores on a $0-5$ ordinal scale (from semantically unrelated to equivalent). Similar benchmarks have been introduced and extended to multiple languages (Agirre et al. 2013; 2014; 2015; 2016; Cer et al. 2017). Sentences in these benchmarks have been collected from news headlines, video and image descriptions, glosses, machine translation evaluation data, tweet news and comments, questions and answers from Q&A forums, and Wikipedia sentences. Annotations are crowdsourced in the form of similarity judgements on a rating scale. The STS benchmark (Cer et al., 2017) included in the GLUE benchmark (Wang et al., 2018) has been translated to German using machine translation systems.[1]

Generally, the STS task evaluates how well a model has learned the semantic space and how semantically meaningful the representations created by the model are. It is widely used for evaluating autoregressive and autoencoding language models. Autoregressive LMs, such as GPT models (Radford et al., 2018), can be tested on producing similar text given a context and autoencoding LMs are tested on creating semantically similar representations for similar texts. Typically, STS is an approach for evaluating conversational LMs, which are in turn employed in the natural language understanding (NLU) components of task-oriented dialogue systems (Yang et al., 2018; Henderson et al., 2019b, 2020; Casanueva et al., 2020; Vulic et al., 2021; Henderson and Vulić, 2021).

To the best of our knowledge, the majority of the benchmarks have been mainly created from written language resources. This results in a sub-optimal benchmark for the evaluation of conversational language models in dialogue systems.

2.2 Conversational Datasets

In order to train, fine-tune or evaluate conversational models for task-oriented dialogue systems, it is crucial to have datasets which are representative of the interaction in task-oriented dialogue. Henderson et al. (2019a) introduce a repository of three large and diverse datasets (Reddit, OpenSubtitles, AmazonQA) for conversational tasks and LM training, each consisting of context–response pairs. Among these, the OpenSubtitles data contains other languages, including German. Prior to this work, Wang et al. (2013) present a dataset of over 12K labeled post–response pairs from the microblog domain. Moreover, Yang et al. (2018) extracted pairs of input–response from a multi-turn open-domain dialogue data, collected by Al-Rfou et al. (2016) from Reddit. Given pairs of related utterances, conversational LMs can be evaluated on capturing the similarity of pairs of utterances by generating semantically similar representations.

There are a few German conversational datasets which are available for research purposes. Among those, the BAS SmartKom corpus is a multi-modal corpus, released in two versions (Schiel et al., 2002; Schiel and Türk, 2006).[2] The data has been recorded in a Wizard-of-Oz setting and is labeled with emotions, gestures, domains, noises, etc., and therefore, it has a wide range of applications including task-oriented dialogue systems. Frommherz and Zarcone (2021) published 113 German dialogues, called CROWDSS[3], collected using the Wizard-of-Oz framework. Data is labeled with dialogue acts and covers one domain. Therefore, it can be used for a variety of NLP tasks. These datasets cannot be directly used as a German STS benchmark. In this paper, we use the audio transcriptions of BAS SmartKom and the dialogues in CROWS-

[1] https://github.com/t-systems-on-site-services-gmbh/german-STSbenchmark

[2] https://www.phonetik.uni-muenchen.de/Bas/BasSmartKomPubliceng.html

[3] https://fordatis.fraunhofer.de/handle/fordatis/198

DSS as our main resources for collecting multi-turn German dialogues in our STS benchmark.

2.3 Conversational Fine-Tuning of LMs

Conversational learning tasks have been introduced to adapt pretrained LMs to conversational models in dialogue systems (Yang et al., 2018; Henderson et al., 2019b, 2020; Casanueva et al., 2020; Vulic et al., 2021; Henderson and Vulić, 2021). The core idea in these tasks is to transform the pretraining or target task into a language understanding task, such as a pairwise STS task and a semantic relatedness task. For this purpose, pairs of queries and responses are created from conversational data and models are trained to score pair items. For instance, Yang et al. (2018), Henderson et al. (2019b) and Henderson et al. (2020) propose a pretraining response selection task (Wang et al., 2013; Al-Rfou et al., 2016; Yang et al., 2018) for learning conversational representations of dialogues. Vulic et al. (2021) introduce CONVFIT, which is a two-stage conversational fine-tuning approach. Similar to previous approaches, first, pretrained LMs are transformed into conversational encoders using the response ranking task. Then, the target intent classification task is treated as a semantic similarity task by pairing utterances in the same intent class as positive pairs and in different classes as negative pairs. Fine-tuning is therefore performed via the STS task. A very recent generative language model, PaLM (Chowdhery et al., 2022), is pretrained on conversations, which is useful in conversational applications and dialogue systems. This model has shown state-of-the-art performance on numerous language understanding tasks.

3 The GiCCS Benchmark

3.1 Data Collection

Creating natural and diverse conversations is a major challenge in the development of task-oriented dialogue systems. It requires manual effort to create such data by crowdsourcing or collecting them from available resources. In this benchmark, we leverage two crowdsourced conversational datasets for German, CROWDSS[4] (Frommherz and Zarcone, 2021) and BAS SmartKom corpora (Schiel et al., 2002; Schiel and Türk, 2006). Both datasets have been collected via the Wizard-of-Oz approach (Budzianowski et al., 2018) to simulate human-machine interaction and contain commonly-used scenarios and domains in dialogue systems.

The CROWDSS dataset contains 113 multi-turn dialogues in the restaurant booking domain and is labeled with dialogue acts. We selected 24 unique dialogues from the dataset starting with the machine's first turn. We randomly split the collected dialogues into two groups of size 12. Keeping all turns would have resulted in a long dialogue history. Moreover, in some cases, the dialogue flow changes from the initial intent after a few turns, which is undesired in our benchmark. Therefore, for the first group, we kept the first three turns of 12 dialogues and for the second group, we kept the first five turns of the 12 dialogues. In a few cases, we corrected some misspelled words or modified the utterances to be more concise. We further extracted transcribed multi-turn dialogues from the BAS SmartKom corpus along with their domain labels for six out of eight domains: cinema, fax, navigation, phone, tourist, and tv.[5] From these, we selected 6 unique multi-turn dialogues for each domain, resulting in 36 unique dialogues in total. Half of the dialogues contained three turns and the second half contained five turns. In some dialogues, we shortened long utterances with multiple sentences by removing irrelevant and unnecessary information from the dialogue, resulting in more focused dialogues.

After the dialogue collection process, we paired the last turn of each dialogue with a set of five handwritten utterances, which were produced by native speakers of German language for this purpose. We chose to hand-write the paired utterances, as pairing randomly-selected sentences with the last turn of each dialogue would have resulted in most sentences being unrelated, which is sub-optimal for benchmarking the models. We thus made sure that the five utterances in the set had different relevancy scores, ranging from unrelated to maximally similar. Following Cer et al. (2017), paired utterances had to satisfy the following criteria to cover the whole range of similarity scores: one paraphrase of the last turn, one sentence that differs in some unimportant details, one sentence that differs in important details, one sentence that shares some details with the last turn but it is not necessarily on the same topic, and one completely unrelated sentence. These similarity judgements on the utter-

[4] https://fordatis.fraunhofer.de/handle/fordatis/198

[5] We extracted data from SK-Home, SK-Public, and SK-Mobil corpora in the following link: http://hdl.handle.net/11022/1009-0000-0001-231F-6.

ances, based on above-mentioned criteria with the main purpose of improving the diversity of the similarity scores, were not used in the main annotation task.

In the end, we obtained 60 dialogues, each paired with five sentences, which resulted in 300 items for the similarity score annotation.

3.2 Data Annotation

Best-Worst Scaling (BWS) (Cohen, 2003; Louviere et al., 2015) is an annotation technique that addresses the limitations of rating scale techniques by employing comparative annotations. In BWS, annotators are presented with n items (n-tuple) at a time and asked which item is the *best*, i.e., highest in terms of the property of interest (for example, *most similar* in our study), and which is the *worst*, e.g., *least similar* in our study. Annotations are then aggregated to obtain real-valued scores of association between the items and the property (Orme, 2009). It has been practically shown that for N items to be annotated, $1.5N$ to $2N$ tuples are sufficient to obtain reliable scores (Louviere et al., 2015; Kiritchenko and Mohammad, 2016). Tuples have to be unique and the items in tuples are distinct. Moreover, each item occurs in approximately the same number of tuples.

In this work, we used BWS to obtain the similarity annotations in GiCCS. We created tuples for each dialogue as follows: From $N = 5$ paired utterances in each dialogue, we generated $2N = 10$ distinct 3-tuples, where each tuple is a random set of three paired utterances. The order of the terms in the 3-tuples is not important, and (following BWS) each term appears in six tuples. We obtained 600 distinct 3-tuples to be annotated.

We set up the annotation task on the crowdsourcing platform Amazon Mechanical Turk (AMT). As requirements for selecting crowd-workers on AMT, we set the approval rate to greater than 98% and the location to Germany. We provided a detailed annotation instruction with examples and asked the annotators to only participate in this study if they were fluent in German. The annotators were presented with two dialogues at a time (one 3-turn dialogue and one 5-turn dialogue), each followed by a 3-tuple for the best- and worst-questions (*which sentence is the most similar to the last utterance in the dialogue? which is the least similar?*). See Appendix A.1 for details on the annotation instructions and a sample of the task presented to the workers. We also included an optional comment section for workers. We collected five different annotations for each 3-tuple.

Dataset	Dialogue turns	α	BW question	Strong agreement
BAS SmartKom	3-turn	0.87	best	99
			worst	100
	5-turn	0.80	best	98
			worst	100
CROWDSS	3-turn	0.63	best	90
			worst	94
	5-turn	0.67	best	95
			worst	97

Table 1: Krippendorff's α and percentage of strong-agreement cases for both source datasets.

3.3 Inter-Annotator Agreement

We computed inter-annotator agreement by considering cases where two annotators provided the same answer to the best- and worst-questions as cases of agreement and cases where two annotators provided different answers to the best- and worst-questions as cases of disagreement. The annotation yielded an acceptable inter-annotator agreement with an overall Krippendorff's α of 0.74 (Artstein and Poesio, 2008). Moreover, Table 1 shows the Krippendorff's α in each source dataset. As can be seen, the overall agreement in BAS SmartKom is higher than CROWDSS.

To provide a better overview of agreements, Table 1 also shows the percentage of items in each portion of the source dataset that had a *strong agreement*. We define strong agreement as cases where at least four out of five annotators selected the same answer in the best and worst questions. Percentages of strong agreement cases are high for both source datasets, which speaks for the validity of the best-worst scaling task. On the other hand, we assume that the slightly lower percentages obtained from the CROWDSS dataset might stem from the fact that CROWDSS has a higher lexical diversity and contains only one domain. This may result in a more difficult comparison between best and worst pairs for the annotators during the annotation process. Results are also in agreement with the lower inter-annotator agreement (α) for the CROWDSS dataset.

3.4 Dataset Preparation

Annotation Aggregation. After the completion of the annotation task, we calculated the final se-

mantic similarity scores for dialogue–utterance pairs from the BWS responses using a simple counting method (Orme, 2009). For each pair, the semantic similarity score is the proportion of times the utterance was chosen as the best minus the proportion of times the utterance was chosen as the worst in the annotation task. This results in similarity scores ranging from -1 to 1, which we normalized to the interval $[0, 1]$. Finally, GiCCS includes 300 items, each containing a domain label, a multi-turn dialogue context, a comparison utterance, and a similarity score between the comparison utterance and the last utterance in the dialogue.

Managing Domain Labels Since the domain of some dialogues didn't perfectly match the dialogues' context (e.g., phone domain in the BAS SmartKom corpora), we have relabeled some of the dialogues from the original labels to obtain a consistent domain labeling; also specified in 3.1. The final dataset includes the following domains: *find_restaurant*, *find_tvProgram*, *find_cinema*, *find_hotel*, *find_touristAttraction*, *find_navigation*.

3.5 The Dataset

Table 2 presents descriptive statistics about GiCCS. We report the average turn length per domain, number of tokens per domain, and the unique count of the lemmatized forms of the words. Tokens were obtained by splitting text based on the whitespaces. To obtain lemmas, i.e., the base form of words that are present in a dictionary, we used the German SpaCy model de_core_news_sm version 3.3.0 (Honnibal et al., 2020).[6]

In order to compare lexical richness between dialogues in each domain, we show root type-token ratio (RTTR) as well as the measure of textual lexical diversity (MTLD; McCarthy and Jarvis 2010) computed with the threshold of 0.72 using the LexicalRichness library (Shen, 2022)[7], as MTLD is more robust to changes in text length. High RTTR and MTLD in both 3-turn and 5-turn dialogues indicate the lexical diversity of GiCCS. In particular, both measures are high in the find_restaurant domain suggesting that this domain may be more complex and challenging for the annotators. This is reflected by lower agreement scores for CROWDSS compared to BAS SmartKom (see Table 1) as most dialogues in the find_restaurant domain are from the CROWDSS dataset.

[6]https://github.com/explosion/spacy-models/releases/tag/de_core_news_sm-3.3.0

[7]https://github.com/LSYS/LexicalRichness

4 Score Reliability Study

We now provide an analysis of the similarity scores obtained with the BWS and of their reliability.

4.1 Score Distribution

Figure 1 shows the distribution of the obtained similarity scores. As expected, the final scores cover a wide range of similarity in the interval $[0, 1]$, which is useful for evaluating LMs on fine-grained scoring and their ability on detecting the nuances in semantics.

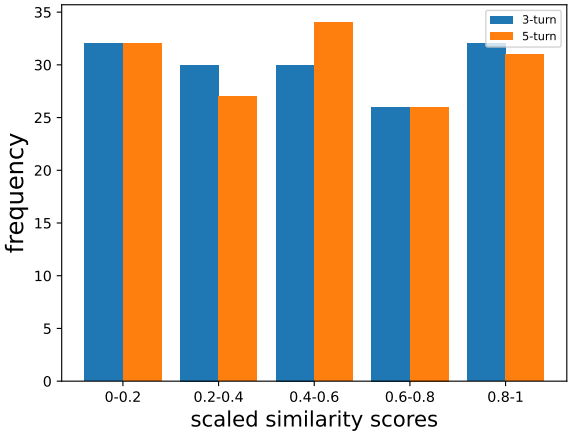

Figure 1: Distribution of the final similarity scores in the interval $[0, 1]$.

A common issue in the rating scales technique is the scale region bias, i.e., annotators have a bias towards a portion of the scale, for instance, towards the middle of the scale. The distribution of the scores in our dataset exhibits that the scale region bias issue was avoided using the BWS technique.

4.2 Split-Half Reliability

Another approach to measure the reliability of the annotations is to assess the reproducibility of the final scores. To measure this, we compute split-half reliability (SHR; Cronbach 1951). To compute SHR, we split the five annotations per tuple to two halves of odd vs. even number of annotations randomly. The first half (group A) includes two annotators, while the second half (group B) includes three annotators. Then, similarity scores of paired utterances are computed based on BWS responses in each half. Finally, the Spearman correlation between the scores obtained by these halves is calculated as an estimate of the annotation reliability.

Dialogue turns	Domain	RTTR	MTLD	Tokens	Lemma	Average turn length
3-turn	find_restaurant	3.04	46.22	377	21	10.47
	find_cinema	1.43	20.34	128	39	10.66
	find_hotel	1.53	15.90	77	35	8.55
	find_navigation	1.48	21.31	82	33	9.11
	find_touristAttraction	1.78	29.60	114	30	12.66
	find_tvProgram	1.81	30.06	165	28	11.00
5-turn	find_restaurant	3.01	36.91	739	47	11.36
	find_cinema	2.08	25.06	297	66	11.88
	find_hotel	1.53	17.56	106	76	7.06
	find_navigation	1.62	21.80	122	27	8.13
	find_touristAttraction	1.76	21.96	132	82	13.20
	find_tvProgram	2.50	32.03	247	51	12.35

Table 2: Descriptive statistics of GiCCS (RTTR = root type-token ratio; MTLD = measure of textual lexical diversity).

We repeat the SHR computation three times and report the average correlations over the repeated runs as shown in Table 3. A high Spearman correlation in both datasets shows that the obtained similarity scores are highly reliable. These results correspond to the percentage of strong agreement cases in Table 1. Slightly lower scores in CROWDSS are due to the fact that the utterances from CROWDSS are less varied in terms of domain diversity and an agreement on the best and worst questions is more challenging for the annotators.

Dataset	Dialogue turns	Spearman
BAS SmartKom	3-turn	0.975
	5-turn	0.970
CROWDSS	3-turn	0.953
	5-turn	0.946

Table 3: Average split-half reliability scores in each source dataset.

4.3 Score Reliability Assessment with Expert Annotation

We conduct an expert evaluation to ensure the final scores per paired utterances match expert expectations. We presented the final dataset to a trained linguist, excluding the final scores, and asked whether the last turn of the dialogue and the paired utterance showed either higher or lower than 0.5 similarity. After assessing the results of the expert evaluation, only two instances out of 300 items didn't match the evaluation (see Table 4). These are interesting cases showing what it means for two utterances in a conversational context to be similar or not: besides a higher or lower degree of lexical overlap, the degree of overlap in the intent behind the utterances - only partial in one case but full in the other - is what motivates the expert score.

5 Experiments

5.1 Evaluating Language Models on GiCCS

We conduct experiments on using GiCCS to evaluate autoencoding and autoregressive multilingual LMs, respectively. More specifically, we consider two types of evaluation tasks for these models. The first task, called **pairwise STS**, is about predicting the similarity score for pairs of utterances, which is a real-valued score between 0 and 1. This task is used for examining autoencoding LMs, such as BERT-based models (Devlin et al., 2019), on creating meaningful conversational representations for utterances. The latter, called **multiple-choice STS**, is about selecting the most similar utterance in a multiple-choice question task. In this task, we evaluate autoregressive models, such as GPT models (Radford et al., 2018, 2019), by considering the dialogue history.

We focus on unsupervised STS, i.e., we evaluate the performance of LMs without training or fine-

Last turn of the dialogue	Paired sentence	Calculated score	Expert evaluation
Die Innenstadt. *Downtown.*	Das Restaurant kann auf dem Land oder in der Stadt sein. *The restaurant may be in the country or downtown.*	0.65	< 0.5
Wie weit ist das? *How far is it?*	Wo befindet sich das Restaurant? *Where is the restaurant located?*	0.35	> 0.5

Table 4: Cases where the expert evaluation does not match the calculated score.

Model	Pearson r	Spearman ρ
STransformers		
distiluse-base-multilingual-cased-v2	0.859	0.855
paraphrase-xlm-r-multilingual-v1	0.849	0.845
paraphrase-multilingual-MiniLM-L12-v2	0.842	0.842
paraphrase-multilingual-mpnet-base-v2	0.830	0.829
distilbert-multilingual-nli-stsb-quora-ranking	0.794	0.814
Encoder		
deepset/gbert-large	0.666	0.680
deepset/gbert-large-sts	0.622	0.679

Table 5: Pearson and Spearman correlation results on all pairs in GiCCS.

tuning them on our data. All studied models are downloaded from the Hugging Face Model Hub.[8]

5.1.1 Pairwise STS Evaluation Task

We examine the following LMs: 1) Multilingual Sentence-Transformers (STransformers) (Reimers and Gurevych, 2019), which are fine-tuned on Natural Language Inference (NLI) and STS tasks, and 2) German encoder-based LMs (Chan et al., 2020). STransformers have been partly trained on spoken text from the English MultiNLI corpus (Williams et al., 2018) and extended to multilingual models using various datasets including conversational data (Reimers and Gurevych, 2020). We conducted experiments on five selected STransformer models, three of them have been trained on paraphrases from conversational data, such as quora and Stackexchange[9]. Selected models can be found in Table 5. German encoder models have been pre-trained partly on conversational corpora such as movie subtitles and one model was further fine-tuned on the German STS benchmark[10] (Cer et al., 2017).

In each model, the utterance embedding is computed from the mean aggregation of its token embeddings. Then, the similarity score is obtained by computing the cosine between the embeddings of utterances in each pair. Following Cer et al. (2017) and Reimers and Gurevych (2019), we measure the performance of the models using the Spearman rank r and Pearson ρ correlations of predicted and gold scores. Table 5 shows the performance results of different models on all dialogues.

As can be seen in Table 5, STransformers outperform pre-trained encoder models. In general, since sentence transformers are finetuned on NLI and STS containing spoken data and are trained to specifically encode sentences, they can better capture the semantic similarity of utterances on sentence level compared to pretrained models. Results of the two encoder models indicate that finetuning LMs on STS, which is a typical task for evaluating language understanding through semantic similarity, may not be always sufficient for improving the model performance on conversations.

5.1.2 Multiple-Choice STS Evaluation Task

To examine autoregressive models, we follow the prompting approach in GPT-3 (Brown et al., 2020) and construct a prompt template for the multiple-choice question task. For this purpose, each multi-turn dialogue (containing the target utterance) is followed by a question and five possible answers. The question is to *select the most similar utterance to the last turn in the dialogue*, and the possible answers are the five utterances paired with the target

[8] https://huggingface.co/models

[9] Please refer to the following link for more information on the training data: https://www.sbert.net

[10] https://github.com/t-systems-on-site-services-gmbh/german-STSbenchmark

Dialog:
Äußerung 1: Ich habe den Nachmittag in Heidelberg frei, möchte einen Bekannten treffen.
Informationen brauche ich über Sehenswürdigkeiten der Stadt. Kann ich bitte einen Plan haben?
Äußerung 2: Die Sehenswürdigkeiten von Heidelberg. Wenn du möchtest, kann ich auch einen Vorschlag machen.
Äußerung 3: Ich möchte gerne Museen, Kloster und Gebäude auch natürlich sehen. Aber zunächst Museen.

Frage: Wählen Sie die Äußerung, die der letzten Äußerung im Dialog am ähnlichsten ist.
Auswahl:
A. Ich habe vor, heute Reiten zu gehen.
B. Ich möchte gern zuerst Klöster sehen.
C. Ich würde gerne Kloster, Museen und Gebäude selbstverständlich auch sehen, aber zuerst Museen.
D. Ich würde gerne das Sportzentrum besuchen.
E. Ich würde gerne die Gebäude sehen, aber ich bin offen für andere Optionen, wie Klöster oder Museen.
Antwort:

Figure 2: A prompt example for the multiple-Choice STS evaluation task.

Model	**3-turn Dialogue**		**5-turn Dialogue**	
	w/ context	**w/o context**	**w/ context**	**w/o context**
mGPT (Shliazhko et al., 2022)	0.133 ± 0.063	0.166 ± 0.069	0.100 ± 0.055	0.067 ± 0.046

Table 6: Zero-shot accuracy results on multiple-choice STS task for 3- and 5-turn dialogues.

sentence in GiCCS, among which the most similar utterance is the correct answer. Figure 2 shows a prompt example in our task. The model takes a multi-turn dialogue followed by the question as the input context. Then, it computes the likelihood of generating each answer sentence, and the sentence with the highest likelihood is selected as the correct answer. Since the STS task is reformulated as a multiple-choice question with only one correct answer, we compute the accuracy of predicting correct answers for all questions. We report the results in the zero-shot setting.

Table 6 shows the evaluation results of a multilingual autoregressive model, mGPT (Shliazhko et al., 2022). We assume that the low performance is due to the fact that the model is mainly trained on written language data and is not focused on conversational training. Providing the dialogue history as the context for 5-turn dialogues resulted in a slightly higher performance compared to an evaluation without the dialogue history. This is not the case for 3-turn dialogues. We speculate that shorter dialogue history can be confusing for the model and increasing the history helps in capturing the context. Moreover, the task setup influences the model performance. Therefore, the prompt can be adapted to condition the LMs on the given task to obtain a better performance.

6 Conclusion

We introduce GiCCS, a first German conversational STS benchmark for evaluating language models on semantic similarity. Each item in the benchmark consists of a domain name, a multi-turn dialogue history, a target utterance paired with the last utterance in the dialogue, and a similarity score between the paired utterances. We leveraged two German dialogue resources, BAS SmartKom and CROWDSS, to collect our data as opposed to machine-translated data. Annotations were crowdsourced using the best-worst scaling technique, which shows a high inter-annotator agreement with an overall Krippendorff's α of 0.74 and reliable annotations with an average split-half reliability score of 0.973 for BAS SmartKom and 0.949 for CROWDSS. Moreover, the dialogue history was presented to the crowdworkers for better similarity judgements based on the conversational context. Final similarity scores cover a wide range of similarities between 0 and 1 introducing a challenge for language models to identify the nuances in semantics of different utterances. Moreover, as the last utterance of each dialogue is paired with five different utterances each with its similarity score, GiCCS can be used for evaluating ranking systems.

Results of evaluating LMs on the STS task using GiCCS shows that pre-training LMs on conversational data brings benefits for the LMs on meaning-

ful representations of conversations.

Overall, GiCCS is a useful resource for evaluating conversational models on capturing similarity in conversational data. Moreover, due to the lack of enough resources for evaluating conversational models in non-English languages, we hope that the annotation procedure described in this work will foster an interest in creating more reliable and high-quality resources similar to GiCCS.

7 Ethical Considerations and Licenses

The crowd-workers on Amazon Mechanical Turk remain anonymous on AMT to adhere to the ethical standards in the community. They were voluntarily recruited, they provided their written informed consent to participate in the study and were allowed to opt out at any point in time .

To create GiCCS, two German dialogue resources, BAS SmartKom and CROWDSS, were used. In the BAS SmartKom resource, texts were partly derived from the BAS SmartKom corpora (corpus PID: 11022/1009-0000-0001-231F-6; Schiel and Türk 2006)[11] and we have received the permission of the copyright holders of the BAS SmartKom corpora to publish the derived text data in our benchmark. CROWDSS corpora is licensed under Attribution 4.0 International (CC BY 4.0). GiCCS is released upon publication of this paper and licensed under CC BY-NC-ND 3.0.[12]

Acknowledgements

This work is supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) through the SPEAKER project (FKZ 01MK19011). This research project was started while Alessandra Zarcone was affiliated with the Fraunhofer IIS. We thank Florian Schiel and the copyright holders of the BAS SmartKom corpora for providing access and permission to publish the derived text data. We thank Sabrina Stehwien and the anonymous reviewers for their constructive feedback.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

[11] http://hdl.handle.net/11022/1009-0000-0001-231F-6

[12] https://doi.org/10.5281/zenodo.7266256

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Steven H. Cohen. 2003. Maximum difference scaling: Improved measures of importance and preference for segmentation.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123–156.

Yannick Frommherz and Alessandra Zarcone. 2021. Crowdsourcing ecologically-valid dialogue data for german. *Frontiers in Computer Science*, 3.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Henderson and Ivan Vulić. 2021. ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3375–3389, Online. Association for Computational Linguistics.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273.

Bryan K. Orme. 2009. Maxdiff analysis : Simple counting , individual-level logit , and hb.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Florian Schiel, Silke Steininger, and Ulrich Türk. 2002. The SmartKom multimodal corpus at BAS. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Florian Schiel and Ulli Türk. 2006. *Wizard-of-Oz Recordings*, pages 541–570. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Ivan Vulic, Pei-Hao Su, Sam Coope, Daniela Gerz, Pawel Budzianowski, Iñigo Casanueva, Nikola Mrksic, and Tsung-Hsien Wen. 2021. Convfit: Conversational fine-tuning of pretrained language models. *CoRR*, abs/2109.10126.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Annotation Instruction

By accepting the task, the workers are directed to an interface where they are presented with a brief

as well as a detailed instruction in German on how to accomplish the task. The detailed instruction contains a description of what the workers will observe in the actual HITS along with one example. In the example, a dialogue history followed by two questions are posed similar to what appears in the actual experiment. The first question expects the participants to choose an utterance from three given options that is most similar to the last utterance of the dialogue. Then, they are asked to choose an utterance that is least similar to the last utterance of the dialogue. In the instruction section, the right answers are already selected. A sample of a dialogue and the two questions are presented in Figure 3.

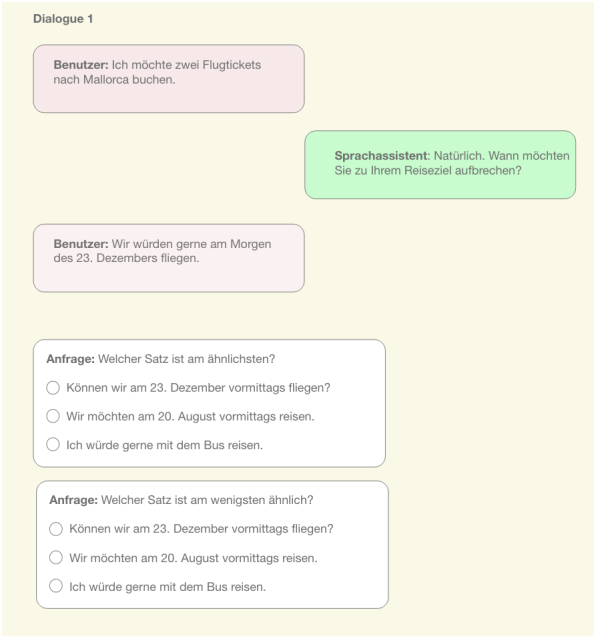

Figure 3: A sample of the annotation task presented on Amazon Mechanical Turk.