Chinese Internet Dialogue Corpus: A High-Quality Dataset from Social Media

Anonymous ACL submission

Abstract

Recently, large-scale dialogue datasets have attracted increasing attention in the research community. Many previous studies have constructed dialogue datasets by gathering massive amount of raw data from social media platforms and converting them into dialogues using rule-based methods. However, the usability of such datasets for training is highly dependent on the quality of the raw data. Unfortunately, most raw data from major social media platforms are unstructured and noisy, making it challenging to generate clean dialogue datasets using only rule-based approaches. To address this issue, we propose a novel transfer method that combines model-based and rule-based techniques to process raw data collected from social media platforms. In addition, we introduce a novel scoring method for evaluating the quality of dialogue datasets. Our experiments find a correlation between our scoring method and human judgments of dialogue quality. Using this method, we further evaluate our proposed dataset and compare it with other existing dialogue datasets. Consequently, we present the Chinese Internet Dialogue Corpus, which contains 3,102,235 short-text dialogues, sourced from Baidu Tieba, a popular Chinese social media platform. The Chinese Internet Dialogue Corpus, including both the code and dataset, will be publicly available soon at https://github.com/anonymous20250123/emnlp2025.

1 Introduction

With the breakthrough progress of Large Language Models (LLMs) in the field of Natural Language Processing (NLP), large-scale, high-quality conversational data has become increasingly critical for advancing dialogue systems. In many early studies, researchers utilized movie scripts or social media posts to construct large-scale datasets, thereby providing training corpora for neural network models, particularly deep neural networks (DNNs) (Henderson et al., 2019). It has been shown that, given sufficiently large-scale corpora, deep neural networks can significantly enhance text generation performance (Sennrich and Zhang, 2019). However, data quality often presents a greater challenge than data quantity (Lison and Tiedemann, 2016; Lison et al., 2018), especially when conversational data is directly scraped from social media platforms, where abundant noise commonly arises. Such data collected from social media platforms has been defined in previous research as unstructured data, referring to text or information lacking a predefined format or logical structure, which makes direct labeling via fixed fields difficult and allows for relatively unconstrained content forms (Lowe et al., 2015).

Extracting and converting dialogue data from such unstructured data proves more challenging than from other types of text. Compared with other types of text data (e.g., review texts), dialogue data needs to be tightly linked to the contextual environment to ensure semantic coherence and a clear subject of discussion. For example, movie review datasets tend to focus on a single object of evaluation, namely, a particular movie, so the context and subject matter are relatively singular and explicit. In contrast, dialogue scenarios often involve multiple participants and several rounds of information exchange. This is especially evident in community discussions featuring nested replies, where the information and semantic references among different levels or sub-replies can frequently become confused, leading to unclear speaker references, fragmented context, or even complete mismatches between the conversation and the topic (Baheti et al., 2018).

To more effectively illustrate the differences in unstructured data between review texts and dialogue texts, we compared these two types of datasets. As shown in Table 1, the frequency of

Source	Language	Category	Posts Number	Nested Posts	Proportion
Weibo	Chinese	dialogue for daily life	$119,988$	$69,413$	57.85%
Douban	Chinese	movie reviews	$1,278,401$	$4,030$	0.3152%
Twitter	English	dialogue for US politics	$970,919$	$339,894$	35.01%
Amazon	English	music instruments reviews	$10,261$	14	0.1364%

Table 1: The difference between review texts and dialogue texts.

nested replies is significantly lower on Douban movie reviews[1] and Amazon music instruments reviews[2] compared to Weibo dialogue for daily life[3] and Twitter dialogue for US politics[4]. This discrepancy can be attributed to the nature of review texts, which typically focus on a single subject or topic. In contrast, social media dialogues often involve multiple participants and multi-level references, naturally resulting in more frequent nested replies. This, in turn, increases the likelihood of ambiguous referents and incoherent semantics.

Against this backdrop, efficiently obtaining higher-quality dialogue data has emerged as a critical challenge in building large language models and various dialogue systems. Existing research has frequently emphasized the importance of data scale for improving model performance (Bengio et al., 2013), yet it has become increasingly clear that the presence of noise can notably diminish training effectiveness and even lead models to generate replies that are inconsistent or incoherent with the given context (Vinyals and Le, 2015). Given that this study focuses on single-turn dialogue scenarios, it is crucial to accurately identify and eliminate issues of ambiguous references caused by unstructured data. Based on the importance of addressing issues of ambiguous references caused by unstructured data, we hypothesize that such efforts can enhance the overall quality of dialogue datasets.

Building on the hypothesis that addressing issues of ambiguous references caused by nested replies can enhance the overall quality of dialogue datasets, our study proposes a mask mechanism for context matching to address the prominent problem of ambiguous speaker references caused by nested replies in dialogue data gathered from social media platforms. This approach aims to substantially improve data quality while maintaining a sufficiently large dataset. We applied this method to Baidu Tieba, a representative community platform, thereby acquiring a large-scale, single-turn short-text dialogue dataset. To evaluate the quality of our dialogue dataset, we identified the need for an automated scoring metric. Given the lack of a scoring metric tailored to the style of daily Chinese conversations, we first developed a novel scoring approach to address this gap. We then validated the effectiveness of this metric in assessing dialogue quality. Subsequently, we applied the scoring metric to our dataset and other datasets, comparing their scores under our scoring framework to evaluate the quality of our dataset. Finally, we released our short-text dialogues, hoping to provide cleaner and more abundant resources for dialogue system research.

In summary, the main contributions of this paper are as follows:

- We propose a masking mechanism to reduce noise and extract high-quality, single-turn dialogue data from Baidu Tieba.
- We propose and validate a dialogue scoring metric that effectively assesses dataset quality.
- We release the Chinese Internet Dialogue Corpus with 3,102,235 short-text dialogues as a resource for dialogue system research.

2 Related Works

In this section, we provide a brief overview of existing dialogue datasets and some commonly used evaluation methods.

2.1 Dialogue Datasets

Existing dialogue datasets used for training large language models can be broadly categorized into

[1] Kaggle:https://www.kaggle.com/datasets/fengzhujoey/douban-datasetratingreviewside-information/data

[2] Kaggle:https://www.kaggle.com/datasets/eswarchandt/amazon-music-reviews/data

[3] Github:https://github.com/SophonPlus/ChineseNlpCorpus

[4] Kaggle:https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets/data

$Utterance$	
$user_p$:	Hey, I only have one day in Tokyo. What are the must-see spots?
$Response$	
$user_a$:	Meiji Shrine, Harajuku, Shibuya, and Skytree – all doable in a few hours.
$user_b$ *response to* $user_a$:	Isn't the Sky Tree located on the opposite side of town compared to others?
$user_c$:	Top attraction in Tokyo right now is the newly opened team labs borderless.
$user_c$:	If you're in Harajuku, don't miss Menchirashi for amazing udon.
$user_d$ *response to* $user_c$:	Hahaha, everyone love Menchirashi!
$user_e$:	Shibuya Crossing is great, especially at night.

Table 2: A typical example of a Baidu Tieba post (translated from the original Chinese text).

two types: manually annotated datasets and web-crawled datasets. Manually annotated datasets are created using various approaches. For example, Persona Chat is constructed by assigning annotators different personas to generate dialogues (Zhang et al., 2018). Empathetic Dialogues involves interactions where annotators take on the roles of speakers and listeners to simulate emotional scenarios (Rashkin et al., 2019). Topical Chat, on the other hand, generates dialogues by providing annotators with Wikipedia content on specific topics rather than assigning roles (Gopalakrishnan et al., 2023).

Web-crawled dialogue data primarily originate from social media platforms. For instance, early datasets include 1.3 million conversations extracted from Twitter (Ritter et al., 2010). Similarly, technical dialogues about Ubuntu were sourced from ubuntu chatrooms (Lowe et al., 2015). In the context of Chinese dialogue datasets, significant efforts have been made, such as the LCCC dataset, which involves cleaning data from major Chinese social media platforms, including weibo, douban and so on (Wang et al., 2020).

While manually annotated dialogue datasets offer high quality, they are costly to produce and relatively small in scale. In contrast, in the era of large language models, web-crawled datasets are more beneficial for improving model performance due to their scale. However, most web-crawled dialogue datasets originate from unstructured social media data, posing challenges for transforming such data into daily dialogue formats (details in section 3.2). To address this issue, this paper proposes a novel method for converting dialogue data from social media platforms.

2.2 Scoring Metrics

Existing metrics for evaluating dialogue data can be divided into referenced metrics and unreferenced metrics.

Referenced metrics, such as BLEU and METEOR, compare generated dialogues against human-generated reference responses (Liu et al., 2016). While these methods provide a benchmark for evaluation, they require human-generated references, which are costly to obtain. Additionally, they are highly sensitive to variations in responses, making them less suitable for evaluating open-ended conversational dialogues. For the question "What is your favorite book?", a referenced metric like BLEU would score "Pride and Prejudice" low if the reference answer is "To Kill a Mockingbird" despite both being valid English classics.

Unreferenced metrics, often evaluate dialogue quality based on sentence attributes such as connectivity and content relatedness (Akama et al., 2020). For the question "What is your favorite book?", responses like "Pride and Prejudice" and "To Kill a Mockingbird" would receive comparable scores due to their contextual relevance and equivalence as valid answers. These methods are low-cost, fast, and less sensitive to response variations, but they are highly sensitive to semantic similarity. Given the scarcity of scoring metrics tailored for Chinese daily conversations, this paper introduces a new scoring metric specifically designed for evaluating Chinese conversational data.

3 Dialogue Conversion

We propose a method to transform raw social media data into conversational datasets by applying a model-based masking mechanism for context matching, followed by rule-based noise filtering to improve data quality.

3.1 Task Definition

Let x represent an utterance and y represent a response to x. The user associated with x is defined as the recipient $user_x$, while the user associated with y is defined as the sender $user_y$. An utterance pair can thus be expressed as $((user_x, x), (user_y, y))$. As with most unstructured data, the raw data from Baidu Tieba also consists of multiple utterances within a single post. Therefore, we introduce a post's content p and its initiator $user_p$.

Due to the unstructured nature of the data, the obtained raw dataset can be conceptualized as resembling a chat room scenario. In such a setting, users may respond to anyone they wish, meaning that a given user may respond to another user who has replied to the post initiator, rather than responding solely to the post initiator. Hence, while every user is a sender of some utterance, the recipient need not be the post initiator. This presents a critical challenge in transforming unstructured data into a dialogue dataset: identifying the appropriate $(user_x, x)$ for each $(user_y, y)$.

Various prior studies have proposed strategies to address this challenge. For instance, when dealing with Weibo, a prominent Chinese social media platform, the problem is simplified by assuming each sender is responding directly to the post initiator (Wang et al., 2013). In other words, all utterances under a given post are assumed to have the same recipient $user_p$ and the same reference content p. They then apply rule-based filtering to remove low-relevance utterance pairs and duplicates. Although this assumption greatly simplifies the utterance pairing process, it simultaneously discards crucial contextual information, such as usernames and timestamps.

In contrast, the Ubuntu Dialogue Corpus adopts a rule-based approach that leverages contextual information such as usernames and timestamps. For example, messages explicitly targeting a particular user (as indicated by a username mention) are regarded as responses to that user (Lowe et al., 2015). Furthermore, if a user interacts exclusively with a single target user, all of that user's unspecified messages are incorporated into the conversation. Additional measures, such as timestamp-based filtering, are also employed.

A simpler approach is taken in the Douban dataset by assuming that the final utterance in each conversation is the correct response to the preceding turns, thereby implicitly utilizing timestamp information (Wu et al., 2017).

In summary, most existing dataset construction methodologies rely solely on rule-based filtering and make use of only a fraction of the available contextual information inherent in unstructured data.

3.2 Mask Mechanism for Context Matching

To address the characteristics of unstructured data, we propose Mask Mechanism for Context Matching (MMCM) approach. As mentioned earlier, our task involves identifying the appropriate $(user_x, x)$ for every $(user_y, y)$, i.e., determining the corresponding recipient for each sender.

Meanwhile, we observe that the unstructured data of the mainstream social media platforms implicitly contains temporal information for each of its posts: the statements of the users are always in chronological order. For example, as shown in Table 2, $user_b$ cannot be replying to $user_c$. Utilizing this property, the sender corresponds to a recipient whose position is earlier in the conversation. Preserving this spatial structure of the post when acquiring raw data is a indirect use of temporal information.

We also observe that in the unstructured data of mainstream social media platforms, when a user wants to reply to another user, it will always be as shown in Table 2, and the response will always be formatted with *response to* $user_x$. However, if $user_c$ appears multiple times in a post, it is impossible to determine which statement of $user_c$ is the corresponding recipient, even if the *response to* $user_c$ is carried in the sender.

Integrating these observations, we construct two matrices: post matrix $\mathcal{P}$ and user matrix $\mathcal{U}$. As shown in Figure 1, $\mathcal{P}$ is computed based on the spatial order of posts to capture semantic similarity. This thesis adopts the unsupervised pre-trained fastText on a large number of Chinese texts to extract word embeddings for each utterance (Joulin et al., 2017). Then, dot multiplication is performed on the word embeddings in order to complete the semantic similarity computation. Combined with the spatial structure feature of the post, $\mathcal{P}$ and $\mathcal{U}$ only need to use the lower triangular part. $\mathcal{U}$ is a mask matrix, each row represents the corresponding spatial location of the sender $user_y$, and performs *response to* $user_x$ regular matching on its response to get the

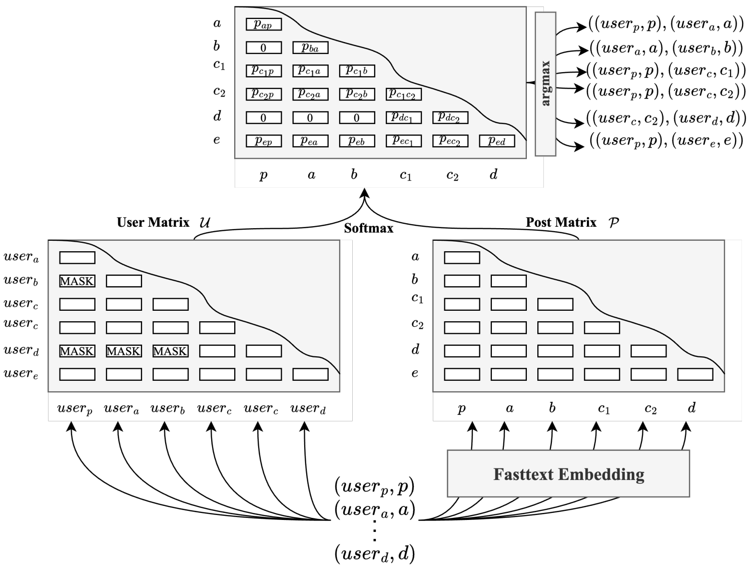
p_{ap}
p_{ba}
p_{c_1p} p_{c_1a} p_{c_1b}
p_{c_2p} p_{c_2a} p_{c_2b} $p_{c_1c_2}$
0 0 0 p_{dc_1} p_{dc_2}
p_{ep} p_{ea} p_{eb} p_{ec_1} p_{ec_2} p_{ed}
argmax
$((user_p, p), (user_a, a))$
$((user_a, a), (user_b, b))$
$((user_p, p), (user_c, c_1))$
$((user_p, p), (user_c, c_2))$
$((user_c, c_2), (user_d, d))$
$((user_p, p), (user_e, e))$
User Matrix $\mathcal{U}$
Softmax
Post Matrix $\mathcal{P}$
MASK
Fasttext Embedding
$(user_p, p)$
$(user_a, a)$
$(user_d, d)$

Figure 1: The MMCM procedures of the post in Table 2. $user_x$ represents the ID of user x. x is the text content of the corresponding user. Since $user_c$ responds twice, there are two text contents of $user_c$, referred to as c_1 and c_2.

object $user_x$ of its reply, masking all users except $user_x$. If the reply object $user_x$ does not exist, then no masking is applied to all users. Combining $\mathcal{P}$ with the user matrix $\mathcal{U}$ for the corresponding position masking operation, a softmax calculation is performed for each row. The response with the highest probability after softmax is then selected as the recipient, completing the construction of the utterance pair.

With MMCM, temporal information is implicitly utilized in the post space structure, while user information is leveraged in the construction of the mask matrix, which makes full use of the context information of the post. Furthermore, because it is a matrix operation and the word embedding adopts fastText, the computation of MMCM only requires CPU and is quite fast.

3.3 Rule-based Noise Filtering

At this stage, various types of conversational noise are eliminated through rule-based filtering, including: (1) blacklist filtering, which primarily targets undesirable content such as profanity and sensitive political topics; (2) filtering of undecodable emojis or characters; (3) removal of meaningless repetitive utterances, such as the numerous "hahahaha" style messages common in the Baidu Tieba corpus; (4) filtering out dialogues containing images, advertisements, and URL hyperlinks; (5) filtering out utterances containing private information such as email addresses, phone numbers, and personal names; (6) removing utterances composed of long strings of digits and/or letters; (7) removing non-Chinese dialogues; (8) deleting utterance pairs that are too short.

The blacklist includes not only explicit profanity and political terms but also their phonetic variants. Any utterance pair containing such terms is discarded. To ensure informativeness, queries must exceed 10 characters and responses at least 8.

4 Dialogue Evaluation Scoring

To validate the effectiveness of the dialogue conversion methodology presented in the section 3, we propose a dialogue evaluation scoring mechanism based on the concepts of semantic relevance and domain adaptiveness. Semantic relevance represents the semantic correlation, while the domain adaptiveness captures aspects of conversational style. Ultimately, the two are combined through a simple summation, as follows:

$$S(x, y) = R(x, y) + D(x, y) \qquad (1)$$

This score will first be validated for effectiveness during the experimental phase, and then applied to our dataset.

4.1 Semantic Relevance

Semantic relevance is widely used in the field of Natural Language Processing (NLP), particularly in tasks such as dialogue generation, information retrieval, and text matching. In both evaluation and generation contexts, semantic relevance allows for the acceptability of multiple valid answers to a single query. For instance, given the request "recommend a good book," different recommendations

Scoring Method	Spearman's ρ	p-value
model score	0.1334	0.0350
semantic relevance	0.0858	0.1763
domain adaptiveness	0.0991	0.1179

Table 3: The result of validation experiment for the scoring metric.

Dataset	Semantic relevance	Data adaptiveness	Model score
Chinese Internet Dialogue Corpus	**0.8647**	0.5165	**0.6906**
E-commerce Dialogue Corpus	0.7796***	**0.5745*** **	0.6770***
Douban Conversation Corpus	0.7781***	0.4994***	0.6388***
Xiaohuangji Conversation Corpus	0.7919***	0.4378***	0.6149***
PTT Gossiping Corpus	0.7831***	0.4687***	0.6259***
Weibo Dialogue Corpus	0.8238***	0.5389***	0.6813**
Qingyun Dialogue Corpus	0.7783***	0.4883***	0.6333***

Table 4: The result of comparison experiment with public datasets. Chinese Internet Dialogue Corpus is our dataset. *, **, and *** indicate p-values derived from t-tests, representing statistical significance levels of < 0.05, < 0.01, and < 0.001, respectively.

may vary in their specific choices or expressions, yet all exhibit high semantic relevance. This flexibility makes semantic relevance well-suited for assessing open-domain dialogue tasks and evaluating model performance.

In numerous previous studies, semantic relevance has consistently served as a key evaluation metric. For example, evaluators are explicitly tasked with determining whether a suitable response is semantically coherent and thematically aligned with the given input (Xu et al., 2018; Ritter et al., 2011). Similarly, semantic relevance is used as contextual guidance in generative models to improve both the correlation and diversity of the generated content (Pei and Li, 2018). MAUDE employs semantic relevance to generate and distinguish between positive and negative samples, thereby modeling the semantic alignment between context and response, and consequently enhancing the accuracy and robustness of unreferenced metrics for dialogue evaluation (Sinha et al., 2020). For these reasons, semantic relevance is integrated as a core component of the dialogue evaluation scoring method proposed in this paper.

In this study, we adopt pre-trained BERT embeddings instead of fasText embeddings to compute semantic relevance. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model designed to understand context by processing text bidirectionally, enabling tasks such as question answering and natural language understanding (Devlin, 2018). BERT's contextual word embeddings better capture deep semantic relationships, making them more appropriate for high-quality evaluation. While fastText is suitable for rapid and lightweight data filtering, BERT-based embeddings allow for more objective and precise assessments of the dataset's semantic integrity.

For each utterance pair $((user_x, x), (user_y, y))$, let $v(x)$ and $v(y)$ denote the BERT-based contextual word embeddings of the utterance x after dialogue conversion and its corresponding response y, respectively. The semantic relevance score is computed as follows:

$$R(x, y) = max(\frac{v(x) \cdot v(y)}{|v(x)| \cdot |v(y)|}, 0) \tag{2}$$

4.2 Domain Adaptiveness

Domain adaptiveness is a technique for extracting data relevant to a specific target domain and has been widely applied to parallel corpora, particularly in the field of machine translation. For instance, a domain adaptiveness-based method that combines bidirectional cross-entropy, semantic embedding similarity, and domain characteristics was introduced to identify high-quality, low-noise data (Junczys-Dowmunt, 2018). To address the scarcity of standard translation corpora for low-resource languages (e.g., Japanese), domain adaptiveness has been used to select low-noise, highly relevant sentence pairs from noisy web-crawled data, thereby providing more reliable training data sources for low-resource translation models (Zhang et al., 2020). Although domain adaptiveness is frequently used for parallel corpora, dialogue data are not parallel corpora, which explains its less frequent application. Drawing inspiration from previous studies in the machine translation field, this paper employs domain adaptiveness as a measure of dialogue quality.

The purpose of using the domain adaptiveness in this study differs from that in the machine translation domain. While domain adaptiveness in machine translation is primarily used for noise reduction, the dialogue data in this paper have already undergone various noise-filtering steps during the dialogue conversion process. Instead, goal is to further refine the selected dialogue data to align more closely with human conversational style. This is similar to how domain techniques are leveraged to extract target domain-relevant data from non-domain-specific corpora (Moore and Lewis, 2010). However, our focus is primarily on ensuring the naturalness and everyday style of the dialogue data.

For our domain adaptiveness, following previous studies, we adopt the cross-entropy difference scoring method. More specifically, we use the dataset from CLUE, associated with the next sentence prediction task, as the in-domain dataset I, and the instruction-tuned dataset as the non-domain dataset N (Xu et al., 2020; Cui et al., 2023). The data style in CLUE is more casual, covers a wider range of topics, and closely reflects everyday human life. In contrast, the instruction-tuned data are more directive in nature, involve fewer topics, and rarely relate to daily human activities. Data derived from social media platforms tend to resemble the style of CLUE more closely than that of the instruction-tuned data.

Based on the two domain datasets, we fine-tune two models M_N and M_I. Let $P_N(x, y)$ and $P_M(x, y)$ represent the predicted probabilities from M_N and M_I respectively, for a pair of utterances $((user_x, x), (user_y, y))$. We then adopt the cross-entropy difference scoring as the domain adaptiveness score, computed as follows:

$$D(x, y) = \frac{\boldsymbol{e}^{P_I(x,y)}}{\boldsymbol{e}^{P_N(x,y)} + \boldsymbol{e} - 1} \qquad (3)$$

Whereas previous studies employing cross-entropy difference scoring generally use generative models, this paper uses a discriminative model and applies exponentiation-based normalization to produce a more smoothly varying scoring curve.

5 Experiments and Results

In this section, we describe the validation experiment for the scoring method and then apply it to our dataset, comparing the results with those from several publicly available datasets.

5.1 Validation Experiment for the Scoring Method

To validate our scoring method, we conducted a correlation analysis between our scoring method and human evaluation. We randomly sampled 250 utterance pairs from a noisy Chinese dialogue dataset derived from Weibo. As described in the section 3.1, the Weibo dataset employs a straightforward approach that assumes all sender-recipient pairs correspond to post content p. Additionally, several rule-based methods are used to filter out low-quality dialogue data.

Next, we recruited four native Chinese-speaking students to evaluate each utterance pair by answering the question: "Do you think this utterance pair could serve as a natural daily conversation?" The students were asked to provide their answers on a five-point Likert scale (from 5: Strongly agree to 1: Strongly disagree) (Likert, 1932). The average score provided by the four students was taken as the human score for each utterance pair.

Subsequently, we calculated scores for the same 250 samples using the dialogue evaluation scoring method described earlier, producing a model score $S(x, y)$.

Since the human score and model score do not follow a joint normal distribution, Spearman correlation analysis was employed. Additionally, as part of an ablation study, we separately computed the Spearman correlation between the semantic relevance $R(x, y)$ and the human score, as well as between the data adaptiveness $D(x, y)$ and the human score.

5.2 Comparison Experiment with Public Datasets

To validate our dataset, we applied the proposed scoring method to six well-known Chinese dialogue datasets and compared their scores with ours. For each dataset, we performed 10 rounds of random sampling, selecting 1,000 utterance pairs per round. Each pair was scored, and the average score per sample set was computed. For multi-turn datasets, only the first exchange was used. The six datasets are as follows:

- E-commerce Dialogue Corpus: This dataset consists of multi-turn dialogues from China's Taobao e-commerce platform, containing 1 million utterances. The average number of dialogue turns is 5.51, and the average sentence length is 7.02.
- Douban Conversation Corpus: As described in the section 3.1, this dataset contains multi-turn dialogues from China's Douban platform. It includes 1 million utterances, with an average of 6.69 dialogue turns per conversation and an average sentence length of 18.56.
- Xiaohuangji Conversation Corpus: This dataset is a single-turn dialogue corpus collected from the Chinese internet, containing 0.45 million utterances.
- PTT Gossiping Corpus: This dataset is derived from the PTT Gossiping forum (a Taiwanese bulletin board system) and consists of single-turn dialogues. It is an unstructured dataset resembling a post-response structure and contains 0.77 million utterances.
- Weibo Dialogue Corpus: As described in the section 3.1, this single-turn dialogue dataset comes from the Chinese social media platform Weibo. It includes 4.43 million utterances.
- Qingyun Dialogue Corpus: This single-turn dialogue dataset is collected from the Chinese internet and contains 0.1 million utterances.

5.3 Results and Analysis

As shown in Table 3, the model score $S(x, y)$ exhibits a significant correlation with human scores ($p < 0.05$), while neither $R(x, y)$ nor $D(x, y)$ individually shows significant correlation with human scores ($p > 0.05$). This result demonstrates the effectiveness of our scoring method, providing a solid foundation for subsequent evaluations on other datasets. Moreover, unlike previous studies that focused solely on semantic relevance, this result highlights the necessity of a domain adaptiveness. For filtering conversational data in daily dialogue, relying solely on semantic relevance may be insufficient. Most importantly, it indicates the feasibility of introducing the domain adaptiveness method into dialogue datasets.

Table 4 presents a comparative analysis of our Chinese Internet Dialogue Corpus against six other datasets using our scoring method. We found that our dataset consistently achieved significantly higher model scores $S(x, y)$ compared to the others. However, in terms of $D(x, y)$, we fell behind the Weibo Dialogue Corpus and the E-commerce Dialogue Corpus. For the Weibo Dialogue Corpus, we hypothesize that its higher $D(x, y)$ score stems from its more diversified community dialogue styles. While collecting Baidu Tieba dialogue data, we restricted our selection to posts from the 12 largest Baidu Tieba community topics, whereas the Weibo Dialogue Corpus did not have such limitations. For the E-commerce Dialogue Corpus, we speculate that its higher $D(x, y)$ score is due to the use of the CLUE dataset during the scoring method fine-tuning phase, which includes a portion of e-commerce domain data. This likely introduced a bias in our domain adaptiveness, favoring e-commerce domain data.

Regarding $R(x, y)$, our dataset significantly outperforms all others, indicating that our designed MMCM mechanism can effectively enhance the semantic relevance of matches.

5.4 Conclusion

This paper addresses the challenge of constructing high-quality dialogue datasets from noisy social media data by proposing a novel transfer method that combines model-based and rule-based techniques. We also introduce an innovative scoring method for dialogue quality evaluation, showing notable correlation with human judgment. Using these advancements, we present the Chinese Internet Dialogue Corpus (CIDC), comprising over 3 million short-text dialogues sourced from Baidu Tieba, setting a new benchmark for Chinese dialogue datasets.

Limitations

While our scoring method proves effective for large-scale automatic evaluation, its correlation with human judgment remains weak, suggesting the need for further refinement to more accurately assess the quality of daily-style Chinese conversations. Moreover, the proposed MMCM mechanism has only been validated on data from Baidu Tieba. Its applicability to other social media platforms, especially those with different linguistic or structural characteristics, remains unverified. Future work should explore its generalizability and investigate potential biases introduced by platform-specific dialogue features.

References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 941–958, Online. Association for Computational Linguistics.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Matthew Henderson, PawełBudzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkˇ sić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. 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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748, Miyazaki, Japan. European Language Resources Association (ELRA).

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Jiaxin Pei and Chenliang Li. 2018. S2SPMN: A simple and effective framework for response generation with relevant information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 745–750, Brussels, Belgium. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

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.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference*, pages 91–103.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Liang Xu, Hai Hu, and Xuanwei Zhang. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. 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 2070–2080, New Orleans, Louisiana. Association for Computational Linguistics.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Dataset Usage and License Notes

In this work, we use several publicly available dialogue datasets for comparative evaluation, including the Douban Conversation Corpus, Weibo Dialogue Corpus, PTT Gossiping Corpus, Xiaohuangji Corpus, Qingyun Dialogue Corpus, and the E-commerce Dialogue Corpus. All of these datasets are properly cited in Section 5.2, along with their respective sources such as academic papers or public repositories.

These datasets are released under terms that allow academic use and redistribution for non-commercial research purposes. Our use strictly adheres to these conditions, and all datasets were used solely for evaluation and benchmarking in a research context. Table 5 summarizes the license and usage conditions for each dataset. Furthermore, the Chinese Internet Dialogue Corpus (CIDC) introduced in this paper is intended for academic research only, and will be released under a research-permissive license to ensure compliance with typical community standards.

To address ethical and safety concerns, we applied extensive rule-based filtering to remove utterances containing personal information (e.g., phone numbers, email addresses, and names), offensive language, political sensitivity, URLs, emojis, and meaningless repetitive content. These procedures are detailed in Section 3.3 and ensure that the final dataset does not contain personally identifiable or harmful content.

The structure and coverage of the CIDC dataset are documented in the main text. It consists of over 3 million high-quality single-turn dialogue pairs extracted from 12 major Baidu Tieba community forums, covering a wide range of daily-life topics and informal conversational styles. The dataset includes utterance-level metadata and contextual linking via MMCM for improved coherence.

We also provide detailed dataset statistics, including the number of total utterance pairs, average sentence lengths, filtering rates, and scoring results across datasets. These statistics are reported in Sections 3.3 and 5.2 and will be further expanded in the following parts of the appendix.

B Topic Distribution and Semantic Diversity

To highlight the thematic richness and diversity of the CIDC dataset, we provide a pie chart in Figure 2 showing the number of dialogue pairs extracted from each of the 12 Baidu Tieba forums included in our collection.

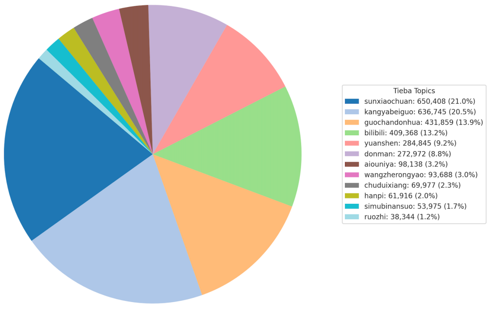

Figure 2: Distribution of dialogue samples across 12 Baidu Tieba communities in the CIDC dataset.

These communities span a broad spectrum of topics, enabling the dataset to capture a wide variety of real-world conversational styles. The topics include:

- **sunxiaochuan, kangyabeiguo, guochandonhua**: Focused on patriotic culture, controversial internet celebrities, and sociopolitical discourse.
- **bilibili, yuanshen, donman, wangzherongyao**: Covering pop culture, anime/games (e.g., Genshin Impact), video sharing, and mobile gaming.
- **chuduixiang, simubinansuo**: Oriented toward emotional expression, anonymous confessions, and parody of psychological counseling.
- **aiouniya, hanpi, ruozhi**: Satirical and humor-driven subcultural communities, often showcasing sarcasm, absurdity, and creative trolling.

This distribution ensures that the CIDC dataset does not merely reflect task-oriented or formal interactions, but also encompasses highly informal, humorous, emotionally charged, and socially nuanced dialogues. Such diversity is crucial for training dialogue models that aim to generalize across different user styles and domains in real-world online environments.

C Utterance and Reply Analysis

To evaluate the lexical richness and structural diversity of CIDC, we analyze both the character-level

Dataset Name	Source Platform	License / Terms	Notes
Douban Conversation Corpus	Douban	No explicit license	Widely used in prior NLP work.
Weibo Dialogue Corpus	Weibo	No explicit license	Widely used in prior NLP work.
PTT Gossiping Corpus	PTT (Taiwan BBS)	Apache License 2.0	Open-source under Apache 2.0.
Xiaohuangji Corpus	Chinese online forums	No explicit license	Openly crawled; used for academic.
Qingyun Dialogue Corpus	Chinese forums	Apache License 2.0	Cited in prior published work.
E-commerce Dialogue Corpus	Taobao / Alibaba	No explicit license	Collected and shared by prior studies.
Amazon Musical Instruments Reviews	Amazon	CC0 (Public Domain)	No restrictions; fully public dataset.
Twitter US Politics	Twitter	CC0 (Public Domain)	No restrictions.

Table 5: Licensing and usage conditions for datasets used in this paper.

lengths and semantic relatedness scores of dialogue pairs. These analyses provide evidence for the naturalness, coherence, and usability of the dataset in training robust conversational models.

Length Statistics. We begin with the basic statistics of the utterance (preceding text) and reply (following text). Table 6 summarizes the minimum, average, and maximum lengths in characters.

Field	Min	Average	Max
Utterance	11	43.39	1032
Reply	9	29.23	180

Table 6: Text length statistics of utterances and replies in CIDC.

Compared to traditional short-text datasets such as Xiaohuangji (with an average reply length under 20 characters), CIDC provides significantly more expressive user utterances and moderately long replies. This allows models trained on CIDC to better learn diverse linguistic structures and context-aware responses, especially in informal and emotionally rich conversations.

Length Distributions. Figure 3 and Figure 4 illustrate the character-level distribution of utterances and replies, respectively. The utterance distribution shows a wide spread, with a heavy concentration between 20–60 characters, while still retaining long-tail samples up to 1000 characters. This reflects the natural variability found in real-world social forums, where some users express themselves briefly and others write detailed narratives or rants.

The reply length distribution peaks around 20–30 characters, which is typical of short-form responses in online discussion. Importantly, replies in CIDC are significantly longer and more semantically meaningful than those in datasets like the Weibo Dialogue Corpus or PTT, where many responses are single phrases or emojis. This richer reply content improves the potential for training generation models that require longer-range coherence.

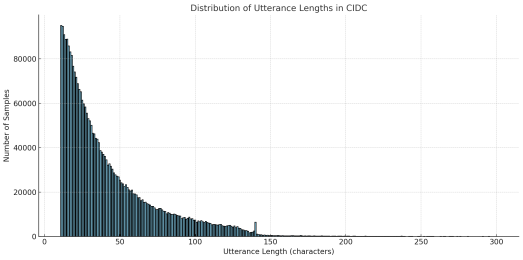

Figure 3: Distribution of utterance lengths (character count).

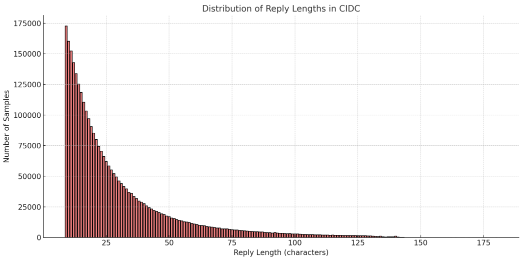

Figure 4: Distribution of reply lengths (character count).

Semantic Coherence. To quantify the semantic alignment between utterances and replies, we compute a relatedness score using a pre-trained semantic similarity model. Figure 5 shows the histogram of these scores. Over 70% of pairs score above 0.8, indicating that CIDC maintains high-quality alignment while preserving natural conversational variance. Unlike some web-mined corpora that include loosely related pairs for coverage, CIDC carefully filters for coherence.

Conclusion. Together, the lexical length diversity and high semantic relatedness demonstrate that CIDC is not only large-scale, but also struc-

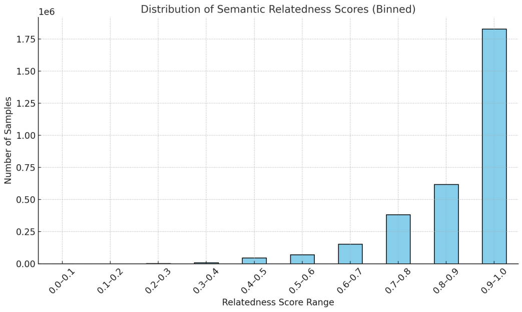

Figure 5: Distribution of semantic relatedness scores between utterances and replies.

turally and semantically rich. These qualities make it highly suitable for training both retrieval-based and generation-based dialogue models, particularly in settings that require nuanced understanding of informal, user-generated content.