CHAmbi: A New Benchmark on Chinese Ambiguity Challenges for Large Language Models

Qin Zhang[1], Sihan Cai[1], Jiaxu Zhao[2], Mykola Pechenizkiy[2], Meng Fang[3]*
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Department of Mathematics and Computer Science, Eindhoven University of Technology
[3]Department of Computer Science, University of Liverpool
qinzhang@szu.edu.cn, caisihan2023@email.szu.edu.cn, {j.zhao, m.pechenizkiy}@tue.nl, Meng.Fang@liverpool.ac.uk

Abstract

Ambiguity is an inherent feature of language, whose management is crucial for effective communication and collaboration. This is particularly true for Chinese, a language with extensive lexical-morphemic ambiguity. Despite the wide use of large language models (LLMs) in numerous domains and their growing proficiency in Chinese, there is a notable lack of datasets to thoroughly evaluate LLMs' ability to handle ambiguity in Chinese. To bridge this gap, we introduce the CHAmbi dataset, a specialized Chinese multi-label disambiguation dataset formatted in Natural Language Inference. It comprises 4,991 pairs of premises and hypotheses, including 824 examples featuring a wide range of ambiguities. In addition to the dataset, we develop a series of tests and conduct an extensive evaluation of pre-trained LLMs' proficiency in identifying and resolving ambiguity in the Chinese language. Our findings reveal that GPT-4 consistently delivers commendable performance across various evaluative measures, albeit with limitations in robustness. The performances of other LLMs, however, demonstrate variability in handling ambiguity-related tasks, underscoring the complexity of such tasks in the context of Chinese. The overall results highlight the challenge of ambiguity handling for current LLMs and underscore the imperative need for further enhancement in LLM capabilities for effective ambiguity resolution in the Chinese language.

1 Introduction

Ambiguity is a prevalent and noteworthy linguistic phenomenon (Piantadosi et al., 2012). Differences in context or individual interpretations will lead to different understandings of ambiguous sentences, complicating the comprehension and communication of information (Piantadosi et al., 2012; Chao and Zipf, 1950). Large language models (LLMs) have made remarkable achievements in natural language processing (NLP) (Brown et al., 2020a; Bang et al., 2023; Brown et al., 2020b), making them essential tools in daily life. Consequently, it is crucial to scrutinize and evaluate the performance of LLMs in handling ambiguity. This will enhance the reliability and effectiveness of LLMs in intricate contexts and promote the development of natural language understanding and generation.

There are a few works pay attention to ambiguity problem (Min et al., 2020; Pavlick and Kwiatkowski, 2019; Jiang and Marneffe, 2022; Liu et al., 2023; Nie et al., 2020). Min et al. construct AmbigNQ, offering multiple potential answers for ambiguous open-domain questions and supplying disambiguation questions corresponding to each answer. They also built a baseline model for generating multiple answers to open-domain questions (Min et al., 2020). Pavlick and Kwiatkowski conduct a thorough examination of disagreement in human judgments on the Natural Language Inference (NLI) task. Their findings reveal that numerous disagreements remain after augmenting the number of annotators and the contextual information provided (Pavlick and Kwiatkowski, 2019). All these efforts contribute to a deeper understanding and resolution of language ambiguity. However, they mainly focus on the ambiguity problem in English. Chinese, as a language significantly distinct from English (Ling and Mahadi, 2016), presents greater intricacy at the lexical and morphemic levels, involving a wider range of ambiguity types. For example, if *A*, *B* and *AB* are all possible words, then *AB* exhibits combinatorial ambiguity, such as "才"(just), "能"(be able to) and "才能"(talent).

In the Chinese context, there is a lack of annotated datasets to comprehensively evaluate the ability of LLMs to identify and resolve ambiguity. To address this gap, we introduce CHAmbi[1], a Chi-

*Corresponding author.

[1]Dataset available at: github.com/aialt/CHAmbi.

nese multi-label disambiguation dataset for Natural Language Inference. The dataset contains 4,991 examples, including 824 ambiguous instances, each annotated with two or more disambiguation labels, representing different interpretations for resolving the ambiguity. Enumerating all possible interpretations of a ambiguous sentence is challenging. Therefore, we define ambiguity using the format of premise-hypothesis pairs in NLI dataset inspired by AmbigQA (Min et al., 2020). Natural language inference is a task to identify whether a hypothesis is true, false, or uncertain given a premise (Hu et al., 2020). Ambiguity in premise and (or) hypothesis may affect the determination of labels, leading to multi-label example. We only need to fully analyze the potential multiple relationships between the premise and the hypothesis, and list a different interpretation for each relationship.

To standardize ambiguity identification, we define seven categories of ambiguity, including spoken ambiguity (accent ambiguity, pause ambiguity), written ambiguity (vocabulary, grammar, semantics, reference ambiguity) and incomplete ambiguity. Through the CHAmbi dataset, we design various tests to evaluate the ambiguity identification and disambiguation abilities of popular Chinese-supporting generative models, including the GPT series model, Baidu large model ERNIE-Bot, the Chinese large model Atom[2], and the Chinese-English bilingual large model ChatGLM2 (Du et al., 2022). Our main contributions are:

- To our knowledge, we present the first publicly available Chinese disambiguation language dataset: CHAmbi. It contains 4991 examples, covering ambiguities such as vocabulary, grammar, semantics, etc.
- We extensively evaluate the ambiguity handling capabilities of popular Chinese-supporting generative models. The results indicate that handling ambiguity is a challenge for existing large language models.
- Leveraging the CHAmbi dataset, we fine-tune the open-source large model ChatGLM2. Experimental results demonstrate an improvement in identifying ambiguity.

2 Related Work

Ambiguity and disambiguity. The ambiguity of language is a longstanding challenge in NLP tasks. Some research emphasizes the presence of ambiguity in various tasks, such as question answering (Shafahi et al., 2019), frame disambiguation (Gangemi et al., 2016), coreference resolution (Poesio and Artstein, 2005).

As the NLP progresses toward higher-level understanding and reasoning, AMBIGQA introduces a new task, which involves providing multiple possible answers to an open domain question that may be ambiguous, and providing a disambiguated question for each answer, and proposes a dataset AMBIGNQ about this task (Min et al., 2020). Xu et al. transform the CPH disambiguation task into a challenging Natural Language Inference (NLI) task, introducing the first Chinese Adversarial NLI Challenge Set (CANLI). The study emphasizes the challenges in CPH because causative and passive forms cannot be distinctly identified through sentence syntactic structures. The poor fine-tuning performance of pre-trained transformer models like RoBERTa on large-scale Chinese NLI benchmark datasets further emphasizes the difficulty in handling such ambiguity problems (Xu and Markert, 2022). By developing a classification of divergent sources with 10 categories, Jiang and de Marneffe find that the inconsistency may come from the uncertainty of sentence meaning, the bias of annotators and task artifacts. In order to detect potentially inconsistent items, researchers explore two modeling methods, including a four-way classification and a multi-label classification. It is found that multi-label classification can better trace back the possible explanations in the data (Jiang and Marneffe, 2022).

Label variation on NLI. NLI is the task of determining the logical relationship between a given premise and a hypothesis, which can be divided into three types: entailment, contradiction and neutral. Entailment means that the premise can imply the truth of the hypothesis. Contradiction means the direct conflict between premise and hypothesis, including opposition and mutual exclusion. Neutral means that there is no clear entailment or contradiction between premise and hypothesis, and they can exist independently (Storks et al., 2020).

The research on label variation in NLI has seen notable contributions. Pavlick and Kwiatkowski (Pavlick and Kwiatkowski, 2019) initiate groundbreaking work in this domain, subsequently, Nie et al. (Nie et al., 2020) expand this by collecting more annotations. Additionally, efforts are made to model label variation (Zhou et al., 2021; Zhang

[2] https://github.com/FlagAlpha/Llama2-Chinese

Example	Rewrites I	Rewrites II	Category
P: 他们正在看电影。 (They are watching a film.) H: 他们正在电影院里。 (They are in the cinema.) L: entailment,contradiction	P: 他们正在电影院里看电影。 (They are watching a film in the cinema.) H: 他们正在电影院里。 (They are in the cinema.) L: entailment	P: 他们正在家里看电影。 (They are watching a film at home.) H: 他们正在电影院里。 (They are in the cinema.) L: contradiction	incomplete
P: 小明邀请了她，小暗也邀请了她，她犹豫了一下，然后接受了他的邀请。 (XiaoMing invited her, and XiaoAn also invited her. She hesitated, and then accepted his invitation.) H: 她接受了小明的邀请。 (She accepted XiaoMing's invitation.) L: entailment,contradiction	P: 小明邀请了她，小暗也邀请了她，她犹豫了一下，然后接受了小明的邀请。 (Xiao Ming invited her, and Xiao An also invited her. She hesitated, and then accepted XiaoMing's invitation.) H: 她接受了小明的邀请。 (She accepted XiaoMing's invitation.) L: entailment	P: 小明邀请了她，小暗也邀请了她，她犹豫了一下，然后接受了小暗的邀请。 (Xiao Ming invited her, and Xiao An also invited her. She hesitated, and then accepted XiaoAn's invitation.) H: 她接受了小明的邀请。 (She accepted XiaoMing's invitation.) L: contradiction	reference
P: 她慢慢地放下了手中的书，转身面对我。 (She slowly put down the book in her hand and turned to face me.) H: 她转过身来，慢慢地放下了手中的书。 (She turned around and slowly put down the book in her hand.) L: entailment,contradiction	P: 她慢慢地放下了手中的书，同时转身面对我。 (She slowly put down the book in her hand and turned to face me at the same time.) H: 她转过身来，同时慢慢地放下了手中的书。 (She turned around and slowly put down the book in her hand at the same time.) L: entailment	P: 她先慢慢地放下了手中的书，然后转身面对我。 (She slowly put down the book in her hand and then turned to face me.) H: 她先转过身来，然后慢慢地放下了手中的书。 (She turned around and then slowly put down the book in her hand.) L: contradiction	incomplete

Table 1: Some ambiguous examples in CHAmbi dataset, where "P, H, L" means "Premise, Hypothesis, Label". We underline the ambiguous premise and (or) hypothesis that lead to multi-label for the example. "Rewrites I/II" is the disambiguation rewrite of the ambiguous premise and (or) hypothesis for each label.

et al., 2021). Some focus on predicting the probability of entailment (Zhang et al., 2016; Chen et al., 2020). Another intriguing approach introduces a fourth "inconsistent" label (Zhang and de Marneffe, 2021). We introduce a novel research approach to NLI. In this method, the task for NLI models is not predicting a single label but forecasting a set of labels. It enhances task complexity and provides a different perspective on understanding relationships between premises and hypotheses. Jiang and de Marneffe (Jiang and Marneffe, 2022) conduct an in-depth analysis of the Multi-NLI (MNLI) dataset, categorizing sources of divergence, with a specific mention of lexical and semantic ambiguity. Our research incorporates a broader range of ambiguity types, contributing to a more comprehensive understanding of the sources of divergence.

3 Dataset Construction

Here, data collection, annotation, and validation process are described. The final dataset contains 4991 NLI examples, of which 824 are ambiguous. Each example has a premise sentence and a hypothesis sentence. Each ambiguous example is annotated with a set of labels, reflecting the ambiguity in the premise and (or) hypothesis. Additionally, it includes ambiguity category and the disambiguation rewrites of the premise and (or) hypothesis for each label, as shown in Table 1.

We employ two methods to collect ambiguous NLI examples: automatic generation and keyword crawling. Automatic generation (Section 3.1) takes advantage of text generation and pattern replication capabilities of LLMs to obtain potentially ambiguous examples. Crawling and manual curation (Section 3.2) involves acquiring ambiguous sentences from various sources as premises, followed by manual formulation of hypotheses. Then we manually annotate and validate collected examples, and obtain high-quality examples and disambiguation rewrites of ambiguous examples.

3.1 Automatic Generation

Inspired by WANLI (Liu et al., 2022), we leverage the generative ability of LLMs to create examples.

The method requires a Chinese NLI dataset as the initial dataset D. We use the CMNLI dataset[3]. This method also requires a strong language model trained on D as a classifier. We use the fine-tuned Roberta-large-Chinese model (Cui et al., 2020, 2019) M on CMNLI as the classifier.

We initially employ data map (Swayamdipta et al., 2020) to automatically identify ambiguous examples in D. Following this, we leverage the pattern replication capability of ChatGPT to over-generate similar new examples. The generated examples may be low quality or lack ambiguity, we apply simple rules to filter out the failed generated examples and further employ model M to filter the more ambiguous examples.

Data Map. A tool based on training dynamics aims to characterize and diagnose the quality of large datasets in NLP research. It leverages the model's behavior on individual examples during training to generate two measures: the confidence in the true class and the variability of this confidence throughout the training process. Using these two measures, we build a data map that divides examples in D into three different regions: easy-to-learn (high confidence, low variability), difficult-to-learn (low confidence, low variability), and ambiguous (high variability) (Swayamdipta et al., 2020). We focus on the ambiguous region and select 26,359 examples with variability $>= 0.3$ as D_{ambi}.

Over generation. For each example x_i in D_{ambi}, we can find 2 nearest neighbors in D_{ambi} according to the last layer representation of the model M. We consider that these three examples share similar reasoning patterns. Combining these three examples with an instruction "*Write a pair of sentences that have the same connection to each other. For example:*", we construct a prompt for ChatGPT, as shown in Table 2. In the prompt, we place x_i as a seed example at the end to enhance the similarity between newly generated examples and seed example. We generate three new examples parsed into premises and hypotheses for each prompt, resulting in D_{gen}. To control costs, D_{gen} is ultimately comprised of only 36,431 examples.

Automatic filtering. We discard the following examples in D_{gen}: 1) the premise and hypothesis are the same, 2) the generated example is a copy of the example in prompt, 3) the premise or hypothesis is a question (unlike assertions, questions have no truth value (Groenendijk and Stokhof, 1984). Therefore, it is theoretically unclear to annotate whether a question is ambiguous.), 4) the premise or hypothesis is shorter than 5 characters. Then we obtain D_{filex}. The number of D_{filex} is large and the cost of manual annotation is high, so we employ model M to predict D_{filex} and further filter more ambiguous examples. If the model predicts two or more labels with a probability > 0.05 for a given example, we consider it as an example of a greater likelihood of ambiguity. In this step, we obtain $D_{filambi}$, which contains 4632 unlabeled premise-hypothesis pairs.

写一对彼此之间有和所给例子相同联系的句子对。例子如下: (Provide a pair of sentences that share similar connections as the given example. Examples:) 1. Sentence 1: 他们说他们在这个地方找到了一个很好的工作。 (They said they found a good job at this place.) Sentence 2: 我在这个地方找到了一个很好的工作。 (I found a good job at this place.) 2. Sentence 1: 他在研究一种新的治疗方法，这种方法可以帮助病人康复。 (He is researching a new treatment that can aid patients in recovery.) Sentence 2: 我听说医生正在研究一种新的治疗方法，这种方法可以帮助病人康复。 (I heard that the doctor is researching a new treatment that can aid patients in recovery.) 3. Sentence 1: 我昨天晚上梦到了一个很奇怪的场景，我在一个陌生的城市里迷路了。 (Last night, I dreamt of a strange scene where I got lost in an unfamiliar city.) Sentence 2: 我从来没有在梦里迷路过。 (I have never gotten lost in a dream before.) 4. Sentence 1:

Table 2: Input prompt template for over generate examples.

3.2 Crawling and Manual Curation

Through the keywords "以下"(following), "歧义"(ambiguity), and "句子"(sentence), we gather a variety of Questions and Answers related to ambiguous sentences from 12tiku.com[4]. We also collect many ambiguous sentences from a paper published in the Journal of Language Research (Huang, 1985). After manual curation, we end up with 388 ambiguous sentences as the premises of the examples. Then we compile hypotheses using simple strategies: 1) a certain interpretation of the premise, 2) a fact inferred from a certain interpretation of the premise, 3) the negation of a fact inferred from a certain interpretation of the premise, and 4) a supplementary interpretation of the premise. For

[3] https://github.com/CLUEbenchmark/CLUE

[4] https://www.12tiku.com/

example, if the premise is "I go to the classroom.", the hypothesis compiled is "I go to the classroom to teach."

3.3 Human Annotation and Validation

During the annotation process, annotators receive detailed guidelines and examples for reference, and we constantly refine and clarify our annotation rules to improve the quality of annotations.

Annotation. Annotators annotate and disambiguate the premise-hypothesis pairs obtained in sections 3.1 and 3.2, with each example annotated by two annotators. Annotators are instructed to analyze whether the premise and (or) hypothesis are ambiguous, analyze relationship between the premise and hypothesis (entailment, contradiction, and neutral), and assign a set of labels or a label to the example. For the examples obtained in Section 3.2, only the premises are ambiguous by default.

To enable annotators to have a deeper understanding of ambiguity and make higher quality annotations, we further categorize ambiguity and require annotators to annotate the ambiguity category. Additionally, annotators have the flexibility to modify the premises and (or) hypotheses according to their understanding, transforming them into ambiguous or higher quality examples. When annotators identify an example as ambiguous and assign a set of labels, they are required to provide the sentence disambiguation rewrite for each label. During the disambiguation rewrites, annotators are asked to eliminate ambiguity with minimal modifications. If an example is of low quality or offensive, annotators can discard it. See Appendix A for details.

Validation. We acknowledge that different people may have different interpretations of the same sentence, leading to a low probability of agreement on the ambiguity of a given example. Therefore, for a batch of example annotated by every two annotators, the two annotators and the third validator validate the annotations together. If an example is judged as ambiguous by only one annotator, but is confirmed as ambiguous after three validators' analysis and discussion, it is retained. During this phase, all validators are required to: 1) Analyze whether the examples labeled as ambiguous are really so, 2) Ensure the completeness of the label set, 3) Assess whether disambiguation rewrites maintain ambiguous, 4) Verify the accuracy of the assigned ambiguity category. These steps contribute to the high quality of the validation process.

3.4 Dataset Analysis

After collection, annotation, and validation, we finally construct CHAmbi dataset. Table 3 provides the statistics of labels in the dataset. The dataset consists of a total of 4,991 NLI examples, among which 824 are ambiguous and have two or more labels. In addition, it contains 792 ambiguous premises and 93 ambiguous hypotheses.

Dataset	E	N	C	E,N	E,C	N,C	E,N,C
CHAmbi	1389	2552	226	474	277	59	14

Table 3: The label distribution of the CHAmbi dataset, where "E, N, C" denotes that the label of example is "entailment, neutral, contradiction".

4 Evaluation of LLMs

With the proposed dataset CHAmbi, we evaluate the capabilities of existing popular language models in two aspects: ambiguity identification and ambiguity resolution. Two tests evaluate ambiguity identification capabilities, including evaluating whether LLMs can recognize ambiguities in premises or hypotheses (Section 4.1), and whether they can recognize ambiguity and perform correct multi-label classification for examples (Section 4.2). Additionally, two tests evaluate disambiguation capabilities, including evaluating whether LLMs can directly generate disambiguation rewrites (Section 4.3), and whether they can recognize the effectiveness of disambiguation rewrites (Section 4.4). For these tests, we only focus on two-label and three-label examples from CHAmbi.

The LLMs we select for evaluation are GPT-3 (davinci), InstructGPT (text-davinci-003), ChatGPT (gpt-3.5-turbo), GPT-4, Baidu large model ERNIE-Bot (ERNIE-Bot-turbo-0704), the open-source large pre-trained model Atom-7b based on Llama2 for Chinese, and the bilingual open-source conversational model ChatGLM2-6b. To better evaluate models' capabilities, we also compare human performance as a baseline. Due to the high time and labor costs, for each test, we randomly select 50 examples and ensure that three participants participated.

4.1 Recognizing Ambiguity

Initially, we consider the ability of LLMs to directly recognize ambiguities in premises or hypotheses. We construct a simple prompt with the format of

Model	Recognizing Ambiguity				Multi-label Classification				
	average accuracy	macro_F1	macro_p	macro_r	micro_F1	micro_p	micro_r	2 labels accuracy	3 labels accuracy
davinci	35.32	26.10	19.84	38.14	67.25	67.91	66.61	33.21	0.00
text-davinci-003	45.93	44.98	45.06	45.21	66.99	73.41	61.61	23.95	0.00
gpt-3.5-turbo	53.34	35.14	39.27	49.69	57.00	68.64	48.74	1.73	7.14
gpt-4	49.27	47.67	51.02	50.84	73.37	71.35	75.51	23.09	35.71
ERNIE-Bot-turbo	54.07	36.60	59.47	50.46	35.43	63.65	24.55	1.11	7.14
atom-7b	53.70	34.94	26.85	50.00	42.50	59.98	32.91	4.44	0.00
chatglm2-6b	49.09	36.45	38.10	46.04	66.62	71.16	62.64	26.42	0.00
humans	63.00	62.58	66.76	64.66	62.04	76.70	52.15	18.37	-

Table 4: Performance of LLMs identifying ambiguity on CHAmbi dataset. For the evaluation metrics in the multi-label classification task, micro_F1, micro_p, and micro_r consider partially correct results, while 2 and 3 labels accuracy account for completely correct results.

以下例子中，上下文可能存在歧义，主张可能存在歧义，这导致上下文和主张之间至多有三种关系:根据上下文得到主张正确，根据上下文得到主张错误，根据上下文无法确定主张正确性。回答上下文和主张之间的几种关系。 (In the following example, the context may be ambiguous and the claim may be ambiguous, which results in at most three relationships between the context and the claim: the claim is correct based on the context, the claim is wrong based on the context, and the claim is uncertain based on the context. Answer several relationships between context and claim.)
1.问题：上下文：她从冰箱里拿出一瓶水，打开盖子。主张：她从冰箱里拿出一瓶饮料，打开瓶盖。上下文和主张之间有几种关系? (Question: Context: She takes a bottle of water from the refrigerator and opens the lid. Claim: She takes out a beverage from the refrigerator and opens the lid. How many relationships exist between the context and the claim?) 回答：2种，即根据上下文得到主张正确、根据上下文得到主张错误。 (Response: Two types, namely, the claim is correct based on the context, and the claim is wrong based on the context.) 2.问题：上下文：他哥又有了一个孩子。主张：他哥至少有一个孩子。上下文和主张之间有几种关系? (Question: Context: His brother has another child. Claim: His brother has at least one child. How many relationships exist between context and claim?) 回答：(Response:)

Table 5: Input prompt template for evaluating whether LLMs can assign correct multi-labels for examples.

"*'[premise/hypothesis]' , is there any ambiguity in the sentence? Just answer 'yes' or 'no'*". Significantly, if the LLM's output resembles "*Due to the incomplete context, it is impossible to judge whether it is present or not. Both explanations are possible.*", we consider the example as ambiguous. This situation corresponds to the "incomplete" category defined during dataset construction. We use accuracy, macro F1, macro precision, and macro recall as the evaluation metrics.

Table 4 presents the comparison of the ambiguity recognition capabilities among LLMs, ERNIE-Bot-turbo achieves the highest average accuracy of 54.07%. In order to ensure the consistent contribution of ambiguity and non-ambiguity to the evaluation result, we employ macro F1, macro precision, and macro recall as evaluation metrics. GPT-4 achieves the highest macro F1 value of 47.67%, followed by InstructGPT and ERINE-Bot-turbo. All LLMs perform below human capabilities, indicating that LLMs have significant room for improvement in recognizing ambiguity without explicit prompts.

4.2 Multi-label Classification

In this evaluation, we further consider whether LLMs can recognize ambiguous examples and classify them correctly. This evaluation is equivalent to using LLMs to perform multi-label classification on data samples. We construct an instruction that emphasizes the existence of ambiguity in the premise or hypothesis, potentially leading to multiple relationships between "context" (premise) and "claim" (hypothesis), and provide two demonstration examples. The detailed prompt is presented in Table 5. All input prompts are based on the same two demonstration examples that are not included in the dataset. Regarding the evaluation metrics, micro F1, micro precision, and micro recall are considered partially correct results, while 2-label accuracy and 3-label accuracy reflect completely correct results.

Model	Generating Rewrites		Recognizing Rewrites			
	edit_F1	manual correctness	average accuracy	macro_F1	macro_p	macro_r
davinci	17.59	0.39	31.87	31.45	31.41	31.87
text-davinci-003	24.15	49.41	51.17	42.29	53.04	51.17
gpt-3.5-turbo	21.31	36.83	63.26	62.59	64.31	63.27
gpt-4	20.60	73.00	63.86	62.96	65.35	63.86
ERNIE-Bot-turbo	16.09	12.71	53.69	47.23	57.23	53.69
atom-7b	13.16	3.01	41.79	41.09	41.38	41.79
chatglm2-6b	10.88	17.30	50.54	49.31	50.59	50.54
humans	36.84	68.00	82.67	82.32	84.45	82.67

Table 6: Performance of LLMs resolving ambiguity on CHAmbi dataset.

以下每个例子中，都给了你一句上下文和一句主张，由于上下文里存在歧义，主张的正确性受到影响。列举出对上下文的两种或三种解读，这些解读会导致该主张正确、错误或不确定。
(In the following example, you are given a context and a claim, the correctness of the claim is affected by the ambiguity in the context. List two or three interpretations of the context that would make the claim true, false, or uncertain.)

上下文: 饭吃完了。(Context: Meal is over.)
主张: 饭没了。 仅仅考虑上下文，分析主张是正确、错误还是不确定？(Claim: No more food. Consider only the context, analyze whether the claim is correct, incorrect, or uncertain.)
我不知道，因为对上下文可以有几种不同的解读：(I don't know because there can be several different interpretations for the context:)
1. 吃完饭了。 则主张是不确定的。(The meal is finished, and the claim is uncertain.)
2. 饭吃光了。 则主张是正确的。(The meal was eaten up, and the claim is correct.)

上下文: 这里有的是化妆用品。(Context: Cosmetic products are available here.)
主张: 这里只有化妆用品。 仅仅考虑上下文，分析主张是正确、错误还是不确定？(Claim: This place only stocks cosmetic products. Consider only the context, analyze whether the claim is correct, incorrect, or uncertain.)
我不知道，因为对上下文可以有几种不同的解读：(I don't know because there can be several different interpretations for the context:)

Table 7: Input prompt template for evaluating whether LLMs can directly generate disambiguation rewrites.

As shown in Table 4, GPT-4 achieves the highest micro F1, reaching 73.37%, but its complete classification accuracy is not satisfactory. For 2 label examples, the highest accuracy for completely correct classification is only 33.21% (achieved by GPT-3). Although GPT-3 performs poorly on other tests, it achieves the best performance here. This demonstrates that LLMs face many difficulties and challenges when dealing with complex and multi-level language understanding and reasoning tasks, resulting in complex and inconsistent capabilities. For 3 label examples, the highest accuracy for completely correct classification is only 35.71% (achieved by GPT-4). Overall, Most models struggle to correctly identify and classify examples with two or three labels.

For this evaluation task, both GPT-4 and GPT-3 outperform humans. They are not as good as humans at recognizing ambiguity (in Section 4.1), but they perform well on the task of "recognizing ambiguity and correctly classifying multiple labels". We believe this may be affected by the task prompt. Another possible reason is that LLMs may be more accurate than humans in discerning the relationship between context and claim.

4.3 Generating Disambiguation Rewrites

This evaluation assesses the ability of LLMs to directly generate disambiguation rewrites. The previous evaluations have revealed LLMs' limited ambiguity identification ability. Therefore, we simplify the evaluation and construct an input prompt template as shown in Table 7. Additionally, to make the task simple, we do not consider examples where both context and claim are ambiguous. We not only use edit F1 (Min et al., 2020) as the evaluation index but also manually evaluate the reasonability of the disambiguation interpretations generated by LLMs. See appendix B for details of manual evaluation.

Table 6 shows the results of each model on two metrics. Notably, the InstructGPT model achieves the highest edit F1 score. The edit F1 metric evaluates the quality of generated disambiguation rewrites. Specifically, it calculates the F1 score between added and deleted words in the gold disam-

问题：{a}可能意味着{d}？只回答对或不对。 (Question: Could 'a' possibly mean 'd'? Just answer yes or no.) 回答：对 (Response: Yes.)
问题：{a}不一定意味着{d}？只回答对或不对。 (Question: Doesn't 'a' necessarily mean 'd'? Just answer yes or no.) 回答：对 (Response: Yes.)
问题：{a}不意味着{d}？只回答对或不对。 (Question: Doesn't 'a' mean 'd'? Just answer yes or no.) 回答：不对 (Response: No.)
问题：{a}仅仅意味着{d}？只回答对或不对。 (Question: Does 'a' just mean 'd'? Just answer yes or no.) 回答：不对 (Response: No.)

Table 8: 4 types of templates for evaluating LLMs whether can recognize the effectiveness of disambiguation rewrites, where {a} denotes the ambiguous sentence and {d} denotes a reasonable disambiguation rewrite.

biguation rewrite and the predicted disambiguation rewrite. We also perform manual evaluation, considering that there can be multiple interpretations of an ambiguous sentence. The ambiguity points identified by annotators may differ from those perceived by LLMs, but both are reasonable. The results indicate that the GPT-4 achieves the highest accuracy of 73%, exceeding the human benchmark of 68%, demonstrating its superior natural language understanding and text generation capabilities. It is noteworthy that most LLMs have manual correctness below 20%, with some as low as 0.39%. This shows that generating reasonable disambiguation rewrites is a significant challenge for most large models.

4.4 Recognizing Disambiguation Rewrites

In this section, we assess whether LLMs can recognize reasonable disambiguation rewrites. The evaluation only focuses on the ambiguous premises or hypotheses in examples. We use four different evaluation templates to evaluate LLMs, as shown in Table 8. An ambiguous sentence may mean some reasonable disambiguation rewrites, which is inevitable, but it does not necessarily mean it, because it is an ambiguous sentence and can be interpreted in different meanings. An ambiguous sentence does not mean some reasonable disambiguation rewrite, which is definitely wrong, but only means some reasonable disambiguation rewrite, which is naturally wrong. To ensure equal contribution of the four templates to evaluation results, we use accuracy, macro F1, macro precision and macro recall as evaluation metrics.

Model	1	2	3	4	all_avg	avg
davinci	26.59	21.35	32.79	46.76	0.73	31.87
text-davinci-003	23.72	0.17	100.00	80.79	0.00	51.17
gpt-3.5-turbo	62.99	90.42	39.55	60.06	0.97	63.26
gpt-4	63.61	95.32	23.44	73.07	1.82	63.86
ERNIE-Bot-turbo	89.41	87.94	13.75	23.66	0.12	53.69
atom-7b	51.74	53.49	38.25	23.69	0.00	41.79
chatglm2-6b	74.48	57.69	31.44	38.54	0.36	50.54
humans	93.33	86.00	65.33	86.00	49.33	82.67

Table 9: Accuracy of LLMs on the four templates of recognizing disambiguation rewrites.

As shown in Table 6, regarding all the questions formed by the four templates for all ambiguous premises or hypotheses, GPT-4 achieves the highest average accuracy (63.86%), macro F1 (62.96%), macro precision (65.35%), and macro recall (63.86%). However, all LLMs perform worse than humans. If LLMs can answer four templates correctly for a ambiguous premise or hypothesis, it signifies that LLMs have a thorough understanding of the ambiguous sentence and the annotated reasonable interpretations. Therefore, we further calculate the accuracy of models in answering all four templates correctly. As shown in Table 9, humans correctly answer all four templates with an average accuracy of 49.33%, whereas LLMs, perform below 2%. This supports the idea that LLMs can analyze ambiguity more deeply when faced with more detailed task prompt. In the absence of cues from complex tasks, LLMs may fail to adequately understand ambiguity.

We also examine the accuracy of models for each template, but the results do not reflect a consistent trend among the models. Additionally, we find inconsistency in model responses. For example, for two disambiguation rewrites "d1" and "d2" of a ambiguous premise "a", the model think that "a" just mean "d1", and also think that "a" just mean "d2". The above results indicate that it is a great challenge for LLMs to accurately recognize reasonable disambiguation interpretations.

4.5 Evaluation of Fine-tuned LLM

We perform specific fine-tuning on ChatGLM2 using the CHAmbi. Table 10 demonstrates that the fine-tuned ChatGLM2 achieves noticeable im-

Model	Recognizing Ambiguity				Multi-label Classification					
	average accuracy	macro_F1	macro_p	macro_r	micro_F1	micro_p	micro_r	1 labels accuracy	2 labels accuracy	3 labels accuracy
chatglm2-6b	49.09	36.45	38.10	46.04	57.88	53.23	63.41	2.42	34.41	0.00
fine-tuned chatglm2-6b	54.43	37.75	61.45	50.88	67.24	66.33	68.17	34.55	45.79	0.00

Table 10: Performance of fine-tuned LLM on identifying ambiguity.

provements in identifying ambiguity across the two evaluations, highlighting the effectiveness of the fine-tuning process. Due to the imbalanced distribution of ambiguous and non-ambiguous examples in the CHAmbi dataset, we employ oversampling. The specific method is to divide 1648 randomly selected non-ambiguous examples and 824 ambiguous examples into train set, val set and test set at a ratio of 7:2:1 during fine-tuning, and then copy the ambiguous examples in each set.

5 Conclusion

We present CHAmbi dataset, the first Chinese multi-label disambiguation dataset, serving as a benchmark for evaluating ambiguity identification and disambiguation capabilities of LLMs. Leveraging this dataset, we conduct detailed evaluations of several state-of-the-art LLMs that support Chinese, revealing the challenges they face in effectively handling ambiguity. Additionally, CHAmbi can serve as a resource for fine-tuning LLM to create an ambiguity detector. In our experiment, the fine-tuned ChatGLM2, with its improved ambiguity identification capabilities, demostrates its potential for handling ambiguous language in real world applications. We hope this new benchmark will serve as a foundation for research in ambiguity identification and disambiguation in Chinese, and foster the development of large language models in this area.

6 Limitations

In our work, we propose the first Chinese disambiguation dataset based on NLI format, and use this dataset to conduct comprehensive evaluation of the ambiguity handling capability of popular LLMs. However, our dataset and evaluation methods have the following shortcomings:

- The dataset contains a limited number of ambiguous examples. We use some methods to over generate a large number of potentially ambiguous examples, but few examples were determined to be truly ambiguous after annotation and validation. This limitation hinders the effective fine-tuning of LLMs to serve as reliable ambiguity detectors.
- We acknowledge that the dataset annotations may be subjective, potentially introducing errors. This aspect requires further exploration in future investigations.
- We have not extensively explore prompt for downstream evaluation tasks. We believe that constructing a better prompt could better exploit and evaluate the genuine capabilities of LLMs in ambiguity handling.

7 Ethics Statement

For CHAmbi, some examples are from public websites, while others are generated by LLMs based on publicly available datasets. Each example in the dataset has been carefully manually annotated and validated, with offensive content systematically removed. Annotation and validation are conducted by postgraduates, who have a good understanding of ambiguity in natural language, ensuring the high quality and ethical standards of the dataset.

Acknowledgments

This research was supported by National Natural Science Foundation of China (62206179), Guangdong Provincial Natural Science Foundation (2022A1515010129), and University Stability Support Program of Shenzhen (20220811121315001).

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yuen Ren Chao and George Kingsley Zipf. 1950. Human behavior and the principle of least effort: An introduction to human ecology. *Language*, 26:394.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Aldo Gangemi, Mehwish Alam, and Valentina Presutti. 2016. Word frame disambiguation: Evaluating linguistic linked data on frame detection. In *LD4IE@ ISWC*, pages 23–31.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Lawrence S. Moss. 2020. Ocnli: Original chinese natural language inference. *Preprint*, arXiv:2010.05444.

Guoying Huang. 1985. Ambiguous phrases in modern chinese. *Studies in Language and Linguistics*, (69-89).

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Lee Pin Ling and Tengku Sepora Tengku Mahadi. 2016. The differences between english and chinese language sentence structure and their impacts to english-chinese machine translation. *UNIVERSITI SAINS MALAYSIA*, 28.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *Preprint*, arXiv:2304.14399.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *CoRR*, abs/1904.12843.

Anshi Shi. 1988. Say ambiguity. *Journal of Chinese Linguistics*, pages 1–16.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *Preprint*, arXiv:1904.01172.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Shanshan Xu and Katja Markert. 2022. The Chinese causative-passive homonymy disambiguation: an adversarial dataset for NLI and a probing task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4316–4323, Marseille, France. European Language Resources Association.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2016. Ordinal common-sense inference. *CoRR*, abs/1611.00601.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. *CoRR*, abs/2109.04408.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2021. Distributed NLI: learning to predict human opinion distributions for language reasoning. *CoRR*, abs/2104.08676.

A Details of dataset construction

A.1 Dataset format

Why we use NLI format? We aim to create a Chinese disambiguation dataset. Enumerating all possible interpretations of a ambiguous sentence is challenging. Therefore, we define ambiguity using the format of premise-hypothesis pairs in NLI datasets. If there are multiple relationships between a premise and a hypothesis (indicating ambiguity), we provide one interpretation for each relationship. This approach eliminates the need to list all possible interpretations of a ambiguous sentence. We only need to fully analyze the potential multiple relationships between the premise and the hypothesis, and list a different interpretation of ambiguous premise or hypothesis for each relationship.

A.2 Annotation details

Guidelines and examples. We provide annotation guidelines and annotation examples for annotators, as shown in Figure 1. We ask annotators to analyze whether premise and (or) hypothesis are ambiguous and assign labels. Then, annotators need to provide disambiguation rewrites for examples judged to be ambiguous, and try to ensure the lowest degree of rewrites. To standardize ambiguity judgment and enhance annotation quality, annotators also need to provide ambiguity category.

Category	Description
Vocabulary	A word has multiple meanings, has multiple sounds, or has multiple parts of speech.
Reference	The object represented by the pronoun or noun in a sentence is ambiguous; The person performing the action is ambiguous.
Grammar	The organizational ordering of sentence components leads to ambiguity, such as subject, predicate, object.
Semantics	Ambiguous connections in meaning between words, phrases or sentences.
Accent	Stress on different words can lead to ambiguity.
Pause	Pauses after different words create ambiguity.
Incomplete	Incomplete sentences make it challenging to precisely understand meanings. For example, humans often make common-sense and situational assumptions from a subjective perspective, leading to incomplete sentences being interpreted in different ways.

Table 11: Ambiguity categories.

Ambiguity categories. In modern Chinese, ambiguity is linguistically categorized into spoken and written forms, with the latter further divided into compositional and lexical ambiguities (Shi, 1988). Compositional ambiguity includes grammatical and semantic combination. Following this classification and other approaches, we ultimately identify seven types of ambiguity: spoken ambiguities include accent ambiguity and pause ambiguity; written ambiguities include vocabulary, grammar, semantics, reference ambiguity; and incomplete ambiguity. See Table 11 for details. The ambiguity in a sentence could be attributed to several

different categories, and we randomly select one for annotation.

Revised Examples. To enhance the quality of the dataset and increase the number of ambiguous examples, annotators are permitted to modify premises and (or) hypotheses during the annotation process based on their understanding, making them ambiguous sentences or higher quality examples. Table 12 illustrates several revised examples.

Before P: 这个问题没有答案。 (This problem has no answer.) H: 这个难题没有解决办法。 (This difficult problem has no solution.) **After** P: 这个题目没有答案。 (This question has no answer.) H: 这个题目没有解决办法。 (This question has no solution.)
Before P: 他们俩都是爱玩游戏的人。 (Both of them love games.) H: 我想他们俩都是喜欢玩游戏的人。 (I think both of them love games.) **After** P: 他们俩都是爱玩游戏的人。 (Both of them love games.) H: 他们俩都是爱玩电子游戏的人。 (Both of them love computer games.)

Table 12: Some revised examples when annotating.

Disambiguation Methods. When judging example as ambiguous and assigning a set of labels, we need to provide disambiguation rewrites of ambiguous premise and (or) hypothesis for each label. We summarize five methods for disambiguating sentences, as shown in Table 13. Annotators can choose an appropriate disambiguation method based on their understanding of ambiguous sentence.

B Details of evaluation

B.1 Generating Disambiguation Rewrites

This evaluation assesses the LLMs' capability to directly generate reasonable disambiguation rewrites. In the experiment, we use not only edit F1 as the evaluation metric but also manually evaluated the rationality of the disambiguation interpretations. To ensure quality, each example is assigned to the three annotators, the label with the most votes becomes the final label (Is disambiguation interpretations reasonable, True or False?). We consider it false if LLMs fail to recognize multiple relationships. If LLMs recognizes multiple relationships, we analyze whether the corresponding disambiguation rewrites are different and reasonable. Different interpretations mean that they convey different meanings. Reasonable interpretations means that the unambiguous disambiguation interpretation is indeed a possible interpretation of the ambiguous premise or hypothesis.

Method	Example
add word	**Before:** 我说服妈妈和你一起去。 (I convinced mom, and go with you.) **After:** 我说服妈妈，要妈妈和你一起去。 (I convinced mom, and ask her to go with you.)
adjust word order	**Before:** 这是一位知识渊博的王老师的学生。 (This is a very knowledgeable Mr. Wang's student.) **After:** 这是知识渊博的王老师的一位学生。 (This is a student of Mr. Wang who is very knowledgeable.)
set contextual background	**Before:** 王师傅也太黑了。 (Master Wang is too dark.) **After:** 王师傅也太黑了，刘师傅只卖5元一斤。 (Master Wang is too dark, master Liu sells only 5 yuan for a catty.)
change sentence structure	Examples are omitted here because the corresponding English cannot convey ambiguity.
change word	Examples are omitted here because the corresponding English cannot convey ambiguity.

Table 13: Disambiguation methods.

B.2 Evaluation of Fine-tuned LLM

We perform specific fine-tuning on ChatGLM2 using the proposed CHAmbi dataset. We use LoRA (Hu et al., 2021) to fine-tune and transform the dataset into Input-Output format, where input and output are similar to the prompt in Section 4.2, except that the demonstration examples is deleted. We use the learning rate of 5e-4, set the batch size to 4 for training and 1 for validation, and train three epochs. We set the maximum sequence length to 700, the dimension of the LoRA low-rank matrix to 16, and the scaling factor of the LoRA low-rank matrix to 32.

Annotation guidelines

Background:

For a natural language inference dataset, the data samples consist of pairs of premises and hypotheses, along with labels indicating the relationships between the premises and hypotheses. Label is a single label and multiple categories. The relationship between the given premise and hypothesis is typically categorized into the following three classes:

- Contradiction: Indicates a direct contradictory relationship between the premise and the hypothesis, implying opposition, refutation, mutual exclusion, or mutual negation.
- Entailment: Indicates that the premise imply the truth of the hypothesis.
- Neutral: Indicates that there is no obvious entailment or contradiction between the premise and the hypothesis, and they can exist independently.

Your task:

Due to the ambiguity in language, it is sometimes impossible to exactly give a single label for the premise-hypothesis pair. Now, we aim to construct a Chinese multi-label disambiguation natural language inference dataset. For each example:

1. Analyze whether there is ambiguity between the premise and hypothesis, and assign label. If there is ambiguity, there will be multi-relationship between the premise and hypothesis.
2. For the example judged to be ambiguous, perform disambiguation rewriting. When rewriting, please ensure the lowest degree of rewriting.
3. Give the ambiguity category.

Analysis steps:

First of all, analyze whether the premise is ambiguous, analyze whether the hypothesis is ambiguous. If you think there is no ambiguity, further combine the premise and hypothesis and analyze whether the premise implies the hypothesis, this implication is uncertain. Or whether the premise implies a negation of the hypothesis, but this negation is also uncertain. This likely indicates ambiguity in the expression of the premise and/or hypothesis, making it challenging to accurately comprehend their meanings. If the premise does not imply a hypothesis, and deny the hypothesis, then there is only a neutral relationship.

Ambiguity categories:

See Table 3.

Please analyze the examples in this document before your annotation.

Examples

P: 运动员的目标是获得金牌。
(An athlete's goal is to win gold medals.)
H: 运动员的目标是打败个人最好成绩。
(An athlete's goal is to beat a personal best.)
Analysis:
First, analyze the premise and hypothesis independently, we can not see ambiguity. Then analyze the relationship between the premise and hypothesis, it seems that they are opposite or neutral. Is an athlete's goal mentioned in the premise and hypothesis the only goal or a goal? Therefore, it is judged that the premise and hypothesis are incomplete, leading to ambiguity.
Labels: neutral, contradiction
Disambiguation1:
P: 运动员的主要目标是获得金牌。
(An athlete's main goal is to win gold medals.)
H: 运动员的主要目标是打败个人最好成绩。
(An athlete's main goal is to beat a personal best.)
Label: contradiction
Disambiguation2:
P: 运动员的一个目标是获得金牌。
(One of the goals of an athlete is to win gold medals.)
H: 运动员的一个目标是打败个人最好成绩。
(One of the goals of an athlete is to beat a personal best.)
Label: neutral
Category: incomplete

P: 准时赴约非常重要。
(It is very important to arrive on time for your appointments.)
H: 准时参加面试非常重要。
(It is very important to arrive on time for your interviews.)
Analysis:
First, analyze the premise and hypothesis independently, we can not see ambiguity. Then analyze the relationship between the premise and hypothesis, it can be determined to be neutral, because these are two objective and independent truths.
Labels: neutral

P: 我的妻子不会和我的女儿一起来参加活动。
(My wife won't come to the event with my daughter.)
H: 我的妻子不会来。
(My wife won't come.)

Examples

Analysis:

First, analyze the premise and hypothesis independently, we can find the premise is ambiguous. If we can not find it, further analyze the relationship between the premise and hypothesis. We can not be sure whether the relationship is entailment or contradiction. Does the wife not come with the daughter? Will the wife come or not? Will the daughter come or not? The premise is ambiguous at the grammatical level.

Labels: neutral, entailment, contradiction

Disambiguation1:

P: 我妻子和女儿不会都来。
(My wife and daughter won't both come.)

H: 我的妻子不会来。
(My wife won't come.)

Label: neutral

Disambiguation2:

P: 我女儿要来，我妻子不会来。
(My daughter will come and wife won't come.)

H: 我的妻子不会来。
(My wife won't come.)

Label: entailment

Disambiguation3:

P: 我的妻子和女儿都来参加活动，但他们不会一起来。
(My wife and daughter will both come, but they won't come together.)

H: 我的妻子不会来。
(My wife won't come.)

Label: contradiction

Category: gramma

P: 新医院靠近州际公路，患者前往那里很方便。
(The new hospital is close to an state highway, making it easy for patients to get there.)

H: 去新医院很方便。
(It is very convenient to go to the new hospital.)

Analysis:

First, analyze the premise and hypothesis independently, we can find the premise is ambiguous. Is 'there' referring to the 'new hospital' or the 'state highway'? Therefore, it is judged that the unclear reference of the premise leads to ambiguity.

Labels: neutral, entailment

Examples

Disambiguation1:

P: 新医院靠近州际公路，方便患者前往医院。

(The new hospital is located near an state highway, making it easy for patients to get the new hospital.)

H: 去新医院很方便。

(It is very convenient to go to the new hospital.)

Label: entailment

Disambiguation2:

P: 新医院靠近州际公路，方便患者前往州际公路。

(The new hospital is located near an state highway, making it easy for patients to get the state highway.)

H: 去新医院很方便。

(It is very convenient to go to the new hospital.)

Label: neutral

Category: reference

P: 这家餐厅是这个城市中最好的。

(This restaurant is the best in the city.)

H: 其他餐厅没有这家餐厅好。

(Other restaurants are not as good as this one.)

Analysis:

First, analyze the premise and hypothesis independently, we can not see ambiguity. Then analyze the relationship between the premise and hypothesis. The hypothesis is not complete, leading to multiple relationships.

Labels: neutral, entailment

Disambiguation1:

P: 这家餐厅是这个城市中最好的。

(This restaurant is the best in the city.)

H: 这个城市的其他餐厅没有这家餐厅好。

(Other restaurants in the city are not as good as this one.)

Label: entailment

Disambiguation2:

P: 这家餐厅是这个城市中最好的。

(This restaurant is the best in the city.)

H: 世界上其他餐厅没有这家餐厅好。

(No other restaurant in the world is as good as this one.)

Label: neutral

Category: incomplete

Figure 1: Guidelines and examples for annotation.