Automatic Identification of Chinese Modality Based on Pre-trained Language Models

Anonymous ACL submission

Abstract

Recognizing the modality of utterances is crucial to NLP tasks that require a deep understanding of semantics and pragmatics, including semantic inference, dialogue systems, and so on. This paper focuses on the automatic identification of Chinese language modality with different machine learning models, including classic models, pretrained transformers, and Large Language Models (LLMs). We conduct experiments on a Chinese dataset that we annotate with four types of modalities. The results show that the fine-tuned BERT model achieves the best performance, with an F1 score of 0.74, significantly outperforming the LLMs and other models. The study reveals the difficulty of the task, and while LLMs have demonstrated exceptional performance across a wide range of NLP tasks, their ability to handle tasks that heavily rely on semantic and pragmatic understanding remains limited, underscoring the need for more efforts in improving the NLP models, including LLMs, on this task.

1 Introduction

Modality expresses speakers' attitudes towards the truth values of the propositions referenced to real-world situations in terms of necessity, possibility, and obligation. By default, the speaker uses non-modal expressions to pass a message that reflects a factual situation in the real world. For example, the sentence "Kate is in her office" describes a fact of the world. On the contrary, the sentence "Kate may be in her office" describes a possibility inferred by the speaker about the state of the world. The latter is in an epistemic modality. In a typical theoretical framework proposed by Palmer Palmer (1986, 2001), there are four main categories of modalities, namely epistemic, evidential, deontic, and dynamic, as shown in Table 1.

Modalities can be expressed in various ways, including auxiliaries, verbs, adverbs, and even zero forms. The non-zero linguistic cues are important for automatically identifying the correct modality of sentences. However, it is not an easy task due to many of them are ambiguous and disambiguation of them requires deep semantic and pragmatic knowledge and also relies on contextual information. For example, 'must' and 'should' can express both deontic and epistemic modalities. In Chinese, modality is also related to sentence-final particles. Even minor differences in the surface form can result in different modalities. An example is shown in (1), where the additional sentence final particle *ba* indicates the uncertainty of the speaker's belief about the corresponding proposition. It interacts with the auxiliary verb *yinggai* (应该 'should') and makes it be interpreted as the epistemic modality. Without the particle, the speaker expresses an obligation that, as college students, they ought to possess the corresponding knowledge, thus, it is in a deontic modality. We can see that the successful identification of language modality is crucial to downstream applications such as dialogue systems to understand speakers' intent in order to make proper reasoning and generate appropriate responses.

(1) 你们应该知道这个(吧)

nimen yinggai zhidao zhe ge (ba)
you should know this (BA)

With 'Ba': 'You probably know this, right?' (epistemic)

Without 'Ba': 'You should know this.' (deontic)

The study of modality has been continued for a long time in different areas and can be traced back to Aristotle in logic studies. However, the study of modality identification in NLP is still underexplored, partially due to the scarcity of available data resources annotated with modalities. Among the few, (Xu and Huang, 2014) proposes a framework for the simultaneous identification and classification of modality, speech acts, and event types of

Modality	Definition	Examples
Epistemic	The speaker expresses their judgments about the factual status of the proposition.	Kate must be at home.
Evidential	The speaker explicitly indicates the evidence for the factual status of a proposition.	He is said to be extremely rich.
Deontic	It expresses obligation or permission towards certain actions/events, emanating from an external source, such as rules, laws, or the desire of the speaker.	John must come in now.
Dynamic	It expresses the subject's ability or willingness to perform certain actions.	John can speak French.

Table 1: The categorization framework of modality by Palmer (2001).

Chinese sentences with traditional statistical models. Their model shows an overall result of about 50%

To boost the study of modality identification, we construct an annotated Chinese dataset based on an existing one (Xu and Huang, 2014). Following Palmer Palmer (1986, 2001)'s framework, we differentiate three types of modality: epistemic, deontic, dynamic, and a non-modality type, statement [1]. Then, we train different models, including traditional feature-based ones (SVM and Multinomial Logistic Regression) and deep neural nets (LSTM and finetuned BERT), for automatically identifying the modality types. We also test four LLMs, including GPT-4o, Llama-3-70b, Baidu Ernie-4.0-turbo-8k, and Deepseek-v3 on the same task. The experimental results reveal that the fine-tuned Chinese BERT model achieves the best performance, with a macro F1 score of 0.74, significantly outperforming other models, including LLMs, marking an improvement of nearly 0.20. This paper contributes in two ways. Firstly, it introduces a valuable dataset of modality-annotated Chinese sentences, which can be used for future studies on modality detection and classification. Secondly, it provides new important insights into the understanding of LLMs in performing such tasks that involve rich semantics and pragmatics.

2 Related Work

2.1 Linguistic Studies on Modality

Linguists have put forward various modality classification frameworks from different perspectives. Philosophical logic laid the foundation for modality research with Aristotle's exploration of necessity and possibility, introducing the concept of alethic modality.

Rescher (1968) expanded modality into eight categories, including alethic, epistemic, temporal, boulomaic, deontic, evaluative, causal, and likelihood. In Chinese linguistics studies, foundational work by Lü (1951) and Tsang (1981) categorized Chinese modal verbs into epistemic and deontic domains, while later studies, such as Huang et al. (2000) and Cui (2003), introduced evaluative and dynamic modalities, emphasizing their subjectivity and connection to factuality.

There are also studies on modality itself from its syntax, semantics, functionality, and cognition aspects. Structuralist grammarians, such as Bloomfield (1933), analyzed modal auxiliaries based on their syntactic roles, while generative grammar studies, like those by Fillmore (1968) and Radford (1988), treated modality as a core syntactic feature alongside tense and aspect. Functional linguists, such as Halliday (1994), examined modality's interpersonal meta-functions, distinguishing modalization (probability) from modulation (obligation). Cognitive approaches, represented by Sweetser (1990), emphasized the conceptual grounding of modality in event schemas.

Among these traditions, Palmer (1986, 2001) provided a comprehensive and unified framework of modality, emphasizing its semantic nature and its linguistic manifestations. Unlike structural or purely syntactic approaches, Palmer's work bridges linguistic theory and practical applications, making it particularly suitable for computational adaptation. Thus, in our study, we follow Palmer's modality classification framework to systematically annotate and classify Chinese modality in computational models.

[1] The evidential modality is treated as a special epistemic modality in our framework, considering that they are similar and sometimes hard to differentiate.

2.2 Computational Studies on Modality

In NLP, modality recognition is related to various key NLP applications, including event detection, factuality prediction, sentiment analysis, and so on. Early work, such as Siegel (1999) and Palmer et al. (2007), investigated modality in event classification and annotation. Siegel (1999) categorized modality into state, activity, accomplishment, achievement, and semelfactive to detect semantic differences in events. He further uses machine learning methods to achieve an automatic identification of the above category based on lexical and syntactic features, and achieves good results. Palmer et al. (2007) categorizes modality into propositional modality(epistemic, evidential) and event modality (deontic and dynamic). The former category relates to the truth value of the proposition, and the latter refers to the event features concerning permission, obligation, and ability. This provides the foundational categorization of modality types. Notably, Saurí and Pustejovsky (2012) integrated modality into factuality prediction frameworks, and proposed an annotation framework with a factuality annotated corpus, FactBank. This enabled computational assessment of event factuality in NLP applications. Xu and Huang (2014) trained an SVM classifier to differentiate 4 types of modalities, 6 types of speech acts, and 13 event types of Chinese sentences, incorporating syntactic, semantic, and temporal features. The overall performance for modality is around 0.34 macro-F1 across all the fine categories. Most recently, Sun et al. (2023) proposed a BERT-based model tailored to deontic modality classification, achieving state-of-the-art results through the fine-tuning technique with a customized dataset.

Existing studies on modality identification contribute by providing valuable linguistic resources and tools. However, they remain limited and lack a comprehensive investigation of Chinese modality using pretrained models and large language models (LLMs). In this study, we aim to fill this gap by providing a Chinese dataset annotated with sentence-level modality types and building models for automatic classification of modality with different machine learning models, including LLMs.

3 Data Construction

The dataset used in this study is constructed based on an existing annotated dataset from Xu and Huang (2014), which consists of 5,612 sentences extracted from the Sinica Treebank 3.0 (Hu et al., 2005). This dataset provides annotations for 4 types of modality (epistemic, deontic, dynamic, and evaluative), 6 types of illocutionary speech acts, and 13 types of events.

We make revisions to this categorization framework for several reasons. Firstly, evaluative is not traditionally recognized as a modality in frameworks such as Palmer (2001)'s and is challenging to annotate due to the fact that Chinese (English as well) has no morphological clues to indicate such information. For example, "he is smart" can be the speaker's personal evaluation or an expression of a well-accepted truth. This differs from other cases like "he is in the office", which has only the interpretation of the expression of a fact. Thus, we remove evaluative modality and simply treat all such examples as statements.

Besides, speech act categories are removed as they are predominantly relevant to spoken language and are rarely observed in written text. Finally, the event types are differentiated among factual events based on the event structures, which are the combination of viewpoint aspect and situation aspect. Since the goal of this study is not to recognize the subtle difference between the event types, we simply consolidate all event types into the "statement" category, representing non-modality/factual events. After this refinement, the dataset contains 4,594 sentences.

Another problem of the existing dataset is genre- and category-biased. The 4,594 sentences mostly come from news, and the category of statement is extremely large compared to others. To address this issue, we supplement the small categories in the dataset with 1,820 sentences extracted from the novel "The Three-Body Problem" (Liu, 2008). Sentences were selected using keyword-based extraction and annotated by a well-trained linguist following the same framework. The data is finally split into training and test sets evenly. The information of the dataset is shown in Table 2.

Modality	Train	Train (Aug)	Test
Epistemic	277	2,452	279
Deontic	237	2,452	236
Statement	2,452	2,452	2,454
Dynamic	240	2,452	239
Total	3,206	9,808	3,208

Table 2: Dataset for Chinese modality classification.

4 Models and Evaluation

We evaluate four machine learning models trained by ourselves and three large language models (LLMs) to assess their performance on the task of Chinese modality classification. The machine learning models include Support Vector Machine (SVM), Multinomial Logistic Regression/Maximum Entropy Model (MaxEnt), Bi-LSTM, and a fine-tuned model based on BERT-base-Chinese (BERT). The LLMs tested are GPT-4o (GPT), Llama-3-70b (Llama), Baidu Ernie-4.0-turbo-8k (Ernie), and Deepseek-v3 (Deepseek).

4.1 Training Data Augmentation

As shown in Table 2, the categories are still imbalanced even after the data supplement, with the "statement" category significantly outnumbering other modalities. This poses a risk of bias in model training, favoring the dominant class. To address this issue, two data augmentation techniques were applied to the training set: oversampling and the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). Oversampling involved duplicating instances of minority categories to increase their representation relative to the majority class. SMOTE was employed to synthesize new instances for minority classes by interpolating between existing samples, thus preserving semantic consistency while diversifying the dataset.

For SMOTE, we preserve the original sentence structure and use pronoun substitution to generate new samples. In details, sentences containing common pronouns including *wo*, (我 'I'), *ni*, (你 'you'), or *ta* (他 'he') in various positions, will be substituted by different pronouns randomly to create new instances. For sentences lacking such pronouns, direct duplication is performed to ensure adequate representation of the minority categories in the training set. The augmented training set finally contains 9,808 sentences [2].

4.2 Model Training

For the feature-based machine learning models (SVM, MaxEnt), we use the feature set proposed by Xu and Huang (2014), including part-of-speech (POS) tags and dependency relations. Those features are extracted with HanLP (He and Choi, 2021). The SVM classifier is trained with a linear kernel and a regularization parameter C of 1.0. The MaxEnt classifier is trained with a regularization parameter C of 10. For the Bi-LSTM model, static word embeddings are initialized randomly and updated during training. The Bi-LSTM layer is finally connected to an output feedforward layer. Optimal hyperparameters are identified using grid search: embedding size of 300, hidden size of 512, number of layers of 1, learning rate of 0.0001, and maximum iterations of 10. For BERT, we use the BERT-base-Chinese model plus a feedforward layer. The feedforward layer has one hidden layer with a size of 512. The training is performed using the cross-entropy loss with a learning rate of 0.0001 and a maximum iterations of 10.

For evaluating LLMs on the same task, we use the following prompt[3].

4.3 Results

We evaluate the performance of models on the Chinese modality classification task in terms of accuracy, precision, recall, and F1-score as shown in Table 3.

As we can see, the feature-based machine learning models (SVM, MaxEnt) exhibit varying degrees of difficulty in handling modalities. Their average F1 score for modality identification is below 0.6, with epistemic and dynamic particularly low, although it is better than the result (0.45) in previous work using the same feature set (Xu and Huang, 2014). This is not surprising due to the fact that the task requires a deep understanding of semantics and the pragmatics of sentences. Bi-LSTM shares similar limitations with an F1 score of 0.55, still failing to accurately identify Chinese modalities.

For LLMs, the macro-F1 scores are similar to those of feature-based models, while the accuracy is relatively lower. For the BERT model, when the data is not augmented, it predicts all the cases to the statement category, failing to achieve a successful classification. As we can see, data augmentation significantly improved BERT's and Bi-LSTM's ability to classify minority categories, with macro-average F1 increasing from 0.22 to 0.74. The gains are especially evident in the epistemic and dynamic categories. On the whole, the bias problem is mitigated but not resolved absolutely by augmentation. We infer that it is due to the simplicity of the augmentation method cannot generate samples to represent useful features, and thus the

[2]The data will be available upon the acceptance of the paper.

[3]See Appendix A.1 for the full English translation of the original Chinese prompt.

Model	Deontic	Epistemic	Statement	Dynamic	MacroAVG	Acc
SVM(unaug)	0.44	0.42	0.90	0.46	0.55	0.82
SVM	0.45	0.39	0.90	0.47	0.55	0.81
MaxEnt(unaug)	0.40	0.36	0.90	0.44	0.53	0.82
MaxEnt	0.47	0.45	0.90	0.46	0.57	0.81
Bi-LSTM(unaug)	0.28	0.30	0.89	0.41	0.47	0.79
Bi-LSTM	0.42	0.44	0.90	0.42	0.55	0.89
BERT(unaug)	0.00	0.00	0.87	0.00	0.22	0.77
BERT	**0.65**	**0.72**	**0.94**	**0.67**	**0.74**	**0.92**
GPT	0.54	0.40	0.86	0.34	0.54	0.75
Ernie	0.44	0.49	0.83	0.25	0.50	0.70
Llama	0.46	0.42	0.80	0.23	0.48	0.67
Deepseek	0.48	0.61	0.92	0.28	0.57	0.84

Table 3: Performance of different models in F1 scores across modality types. The results of models with and without training data augmentation are provided.

sparseness problems still exist, or the extracted features cannot pinpoint the subtle difference between the modal categories.

Overall, the fine-tuned BERT after data augmentation outperforms all other models, including LLMs, across all modality categories, achieving 0.74 in the Macro-F1 measure. This shows that Chinese modality classification based on Palmer's framework is differentiable using a fine-tuned pretrained BERT model. Meanwhile, LLMs currently still have room for improvement on the task.

5 Error Analysis

To delve deeper into the classification results of various models, we examine their confusion matrix heatmaps as shown in Table 1.

5.1 Misclassifications of modalities to statements

We can see that, even with augmentation techniques, SVM, MaxEnt, and Bi-LSTM show a strong bias towards the major category, statement.[4] However, confusion among different modalities is relatively rare. For finetuned BERT, it shows a true advantage over other models, reducing the effect of bias. LLMs show a quite different pattern. They tend to misclassify statement sentences into modality categories. GPT mistakes many statement instances for epistemic, Ernie to deontic, and Llama to deontic and epistemic. Deepseek mistakes statements to deontic. On the other hand, LLMs show a better recall rate for deontic and epistemic modality, while for dynamic modality, LLMs have trouble recognizing them effectively.

Specifically, traditional models, including SVM and MaxEnt, show a strong bias towards the major class, *statement*. This explains their higher accuracy but similar F1 scores compared to LLMs. Modalities such as epistemic and dynamic are particularly challenging, given their reliance on subtle and context-dependent markers.

(2) 恐怕是世界上其他国家所不能比的。

Kongpa shi shijie shang qita guojia suo
I'm.afraid is world on other country REL
buneng bi de
NEG.can compare DE

"I'm afraid this is something other countries in the world cannot compare to." (epistemic)

Predicted: Statement **Actual:** Epistemic

This sentence contains the hedging expression Kongpa (恐怕, "I'm afraid"), which conveys subjective uncertainty—a hallmark of epistemic modality. However, the model fails to detect this modality marker and classifies the sentence as a neutral statement.

(3) 他有责任保护人民。

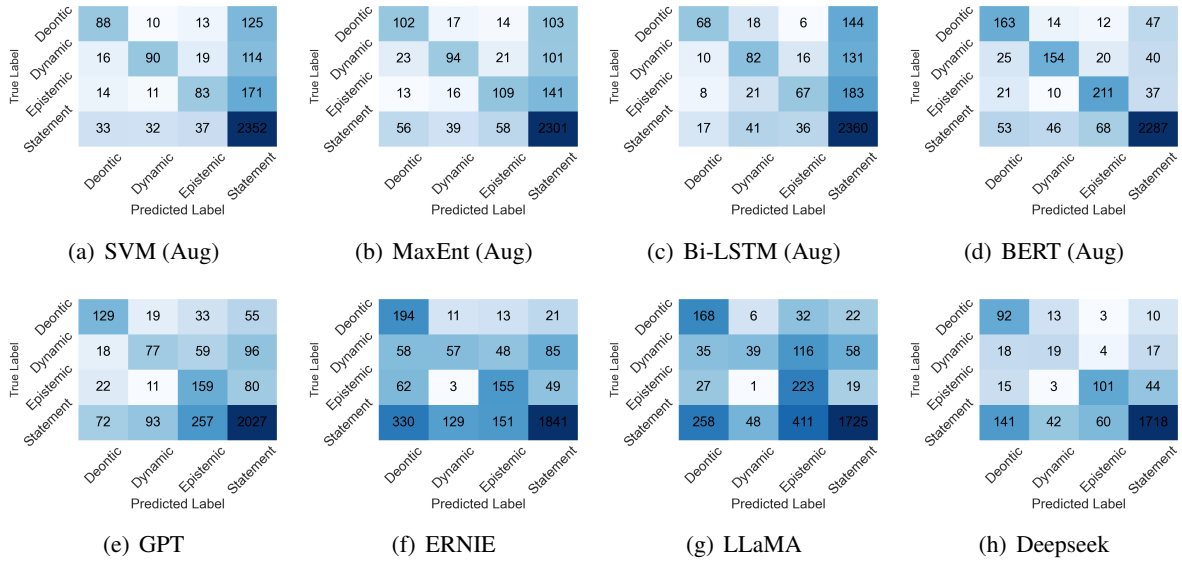
Ta you zeren baohu renmin
he has responsibility protect people

"He has a responsibility to protect the people."

Predicted: Deontic **Actual:** Statement

Although the expression *you zeren* (有责任 'has

[4]For full confusion matrices including unaugmented model outputs, see Appendix A.2.



(a) SVM (Aug) (b) MaxEnt (Aug) (c) Bi-LSTM (Aug) (d) BERT (Aug)

(e) GPT (f) ERNIE (g) LLaMA (h) Deepseek

Figure 1: Confusion matrix of the classification results of all models.

a responsibility') is commonly associated with deontic modality, here it serves to describe the subject's role or function in a factual manner. The model's misclassification reflects a challenge in distinguishing between fact-stating and obligation-imposing usage in similar syntactic structures.

The fine-tuned BERT model demonstrates substantial improvement over traditional models, Bi-LSTM, and LLMs, particularly in identifying epistemic and dynamic modalities. Using contextual embeddings and fine-tuning on a modality-annotated training set, BERT achieves a promising result with balanced precision and recall across categories. These results underscore the advantage of contextual embeddings when addressing the nuanced distinctions inherent in Chinese modality.

5.2 Misclassifications of Dynamics in LLMs

LLMs have a lower error rate for misclassifying modality categories into statements. However, they struggle with certain modalities, particularly dynamic modality. Misclassification often arises from shared markers of statements with modality types, reflecting the inherent complexity of identifying subjective propositions.

The confusion matrices reveal persistent challenges, particularly for dynamic modality, which is often misclassified as statements or epistemic modality across all models. This indicates that the subtle linguistic cues characterizing dynamic modality, such as contextual dependencies and subjective intent, remain difficult for current LLMs to capture. This is especially the case when it comes to complex sentences. Below are illustrative examples:

(4) 我们可以想象出当时贵族的奢华。
Women keyi xiangxiang chu dangshi
we can imagine out then
guizu de shehua
aristocracy GEN luxury
"We can imagine the luxury of the aristocracy at that time." (dynamic)
Predicted: Statement **Actual:** Dynamic

In this example, the modal verb *keyi* (可以 'can') indicates the subject's cognitive ability, reflecting dynamic modality. The model misclassifies it into statements, failing to recognize it.

(5) 我们仍不会接受。
Women reng bu hui jieshou
we still not will accept
"We still will not accept it." (deontic)
Predicted: Dynamic **Actual:** Deontic

The modal marker *hui* (会 'will') often conveys epistemic or dynamic meaning depending on context. Here, *hui* expresses deontic refusal of permission or possibility, reinforced by the negation *bu* (不 'not'). The pragmatic context reflects an external constraint rather than internal intention, aligning with deontic modality. The model likely misclassified this due to the future orientation and subjectivity implied by *jieshou* (接受 'accept').

5.3 Distribution and Implication of Errors

Among the various modality categories, two notable findings emerge. Firstly, fine-tuned BERT achieves a recall of 0.78 in detecting epistemic modality, demonstrating its capacity to capture linguistic expressions associated with possibility and uncertainty. This improvement reflects the model's ability to contextualize polysemous markers such as *hui* (会 'might') and *yinggai* (应该 'should'), which often carry epistemic meaning depending on their syntactic and semantic surroundings. For dynamic modality, BERT achieves an F1-score of 0.67, with a recall of 0.62. Despite this improvement, challenges remain in handling sentences with ambiguous markers or limited contextual cues. Dynamic modality, often expressed through markers like *yao* (要 'want to') or *yuanyi* (愿意 'willing to'), requires the model to distinguish subjective intent from factual or epistemic statements.

Another finding is that many misclassifications arise from polysemous modal markers or sentences where modality is ambiguous without additional context. Below is a representative example:

(6) 是否会造成国际股市慢速盘跌的走势值得观察。

Shifou hui zaocheng guoji
whether will cause international
gushi mansu pandie de zoushi
stock.market slow decline GEN trend
zhide guancha
worth observing

"Whether it will cause a slow decline in international stock markets is worth observing."

Predicted: Epistemic **Actual:** Statement

The modal marker *hui* (会 'will') frequently signals epistemic modality, indicating likelihood or possibility. However, in this sentence, hui appears within a subordinate clause functioning as the nominalized object of the main clause *zhide guancha* (值得观察 'is worth observing'). Thus, the sentence primarily makes a factual statement without explicit modality. The misclassification suggests the model overlooks clause embedding and syntactic structure when detecting modality. These examples further demonstrate the intricate relationship between syntax, semantics, and pragmatics in Chinese modality recognition.

Overall, the fine-tuned BERT model sets a strong foundation for automatic modality classification in Chinese, demonstrating its potential to outperform traditional and LLM-based approaches. However, the nuanced nature of modality, particularly in polysemous markers and context-dependent expressions, continues to challenge even the most advanced models.

6 Discussion: Challenges, Misclassifications, and Linguistic Considerations

The experimental results reveal limitations in traditional machine learning models, Bi-LSTM, and LLMs for capturing the complexity of Chinese modality. While the manually designed features relied on by traditional models and Bi-LSTM are effective in controlled scenarios, they struggle with nuanced contextual dependencies and ambiguous modality markers. LLMs also face challenges when domain-specific training data is limited. This stresses the need for modality-specific architectures and pre-training strategies that better explain linguistic subtleties and contextual ambiguity.

Fine-tuned BERT demonstrates substantial improvements in modality classification. It shows improvements not only in recognizing modality markers but in handling complex or ambiguous modality markers. These findings highlight the potential of fine-tuned pre-trained models for the Chinese modality classification.

6.1 Linguistic and Computational Challenges

6.1.1 Linguistic Challenges

The first challenge of classifying Chinese modality is polysemy. Chinese modal markers like *yao* (要 'want/must') and *gan* (敢 'dare') have various meanings depending on syntactic and semantic context. For example, *yao* may signify volition (要完成任务 'want to complete the task') or obligation (要遵守规定 'must follow the rules'). Disambiguating such markers remains challenging, particularly in sentences lacking explicit contextual cues.

The second challenge comes from the identification of the modality holders. Chinese modality is not explicitly tied to the subject and can shift dynamically within context. For example, in "有人说你应该负责" ("Someone says you should take responsibility"), the epistemic and deontic modalities are associated with different entities. Misidentifying these shifts can lead to classification errors.

Another challenge is from complex sentences. Nested and multi-layered sentence structures com-

plicate the interpretation of modal markers. For instance, in "如果他能按时完成任务，我们可能会考虑提拔他" ("If he can complete the task on time, we might consider promoting him"), the interaction between *keneng* (可能 'might') and the conditional clause adds layers of complexity. Models without explicit syntactic parsing struggle to capture these interactions accurately.

6.2 Computational Challenges

The computational challenges of the modality identification model come from two aspects. One is the imbalance in datasets. Current datasets are often dominated by statements, leading to classification bias. Less frequent modalities, like dynamic markers, are often misclassified due to subtle linguistic cues and data imbalance.

Another challenge found in this study is about LLMs. LLMs like GPT excel in processing factual statements but struggle with subjective or context-sensitive modalities, such as epistemic and dynamic expressions. Without fine-tuning on modality-specific data, these models often fail to capture pragmatic nuances or implicit meanings.

6.3 Bridging Linguistic Theory and Computational Practice

Addressing challenges in Chinese modality classification requires bridging linguistic theory and computational methodologies. Insights from linguistic research, such as the interaction between modality and clause structures, can guide computational model design.

For example, incorporating theoretical insights into the interaction of modality with tense-aspect systems can enhance the interpretability and accuracy of computational models. Future work could also expand datasets with the less frequent modalities and complex structures to handle hierarchical modality.

Fine-tuned BERT models have demonstrated this potential by integrating modality-specific annotations and leveraging syntactic-semantic features.

7 Conclusion

This study shows the effectiveness of fine-tuned BERT models for Chinese modality classification. By leveraging contextual representations and enriched linguistic features, these models significantly improve classification accuracy, outperforming traditional models and even advanced LLMs like GPT in specific aspects.

Despite these advancements, challenges such as polysemy, implicit modality expressions, and nested structures persist. Future research should prioritize integrating discourse-level features, diversifying datasets, and refining models with linguistic insights.

In conclusion, this study bridges linguistic theory and computational practice, providing a robust foundation for automatic modality identification in Chinese. These findings lay the groundwork for developing more interpretable and linguistically informed AI systems, advancing both theoretical understanding and practical applications.

Limitations

This study is limited by several factors. Firstly, our dataset is imbalanced across modality types, with the statement category dominating the dataset. Though we conducted data augmentation, the imbalance still affects the model performance of traditional statistical models and Bi-LSTM. Second, though fine-tuned BERT shows improvement, and we added syntactic and semantic features in Bi-LSTM and traditional statistical models, they still struggle with polysemy, implicit modality, and context-dependent holder shifts. Third, our classification is sentence-level, lacking discourse-level modeling that could better capture hierarchical modality and dependencies. Lastly, comparisons with large language models like GPT were done without modality-specific fine-tuning, which may underestimate their potential.

References

Leonard Bloomfield. 1933. *Language*. Henry Holt, New York.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Xiliang Cui. 2003. Event modality and the assertion system in chinese. In *Grammar Research and Exploration (Twelve)*, pages 331–347. Commercial Press, Beijing.

C.J. Fillmore. 1968. The case for case. In E. Bach and R. Harms, editors, *Universals in linguistic theory*, pages 1–88. Holt, Rinehart & Winston, New York.

M.A.K. Halliday. 1994. *An Introduction to Functional Grammar*. Edward Arnold, London.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Hu, Chien-Jung Chen, Tzu-Hsiu Lee, Wei-Tsong Lee, Szu-Yu Hsieh, and Yung-Fu Li. 2005. Sinica treebank 3.0. Accessed: 2024-11-13.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhaoming Gao, and Kuang-Yu Chen. 2000. Sinica treebank design criteria, annotation guidelines, and online interface. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, pages 29–37.

Cixin Liu. 2008. *The Three-Body Problem*. Chongqing Press, Chongqing, China.

Shuxiang Lü. 1951. Various modalities. *Progressive Youth*, 240:54–64.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 896–903, Prague, Czech Republic. Association for Computational Linguistics.

F.R. Palmer. 1986. *Mood and Modality (1st edition)*. Cambridge University Press, Cambridge.

Frank Robert Palmer. 2001. *Mood and Modality*, second edition edition. Cambridge University Press, Cambridge.

A. Radford. 1988. *Transformational Grammar*. Cambridge University Press, Cambridge.

N. Rescher. 1968. *Topics in Philosophical Logic*. Reidel, Dotrecht.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Eric V. Siegel. 1999. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 112–119, College Park, Maryland, USA. Association for Computational Linguistics.

Jingyun Sun, Shaobin Huang, and Chi Wei. 2023. A bert-based deontic logic learner. *Journal of Engineering Applications of Artificial Intelligence*.

Eve Sweetser. 1990. *From etymology to pragmatics*. Cambridge University Press, Cambridge.

Chui Lim Tsang. 1981. *A Semantic Study of Modal Auxiliary Verbs in Chinese*. Ph.D. thesis, UMI.

Hongzhi Xu and Chu-Ren Huang. 2014. Annotate and identify modalities, speech acts and finer-grained event types in chinese text. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, pages 157–166. Association for Computational Linguistics.

A Appendix

A.1 Prompt for Modality Classification

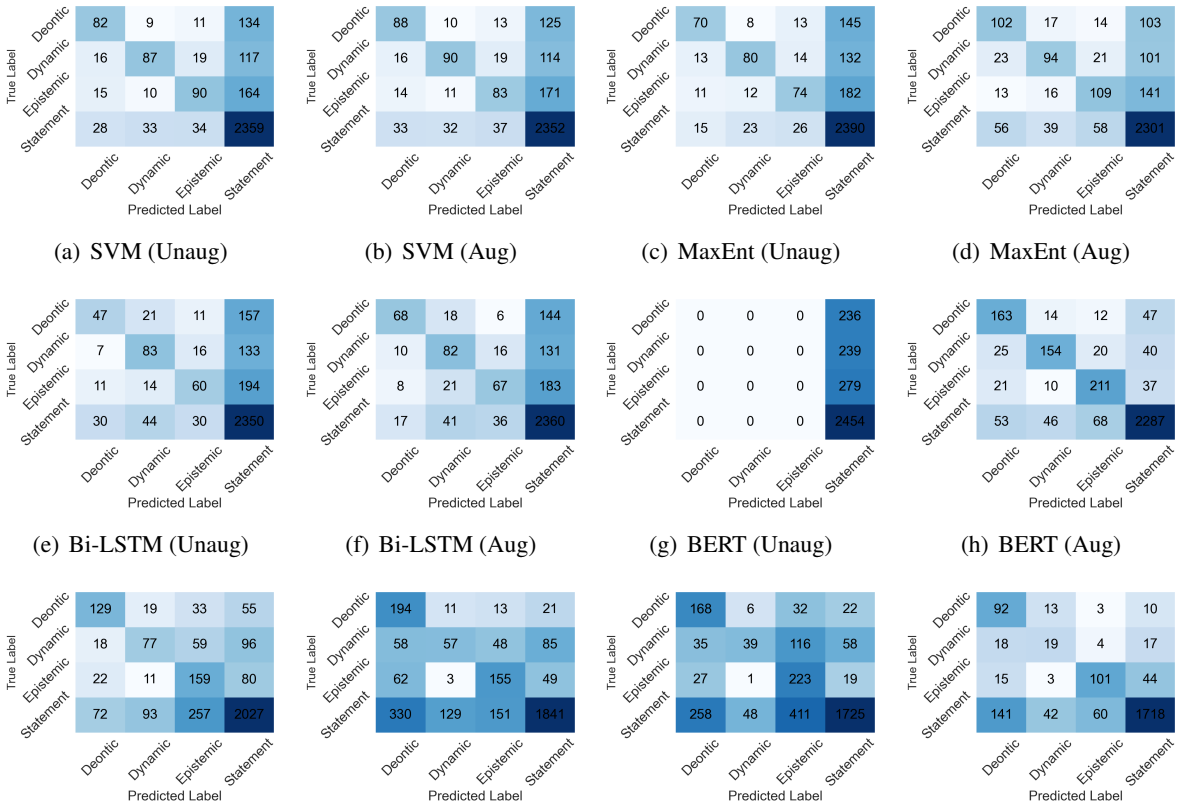
You are a linguist specializing in the Chinese language. Your task is to identify the modality of the given sentence. Modality refers to the speaker's attitude towards the proposition in the sentence. The classification and definitions of modalities (based on Palmer (2001)'s definition) are as follows:

1. *Statement: Simple declarative sentences without modal meaning, included for comparative purposes.*
2. *Epistemic: Modality expressing the speaker's judgment about the truth or likelihood of a proposition.*
3. *Deontic: Modality related to the speaker's directives, obligations, or permissions.*
4. *Dynamic: Unlike other modalities, dynamic modality expresses the subject's ability to change the development of the real world (e.g., taking action, changing states, etc.).*

Please classify the modality of the following sentence according to the definitions above and provide the answer in the following format: `modality:<1|2|3|4>`*, where "1|2|3|4" corresponds to the modality category number.*

A.2 Full Confusion Matrices

A.3 Detailed Classification Results for All Models



(a) SVM (Unaug) (b) SVM (Aug) (c) MaxEnt (Unaug) (d) MaxEnt (Aug)

(e) Bi-LSTM (Unaug) (f) Bi-LSTM (Aug) (g) BERT (Unaug) (h) BERT (Aug)

(i) GPT (j) ERNIE (k) LLaMA (l) Deepseek

Figure 2: Full confusion matrix for all models.

Model	Modality Type	Precision	Recall	F1-score	Support
SVM	Deontic	0.583	0.373	0.455	236
	Dynamic	0.629	0.377	0.471	239
	Epistemic	0.546	0.297	0.385	279
	Statement	0.852	0.958	0.902	2,454
	Macro Avg	0.652	0.501	0.553	3,208
	Weighted Avg	0.789	0.815	0.792	3,208
MaxEnt	Deontic	0.526	0.432	0.474	236
	Dynamic	0.566	0.393	0.464	239
	Epistemic	0.540	0.391	0.453	279
	Statement	0.870	0.938	0.902	2,454
	Macro Avg	0.625	0.538	0.574	3,208
	Weighted Avg	0.793	0.812	0.799	3,208
Bi-LSTM	Deontic	0.660	0.288	0.401	236
	Dynamic	0.506	0.343	0.409	239
	Epistemic	0.536	0.240	0.332	279
	Statement	0.837	0.962	0.895	2,454
	Macro Avg	0.635	0.458	0.509	3,208
	Weighted Avg	0.774	0.803	0.774	3,208
BERT	Deontic	0.622	0.691	0.655	236
	Dynamic	0.688	0.644	0.665	239
	Epistemic	0.678	0.756	0.715	279
	Statement	0.949	0.932	0.940	2,454
	Macro Avg	0.734	0.756	0.744	3,208
	Weighted Avg	0.882	0.877	0.879	3,208
GPT	Deontic	0.535	0.547	0.541	236
	Dynamic	0.385	0.308	0.342	250
	Epistemic	0.313	0.585	0.408	272
	Statement	0.898	0.827	0.861	2,450
	Macro Avg	0.426	0.453	0.430	3,208
	Weighted Avg	0.782	0.746	0.759	3,208
Ernie	Deontic	0.301	0.812	0.439	239
	Dynamic	0.285	0.230	0.254	248
	Epistemic	0.422	0.576	0.487	269
	Statement	0.922	0.751	0.828	2452
	Macro Avg	0.386	0.474	0.402	3208
	Weighted Avg	0.785	0.700	0.726	3208
LLaMA	Deontic	0.344	0.712	0.464	236
	Dynamic	0.415	0.156	0.227	250
	Epistemic	0.285	0.820	0.423	272
	Statement	0.946	0.704	0.807	2,450
	Macro Avg	0.398	0.478	0.384	3,208
	Weighted Avg	0.804	0.672	0.704	3,208
Deepseek	Deontic	0.346	0.780	0.479	118
	Dynamic	0.247	0.328	0.281	58
	Epistemic	0.601	0.620	0.610	163
	Statement	0.960	0.876	0.916	1,961
	Macro Avg	0.539	0.651	0.572	2,300
	Weighted Avg	0.885	0.839	0.856	2,300

Table 4: Precision, recall, and F1-score for each modality category and overall scores for different models evaluated on the test sets.