PEACE - An In-Domain and Cross-Domain Chinese Proposition Classification Evaluation Benchmark for Natural Language Processing

Anonymous ACL submission

Abstract

A large number of propositions with rich expressions exist in natural language. Correct classification of propositions is helpful to natural language understanding and reasoning. However, most of the existing researches are limited by logical constants, while most propositions in natural languages are implicit. And there is a lack of complete proposition classification system, resources and research on cross-domain tasks. We propose the concept of implicit proposition which is more suitable for NLP application scenarios. And we present PEACE, for in-domain and cross-domain proposition classification tasks, covers all tasks related to proposition classification, among which the task of categorical proposition classification is put forward for the first time, which is a large-scale proposition classification data set with implicit propositions. It contains over 45k sentences, multi-level classes and 5 different domains. We use PEACE as a benchmark dataset and propose a series of proposition classification tasks. We use multiple popular machine learning methods as our baseline methods and run experiments on each task. The results show that the existing pre-training models can classify all kinds of propositions relatively well, but the cross-domain tasks of non-modal proposition classification is still challenging. We release this benchmark with the hope of advancing research in natural language understanding, reasoning, and generation.

1 Introduction

Propositions are defined as the meaning of declarative sentences in linguistics and logic(McGrath and Frank, 2020). Different classes of propositions imply different logical relations, which is a crucial part of Natural Language Understanding (NLU)(Zhou, 1993).

Concretely, propositions are divided into non-modal propositions and modal propositions according to whether they make deterministic judgments about objects(McGrath and Frank, 2020; Fan et al., 2016). In logic, non-modal proposition is divided into simple proposition and compound proposition. The categorical proposition in simple proposition can be further divided into Affirmative universal (A), nEgative universal (E), affIrmative particular (I) and nOgative particular (O)[1]. Compound proposition can be divided into conjunctive proposition, hypothetical proposition and disjunctive proposition(Fan et al., 2016). Modal propositions can be divided into alethic modality, epistemic modality, deontic modality, dynamic modality, and evaluative modality according to the different situational constraints(Wu, 2021). For the convenience of research, logic sums up the corresponding logical constants through the different logical features of each classification. These logical constants are usually regarded as the basis of proposition classification. Although the appearance of logical constant facilitates the further study of logic, the sentence forms of natural language we use in our daily life are often flexible and diverse, which are not as rigorous as the logical constant defined by logic.

In terms of NLP, proposition classification is relevant for many downstream tasks. Classify modal proposition is helpful to automatic detection of intention, uncertainty, behavior and so on(Vincze et al., 2008; Zerva et al., 2017; Prieto et al., 2020).

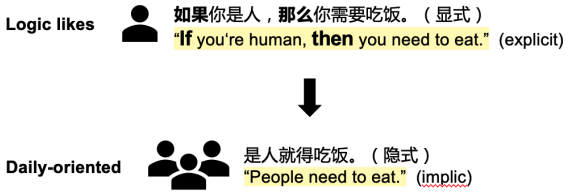

Figure 1: An example of explicit propositions with logical constants in logic and implicit propositions that are more inclined to be used in everyday language.

[1]To be more correct, A and I letters came from the Latin affirmo, and E and O from the Latin nego.

Classifying non-modal proposition helps to identify logical symbols, extract logical relationships in sentences, which is further helpful for natural language reasoning(Wang et al., 2021; Jiao et al., 2022; Huang et al., 2021). Furthermore, classifying categorical propositions helps to test the inferential validity of syllogisms[2].

In recent years, many researchers began to pay attention to how to transfer the research idea of parsing sentence logic from linguistics to NLP. Liu et al. (2021) introduced the concept of proposition into Chinese NLP community and provided an explicit dataset for non-modal propositions. Rubinstein et al. (2013) proposed the event-based modal detection task and provided the corresponding corpus for it. Pyatkin et al. (2021) further detected and classified modal expressions based on them. However, none of these solutions has a systematic and comprehensive corpus resource for the community. Meanwhile, human beings can transfer proposition classification knowledge acquired from one domain to another, and the ability of domain transfer is also an important embodiment of artificial intelligence. We think it is necessary to further investigate the task of cross-domain proposition classification.

Different endeavors suffer from one (or more) of the following types of deficiencies with respect to their expressivity and coverage. First, Most researchers directly apply the proposition concept of formal logic to nlp, but the propositions we use in everyday language are not as rigorous as the words and structures of sentences defined by logic. The proposition that follows the definition of complete logic, As shown in Figure 1, we prefer to use implicit propositions in everyday natural languages. Therefore, it is necessary to break through the limitation of sentence patterns, redefine the propositions in NLP, and expand the scope of proposition research to include implicit propositions. Second, there is still no work that makes a complete inquiry into all types of propositions. For categorical propositions, to the best of our knowledge, there are currently no classification tasks and datasets for such propositions, and categorical propositions can directly aid in syllogistic inference tasks. For modal propositions, previous studies mostly focused on the disambiguation of modal keywords, and only paid attention to a few keywords. Although the research of Pyatkin et al. (2021) expanded modal keywords to some extent, they were still triggered by keywords and still belonged to explicit propositions. Third, further cross-domain research needs to be improved. It is natural for humans to generalize the domains of proposition classification, but the domain generalization ability of machines on these tasks has not been investigated in detail. We believe that the AI community needs an extensive investigation for the tasks of classifying cross-domain propositions.

All in all, we think that proposition classification needs a systematic and comprehensive evaluation method, which includes all the classification tasks of propositions with practical application value. We should not only consider explicit propositions and implicit propositions, but also further explore their classification in various domains.

In this paper, we propose PEACE, a manually annotated dataset with multiple benchmark tasks for proposition classification. We transfer the classification system of propositions from linguistics to NLP, and make corresponding adjustments according to the understanding of natural language semantics. Based on this, we propose a multi-level classification system. We use SVM, TextCNN, TextRNN, DPCNN and Bert, RoBERTa to explore the experimental performance of proposition classification tasks and obtain corresponding baseline results, which provide the corresponding reference for the future methods of proposition classification. The main contributions of this paper can be summarized as follows:

- We put forward the concept of implicit proposition, get rid of the restriction of logical constant on proposition, expand the object of proposition classification to the actual natural language, and make the task more suitable for NLP application scenarios.

- PEACE covers all tasks related to proposition classification, among which the task of categorical proposition classification is put forward for the first time, and the datset we built is the first large-scale proposition classification data set with implicit propositions.

- We further explore the task of cross-domain proposition classification, and give a set of strong baseline methods, which can provide a reference for the future exploration of in-domain and cross-domain Chinese proposition classification.

[2]the most basic form of reasoning.

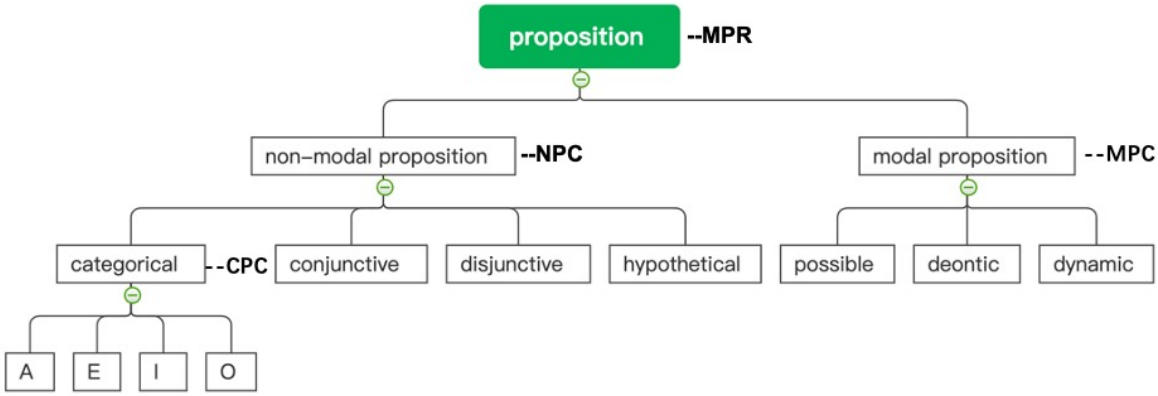

Figure 2: The proposed framework of proposition classification and the corresponding relationship with our tasks. MPR denotes Modal Proposition Recognition, NPC denotes Non-modal Propositions Classification, CPC denotes Categorical Proposition Classification, MPC denotes Modal Propositions Classification. See section Benchmark Tasks for more detailed explanation.

2 Benchmark Tasks

In order to explore the classification of propositions, we combined logic to explore the whole proposition system. Figure 2 presents the complete taxonomy that we propose for proposition classification in NLP. The examples of each classes are shown in Appendix. Based on the classical classification of logic, it has been simplified to make it intuitive and easy for NLP practitioners and non-logicians to use. In this section, we discuss the redefinition of various propositions in NLP and the specific description of each task.

2.1 Redefinition

Linguistics and logic define proposition as the meaning of declarative sentence, which has nothing to do with language form(Portner, 2009; Fan et al., 2016). Different propositions contain different logical keywords, which can express different logical relationships. When propositions were introduced into NLP, Liu et al. (2021) used these keywords as the basis of proposition classification, but we think that the basis of proposition classification should be semantic logic rather than language form, and the concept of implicit propositions should be introduced. We show the redefinition of categorical proposition and conjunctive proposition as follows, and the redefinition of complete system classification is detailed in the appendix.

categorical proposition Make a direct and unconditional judgment on whether an object logically contains a certain attribute or belongs to a certain category.

conjunctive proposition A compound proposition, which reflects that several situations or properties of objects exist at the same time, and logically has a conjunctive relationship.

2.2 MPR: Modal Proposition Recognition

Logic first divides propositions into modal propositions and non-modal propositions. However, modal propositions have not been paid enough attention in previous studies of NLP, which leads to many modal propositions being labeled as non-modal propositions by mistake. Therefore, we think it is necessary to identify modal propositions to distinguish modal propositions from non-modal propositions. The task is to predict whether a given sentence is a proposition and whether it is a modal proposition. We use the three-way(*not-proposition/non-modal/modal*) class split, and use only sentence-level labels.

2.3 NPC: Non-modal Propositions Classification

The ProPC(Hu et al., 2021) provides a dataset for in-domain and cross-domain non-modal propositions classification. The sources are Baidu Encyclopedia and news, medical, law and finance, but it doesn't pay attention to modal proposition, and the news domain may contain other domains. Therefore, we think it is necessary to re-label and re-divide the dataset to test its performance on non-modal propositions classification. See the third section for the specific treatment of the data. The task is to predict the classification of a given non-modal proposition.

2.4 CPC: Categorical Proposition Classification

Syllogism is the most common form of deductive reasoning, among which categorical syllogism is the most common syllogism. The validity of categorical syllogism can be judged by the classification of categorical propositions(Fan et al., 2016). Therefore, we believe that identifying and classifying categorical propositions is of certain significance to the subsequent natural language reasoning. The task is to predict the classification of a given categorical proposition.

2.5 MPC: Modal Propositions Classification

In logic, modal propositions are divided into alethic modal, cognitive modal, deontic modal and dynamic modal(Wu, 2021). The alethic modal is based on the state of the objective world, and the cognitive mode is based on personal subjective experience and knowledge to judge the proposition. Applying to NLP, we find that both the alethic modal and the cognitive modal are related to possibility and difficult to distinguish from each other in semantic understanding, as in the following two examples: 1) Water will **inevitably** turn into steam when it is heated to the boiling point.(*alethic modal*). 2) I think Mr. Wang **must** be at school.(*cognitive modal*). Therefore, we think they can be classified as possible modal.

Pyatkin et al. (2021) put forward a modal hierarchical classification system based on modal event detection, but its proposed classification system is too fine-grained, just like the alethic modal and cognitive modal mentioned above. Although the article explains the difference between them by "by state of word" and "by state of knowledge", we think that the cost of manual annotation is still very high, and the connection between such fine-grained classification and downstream tasks is not very close. Therefore, we think it is enough to divide modal propositions into possible modal, deontic modal and dynamic modal. The task is to predict the classification of a given modal proposition.

3 Dataset Construction

3.1 Data Acquisition

3.1.1 None-modal propositions.

A suitable dataset for the classification of None-modal propositions should have canonical statements and clearly stated semantics. ProPC (Hu et al., 2021) is a dataset constructed for non-modal propositions, including explicit and implicit propositions. Its data source is Baidu Encyclopedia, which is an open and free online encyclopedia with complete sentence structure and clear meaning. There are 15,000 annotated statements in ProPC, including 1,000 each in the domains of finance, law, medical, and news. We use ProPC (Hu et al., 2021) as our data source and remove the sentences containing modal keywords to modal propositions dataset and conducted a secondary check on the annotation results. Besides, in order to match the modal dataset, we split the news corpus into four other fields.

3.1.2 Modal propositions.

Most statements in the encyclopedia are deterministic statements describing facts. The sentences in novels, blogs and publications are complete and there are more modal events. We extracted this part of sentences from CCL corpus(ZHAN, 2003) as as our original corpus source. We selected 12 typical modal words(Wu, 2021) as trigger words with 1,000 statements of each, so a total of 12,000 statements were obtained as comprehensive corpus. In the same way, we tried our best to find 100 statements corresponding to each keyword in the three domains of finance, law, and medical, with a total of 1,200 statements of each domain. Furthermore, we added sentences from ProPC which is modal proposition after the secondary check. It is possible that multiple keywords with different modalities be matched in a sentence, so we re-label these sentences separately.

3.1.3 Categorical propositions.

Categorical propositions belong to simple propositions (that is, propositions that contain no other propositions themselves)(Fan et al., 2016), while statements in encyclopedia are mostly nested logical relations and mostly positive semantics. In order to make the corpus contain sufficient statements of various types as much as possible, we added the statements from LogiQA(Liu et al., 2020), which is collected from publically available questions of the National Civil Servants Examination of China. To obtain sufficient categorical propositions and negative-semantic statements, we delete the statements containing the logical keywords of compound propositions and filter some statements with negative words in the predicate part. After manual secondary filtering, we finally identified 10,000 statements. Categorical propositions belong to non-

modal propositions. For the corpus of the vertical domain, we use this part of the sentences in the non-modal proposition corpus for labeling, while there are few sentences that can be labeled as categorical propositions in this part of the corpus. After investigating more corpus sources, we found that the financial sentences in the news can be marked with sufficient statements of categorical proposition. In order to add the diversity of comparison, 1,000 sentences of finance domain are extracted from THUnews(Yluthu, 2016) to form categorical dataset.

3.2 Data Annotation

There are 3 stages of annotation for modal and categorical propositions statements: annotation training, trial annotation and formal annotation.

Before organizing the annotation, we analyzed and labeled some extracted statements, and completed the annotation standard for proposition classification based on natural language processing, and identified 120 statements as the benchmark for annotation training and trial annotation. For the classification of propositions, as a text may not express it's logical relation dependent on the logical keywords, or are not expressed by some keywords at all. Thus, the basic principle that guides the annotation of proposition classification is: look not at keywords but at semantic logic.

The annotation was conducted by 6 undergraduate and graduate students, among which 3 annotator were responsible for modal propositions, 3 annotator for categorical propositions. They first received the annotation training, which provided the project background introduction and annotation specification explanation, and carried out the annotation demonstration of 20 sentences to further explain the specific process of annotation.

In the trial annotation part, 100 statements with identified labels above (50 for modal propositions and 50 for categorical propositions) were used to test the annotators' understanding of the proposition specification, and 3 annotators in each group used the corresponding 50 statements respectively to annotate. The consistency test results show that there was good consistency between the two groups of annotators (Fleiss Kappa(Fleiss and Joseph, 1971)=0.7278 and 0.652), and the accuracy of each two annotator reached more than 80% (compared with the identified labels). We also conducted a pairwise cross-validation. The statements labeled by each annotator has duplicate annotation statements with each other annotators (the number of repetitions in each part is equal), and the agreement rate between each two annotator is also greater than 80%. All these indicate that the trained annotator has understood the annotation specification and can proceed to formal annotation.

Type	EK	EI	FN	Law	Med
Category	3,794	120	42	234	183
Conjunctive	1,569	367	258	218	262
Hypothetical	1,082	56	85	81	28
Disjunctive	213	10	1	4	4
Not	2,542	258	389	188	239
Total	9,201	812	776	726	717

Table 1: The overview of propositions in non-modal dataset.

In the formal annotation part, we have also double-checked the labeling results of each annotator every two days to ensure the quality of the dataset. The resulting set of annotations for a total of 44,968 statements consists of 12,232 annotated statements for non-modal proposition, 20,782 for modal proposition , 11,954 statements for categorical proposition. In this paper, we use "C&E" denotes Comprehensive and Encyclopedia data, "C&L" denotes Encyclopedias and LogiQA data, "FN" denotes Finance data, "Law" denotes Law data, "Med" denotes Medical data. "A" denotes Universal affirmative proposition, "E" denotes Universal negative proposition, "I" denotes Particular affirmative proposition, "O" denotes Particular negative proposition.

3.3 Dataset Analysis

None-modal propositions. There are 4 distinct classifications in ProPC dataset. In detail, we sorted out five files of statistical data, including "EK": Encyclopedia data with logic keywords(not all explicit propositions), "EI": Encyclopedia data which contains implicit propositions and conforms to natural language distribution, "FN": finance data, "Law": law data, "Med": medicine data, "Not": a sentence which is not a proposition. The overall distribution of propositions in non-modal dataset shows in Table 1 indicates that non-modal propositions exist more widely, especially in Chinese language, and from Figure 3, we found that non-modal implicit propositions accounted for a considerable proportion in each domain. Under natural distribu-

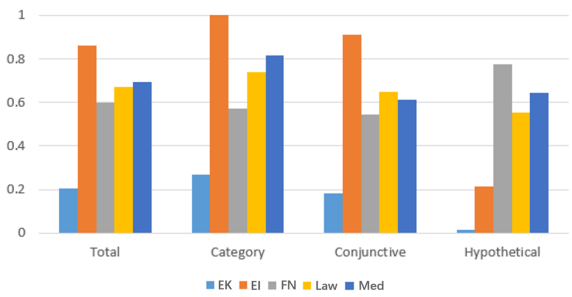

Figure 3: Implicit propositions in non-modal[3]. Among them, the disjunctive proposition has no implicit form.(Huang, 1991)

Type	C&E	FN	Law	Med
Possible	5,006	289	269	639
Deontic	5,018	625	827	451
Dynamic	3,911	316	135	260
Not	2,640	134	162	100
Total	16,575	1,364	1,393	1,450

Table 2: The overview of propositions in modal dataset.

tion, implicit hypothetical proposition proportion is smaller than other two types.

Modal propositions. For modal dataset, there are three distinct classifications and 4 domains: Comprehensive and Encyclopedia data, finance data, law data and medicine data. Except for sentences containing keywords, which are explicit, we also selected implicit modal propositions from sentences that do not contain keywords in ProPC dataset for each domain. During annotation, we notice that it exists cross-modal keywords.

From Table 2, we observed that sentences containing keywords account for a large proportion of modal propositions and the Not means the sentence is not a modal proposition but it doesn't mean that these sentences are not propositions which need a further annotation. And comparing the Total and modal proposition number in C&E, implicit modal propositions proportion is relatively small.

Numerous statistics of cross-modal keywords were given in Appendix Table 6 and specific cross-modal keywords distribution were given in Appendix Figure 5. It found that every keyword has cross-modal meanings and there are some keywords with obvious cross-modal meanings.

[3]Implicit propositions account for almost 100% of the categorical propositions in EI, is because EI data itself is encyclopedia with implicit data, and people's daily expression of nature propositions tends to be non-standard logic forms, that is, implicit forms.

Type	C&L	FN	Law	Med
A	7,562	782	733	751
E	1,341	125	162	56
I	861	89	86	177
O	236	4	19	16
Total	10,000	1,000	1,000	1,000

Table 3: The overview of propositions in categorical dataset.

Categorical propositions. There are 4 distinct classifications in categorical propositions, including sentences from Encyclopedias and LogiQA(Liu et al., 2020) publications, finance sentences from THUnews(Yluthu, 2016), Law data and Medical data from ProPC(Hu et al., 2021).

From Table 3, the A occupies the largest proportion in categorical dataset which indicated that among simple propositions, universal affirmative propositions are the easiest to appear, and have the characteristics of simplicity and directness.

4 Experiments

4.1 Baseline Methods

In order to explore and analyze the performance of machines in PEACE, we selects eight models from the five popular text classification methods as follow:Rule-based method, Majority, SVM, TextCNN, TextRNN, DPCNN, BERT and RoBERTa. See appendix for the detailed description of the models.

4.2 Experiments Setup

Predicting proposition classification is a multi-class classification problem. Given a sentence-level text $\mathcal{T}$and a set of labels $\mathcal{L} = (l_1, l_2, \ldots, l_n)$ for proposition classification, we hope to learn a mapping $\mathcal{C} : \mathcal{T} \mapsto \mathcal{P}(\mathcal{L})$.

4.2.1 In-Domain Evaluation

We select the data of C&E(C&L in CPC task), which is the easiest to obtain and label, and has the largest number, as the datasets for each in-domain task exploration, and scrambled the data according to the ratio of 8: 1: 1 and divided it into training, verification and test set. For NPC task, because the sentences with logical keywords are easier to obtain and label, we want to test the performance of the model trained by explicit proposition on naturally distributed data. Therefore, we have designed the experiment of the model on EI dataset (554

Model	MPR	NPC		MPC	CPC
	C&E	EK	EI	C&E	C&L
Rule-based	0.34	0.60	0.49	0.48	0.08
SVM	0.69	0.49	0.24	0.71	0.77
TextCNN	0.44	0.56	0.22	0.36	0.74
TextRNN	0.44	0.56	0.22	0.36	0.74
DPCNN	0.44	0.56	0.22	0.36	0.74
BERT	**0.86**	**0.91**	0.48	**0.83**	**0.95**
RoBERTa	**0.86**	0.90	**0.54**	**0.83**	**0.95**
Majority	0.44	0.56	0.22	0.36	0.74

Table 4: The weighted average F1-score of in-domain tasks with different models.

sentences of data with natural distribution obtained by additional labeling), that is, training a classifier $\mathcal{C}(EK, EI)$.

4.2.2 Cross-Domain Evaluation

To perform cross-domain evaluation, we use C&E data as $\mathcal{T}_{source}$, and the data of Financial, Law and Medical domains as the $\mathcal{T}_{target}$. We treat $\mathcal{T}_{source}$ as available data and $\mathcal{T}_{target}$ as an incoming dataset from a novel domain. In order to explore the generalization ability of the model, we use the model trained on $\mathcal{T}_{source}$ to test it directly on $\mathcal{T}_{target}$ data, that is, training a classifier $\mathcal{C}(source, target)$. To further explore the migration ability of the model, we first train the classifier on $\mathcal{T}_{source}$ and then finetune it on $\mathcal{T}_{target}$, that is, training a classifier $\mathcal{C}(finetune, target)$. We choose Bert, who performs well on the in-domain task, as the model of this part, and marked as BERT-trans in the experiments.

During our experiments, we use F1-score as the main evaluation metric, weighted across all classes.

5 Results and Discussion

We evaluate the performance of the model on four tasks by in-domain and cross-domain settings. We present performance on the main benchmark tasks in Table 4 and Table 5.

5.1 General Trends

Aside from the settings of in-domain and cross-domain, we observe some general trends. Among all the tasks, the CPC task has the highest score on the whole, which may be due to the more obvious characteristics of the categorical propositions, and the extremely unbalanced data distribution. The A classification

Task	Model	Domains		
		Finance	Law	Medical
MPR	Rule-based	0.27	0.30	0.28
	SVM	0.60	0.63	0.64
	TextCNN	0.49	0.51	0.53
	TextRNN	0.49	0.51	0.53
	DPCNN	0.49	0.51	0.53
	BERT	0.65	0.79	0.68
	RoBERTa	0.62	0.75	0.48
	BERT-trans	**0.76**	**0.81**	**0.71**
	Majority	0.49	0.51	0.53
		Finance	Law	Medical
NPC	Rule-based	0.14	0.14	0.22
	SVM	0.50	0.38	0.56
	TextCNN	0.11	0.44	0.38
	TextRNN	0.11	0.44	0.38
	DPCNN	0.11	0.44	0.38
	BERT	0.42	0.64	0.48
	RoBERTa	0.32	0.67	0.48
	BERT-trans	**0.77**	**0.74**	**0.78**
	Majority	0.11	0.44	0.38
		Finance	Law	Medical
MPC	Rule-based	0.55	0.46	0.50
	SVM	0.44	0.66	0.57
	TextCNN	0.51	0.67	0.47
	TextRNN	0.51	0.67	0.47
	DPCNN	0.51	0.67	0.47
	BERT	0.78	**0.83**	0.78
	RoBERTa	0.79	0.82	0.80
	BERT-trans	**0.81**	**0.83**	**0.83**
	Majority	0.51	0.67	0.47
		Finance	Law	Medical
CPC	Rule-based	0.10	0.04	0.06
	SVM	0.80	0.74	0.68
	TextCNN	0.78	0.62	0.64
	TextRNN	0.78	0.62	0.64
	DPCNN	0.78	0.62	0.64
	BERT	**0.97**	0.89	0.85
	RoBERTa	**0.97**	0.87	0.85
	BERT-trans	0.90	**0.97**	**0.93**
	Majority	0.76	0.62	0.64

Table 5: The weighted average F1-score of cross-domain tasks with different models.

alone accounts for about 76%. Among all the methods, Rule-based methods are not good for all tasks, which shows that it is not feasible to identify proposition types only by logical connectives, and the pre-training language model has obviously achieved better experimental results, and has a better performance than other methods in all tasks.

5.2 In-Domain Evaluation

We observe that: On the whole, the models based on pre-training can produce better overall scores than those based on neural networks. Detailed experimental data will be placed in the appendix. By looking at the performance of the model in each category, we found that the model based on neural network predicted all the data to the category with the largest amount of data. We speculate that this may be caused by the imbalance of our data. In NPC, the F1-score obtained by the EK-trained model predicting EI is much lower than EK. Because the features of the sentences extracted by EK according to logical keywords are relatively obvious, while EI is more implicit. The large differences cause poor effect of using the EK-trained model to test EI.

Comparing the F1-scores of C&Es, EK and C&L, it can be found that the C&E result of MPC is slightly worse, which may be related to the fact that the same word has multiple modal meanings in MPC. The logic words in other tasks only belong to one class.

5.3 Cross-Domain Evaluation

To evaluate generalizability and transferability, we analyze the results for the $\mathcal{C}(source, target)$ and $\mathcal{C}(finetune, target)$ settings. Table 5 shows the results. We observe that pre-trained models achieve better average F1-scores in the $\mathcal{C}(source, target)$ settings than the majority baseline. This indicates that although the task of proposition classification has domain-specific nature, the classification knowledge learned in different domains can be extended to new domains to some extent.

In the NPC tasks, the Rule-based method under cross-domain setting is significantly lower than that under in-domain setting, which may be due to the fact that those data under in-domain setting are mostly explicit propositions guided by logical keywords, while the data under cross-domain setting conform to the natural distribution, which also reflects the difficulty of applying the classifier of explicit propositions to implicit propositions.

In the CPC task, the effect of the models on Finance is better than that of Law and Medical, which may be due to the difference of datasets. Figure 4 shows the proportion of different classes in various domains in each task. Because the source of Finance data in the CPC task is news, and the source of Medical and Law data are question-and-answer corpus, the sentence structure of Finance is more regular and unified, and the data in CE is also close, while the sentences in Law and Medical are more colloquial. The experimental results in this part also indicates that colloquial data such as question-and-answer are more challenging.

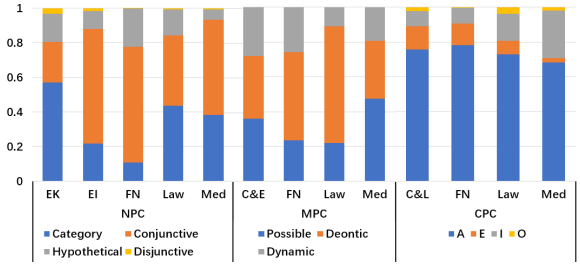

Figure 4: The proportion of different classes in various domains in each task.

Comparing the two settings of $\mathcal{C}(source, target)$ and $\mathcal{C}(finetune, target)$, we can evaluate the transferability of the model. We observe that $\mathcal{C}(finetune, target)$ performs better than or on par with $\mathcal{C}(source, target)$. Among all tasks, the result of NPC task is generally low, which implies that the feature of non-modal proposition are more difficult to learn and more difficult.

6 Conclusions

In this paper, we present four proposition classification benchmark tasks which are based on solid theoretical foundations, yet are adapted to fit the needs of NLP practitioners: modal proposition recognization, non-modal propositions classification, modal propositions classification and categorical proposition classification. To enable research on these tasks, we introduce PEACE, a novel dataset with a variety of manual annotations on proposition classification, consist of explicit and implicit proposition drawn from different sources. We use Rule-based, Majority, SVM, TextCNN, TextRNN, DPCNN and BERT, RoBERTa as our baseline methods to run experiments on each of the tasks. Results of our experiments indicates that the existing pre-training models can classify all kinds of propositions relatively well, but the cross-domain tasks of non-modal proposition classification is still challenging. In the future, we will continue to expand the size of the dataset, optimize the model, and explore more methods for domain generalization.

References

Alexander Rakhlin. 2017. Convolutional neural networks for sentence classification. https://github.com/alexander-rakhlin/CNN-for-Sentence-Classification-in-Keras.

Chang M. W. Lee K. Toutanova K. Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yaohua Fan, Liping Wang, Ping Chu, and Yue Yu. 2016. *Formal Logic*. East China Normal University Press.

Fleiss and L. Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mengyang Hu, Pengyuan Liu, Lin Bo, Yuting Mao, Ke Xu, and Wentao Su. 2021. Propc: A dataset for in-domain and cross-domain proposition classification tasks. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 53–64. Springer.

Shiping Huang. 1991. On the hidden form of logical constant. *Journal of Jianghan University (SOCIAL SCIENCE EDITION) 4*.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. *arXiv preprint arXiv:2103.14349*.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. Merit: Meta-path guided contrastive learning for logical reasoning. *arXiv preprint arXiv:2203.00357*.

R. Johnson and Z. Tong. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Lu Liu, Shiya Peng, Chen YU, and Dong YU. 2021. Automatic recognition and analysis of explicit propositions in natural language. *Journal of Chinese Information Science*, 35(2):41–51.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Matthew McGrath and Devin Frank. 2020. Propositions. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2020 edition. Metaphysics Research Lab, Stanford University.

Paul Portner. 2009. *Modality*. OUP Oxford.

Mario Prieto, Helena Deus, Anita De Waard, Erik Schultes, Beatriz García-Jiménez, and Mark D Wilkinson. 2020. Data-driven classification of the certainty of scholarly assertions. *PeerJ*, 8:e8871.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. *arXiv preprint arXiv:2106.08037*.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Dan Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 workshop on annotation of modal meanings in natural language (WAMM)*, pages 38–46.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. Logic-driven context extension and data augmentation for logical reasoning of text. *arXiv preprint arXiv:2105.03659*.

Yajun Wu. 2021. *Semantic Ontology research of Chinese modal words*. Ph.D. thesis, Southwest University.

Yluthu. 2016. Yluthu/thunews: Spider for crawling http://news.tsinghua.edu.cn/. https://github.com/yluthu/THUNews.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Chrysoula Zerva, Riza Batista-Navarro, Philip Day, and Sophia Ananiadou. 2017. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*, 33(23):3784–3792.

GUO Rui CHEN Yirong ZHAN, Weidong. 2003. The ccl corpus of chinese texts: 700 million chinese characters, the 11th century b.c. - present, available online at the website of center for chinese linguistics (abbreviated as ccl) of peking university. http://ccl.pku.edu.cn:8080/ccl_corpus.

Liquan Zhou. 1993. Formal logic and natural language. *philosophy research*, (12):29–35.

7 Appendix

7.1 Specific Redefinition

7.1.1 non-modal proposition

A deterministic judgment on the actual existence or non-existence of things.

categorical proposition Make a direct and unconditional judgment on whether an object logically contains a certain attribute or belongs to a certain category.

According to the difference of joint term and quantity term, the property proposition can be divided into the following four types:

A: Universal affirmative proposition, a proposition that semantically concludes that all objects of a class of things have certain properties.[4] The explicit proposition has the form: "*All* S *are* P".

E: Universal negative proposition, a proposition that semantically concludes that all objects of a class of things have no certain properties. The explicit proposition has the form: "*No* S *is* P".

I: Particular affirmative proposition, a proposition that semantically concludes that some objects of a class of things have certain properties. The explicit proposition has the form: "*Some* S *are* P".

O: Particular negative proposition, a proposition that semantically concludes that some objects of a class of things have no certain properties. The explicit proposition has the form: "*Some* S *are not* P".

conjunctive proposition A compound proposition, which reflects that several situations or properties of objects exist at the same time, and logically has a conjunctive relationship.

disjunctive proposition A compound proposition, which reflects the existence of at least one condition or attribute of an object, and logically has a disjunctive relationship.

hypothetical proposition A compound proposition, which contains a previous or tentative explanation, and logically has a conditional relationship.

7.1.2 modal proposition

Modality refers to the speaker's modification of state of affairs, which is used to express the concepts of possibility, inevitability, promise, obligation and ability.

possible modal The speaker makes a decision about the likelihood that the central meaning expressed by the proposition will occur.

[4] singular proposition refers to only one object, so it can generally be regarded as a universal proposition.

Modality	Keyword	Pos	Deo	Dyn
Possible	必然 inevitable	550	35	34
	可能 possible	888	69	58
	偶然 accidental	87	16	42
	一定 definite	718	41	80
	应该 should	188	413	33
Deontic	必须 must	37	1,165	13
	可以 can	175	857	481
	应当 should	38	897	8
Dynamic	敢 dare	9	13	318
	会 can	834	283	543
	能 can	986	531	1,838
	要 want	112	652	412
Total	-	4,922	4,972	3,860

Table 6: The cross-modal keyword proposition in C&E.

deontic modal The speaker allows or gives instructions that make actions, states, and events possible or to be performed

dynamic modal Focus on the subject, participant's ability or willingness, not the speaker's point of view or attitude.

Task	C&E	FN	Law	Med
MPR	23,948	2,006	1,957	2,067
NPC	6,659(EK) 553(EI)	386	537	477
MPC	13,935	1,230	1,231	1,350
CPC	10,000	1,000	417	537

Table 7: The overview of propositions in tasks' dataset.

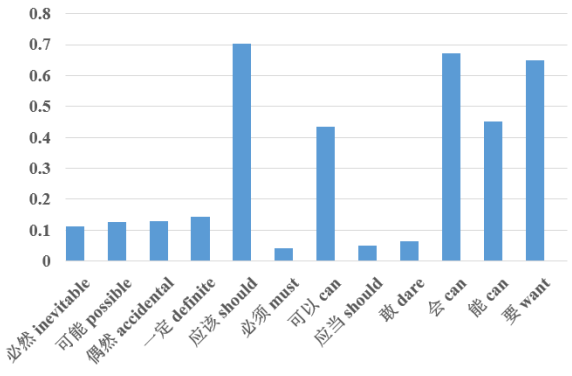

Figure 5: Cross-modal Keyword Proportion in C&E

7.2 Baseline Methods

7.2.1 Rule-based

This method simply matches statements based on the logical keywords corresponding to the different

Task	Domain	Train	Dev	Test
MPR	Total	23,981	2,998	2,998
	C&E	19,156	2,394	2,394
	Finance	19,156	2,394	2,006
	Law	19,156	2,394	1,957
	Medical	19,156	2,394	2,067
NPC	EK	5,326	666	666
	EI	5,326	666	553
	Finance	5,326	666	386
	Law	5,326	666	537
	Medical	5,326	666	477
MPC	C&E	11,147	1,393	1,394
	Finance	11,147	1,393	1,230
	Law	11,147	1,393	1,231
	Medical	11,147	1,393	1,350
CPC	C&L	8,000	1,000	1,000
	Finance	8,000	1,000	1,000
	Technology	8,000	1,000	1,000
	Law	8,000	1,000	417
	Medical	8,000	1,000	537

Table 8: The amount of train, dev and test data in tasks.

classes, which establishes the corresponding template for each class by regular expression and then matches the statements one by one.

7.2.2 Majority

Majority classifier shows the results that just predicting the majority class.

7.2.3 Statistical Machine Learning

SVM(Johnson and Tong, 2017): support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

7.2.4 Neural Network

TextCNN(Alexander Rakhlin, 2017): Several convolution kernels of different sizes are used to extract the key information in sentences, and has strong parallel computing capability, can quickly realize feature extraction. *TextRNN*(Liu et al., 2016): Based on bidirectional LSTM, which is good at capturing longer sequence information. There are two parts of information for its input on each time step: reserved information of the previous time step, and the original information corresponding to the current time step. *DPCNN*(Johnson and Tong, 2017): Deep Pyramid Convolutional Neural Networks for Text Categorization, is a network based on word-level level, which can extract long-distance text dependencies by deepening the network.

7.2.5 Pre-training Language Model

BERT(Devlin, 2018): Bidirectional Encoder Representations from Transformers, is one of the most popular pre-training method in recent years. As a new language model, it performs well in various text tasks such as question answering, named entity recognition, text classification and so on. *RoBERTa*(Liu et al., 2019): Robustly Optimized BERT Approach, optimized the general BERT model details and training strategies, and used a larger training dataset, which can achieve better results in many NLP downstream tasks.

7.3 Related Work

Proposition, as the smallest unit of logical reasoning in logic, has been deeply studied in logic and linguistics, but it has not been fully explored in NLP. Liu et al. (2020) constructed a reading comprehension dataset based on the standardized test questions of human logical reasoning in China National Civil Service Examination. Yu et al. (2020) constructed a reading comprehension dataset based on the logical reasoning questions of standardized postgraduate exams. Although these datasets were constructed without further analysis from the perspective of proposition, many researchers applied logic knowledge in challenging these tasks. Wang et al. (2021) proposed to use logical symbols and expressions to improve the logical reasoning ability of machines. Huang et al. (2021) builds a logical diagram by extracting logical elements and discourse relations to model the logical structure. The logical structures in the above studies can be understood as logical elements and logical relationships, but they are only based on their corresponding tasks and datasets, and there is no systematic and complete introduction of logical structures into NLP.

Liu et al. (2021) first introduced the concept of Chinese proposition into NLP, and proposed explicit proposition identification and element extraction tasks, and constructed the corresponding datset. Hu et al. (2021) introduced the concept of implicit proposition and offered more domains, constructed ProPC, and made preliminary attempts on in-domain and cross-domain proposition classification. Pyatkin et al. (2021) proposed modal sense hierarchical classification based on modal event detection and verified the improvement of modal event detection task by modal classification.

Class	Example
Categorical	
A	所有的鲸鱼都是哺乳动物。（显式） *All* whales *are* mammals. (explicit) 正方形是矩形。（隐式） Square is rectangular. (implicit)
E	所有的行星都不是恒星。（显式） *All* the planets *are not* stars. (explicit) 人都不能知道自己去世之后的事情。（隐式） People can't know what happens after they pass away. (implicit)
I	有的旁观者清。（显式） *Some* bystanders *are* aware of the situation. (explicit) 当局者迷。（隐式） The player can not see most of the game. (implicit)
O	有些人不珍惜生命。（显式） *Some* people *don't* value life. (explicit) 人不珍惜生命。（隐式） People don't value life. (implicit)
Non-modal	
Categorical	所有的人都是贤良的。（显式） *All* people *are* virtuous.(explicit) 人皆贤良。（隐式） Everyone is virtuous.(implicit)
Conjunctive	小张不仅学问多而且很好学。（显式） Xiao Zhang is *not only* knowledgeable *but also* studious. (explicit) 小张别说学问多了，压根就不好学。（隐式） Xiao Zhang, not to mention knowledgeable, doesn't like studying at all. (implicit)
Hypothetical	如果要当一名合格的学生，那么就要好好学习。（显式） *If* you want to be a qualified student, *then* you have to study hard. (explicit) 要当一名合格的学生就要好好学习。（隐式） To be a qualified student, one have to study hard. (implicit)
Disjunctive	小王要么好好学习了，要么成不了好学生。（显式） Xiao Wang will *either* study hard *or* he won't be a good student.(explicit)
Modal	
Possible	水加温到了沸点必然变成水蒸气。（显式） When water is heated to the boiling point, it *must* become water vapor. (explicit) 水加温到了沸点就得变成水蒸气。（隐式） When water is heated to the boiling point, it will change into water vapor. (implicit)
Deontic	你应该好好学习才行。（显式） You *should* study hard. (explicit) 你得好好学习才行。（隐式） You have to study hard. (implicit)
Dynamic	我要吃两碗饭。（显式） I *want* two bowls of rice. (explicit) 给我吃两碗饭。（隐式） Give me two bowls of rice. (implicit)

Table 9: The examples of each class of proposition classification.