Large Language Models Put to the Test on Chinese Noun Compounds: Experiments on Natural Language Inference and Compound Semantics

Le QIU and Emmanuele Chersoni and He ZHOU and Yu-yin HSU
Department of Language Science and Technology, The Hong Kong Polytechnic University
11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China
lani.qiu@connect.polyu.hk,
{emmanuele.chersoni,he.zhou,yu-yin.hsu}@polyu.edu.hk

Abstract

Noun compounds are generally considered an open challenge for NLP systems, given to the difficulty of interpreting the implicit semantic relation between modifier and head, although the advent of Large Language Models (LLMs) recently led to remarkable performance leaps. However, most evaluations have been carried out on English benchmarks.

In our work, we test LLMs on compound semantics understanding in Chinese, adopting two different evaluation scenarios: an extrinsic evaluation in a Natural Language Inference task, and an intrinsic evaluation in which models are directly asked to predict the semantic relation linking the two constituents.

Our results show that the bigger and more recent LLMs are able to surpass supervised baselines in the inference task, especially when tested under the few-shot setting. In the more challenging task of selecting the correct interpretation of the compounds out of a fine-grained typology of semantic relations between head and modifier, the best Chinese LLM (Qwen-plus) manages to select the correct option in about one third of the cases.

1 Introduction

Noun-noun compounds are ubiquitous in natural languages, and they notoriously represent a challenge for NLP applications due to the ambiguity of the implicit semantic relation linking the two nouns, the modifier and the head (Nakov, 2008b; Libben, 2014). The correct interpretation of a compound may be essential for the correct understanding of the semantics of a sentence, and for the appropriateness of an automatic translation: when an English speaker hears about a *carrot cake*, s/he should understand that the cake *is made* of carrot; when a Chinese speaker hears about a 爱情故事 (*love story*), s/he should understand that the story is *about* love. Significantly, native speakers are able to identify similar relationships even in compounds that have never been met before (Van Jaarsveld and Rattink, 1988), with entities having similar semantic features, which explains why compounding is a very productive mechanism for creating novel words. NLP evaluation generally focused on eliciting a plausible paraphrasing of a noun compound from the models, typically in the form of a verb phrase (e.g. *flu virus* → *virus that causes flu*) (Nakov, 2008a; Butnariu et al., 2009; Hendrickx et al., 2013; Shwartz and Waterson, 2018; Shwartz, 2019; Coil and Shwartz, 2023; Rambelli et al., 2024), and mainly using English as the language of study.

In our work, we test the understanding of Chinese noun compound semantics in current LLMs. Our evaluation is first carried out in an *extrinsic* task, where models are required to grasp the meaning of the compound to perform natural language inference (NLI) (Bowman et al., 2015); and then in an *intrinsic* task, where they are asked to select a semantic relation from a limited inventory, representing the link between modifier and head. We observed that, while the best Chinese LLMs and GPT-4 perform similarly for the NLI task (even beating supervised models when prompted with few-shots), the former are more accurate in selecting specific, human-like semantic interpretations, with Qwen-plus achieving the top performance. [1]

2 Related Work

Some studies in the Chinese NLP literature tackled the challenge of interpreting noun compounds (Wang et al., 2010; Gu et al., 2016; Wang et al., 2016), but they share the main limitation that their evaluation datasets were not made available. The study of Liu et al. (2022) adopted a hybrid approach to the interpretation problem, first employing a classifier to identify the relations of compound nouns,

[1] Code and data are available at https://github.com/Laniqiu/zh.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 1–7

and later utilized a paraphrasing model to interpret those that were labeled with an arbitrary undefined relation. Liu and colleagues did release their benchmark, a dataset in the life service domain containing 1,478 compounds with annotated relation labels. However, the number includes different types of compounds, such as for example adjective-noun compounds, so that the total number of actual noun-noun compounds available for evaluation is relatively small. Moreover, the labels in the dataset refer to the type of meaning carried by the modifier, rather than to the relationship between the modifier and the head: for example, in 生日礼服 (*birthday dress*), the labeled relation is `time`, as the modifier indicates the occasion on which the dress is worn.

Using the noun compounds in Liu and colleagues' data, Zhou et al. (2024) adopted a template-based approach to generate a NLI dataset, where the premise always contains a noun compound and the hypothesis label (`entailment`, `neutral` or `contradiction`) depends on the correct understanding of the compound meaning. Using a total of 66 templates on 625 of the compounds from Liu et al. (2022), they obtained an evaluation dataset of 3,740 premise-hypothesis pairs. Some examples of the dataset items are shown below:

(1) 前提: 运动员有一个不锈钢饭盒。
假设: 不锈钢是饭盒的制作材料。
类别: 蕴含
Premise: The athlete has a stainless steel lunch box.
Hypothesis: Stainless steel is the material that the lunch box is made of.
Category: Entailment

(2) 前提: 清洁工昨天吃了巧克力蛋糕。
假设: 清洁工吃的蛋糕里没有巧克力。
类别: 矛盾
Premise: The janitor ate chocolate cake yesterday.
Hypothesis: There was no chocolate in the cake that the janitor ate.
Category: Contradiction

Although their study only tested relatively small models (i.e. Qwen and Chinese Alpaca in their 7B parameter versions), they found that such models already perform competitively with fine-tuned encoders (i.e. BERT and RoBERTa).

3 Experimental Settings

3.1 Evaluation Datasets

We ran our LLM evaluation on two datasets. The first one is **NCNLI**, a NLI dataset introduced by Zhou et al. (2024): it includes 3,740 premise-hypothesis pairs, 1,564 labeled as 'entailment', 1,092 as 'contradiction' and 1,084 as 'neutral'.

The second one is a newly-constructed dataset for noun compound interpretation in Chinese. The data are noun compounds extracted from the *New Era People's Daily Corpus* (Huang and Wang, 2019), after applying POS Tagging with Jieba. By definition, a noun-noun compound consists of two nouns standing next to each other. A preliminary list of such compounds was automatically extracted, and then filtered by one of the authors (a native speaker of Mandarin Chinese with a PhD in Computational Linguistics) to exclude cases of POS ambiguity and tagger error. This left us with 2,083 compounds in total. Henceforth we refer to this dataset as **NEPD**, to indicate the original source of the data.

To determine the compounding relation of each word, we recruited three graduate students in Chinese linguistics for the annotation. Specifically, we predefined 11 semantic categories of compounding relations: *CAUSE*, *MAKE*, *HAVE*, *USE*, *BE*, *IN*, *FOR*, *FROM*, *ABOUT*, *AND*, *OR*[2], using the hierarchy constructed by Liu and Liu (2019). Prior to annotation, annotators received training on the guidelines and examples. Each annotator was asked to assign one or more semantic relations to each compound. If none of the predefined categories were deemed as appropriate, the annotators were instructed to select the *OTHER* label.

Each compound was annotated by three annotators, and their input was reviewed by a more experienced linguist and annotator (one of the authors of this study) for additional quality checking. We assigned each compound the majority relation, that is, the relation on which at least two of the annotators agreed. To assess consistency between annotations, we used the Jaccard similarity coefficient to measure the overlap between pairs of annotators: this metric calculates the percentage of labels selected by both annotators out of all labels selected by either one of them. On average, we obtained a coefficient value of 0.412, indicating a moderate level of agreement in the task.

[2]Definition for each category will be given in the prompt.

Compounds for which no dominant relation could be identified (i.e. those for which the three annotators chose three different relations) were discarded. As a result, the final dataset comprised 1,514 compounds. Some examples are in Table 1, while relation frequencies can be seen in Table 2.

Compounds	Relation (s)
岛国 (*island country*), 水草 (*water plant*)	IN
风雨 (*wind and rain*), 书画 (*painting and calligraphy*)	AND
中国画 (*Chinese painting*), 民间舞 (*folk dance*)	FROM, ABOUT

Table 1: Example compounds with full agreement (first 2 rows) and partial agreement (the last row). Agreement statistics can be found in Table 6 of the Appendix.

Relation	Frequency	Majority
CAUSE	133	30
MAKE	454	116
HAVE	774	162
USE	137	27
BE	476	79
IN	679	178
FOR	1501	430
FROM	367	71
ABOUT	1327	356
AND	193	56
OR	76	9
OTHER	212	24

Table 2: Frequency of semantic relations in the NEPD data (note that compounds can be annotated with multiple relations) and their frequency as majority relation.

3.2 Models and Settings

We tested a pool of smaller (i.e. around a 7 billion parameter size) and larger Chinese LLMs on both task: **Qwen-7B** (Bai et al., 2023), **Chinese Alpaca 7B** (Cui et al., 2023), **DeepSeek-7B** (Bi et al., 2024) and **Qwen2.5-7B** (Yang et al., 2024) (all of them in their instruction-tuned versions) were tested on our server, while **Qwen-plus** and **DeepSeek-chat** were queried via the online interfaces. Additionally, **GPT-4o-mini**[3] for the sake of comparison with one of the most capable and popular Western models. The prompts we crafted can be found in the Appendix.

For comparison with pretrained supervised models on the NLI task, we reimplement the Chinese BERT- (Devlin et al., 2019; Cui et al., 2019, 2020) and RoBERTa-based (Liu et al., 2019; Cui et al., 2019, 2020) baselines from Zhou et al. (2024).

[3] https://platform.openai.com/docs/models/gpt-4o-mini.

3.3 Metrics

For the NLI task, we evaluate models in terms of standard **Accuracy** and **F1-Macro** score. For compound interpretation, we use both **Accuracy** (the number of times the model output exactly the majority relation for the target compound, divided by the total number of samples) and **R- Rank** (Camacho-Collados et al., 2018), defined as:

$$R - rank = \frac{1}{n}\sum_{i=1}^{n} rank_i \quad (1)$$

where n is the total number of samples, while $rank_i$ is the rank of the majority relation for the i-th compound sample. Since the correct relation may not always appear in the prediction list, we add **Hit Ratio** (Alsini et al., 2020) as a supplementary metric. Hit@k (or hit ratio @k) is the proportion of test cases in which the correct item appears within the top k positions in the model ranking.

4 Results

A general summary of the results can be seen in Table 3, including both the scores for the LLMs on the two datasets (3a) and the performance of the fine-tuned baselines on the NCNLI data (3b). On NCNLI, the best LLMs in a zero-shot setting are close to the fine-tuned RoBERTa baseline: only the more recent models of the Qwen family are able to consistently surpass it. Perhaps surprisingly, the smaller Qwen2.5-7B model is the one getting the highest accuracy and F1-score in this setting. A different trend becomes visible under the a few-shot setting: while smaller models seem to be inconsistent, bigger LLMs show clear gains from exposure to task examples. GPT-4o-mini, Qwen-plus and DeepSeek-chat all see noticeable boosts, and particularly GPT-4o-mini, which achieves the best score overall. Among the small models, Qwen2.5-7B keeps being competitive, but does not have any gain from few shots. In other words, only bigger models seem to be able to consistently perform in-context learning from the additional examples.

The interpretation task, as expected, is more challenging, given the high number of semantic relations to choose from and the subtle nature of the compound intepretation. A noticeable figure is the

	NCNLI, 0-shot		NLI, 3-shot		NEPD		
	Acc	F1	Acc	F1	Acc	R-rank	Hit@5
Alpaca 7B	71.35	64.54	52.14	42.00	2.01	8.29	18.54
Qwen 7B	64.15	51.44	64.74	56.05	6.74	11.08	2.03
DeepSeek-7B	55.98	45.73	63.84	50.57	9.91	7.97	38.78
Qwen2.5-7B	77.90	78.26	76.52	77.17	13.28	10.25	15.17
GPT-4o-mini	68.29	69.35	**80.41**	**80.24**	8.23	5.08	69.22
DeepSeek-chat	71.00	70.58	74.64	74.59	23.38	5.83	60.74
Qwen-plus	76.67	76.91	79.03	79.67	**36.13**	**4.42**	**73.34**

(a) Results with LLM prompting.

	Acc	F1
BERT	57.70	52.59
RoBERTa	72.94	71.49

(b) NCNLI results with baselines from Zhou et al. (2024), fine-tuned on 50k examples from the OCNLI dataset (Hu et al., 2020).

Table 3: Evaluation results. Only valid outputs counted for the metrics.[4] All scores are reported as the average of 3 runs and reported in %, except for R-rank. Best scores on each dataset are in **bold**.

value of the hit ratio@5: it is clear that small models have a hard time in this task, as they all fall short of ranking the correct semantic relation in the top 5 in most cases, and have accuracy scores mostly in single digits (notice that a random baseline with uniform probability distribution would get around 8.3% of correct answers). A remarkable improvement in the hit ratio can be seen with bigger models, but with an important distinction: while GPT-4o-mini manages to include the right answers at the top of the rank in most cases (over 69%), its accuracy and R-rank are not significantly better than those of the smaller models. This suggests that, while GPT-4o-mini does a better job in selecting plausible semantic relations for the intepretation of a compound, it still struggles in identifying the correct ones within a pool of plausible options.

On the other hand, the two Chinese competitors achieve better scores for those metrics. Qwen-plus is particularly impressive, managing to align with humans on the most plausible relation in about one third of the cases and to achieve a very low R-rank value, suggesting that the correct option is almost always close to the top of the rank. If we consider smaller and bigger models separately, it is interesting to notice that in both "categories" a model of the Qwen family emerges as the best performing one across the two tasks. As it can be seen from Table 4, bigger models tend to predict more frequent semantic relations more accurately, with weak-to-moderate positive correlations between accuracy and relation frequency; two relatively rare relations such as OR and FROM are more challenging for most LLMs (average accuracy < 3%, cf. Table 5).

[4]For the NLI task, Alpaca 7B rejected 11.41% of samples in both settings, Qwen 7B rejected 0.09% (zero-shot) and 0.02% (few-shot); others showed no rejections. For relation interpretation, rejection rates were 1.43% for Alpaca 7B, 0.48% for Qwen 7B, and 0.07% for Qwen-plus, and none for others.

Model	Correlation
Alpaca 7B	-0.30
Qwen 7B	0.12
DeepSeek-7B	-0.03
Qwen2.5-7B	0.20
GPT-4o-mini	0.19
DeepSeek-chat	0.39
Qwen-plus	0.41

Table 4: Spearman correlation index between relation frequency (as majority relation) and accuracy.

Relation	Average Performance		
	Acc	R-rank	Hit@5
CAUSE, MAKE	> 40%	< 2	> 60%
OR, FROM	< 3%	≈ 4	< 15%

Table 5: Hardest (top) and easiest (bottom) relation categories on average.

5 Conclusions

In our work, we evaluated LLMs on noun compounds semantics using two different tasks: a NLI task where the understanding of the inference depends on the correct interpretation of the compound, and a task focusing on identifying the specific semantic relation existing between modifier and head, for which we collected a new dataset.

As for NLI, we found LLMs to be already improving over the performance of pretrained models, with larger LLMs taking the most advantage from task examples in the few-shot setting. Selecting the correct relation in the interpretation task is more challenging, given the ambiguity of noun compounds and the greater number of classes to choose from. As in the NLI task, we observed the most consistent performance from Qwen models, with Qwen-plus being the most aligned with human intuitions of compound semantics.

Limitations

An important limitation of the study lies in our dataset for compound interpretation, since we aimed at recruiting annotators with a high level of expertise (PhD students in linguistics) and, as a consequence, we had a relatively low number of annotators (3) for each dataset instance. Although the annotations were quality checked by one of the authors, who has expert level knowledge of the subject, our choice might have favored some idiosyncratic interpretation of the compounds.

As a term of comparison for Western LLMs we used GPT-4o-mini, which proved to be a cost-efficient and performance-effective option. However, we did not have the time to test more recently-released models, such as GPT-5.

Acknowledgements

EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222).

References

Areej Alsini, Du Q Huynh, and Amitava Datta. 2020. Hit Ratio: An Evaluation Metric for Hashtag Recommendation. *arXiv preprint arXiv:2010.01258*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek LLM: Scaling Open-source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of EMNLP*.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid O Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Jordan Coil and Vered Shwartz. 2023. From Chocolate Bunny to Chocolate Crocodile: Do Language Models Understand Noun Compounds? In *Findings of ACL*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*.

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*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese Llama and Alpaca. *arXiv preprint arXiv:2304.08177*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Min Gu, Yanhui Gu, Fang Xu, Bin Li, Bin Zhao, and Weiguang Qu. 2016. Research on Chinese Noun+Noun Compounds Semantic Classification and Automatic Interpretation. *ICIC Express Letters. Part B, Applications: An International Journal of Research and Surveys*, 7(1):173–179.

Iris Hendrickx, Preslav Nakov, Stan Szpakowicz, Zornitsa Kozareva, Diarmuid O Séaghdha, and Tony Veale. 2013. SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. *Proceedings of SemEval*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of EMNLP*.

Shuiqing Huang and Dongbo Wang. 2019. Construction, Performance and Application of New Era People's Daily Segmented Corpus (I) – Construction and Evaluation Corpus. *Library and Information Service*, 63(22):5–12.

Gary Libben. 2014. The Nature of Compounds: A Psychocentric Perspective. *Cognitive Neuropsychology*, 31(1-2):8–25.

Jingping Liu, Juntao Liu, Lihan Chen, Jiaqing Liang, Yanghua Xiao, Huimin Xu, Fubao Zhang, Zongyu Wang, and Rui Xie. 2022. Noun Compound Interpretation with Relation Classification and Paraphrasing. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8757–8769.

Pengyuan Liu and Yujie Liu. 2019. Semantic Relations Hierarchy and Knowledge Base Construction of Chinese Basic Noun Compounds. *Journal of Chinese Information Processing*, 33(4):20–28.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Preslav Nakov. 2008a. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 103–117. Springer.

Preslav Nakov. 2008b. Paraphrasing Verbs for Noun Compound Interpretation. In *Proceedings of the LREC Workshop on Multiword Expressions*.

Giulia Rambelli, Emmanuele Chersoni, Claudia Collacciani, and Marianna Bolognesi. 2024. Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds. In *Proceedings of ACL*.

Vered Shwartz. 2019. A Systematic Comparison of English Noun Compound Representations. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Vered Shwartz and Chris Waterson. 2018. Olive Oil Is Made of Olives, Baby Oil Is Made for Babies: Interpreting Noun Compounds Using Paraphrases in a Neural Model. In *Proceedings of NAACL*.

Henk J Van Jaarsveld and Gilbert E Rattink. 1988. Frequency Effects in the Processing of Lexicalized and Novel Nominal Compounds. *Journal of Psycholinguistic Research*, 17:447–473.

Meng Wang, CR Huang, Shiwen Yu, Bin Li, and 1 others. 2010. Chinese Noun Compound Interpretation based on Paraphrasing Verbs. *(Journal of Chinese Information Processing)*, 24(6):3–9.

Meng Wang, Lulu Wang, Na Tian, and Bin Li. 2016. Automatic Interpretation of Chinese Noun Compounds Based on Word Similarity. *ICIC Express Letters, Part B: Applications*, 7(6):1215–1221.

An Yang, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and 1 others. 2024. Qwen2.5 Technical Teport. *arXiv preprint arXiv:2412.15115*.

He Zhou, Yu Yin Hsu, and Emmanuele Chersoni. 2024. Evaluating Chinese Noun Compound Interpretation in Natural Language Inference. In *Proceedings of the Chinese Lexical Semantics Workshop (CLSW 2024)*.

Appendix

Prompts

5.0.1 NCNLI

For the NLI task, the prompt follows a fixed template for both settings:

> 请将下面前提（P）和假设（H）之间存在的逻辑推理关系分为以下类别之一：蕴含,矛盾或中立。只需回答类别。
> P: xxx H: xxx
>
> Please identify the semantic relation between the premise (P) and the hypothesis (H) and respond with one of the following semantic relations: Entailment, Contradiction or Neutral. Return their relation only.
> P: xxx H: xxx

In the few-shot setting, we concatenate a representative example for each category to the template above, while ensuring that none of them overlaps with the evaluation set.

> 前提: 运动员有一个不锈钢饭盒。
> 假设: 不锈钢是饭盒的制作材料。
> 输出: 蕴含
>
> **P**: The athlete has a stainless steel lunch box.
> **H**: Stainless steel is the material that the lunch box is made of.
> **Output**: Entailment
>
> 前提: 清洁工昨天吃了巧克力蛋糕。
> 假设: 清洁工吃的蛋糕里没有巧克力。
> 输出: 矛盾
>
> **P**: The janitor ate chocolate cake yesterday.
> **H**: There was no chocolate in the cake that the janitor ate.
> **Output**: Contradiction
>
> 前提: 科学家最喜欢的是椒盐牛蛙。
> 假设: 科学家只吃过椒盐口味的牛蛙。
> 输出: 中立
>
> **P**: The scientist's favorite is salt-and-pepper bullfrogs.
> **H**: The scientist has only eaten bullfrogs with a salt-and-pepper flavor.
> **Output**: Neutral

5.0.2 NEPD

For the NEPD task, the prompt is derived from a concise summarization of the guidelines and instructions we provided to human annotators.

> 给定一个中文复合词语，该词语由两个名词复合构成，请对其名词成分之间的语义关系进行分类，共11个预定义类别，分别为：CAUSE（表示因果关系）、MAKE（表示组成）、HAVE（表示拥有、具备）、USE（表示

使用、利用工具或手段）、BE（表示说明和补充）、IN（表示空间上的包含关系）、FOR（表示目的、用途）、FROM（表示来源）、ABOUT（表示主题或相关内容）、AND（表示并列、组合关系）、OR（表示选择或替代关系）。请仅返回最可能的类别名称，并按可能性从高到低排序。若该词语不属于上述任何类别，请返回OTHER。
复合词: xx

Given a Chinese compound word formed by two nouns, classify the semantic relationship between its noun components into one of 11 predefined categories: CAUSE (indicating causal relation), MAKE (indicating composition), HAVE (indicating possession or having), USE (indicating usage or utilization of tools or means), BE (indicating explanation or description), IN (indicating spatial inclusion), FOR (indicating purpose or function), FROM (indicating source, origin, or starting point), ABOUT (indicating topic or related content), AND (indicating coordination or combination), OR (indicating choice or alternative). Please return only the most likely category names, ordered from highest to lowest likelihood. If the compound word does not belong to any of the above categories, return OTHER.
Compound: xx

Dataset Statistics

	Count
Fully agreed	458
Partially agreed	1,056
Tied	569
Total	2,083

Table 6: Annotation agreement statistics.

We categorized the annotation results into three groups based on the level of inter-annotator agreement among three annotators for each compound. 'Fully agreed' indicates consensus among all three annotators; 'partially agreed' indicates consensus between two annotators with the third disagreeing; and 'tied' indicates three distinct judgement with no agreement between any pair. In addition, we calculated that each annotator provided an average of 2.10 relations, indicating that the annotation task involves a considerable level of difficulty.