Disambiguate Words like Composing Them: A Morphology-Informed Approach to Enhance Chinese Word Sense Disambiguation

Yue Wang[1,2], Qiliang Liang[1,3], Yaqi Yin[1,2], Hansi Wang[1,2], Yang Liu[1,2]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University
[3]School of Electronics Engineering and Computer Science, Peking University
wyy209@pku.edu.cn, lql.pkucs@gmail.com, yyqi@stu.pku.edu.cn,
wanghansi2019@pku.edu.cn, liuyang@pku.edu.cn

Abstract

In parataxis languages like Chinese, word meanings are highly correlated with morphological knowledge, which can help to disambiguate word senses. However, in-depth exploration of morphological knowledge in previous word sense disambiguation (WSD) methods is still lacking due to the absence of publicly available resources. In this paper, we are motivated to enhance Chinese WSD with full morphological knowledge, including both word-formations and morphemes. We first construct the largest releasable Chinese WSD resources, including the lexico-semantic inventories **MorInv** and **WrdInv**, a Chinese WSD dataset **MiCLS**, and an out-of-volcabulary (OOV) test set. Then, we propose a model, **MorBERT**, to fully leverage this morphology-informed knowledge for Chinese WSD and achieve a SOTA F1 of 92.18% on MiCLS dataset. Finally, we demonstrated the model's robustness in low-resource settings and generalizability to OOV senses. These resources and methods may bring new insights into and solutions for various downstream tasks in both computational and humanistic fields[1].

1 Introduction

Word sense disambiguation (WSD) aims to identify the sense of a polysemous word in a specific context. It has become critical for accurate language understanding and has proven effective in various downstream tasks, including Information Extraction (Barba et al., 2021b), Test Summarization (Kouris et al., 2021) and Machine Translation (Pu et al., 2018; Campolungo et al., 2022), etc. Previously, mainstream methods typically regarded the target word as an indecomposable lexico-semantic unit, utilizing various interword and word definition information (Blevins and Zettlemoyer, 2020; Barba et al., 2021a,b), to improve the model's performance. These methods achieved remarkable results, surpassing the F1 score of 80% on standardized English datasets (Raganato et al., 2017), which is the estimated human performance (Navigli, 2009). Recently, large language models (LLMs) have demonstrated remarkable performance across various tasks, but they still struggle to make improvements on WSD. Even ChatGPT, one of the top-performing models, only achieves an F1 score of 73.30% (Kocoń et al., 2023). However, when applied to Chinese, these top-performing models (without morphological knowledge) fail to reach an F1 value of 80% (Yan et al., 2023). This decrease in accuracy may be attributed to the characteristics of the Chinese language.

In a parataxis language like Chinese, word meanings are highly correlated with morphological knowledge, which includes word-formations and morphemes as two aspects (Cao, 2001). Word-formations designate how characters (as morphemes) are combined to compose lexical semantics by specific patterns of morphemes. While word-formation has been proven to be effective in Chinese WSD (Zheng et al., 2021b), the role of morphemes may be equally significant for the task.

In Chinese, morphemes are defined as the smallest semantic and sound-bearing unit (Zhu, 1982) and play a fundamental role in the process of word formation with their characteristics (Yin, 1984; Xu, 1990). Despite the vast and dynamic set of Chinese would-be words, morphemes remain relatively stable in number and meaning (Yuan and Huang, 1998). An overwhelming majority of them maintain their meanings in the process of word formation (Yuan and Huang, 1998), and as a result, the composition of morpheme senses can largely represent the meaning of words (Fu, 1981). Combining morphemes under word-formations can yield even greater efficacy. As shown in Figure 1, the polysemous word "满月" holds two senses: "满

[1]Data and code for this paper are available at https://github.com/COOLPKU/MorBERT.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15354–15365

月$_1$" and "满月$_2$", which are composed of different morphemes ('满$_3$月$_2$" vs. "满$_4$月$_1$') under different word-formation rules (*Verb-Object* vs. *Modifier-Head*). Such semantic compositionality of morphemes, with intuitive clarity, can therefore enhance WSD and improve its performance. What's more, this motivated strategy may bring about further advantages that largely benefit out-of-vocabulary (OOV) issues previously entangled with WSD.

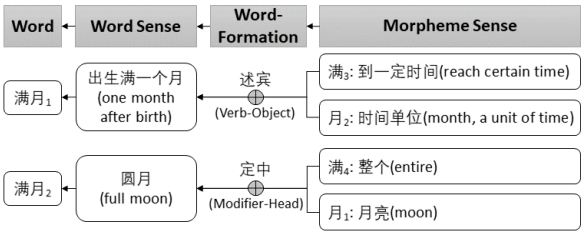

Figure 1: The word formation process of the polysemous word "满月", with word-formation and morpheme senses specified, can denote the word's meaning.

In recent years, researchers have explored the application of morphological knowledge. Zheng et al. (2021b) leveraged word-formation in Chinese WSD and achieved a high F1 of 87.62% on FiCLS. This is largely attributed to the introduction of word-formation knowledge. While the importance of word-formations in Chinese WSD has been explored and yielded notable improvements, the introduction and application of morphemes are still neglected due to the absence of publicly available resources and annotations. With morphology-informed resources and models, the potential of full morphological knowledge can be stimulated, which may further enhance Chinese WSD and improve its performance.

This specific strategy may bring about further advantages, which can largely benefit OOV sense representations and disambiguation. In real Chinese corpora, the OOV issues are quite common, and even more popular than in other languages[2]. For example, the word "千万(*ten million/be sure to*)" may mean "*ten million*" in texts, representing a direct combination of the morphemes "千$_1$ (*thousand*)" and "万$_1$ (*ten thousand*)", but it is unambiguous in the Contemporary Chinese Dictionary (CCD)[3] with only one sense "*be sure to*". These "unambiguous" words also need disambiguation and can therefore enhance WSD by extending the coverage and accuracy of word sense representation (Loureiro and Camacho-Collados, 2020). Currently, existing Chinese WSD datasets cannot cover or handle such cases. This urges for a new kind of WSD resources that may underline and address these problems.

Facing these scenarios of Chinese and inspired by the process of word formation, in this paper, following the previous works (Zheng et al., 2021b), we are motivated to further leverage morpheme features and fully explore the potential of morphological knowledge to enhance WSD. We first construct two lexico-semantic inventories **MorInv** and **WrdInv**, a **M**orphology-**i**nformed **C**hinese **L**exical **S**ample dataset(**MiCLS**), and an OOV test set. Then, we propose **MorBERT** to explicitly incorporate Chinese morphological knowledge into a BERT-based WSD model. Experimental results show that our method brings consistent and substantial performance improvements to Chinese WSD with high accuracy in morpheme prediction. Analysis shows that our method can be greatly generalized to low-resource settings and OOV senses. These resources and methods may bring new insights into and solutions for various downstream tasks in both computational and humanistic fields.

In summary, this paper is committed to fully leveraging morphological knowledge to enhance Chinese WSD. The main contributions are as follows:

(1) We provide large-scale releasable resources for Chinese WSD for the first time, including two lexico-semantic inventories (MorInv and WrdInv), a Chinese WSD dataset (MiCLS), and an OOV sense test set. These constitute the largest releasable Chinese WSD resources so far as we know.

(2) We incorporate morpheme features on the foundation of formBERT (Zheng et al., 2021b) and propose MorBERT to fully leverage Chinese morphological knowledge and achieve a SOTA F1 of 92.18% on MiCLS dataset.

(3) This morphology-informed strategy can greatly meet the demands of low-resource scenarios and OOV sense challenges, such as in Chinese, demonstrating its robustness and generalizability to the WSD task.

2 Related Work

WSD Methods: Recent supervised neural WSD methods achieved remarkable performance by

[2]These OOV senses may also adopt the same word types as unambiguous in dictionaries, making them ambiguous in real corpora.

[3]The most authoritative and influential Chinese dictionaries, published by the Commercial Press.

leveraging lexical knowledge-bases, incorporating definitional (Luo et al., 2018; Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021a), relational (Bevilacqua and Navigli, 2020; Barba et al., 2021b; Zhang et al., 2022), and conceptual (Raviv and Markovitch, 2021) knowledge. Su et al. (2022) further improved performance on rare and zero-shot senses by Z-reweighting. These methods exhibit lower accuracy in Chinese owing to the neglect of Chinese morphology, which plays a pivotal role in its lexico-semantics. Zheng et al. (2021b) leveraged word-formations for Chinese WSD and achieved further improvement, but ignored the incorporation of morphemes due to the absence of publicly available resources and annotations.

Chinese WSD Resources: Compared to English, resources for Chinese WSD are still relatively lacking. SemEval-2007 task 5 (Jin et al., 2007) and Chinese Word Sense Tagging Corpus (STC) (Wu and Yu, 2006) used the corpora of People's Daily[4] (PD, in months of 1998 & 2000) and annotated senses for 40 and 966 polysemous words, respectively. They are currently the most commonly used Chinese WSD datasets, but are limited to low coverage, and the corpora lack timeliness. Word Sense Annotated Corpus for Teaching Chinese as Second Language (Wang et al., 2007) and FiCLS dataset (Zheng et al., 2021b) annotated senses for 1181 polysemous words in Chinese textbooks and 7064 polysemous words in Chinese Wikipedia, respectively, with large scale and corpora exhibiting strong timeliness. However, both datasets are annotated based on CCD and are therefore not publicly available due to copyright issues. Modern Chinese Word Sense Annotated dataset (Yan et al., 2023) annotated senses for 1083 polysemous words in 19082 sentences collected from different sources. It is now publicly available, but the size is still relatively small and without providing morphological knowledge. Due to restrictions on intellectual property, Chinese WSD still faces the challenges of obtaining large-scale, high-quality datasets. It is necessary to construct resources for public use (and adaption to computing applications) through methods such as paraphrasing to facilitate sharing with both the academic and industrial communities.

Chinese Morphology-Related Resources: Existing Chinese morphology-related resources focus on morpheme definition and morpheme-based word-formation features. For morpheme inventory, Yuan and Huang (1998) introduced a morpheme knowledge-base by manually describing 17,470 morphemes for 6,763 characters used for word-formation analysis. Kang et al. (2004) further connected such morphemes to Cilin (Mei et al., 1983), an influential Chinese thesaurus. Lin and Liu (2019) described morphemes with part-of-speech (PoS) and inter-morpheme relations, covering a comprehensive set of 20,855 morphemes for 8,515 characters originating from CCD. Based on it, for morphology-informed datasets, Zheng et al. (2021a) manually annotated word-formations and morphemes for 45,311 words for definition generation (DG). These resources can serve as the basis for the construction of our resources for public use.

3 Resources

Current Chinese WSD resources face challenges in achieving high coverage and restrictions on public use due to intellectual property.In light of these, we aim to construct large-scale, releasable resources for WSD, which can also benefit handling the OOV problems. In this section, we first introduce morpheme inventory (MorInv) and word inventory (WorInv) as the lexico-semantical basis. Then, we construct the MiCLS dataset for WSD. Additionally, we provide a small OOV test set for testing chances of generalizability to OOV senses.

3.1 Chinese Lexico-Semantic Inventories

For the lexico-semantic inventories, we provide both **MorInv** and **WrdInv**, which include morphemes and words with their parts-of-speech (PoS) and sense definitions designed for computing applications, respectively.

Following the previous works (Lin and Liu, 2019; Zheng et al., 2021a), we extracted Chinese morphemes and disyllabic words with their PoS and senses from CCD as the basis to construct releasable resources. It is also noted that dictionaries primarily serve humans, and the sense definitions can be too complex and potentially distractive for computing applications. To tackle these issues, we adopted ChatGPT[5] to further paraphrase (and simplify) the senses. Details of the prompt for paraphrasing are shown in Appendix A.

[4] http://paper.people.com.cn/, China's official newspaper.

[5] All "ChatGPT" in this paper refers to GPT-3.5-turbo-0613.

To ensure data quality, three mother-tongue reviewers manually checked the paraphrased senses and revised the inappropriate ones. We also identified tens of semantic categories for representative patterns (eg. place names, chemical compounds) and designed their corresponding sense templates according to one of the best paraphrased results. This simplified and systematized the sense definitions in MorInv and WrdInv, which is beneficial for computing applications.

The final MorInv contains 20,856 morphemes for 8,516 characters, and WrdInv contains 52,115 senses for 41,479 words, of which 8684 words are polysemous (with 19320 senses covered).

3.2 Chinese WSD Datasets

3.2.1 The MiCLS Dataset

Based on the above inventories, we introduce a **M**orphology-**i**nformed **C**hinese **L**exical **S**ample WSD dataset (**MiCLS**), which includes annotations of word sense and morphological knowledge in context sentences. Each MiCLS entry consists of: (1) a target word, (2) sense definition for the word, (3) word-formation of the word, (4) sense definition for morphemes within the word, (5) a context sentence.

For the context sentences, we initialed by filtering the previous FiCLS dataset (Zheng et al., 2021b) and kept the portion targeting disyllabic words (and replaced the sense definitions with the paraphrased ones in WrdInv), with 22146 entries gathered, covering 7301 senses for 3929 polysemous words. However, it has a low coverage of only 45.24% polysemous words and no monosemous words (both of which may potentially have OOV senses in corpora) in WrdInv.

To ensure high coverage, we further expanded the dataset by crawling context sentences of all words in WrdInv from the corpora of PD (from year 2018 to 2020), and Zaojv Net[6]. In each corpus, we extracted context sentences 10 times the number of senses for each word in WrdInv. Subsequently, we used three large language models (LLMs): ChatGPT, Qwen (Bai et al., 2023), and ChatGLM (Du et al., 2022)[7], to annotate word senses. The models are presented with a target word, a context sentence, and a sense definition, and then asked to determine whether the meaning of the target word in the context matches this sense. Considering LLMs' sensitivity to prompts, each sample was subjected to three carefully designed prompts. Therefore, each sample went through 9 voting processes by LLMs. Details of the prompts for annotation are shown in Appendix B.

We collected the voting results and retained samples that met both conditions: 1) receiving at least 7 votes of "yes" and 2) for the sentence, the sense gets the highest number of "yes". Considering a balance among senses, for each sense, we retained a maximum of 5 contexts. We prioritized keeping the ones with the highest votes. To ensure data quality, each sample getting less than 8 votes was manually checked by two annotators and would be retained only if both of them voted "yes". The inter-annotator agreement of manual annotation reached 94.30%. Note that the inner-annotator agreement is relatively high due to the preliminary annotation of the LLMs, with over 90% positive samples. This new procedure combining LLM annotation and manual checking improves the efficiency of this activity while ensuring its quality, thus facilitating the construction of large-scale datasets.

After annotating word senses in the context sentences, we then obtained word-formation and morpheme sense annotations from the DG dataset proposed by (Zheng et al., 2021a) and replaced the original morpheme senses defined in CCD with the paraphrased ones in MorInv.

The final dataset contains 174,433 entries, covering 42,879 senses for 36,057 words, of which 8126 are polysemous (with 14948 senses covered), totaling 93.57% of all polysemous words. It is the largest releasable Chinese lexical sample WSD dataset so far as we know. The coverage of data from each source is shown in Table 1. Table 2 shows examples of entries with the target word "满月" in MiCLS.

Source	Entries	Words	Senses	Poly. Words	Poly. Senses
FiCLS	22,146	3,929	7,301	3,929	7,301
People's Daily	109,030	28,052	30,452	6,313	8,713
Zaojv Net	43,257	18,435	19,620	4,623	5,808
All	174,433	36,057	42,879	8,126	14,948
Coverage	-	86.93%	81.53%	93.57%	77.37%

Table 1: The coverage of data from different sources. Poly. Words represent the number of polysemous words. Poly. Senses represent the number of senses of polysemous words covered in MiCLS.

[6] https://zaojv.com/, an online sentence-making dictionary.

[7] We used Qwen-max for Qwen and ChatGLM-turbo for ChatGLM.

Word	Word sense	Formation	Morpheme1	Morpheme2	Context
满月	出生满一个月 one month after birth	述宾 Verb-Object	到一定时间 reach a certain time	时间单位 month	我女儿不到～就十来斤了 My daughter was over 5 kg before ～.
	圆月 full moon	定中 Modifier-Head	整个 entire	月亮 moon	玉盘似的～在云中穿行 The jade-like ～moves through the clouds.

Table 2: Examples of entries with the target word "满月" in MiCLS.

Word	Context	Real sense	Senses in MorInv
千万	营收超～的设计企业达8家 Eight design companies have a revenue exceeding ～	一千个一万 ten million	1.务必 1.be sure to
作对	有人以月为题吟诗～ Some people recite poems and ～with the topic of moon	创作对联 write couplets	1.与人为难；2.成为配偶 1.embarrass someone; 2. become spouses

Table 3: Examples of entries in the OOV test set. The real sense in context and senses in MorInv of the target word are additionally shown to facilitate understanding.

3.2.2 The OOV Test Set

Noticing previous datasets' neglect of OOV senses and their necessity for disambiguation, to address this issue, we further introduce an OOV test set for testing the generalizability of Chinese WSD models. Each entry consists of only: (1) a target word, (2) a context sentence. Note that the real sense of the target word in the context sentences is not included in MorInv as previous.

To generate this test set, We randomly selected 1000 contexts rejected by the LLMs in the above procedure, which received no more than 3 votes for either sense in MorInv. Through manual checking, we collected sentences containing OOV senses and constructed a small-scale test set.

The final test set contains 173 entries, covering 148 OOV senses for 147 words (existed in WrdInv), where the same word type may possess different OOV senses. Table 3 shows examples of entries with words "千万" and "作对" in OOV test set.The real sense and MovInv senses of the target word are additionally shown to facilitate understanding.

4 Methodology

4.1 Task Formulation

As stated in Section 1, Chinese words in dictionaries may potentially have OOV senses, which many traditional MSD models cannot handle. In light of this, we formulate WSD as a sentence-level binary classification task, which has been proven to effectively leverage sense definitions in BERT-based WSD methods (Huang et al., 2019). Specifically, given a target word w, its context sentence c, and its component morphemes m_1, m_2, we construct an instance quintuple (w, c, m_1, m_2, d) using a sense definition d of the target word. In this way, a positive triplet contains the correct sense definition with its label $y^* = 1$, while a negative triplet contains the wrong one with $y^* = 0$. We flatten the context, morpheme definitions, and word definition into a character sequence with the BERT-specific prediction token [CLS] and the sentence boundary indicator [SEP][8]. A classifier f is responsible for mapping the prediction token representation $\mathbf{h}$ to the label distribution, and the label of the triplet is predicted as:

$$p(y|w, c, m_1, m_2, d) = f(\mathbf{h}),$$
$$\hat{y} = arg\max_{y} p(y|w, c, m_1, m_2, d).$$

Our goal is to minimize the negative log-likelihood of the ground-truth label y^*:

$$\mathcal{L}_{wsd} = -\log p(y^*|w, c, m_1, m_2, d).$$

4.2 MorBERT: Leveraging Morphological Knowledge

Following FormBERT (Zheng et al., 2021b), we propose MorBERT to leverage full morphological knowledge by further incorporating morpheme features seamlessly into the BERT-based model and formation prediction. The overall architecture of MorBERT is shown in Figure 2. Specifically, given the target word $w = (ch_1, ch_2)$, its component morphemes m_1, m_2, and its word-formation annotation fm^* for the ground-truth definition d in the context c, we learn morpheme embeddings $\mathbf{m_1}$, $\mathbf{m_2}$ using BERT, and a formation embedding $\mathbf{fm}^*$ via a matrix $\mathbf{W_{fm}}$ for each formation type. The

[8]We add weak supervisions in the context and the definition to hint the target word following Huang et al. (2019)

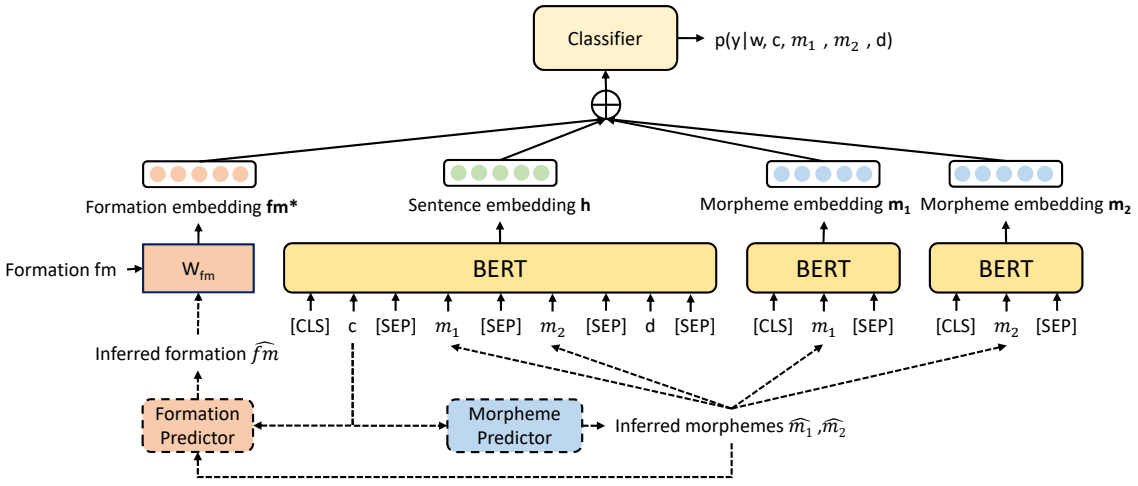

Figure 2: Illustration of the proposed MorBERT with Morpheme Perdictor (MP) and revised Formation Predictor (FP). The dashed line indicates that, during inference, the inferred formation and morphemes based on the context will be exploited to generalize to scenarios without these annotations.

obtained embeddings $\mathbf{m_1}$, $\mathbf{m_2}$, and $\mathbf{fm}^*$ are then combined with $\mathbf{h}$ to produce the label probability distribution:

$$p(y|w, c, m_1, m_2, d, fm^*) = f(\mathbf{h} + \mathbf{m_1} + \mathbf{m_2} + \mathbf{fm}^*).$$

By incorporating the full morphological knowledge of morphemes and word-formations, MorBERT is better informed of how the characters (as morphemes) interact and what they mean in the target word to better distinguish senses. However, morpheme annotations are expensive to acquire and can be unavailable in other resources. Thus, we introduce an auxiliary Morpheme Predictor (MP) to automatically annotate morpheme senses for each character in the target word.

Considering the similarity between WSD and morpheme prediction, we adopt BEM[9] (Blevins and Zettlemoyer, 2020), one of the top-performing WSD models that formulate WSD as a multiclass classification task, as the model for MP. For each character $ch_i (i = 1, 2)$ in the target word:

$$\hat{m_i} = f(ch_i, c).$$

where $f(\cdot)$ is a BEM model pre-trained on MiCLS. To avoid test data leakage, the dataset is split in the same way as WSD. To adapt the morpheme prediction task, we split each MiCLS entry into two entries, each targeting one character in the target word.

MorBERT also inherited a Formation Predictor (FP) from FormBERT to predict word-formations for unannotated words. Morpheme features have been proven effective in the subtask of word-formation prediction (Zheng et al., 2021c), and we leveraged them in the revised FP:

$$p(fm|w, c, m_1, m_2) = g(w, c, m_1, m_2),$$
$$\hat{fm} = arg\max_{fm} p(fm|w, c, m_1, m_2).$$

where $g(\cdot)$ is a MLP formation predictor. During training, where the word-formations are available, a formation prediction objective is added for training the predictor:

$$\mathcal{L}_{fp} = -\log p(fm^*|w, c, m_1, m_2)$$

This objective is combined with the original sense disambiguation loss with a weighting factor λ.

With well-trained MP and revised FP, this framework can generalize to data without morpheme annotations and word-formation annotations.

5 Experiments

5.1 Experimental Settings

Datasets: We split the MiCLS dataset described in Section 3.2.1 into training, validation, and test sets by 8:1:1. Note that MiCLS only annotated the ground-truth sense, so for an entry with the target word having n senses, we split it into n entries, each corresponding to one sense. We label 1 for

[9]Other top-performing WSD models, such as EWISER (Bevilacqua and Navigli, 2020) and ConSeC (Barba et al., 2021b), require special word features that are unavailable for morphemes.

the entry corresponding to the ground-truth sense and label 0 for the other entries.

Baselines: Besides random and most frequent sense (MFS) as the default baseline, we also implement strong baselines with features available in MiCLS, including BEM (Blevins and Zettlemoyer, 2020) and FormBERT (Zheng et al., 2021b), and use the same settings as our model for a fair comparison. Note that BEM formulates WSD as a multi-class classification task, and the model must choose one of the senses. To deal with OOV senses, we added the option "NULL" for each entry, demonstrating that the sense of the word in the context is different from any sense in the inventory. When calculating accuracy, we assign weights based on the sense number of the target word.

Experimental configuration We adopt BERT-wwm-ext (Cui et al., 2021) as the base model. The MP is a BEM model where each encoder is initialized with BERT-base-Chinese (Cui et al., 2020). The revised FP is a 2-layer feedforward network with a hidden size of 768 and ReLU as the activation function. The formation prediction objective weight λ is 0.25. For the baselines, we directly follow the settings in their original papers. Other detailed configurations are shown in Appendix C.

5.2 Results

Model	All	N.	V.	Adj.	Adv.	Func.
Random	57.42	58.32	57.54	55.14	49.17	53.69
MFS	77.60	80.62	76.50	71.73	65.99	68.84
BEM	87.41	88.44	87.88	85.05	76.33	79.71
FormBERT	91.90	93.04	92.08	88.21	83.45	86.02
[MP]	94.36	95.37	94.08	92.84	87.22	91.36
MorBERT	92.18	**93.19**	92.47	88.77	83.83	**86.07**
MorBERT w/MP	**92.19**	93.18	**92.52**	**88.85**	**84.00**	85.89

Table 4: Evaluation results (F1) on MiCLS. [MP] is the precision of the Morpheme Predictor (MP). MorBERT w/MP denotes MorBERT using the MP without annotated morphemes in the valid and test sets.

Table 4 shows the overall F1 results on MiCLS across five PoS categories: noun, verb, adjective, adverb, and functional (including conjunction, preposition, pronoun, etc.) words. Note that PoS is not included during training. We only use it for a parallel comparison with previous works. From the table, we have the following observations:

(1) By leveraging full morphological knowledge, our MorBERT achieves a SOTA F1 of 92.18% and surpasses all baseline models across nearly all PoS. Reaching an improvement at above 90% accuracy is quite challenging, as it has approached the inter-annotator agreement of manual annotation, which is 94.30%[10]. This validates that morphology-informed knowledge can enhance Chinese WSD consistently and comprehensively and is approaching the level of human experts.

(2) Although MorBERT w/MP has no ground-truth information in the valid and test sets, it still reaches the accuracy of MorBERT and even surpasses it in some PoS, which confirms the generalizability of our method. We speculate that the slight advantage over MorBERT can be attributed to the significantly high 94.34% accuracy of the morpheme prediction.

(3) Concerning the performance on different PoS, most models perform the worst on Chinese adverbs. This can be explained by the higher ambiguity of adverbs, which is evident as MFS on adverbs is also the lowest. The high granularity of adverbs in CCD, which was adopted by WordInv, also contributes to this low F1. As shown in Table 5, some of their senses can be similar and need extra context for disambiguation.

Word	Word sense	Context
没有$_6$	“已然”的否定 haven’t	他没有回家。他妈妈很着急。 He hasn’t gone home. His mother is worried.
没有$_7$	“曾经”的否定 didn’t	他没有回家。他去了学校。 He didn’t go home. He went to school.

Table 5: Examples where adverbs hold similar senses and need extra context for disambiguation.

(4) It can also be observed that the F1 of MorBERT with MP is slightly below FormBERT on functional words, which might be attributed to the relatively low accuracy of MP on functional words. Further observation reveals that up to 71.42% entries with wrong morpheme prediction have both morphemes predicted incorrectly. These entries with both morphemes wrongly predicted are more likely to be wrongly predicted in WSD too.

6 Analysis

6.1 Analysis on the High F1 Score

The F1 score of 92.18 is significantly higher than that of top-performing WSD models on other

[10]The agreement is relatively high due to the preliminary annotation of the LLMs, with most samples positive, as stated in Section 3.2.1

datasets. We believe that the relatively high F1 of the models on MiCLS dataset value may be related to the following factors:

(1) The introduction of monosemy words: We first propose that monosemous words in dictionaries also need disambiguation due to the potential existence of OOV sense and add these words to the dataset. Most of these words in context hold the sense in dictionaries, which may result in a higher F1 for disambiguation.

(2) Better sense definitions for computational use: Previous Chinese WSD datasets primarily used senses in dictionaries for the sense inventory. As dictionaries primarily serve humans, the sense definitions can be too complex and potentially distractive for computing applications. Our paraphrased senses remove these unnecessary details for better suiting computation use.

(3) Larger scale of dataset: MiCLS is the largest Chinese WSD dataset so far as we know and a larger scale of dataset always leads to higher accuracy. We further trained MorBERT and the baseline models on a subset of FiCLS, which contains 22146 entries and has a similar data structure as MiCLS. Results are shown in the table 6:

Model	F1
BEM	74.51
FormBERT	78.45
MorBERT-w/MP	79.01

Table 6: Results on a subset of FiCLS dataset.

Note that FiCLS do not contain morpheme annotations, so we use MorBERT w/MP for MorBERT. MorBERT still exhibits the best performance among the three models. However, the overall F1 decreases due to the size of the dataset.

6.2 Analysis in Low-resource Settings

To better understand the overall results, we divide the test set into two subsets: (1) entries with the most frequent definition (MFD) of the target word; and (2) entries with less frequent definitions (LFD) of the target word. We compare MorBERT with and without MP against BEM and FormBERT. As shown in Table 7, the models show similar accuracy on MFD, with BEM as the highest. This is not surprising, as BEM tends to predict the most frequent senses.

However, in LFD, both MorBERT with and without MP introduce improvements over the baselines,

Model	MFD	LFD	OOV
BEM	**93.72**	65.52	0.00
FormBERT	92.33	82.44	64.74
MorBERT	92.56	83.11	-
MorBERT w/MP	92.54	**83.23**	**71.67**

Table 7: Evaluation results on the MFD, LFD subsets of the test set, and the OOV test set. Note that we use precision for OOV since there are no positive samples.

which validates that our method is effective and robust in low-resource settings. Especially BEM, MorBERT surpasses it by 17.71 points. We speculate that this is because BEM treats words as indivisible atomic units, making it more susceptible to the influence of the most common sense when representing words. By leveraging full morphological knowledge, MorBERT can get a deeper understanding of words at a smaller unit level. Therefore, for LFD, the model can learn additional information from other words containing the same morphemes. For example, the word sense "地道$_4$(*underground-road; tunnel*)" appears only once in the training set, but the model can infer the morpheme "道$_1$(*road*)" from other words like "暗道(*hidden-road*)" with similar context and thereby understand the sense of "地道$_4$(*underground-road; tunnel*)".

6.3 Analysis of Generalizability on OOV Senses

To test the generalizability in OOV senses, we further tested the models on the OOV test set. Since the test set does not contain any positive samples, we use precision for evaluation, and a prediction is considered correct only when all senses of the target word are predicted as 0. As shown in Table 7, MorBERT with MP achieves an accuracy of 71.67%, surpassing FormBERT by 6.93 points. This can be due to the fact that OOV senses may have the same word-formation as one of the senses in WrdInv. Take the word '爆炒(stir-frying/intensely hyping up)' in the sentence '北京顾客喜欢爆炒(Beijing customers like stir-frying)' as an example, the real OOV sense 'stir-frying' and the sense in WrdInv 'intensely hyping up' are both composed of the word-formation 'Adverb-Verb'. In cases like this, formation prediction in FormBERT may not provide any extra information. Therefore, FormBERT wrongly predicts the sense in WrdInv 'intensely hyping up' as the sense of '爆炒' in the context. However, the morpheme predictor in MorBERT can predict the right morphemes '爆(fierce)'

and '炒(fry)' in the context. With this extra morpheme information, MorBERT thus rightly disambiguates '爆炒' in the context. This validates that MP and MorBERT can be highly generalizable to OOV senses, which may bring new insights into and solutions for various downstream tasks in both computational and humanistic fields, including but not limited to new word prediction, dictionary compilation, etc.

On the other hand, compared to the results on senses in WrdInv (92.18% F1), the accuracy of the model on OOV senses still shows a certain decrease. We speculate that there are mainly two factors:

(1) Some word senses in WrdInv are derived from OOV senses with high semantic transparency, and they are composed of the same word-formation and morphemes. For example, for the word "铜板(*copper coin/copperplate*)', sense in WrdInv "*copper coin*" and OOV sense "*copperplate*" are both composed of the word-formation "Modifier-Head" and morphemes "铜$_1$(*copper*)" and "板$_1$(*a piece of a hard object*)". In these cases, morpheme prediction may not provide extra information;

(2) Constrained by the derivation and evolution of morpheme senses in new word senses, some morphemes cannot be found in WrdInv. For example, the word "杠杠(*thick straight line/ very good*)" has only one sense in WrdInv, "*thick straight line*". The morpheme "杠(*good*)" means "good" in the OOV sense, "*very good*", but this morpheme sense is not included in WrdInv. This indicates that there are deficiencies in the semantic space division of existing morphemes, as it cannot cover the morpheme senses that may be derived from new words or senses.

After morpheme prediction and WSD, it is possible to infer the new morpheme senses through computational methods, thus assisting with new word prediction as well as dictionary compilation.

7 Conclusion

In this paper, we propose to enhance Chinese WSD with full morphological knowledge, including word-formation and morphemes. We first construct large-scale and releasable WSD resources, including lexico-semantic inventories MorInv and WrdInv, Chinese WSD dataset MiCLS, and OOV test set. Then, following the previous FormBERT, we propose MorBERT to fully leverage morphological knowledge and achieve a SOTA F1 of 92.18% on MiCLS dataset. Analysis reveals that our model is characteristic of robustness in low-resource settings and generalizability to OOV senses. This may bring new insights into and solutions for various downstream tasks in both computational and humanistic fields.

In the near future, we plan to continuously enlarge the WSD datasets, expand them to other similar languages for further improvements in WSD, and apply them to the aforementioned downstream tasks, thereby contributing to a deeper understanding of languages.

8 Limitations

We have constructed the largest releasable Chinese WSD resources so far as we know and proposed MorBERT to fully leverage morphological knowledge, achieving a SOTA result. However, there is still room for improvement:

- Outside the single-character morphemes covered in **MorInv**, there still exists a small percentage of unexplored multi-character ones.
- Outside the disyllabic words covered in **WrdInv**, monosyllabic words, as free morphemes (viz. not bound morphemes), are already covered in **MorInv**. However, there still exists a small percentage of unexplored multisyllabic ones, which can be further decomposed into monosyllabic or disyllabic units through binary divisions.
- Current **MiCLS** only covers disyllabic words and can be extended to monosyllabic and multisyllabic words in the near future.
- The **OOV** test set is still relatively small and needs to be further expended.
- Current **MorBERT** only considers cases of disyllabic words. They may be adapted to further consider monosyllabic and multisyllabic words in the near future.

We will continuously enlarge the resources and broaden the model to encompass the issues challenging Chinese WSD.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (No. 62036001) and the National Social Science Foundation of China (No. 18ZDA295).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Wei Cao. 2001. *Modern Chinese Semantics*. Academia Press, Shanghai, China.

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 the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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

Huaiqing Fu. 1981. The relationship between word meaning and morpheme meaning that constitutes a word. *Dictionary Study*, 1:98–110.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 19–23, Prague, Czech Republic. Association for Computational Linguistics.

Shiyong Kang, Yi Li, and Daogong Sun. 2004. Construction of Chinese system corpus and dictionary compilation. In *Proceedings of the 2004 Workshop on Lexicography and Digitalization*, pages 145–151. Shanghai Dictionary Society.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2021. Abstractive Text Summarization: Enhancing Sequence-to-Sequence Models Using Word Sense Disambiguation and Semantic Content Generalization. *Computational Linguistics*, 47(4):813–859.

Zi Lin and Yang Liu. 2019. Implanting rational knowledge into distributed representation at morpheme level. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), pages 2954–2961.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3514–3520, Online. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411, Brussels, Belgium. Association for Computational Linguistics.

Jiajv Mei, Yiming Zhu, Yunqi Gao, and Hongxiang Yin. 1983. *Chinese Cilin*. Shanghai Lexicographical Publishing House, Shanghai, China.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Ariel Raviv and Shaul Markovitch. 2021. Concept-based approach to word-sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 807–813.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Rare and zero-shot word sense disambiguation using Z-reweighting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4713–4723, Dublin, Ireland. Association for Computational Linguistics.

Jing Wang, Lijiao Yang, Hongfei Jiang, Jingjie Su, and Jingling Fu. 2007. A word sense annotated corpus for teaching Chinese as second language. *Journal of Chinese Information Processing*, 31(01):221–229.

Yunfang Wu and Shiwen Yu. 2006. The principles and methods of sense discrimination for Chinese language processing. *Applied Linguistics*, 02:126–133.

Shu Xu. 1990. *Morpheme*. People's Education Press, Beijing, China.

Fukang Yan, Yue Zhang, and Zhenghua Li. 2023. 基于网络词典的现代汉语词义消歧数据集构建(construction of a Modern Chinese word sense dataset based on online dictionaries). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 43–53, Harbin, China. Chinese Information Processing Society of China.

Binyong Yin. 1984. Quantitative study on Chinese morphemes. *Studies of the Chinese Language*, (05):338–347.

Chunfa Yuan and Changning Huang. 1998. Research on Chinese morphemes and word formation based on morpheme database. *Chinese Teaching in the World*, 2:7–12.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021a. Decompose, fuse and generate: A formation-informed method for Chinese definition generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online. Association for Computational Linguistics.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021b. Leveraging word-formation knowledge for Chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hua Zheng, Yaqi Yan, Yue Wang, Damai Dai, and Yang Liu. 2021c. 基于词信息嵌入的汉语构词结构识别研究(Chinese word-formation prediction based on representations of word-related features). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 386–397, Huhhot, China. Chinese Information Processing Society of China.

Dexi Zhu. 1982. *Grammar Lectures*. The Commercial Press, Beijing, China.

A Prompt for Paraphrasing

请用20字以内转写字"体"的如下词典释义，保持语义与原释义一致，去除典故、举例和具体的细节，原释义如下：
Please paraphrase one of the sense definitions of the character "体(body)" within 20 characters, keep the semantics same with the original definition, remove allusions, examples and details. The original definition is as follows:
身体，有时指身体的一部分。
Body, sometimes refers to a part of the body .
转写后的释义：
Paraphrased definition:
身体或身体的一部分。
Body or a part of the body .

请用20字以内转写词"王国"的如下词典释义，保持语义与原释义一致，去除典故、举例和具体的细节，原释义如下：
Please paraphrase one of the sense definitions of the word "王国(kingdom)" within 20 characters, keep the semantics same with the original definition, remove allusions, examples and details. The original definition is as follows:
以国王为国家元首的国家，如丹麦王国、尼泊尔王国。
Countries with a king as the head of state, such as the Kingdom of Denmark and the Kingdom of Nepal.
转写后的释义：
Paraphrased definition:
国家的元首是国王的国家。
A country where the head of state is a king.

Figure 3: Example of the prompts for paraphrasing and their results for paraphrasing one of the sense definitions of "体" and "王国", which are "体$_1$" and "王国$_1$" in CCD.

We adopt ChatGPT (GPT-3.5-turbo-0613) to paraphrase the sense definitions of morphemes and words. Figure 3 shows examples of the prompts for paraphrasing.

B Prompts for LLM Annotation

Considering LLMs' sensitivity to prompts, we designed three prompt templates for LLM annotation. Figure 4 shows examples of the three prompts templates.

C Experimental Configuration

Our BERT model consists of 12 layers with 768 hidden units, with a learning rate of 5e-5, a batch size of 32, a dropout rate of 0.1, and a max sequence length of 128. The MP is optimized with Adam (Kingma and Ba, 2017), with the initial learning rate of 1e-5, warm-up phase of 10,000 steps, context batch size of 4, morpheme sense batch size of 32, and trained up to 20 epochs. The revised FP is optimized with AdamW, with a learning rate of 4e-4 and a batch size of 32.

Our experiments are conducted on an RTX A5000 GPU with 24GB memory.

Prompt 1

句子"玉盘似的满月在云中穿行"中词"满月"的意思是否是"圆月"？请回答是或否。
In the sentence "The jade-like full moon moves through the clouds", does the word "满月 (full moon) " mean "圆月 (round moon)"? Please answer 'yes' or 'no'.
回答：是
Answer: Yes

Prompt 2

句子"玉盘似的满月在云中穿行"中词"满月"是否可以被解释为"圆月"？请回答是或否。
In the sentence "The jade-like full moon moves through the clouds", can the word "满月 (full moon) " be interpreted as "圆月 (round moon)"? Please answer 'yes' or 'no'.
回答：是
Answer: Yes

Prompt 3

判断句子"玉盘似的满月在云中穿行"中词"满月"是否为"圆月"的意思？请回答是或否。
Determine whether the word "满月 (full moon) " in the sentence "The jade-like full moon moves through the clouds" means "圆月 (round moon)"? Please answer 'yes' or 'no'.
回答：是
Answer: Yes

Figure 4: Examples of the three prompt templates for LLM annotation and their results for annotating an entry targeting "满月".