Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation

Hua Zheng[*], **Lei Li**[*], **Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu**[†]
Department of Computer Science and Technology, Peking University
Key Lab of Computational Linguistics (MOE), Peking University
{zhenghua,daidamai,chendeli,tianyu0421}@pku.edu.cn
lilei@stu.pku.edu.cn
{xusun,liuyang}@pku.edu.cn

Abstract

In parataxis languages like Chinese, word meanings are constructed using specific word-formations, which can help to disambiguate word senses. However, such knowledge is rarely explored in previous word sense disambiguation (WSD) methods. In this paper, we propose to leverage word-formation knowledge to enhance Chinese WSD. We first construct a large-scale Chinese lexical sample WSD dataset with word-formations. Then, we propose a model **FormBERT** to explicitly incorporate word-formations into sense disambiguation. To further enhance generalizability, we design a word-formation predictor module in case word-formation annotations are unavailable. Experimental results show that our method brings substantial performance improvement over strong baselines.[1]

1 Introduction

Word sense disambiguation (WSD) aims to identify the sense of a polysemous word in a specific context, which benefits multiple downstream tasks (Hou et al., 2020). With copious sense-annotated data (Raganato et al., 2017), neural WSD methods achieve superior performance by leveraging definitional and relational features in knowledge bases (KB) (Luo et al., 2018a; Huang et al., 2019; Bevilacqua and Navigli, 2020).

In parataxis languages like Chinese, word meanings are highly correlated with word-formations (Li et al., 2018), which have not been explored in WSD thus far. Specifically, word-formations designate how characters interact to construct meanings. As shown in Figure 1, “征文$_1$" with the Modifier-Head formation means *solicited paper*, where “征" (*solicit*) modifies “文" (*paper*); “征文$_2$" with the Verb-Object formation means *solicit paper*, where “征" (*solicit*) operates on “文" (*paper*). On the flip side, word-formations can be inferred from the characters (Zhu, 1982). For instance, a character combination of *adjective-noun* is highly probable to have a Modifier-Head formation. Thus, after correct inference, word-formations can help to disambiguate polysemous words by indicating how characters interact in each sense.

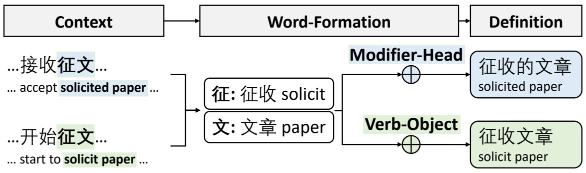

Figure 1: The contexts indicate that the word “征文" holds two senses constructed by different word-formations, which can be used to enhance WSD.

In this paper, we propose to leverage word-formation knowledge to enhance Chinese WSD. We first construct a large-scale Formation-informed Chinese Lexical Sample WSD dataset (**FiCLS**). Then, we propose a model **FormBERT** to explicitly incorporate word-formations into sense disambiguation. To enhance generalizability, we design a word-formation predictor module to predict word-formations for unannotated data. Experimental results show that our method brings substantial performance improvement on WSD with a high accuracy on formation prediction, which remains consistent in low-resource settings.

2 Related Work

WSD methods and resources: Recent supervised neural WSD methods achieve superior performance by leveraging lexical KB, e.g., incorporating definitional (Luo et al., 2018a,b; Huang et al., 2019; Hadiwinoto et al., 2019; Blevins and Zettlemoyer, 2020) and relational knowledge (Kumar et al., 2019; Bevilacqua and Navigli, 2020). However, these methods require copious sense-

[*] Equal Contribution

[†] Corresponding author.

[1] The code is available at https://github.com/TobiasLee/FormBERT.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 918–923
November 7–11, 2021.

Word-Formation	Example	%
Parallel	文体 (literary-physics)	34.40
Modifier-Head	引文 (cited-paper)	18.72
Verb-Object	发文 (publish-paper)	14.66
Adverb-Verb	博引 (widely-cite)	9.09
Single Morpheme	葡萄 (grape)	5.81

Table 1: Top 5 word-formations and examples. % denotes the instance percentage.

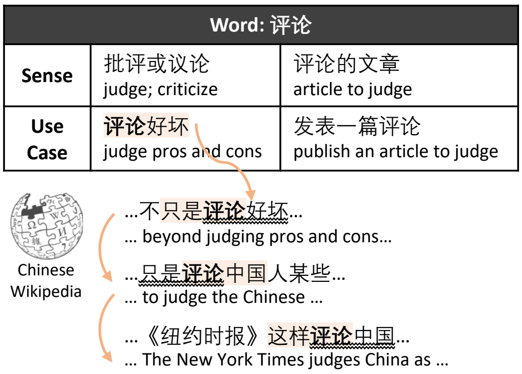

Figure 2: A simplified example of context augmentation for a sense of “评论" with a window size of 4. The underlined sequence is the matched pattern, and the sequence in orange is the sliced new matching pattern.

annotated datasets (Raganato et al., 2017), which are difficult to obtain in Chinese. Thus, previous Chinese WSD datasets (Niu et al., 2004; Jin et al., 2007; Agirre et al., 2009; Hou et al., 2020) are small in vocabulary size (less than 100 words except for Agirre et al., 2009), and it is uneasy to combine these datasets to enlarge their size, since they differ in format, sense inventory and construction guidelines.

Word-Formation knowledge: Instead of combining roots and affixes, Chinese words are constructed by characters using word-formations (Zhu et al., 2019). Word-formations have shown to be effective in multiple tasks like learning embeddings for parataxis languages (Park et al., 2018; Li et al., 2018; Lin and Liu, 2019; Zheng et al., 2021a,b). However, these works lack a clear distinction among different word-formations which require manual annotations.

3 The FiCLS Dataset

The construction of FiCLS includes two phases: collecting a base dataset and annotating word-formations. Each FiCLS entry consists of (1) a word, (2) a sense definition, (3) a word-formation, and (4) a context sentence.

3.1 Chinese WSD Dataset

We first construct a Chinese lexical sample WSD base dataset. We build the sense inventory based on the 5th edition of the Contemporary Chinese Dictionary (CCD) published by the Commercial Press,[2] one of the most influential Chinese dictionaries. Compared with other widely-used Chinese lexical KBs, CCD contains definitions that are more complete and native than HowNet sememes (Dong and Dong, 2006) and the translated Chinese WordNet (Wang and Bond, 2013). CCD contains 62,241 words, of which 22.32% are polysemous. To perform context augmentation, we collect only polysemous words that are labeled with use cases (short sequences containing the target sense), and obtain a total of 7,064 polysemous words (20,382 senses).

Considering the distributional hypothesis (Harris, 1954) that "similar distributions indiate similar meanings", we use matching patterns to expand the use cases into longer contexts via context augmentation using the Chinese Wikipedia corpus.[3] As shown in Figure 2, we slice each use case into matching patterns of window size {3,4,5} containing the target sense. Each pattern is used to match a longer sequence in the corpus as the new context. To enhance data diversity and balance, the new context will be sliced to produce new matching patterns, which repeats for at most 30 contexts per sense. The augmentation yields 145,964 entries in total, where the average length and number of contexts per sense are 53.04 and 7.16, respectively.

To ensure data quality, three mother-tongue reviewers manually check the contexts in three mutually-exclusive subsets of the data. Each reviewer is given a context and a definition to judge whether the context matches definition as a simple binary choice question. The whole revision takes 243 hours, where each reviewer checks about 600 entries in an hour. The final dataset contains 121,655 entries, which is the largest Chinese lexical sample WSD dataset so far as we know.

3.2 Word-Formation Annotations

We perform human annotation on the base dataset to obtain word-formations. Following Liu et al. (2018), we adopt 16 Chinese word-formations. Our annotators are professors and postgraduates ma-

[2] https://www.cp.com.cn

[3] https://dumps.wikimedia.org/zhwiki/20200920/

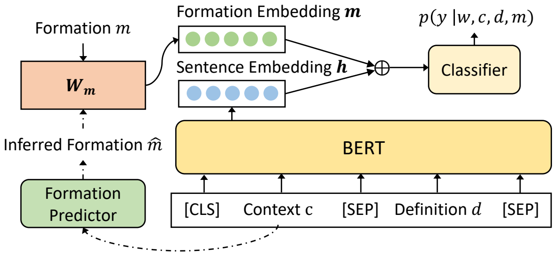

Figure 3: Illustration of the proposed FormBERT with FP. The dashed line indicates that, during inference, the inferred formation based on the context will be exploited to generalize to scenarios without formation.

jor in Chinese linguistics. Given the sense definition, they annotate each sense with its word-formation. Each entry is cross-validated by three independent annotators and reviewed by one. With a detailed guideline available, the inter-annotator kappa (Fleiss and Cohen, 1973) is 92.61. Table 1 shows the top 5 word-formations in instance percentage. We provide the detailed annotation guideline and pipeline in Appendix A.

4 Methodology

4.1 Task Formulation

We formulate WSD as a sentence-level binary classification task, which has been proved to effectively leverage definitions in BERT-based WSD methods (Huang et al., 2019). Specifically, given a target word w and its context sentence c, we construct an instance triplet (w, c, 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 and definition into a character sequence with the BERT-specific prediction token `[CLS]` and the sentence boundary indicator `[SEP]`.[4] 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:

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

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

$$\mathcal{L}_{\text{wsd}} = -\log p\left(y^* \mid w, c, d\right).$$

4.2 FormBERT with Formation Predictor

We first propose FormBERT to incorporate word-formations seamlessly into the BERT-based model (Devlin et al., 2019). Specifically, given the target word w and its word-formation annotation m^* for the ground-truth definition d in the context c, we learn a formation embedding $\mathbf{m}^*$ via a matrix $\mathbf{W_m}$ for each formation type. The obtained formation embedding $\mathbf{m}^*$ is then combined with $\mathbf{h}$ to produce the label probability distribution:

$$p(y \mid w, c, d, m^*) = f(\mathbf{h} + \mathbf{m}^*).$$

By incorporating the word-formations, FormBERT is better informed of how the characters interact in the target word to better distinguish senses. However, word-formations are expensive to acquire and can be unavailable in other datasets. Thus, we introduce an auxiliary formation prediction task, motivated by the fact that word-formations can be inferred from the characters, as stated in Section 1:

$$p(m \mid w, c) = g(w, c),$$
$$\hat{m} = \arg\max_{m} p(m \mid w, c),$$

where $g(\cdot)$ is a MLP formation predictor. Note that we do not utilize the BERT embeddings of the context since the embeddings fuse external information from the definition, which can be wrong in the negative triplet. The inferred formation $\hat{m}$ can thus be exploited as a supplementary formation feature. Figure 3 gives an overview of FormBERT with FP. During training, where the word-formations are available, a formation prediction objective is added for training the predictor:

$$\mathcal{L}_{\text{fp}} = -\log p(m^* \mid w, c).$$

This objective is combined with the original sense disambiguation loss with a weighting factor λ. With a well-trained FP, our framework can generalize to data without word-formation annotations.

5 Experiments

5.1 Experimental Settings

Datasets: We split FiCLS described in Section 3 into training, validation and test sets by 8:1:1, as shown in Table 2. Note that the validation and test sets have the same number of positive and negative entries, as stated in Section 4.1.

Baselines: Besides BERT (Devlin et al., 2019) and most frequent sense (MFS) as default baselines, we

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

Split	#Words	#Senses	#Entries	Context Length	Definition Length
Train	6,989	18,615	95,698	52.32	8.88
Valid	4,004	7,368	12,500	52.45	8.92
Test	3,930	7,307	12,500	52.45	8.83

Table 2: Statistics of FiCLS. The length is calculated as the average number of Chinese characters.

Method	Valid	Test				
		Noun	Verb	Adj.	Adv.	All
MFS	34.39	35.23	34.49	33.25	36.65	34.99
BERT	71.21	74.68	71.10	72.05	64.29	71.78
GLU	71.24	74.80	70.89	71.60	63.79	71.65
GlossBERT	84.55	82.94	81.95	82.59	81.88	84.51
BEM	72.06	73.32	72.58	74.64	66.22	72.17
FormBERT	**87.34**	**88.74**	87.07	**88.59**	81.41	87.35
FormBERT w/ FP	87.33	88.71	**87.67**	88.52	**83.07**	**87.62**

Table 3: Evaluation results (F1) on FiCLS. Best results are shown in **bold**. FormBERT w/ FP denotes FormBERT using the formation predictor without annotated word-formations.

implement strong baselines with features available in FiCLS, including GLU (Hadiwinoto et al., 2019), GlossBERT (Huang et al., 2019) and BEM (Blevins and Zettlemoyer, 2020), and use the same settings as our model for a fair comparison.

Experimental Configurations: We adopt BERT-wwm-ext (Cui et al., 2020) as the base model. Our BERT model consists of 12 layers with 768 hidden units. The formation predictor module is a 2-layer feedforward network with a hidden size of 768 and ReLU as the activation function. We use AdamW with a learning rate of 4e-4 and set the batch size to 32. We set the formation prediction objective weight λ as 0.5 based on the validation performance. All hyper-parameters for training the model are tuned based on the validation performance, as listed in Table 4. For the baselines, we directly follow the settings in their original papers. Specifically, 1) in BERT, we use the hidden states of the target word for predictions; 2) GlossBERT is formulated the same as our model; 3) GLU is based on BERT with an additional gated linear unit for transformation of hidden vectors; 4) BEM is based on BERT with bi-encoders for contexts and definitions. Since all baselines are BERT-based, we use the same hyper-parameter settings as our model for a fair comparison and select the checkpoint based on the best validation performance. Our experiments are conducted on 4 RTX 2080Ti GPUs with 11GB memory.

Hyper-parameter	Value
BERT Learning Rate	5e-5
Formation Predictor Learning Rate	4e-4
Batch Size Per Device	32
Dropout Rate	0.1
Max Sequence Length	128
Formation Prediction Loss Weight	{0.1, 0.2, 0.5, 1.0}

Table 4: Hyper-parameters of the experiments.

Method	LFD	MFD	Zero-shot	Few-shot
GlossBERT	83.89	85.15	76.69	84.53
BEM	63.23	86.58	48.54	65.11
FormBERT	85.81	89.60	82.42	86.01
FormBERT w/ FP	**85.93**	**90.01**	**82.65**	**86.25**

Table 5: Evaluation results (F1) on the MFD, LFD, zero-shot and few-shot subsets of the test set.

5.2 Evaluation Results

Table 3 shows the overall F1 results on FiCLS across 4 main parts-of-speech (PoS). Note that we only label PoS for the test set for a parallel comparison with previous works (Blevins and Zettlemoyer, 2020), and the PoS is not included during training.

From Table 3, we have the following observations: (1) By leveraging word-formation knowledge, our FormBERT achieves substantial improvement by 2.84 F1 points more than GlossBERT, which validates that word-formations can effectively enhance Chinese WSD. (2) Although FormBERT w/ FP has no ground-truth word-formation annotations, it achieves comparable results with FormBERT, which confirms the generalizability of our method. We speculate that the slight advantage over FormBERT can be owing to (i) the significantly-high 93.29 accuracy of word-formation predictions, and (ii) the implicitly regularized context embeddings from the formation prediction objective. (3) Concerning the performance on different PoS, most models perform the worst on adverbs. This can be explained by the high granularity of adverbs in the CCD sense inventory, e.g., the adverb “一头" consists of 8 senses, 6 of which denote the similar meaning of “directly".

5.3 Analysis

Generalizability of FP: To test the generalizability of FP, we evaluate it on an additional set of 500 senses of polysemous words that are unavailable during training. Results show that FP achieves a high accuracy of 92.80, which validates that FP can be highly generalizable to other datasets. Note that we do not apply our method on previous datasets since they differ in sense inventory and construction guidelines, as stated in Section 2.

FormBERT in low-resource settings: To better understand the overall results, we divide the test set into four subsets: (1) entries with the most frequent definition (MFD) of the target word, (2) entries with the less frequent definitions (LFD) than MFD, (3) zero-shot entries of unseen definitions during training, and (4) few-shot entries of definitions appearing less than five times during training. We compare FormBERT with and without FP against GlossBERT and BEM, as shown in Table 5. Results indicate that, by leveraging word-formations, both FormBERT with and without FP introduce consistent improvement over the baselines, which validates that our method is effective and robust even in low-resource settings.

6 Conclusion

In this paper, we propose to enhance Chinese WSD with word-formation knowledge. We first construct a large-scale formation-informed dataset. Then, we propose FormBERT to incorporate the word-formations into BERT and design a formation predictor to ease the reliance on annotated data. Experimental results validate the effectiveness of leveraging word-formations for Chinese WSD.

Acknowledgments

We would like to thank all the anonymous reviewers and annotators for their helpful advice on various aspects of this work, including Fuqian Wu, Ming Liu, Yaqi Yin, Yue Wang, etc. 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

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Andrea Marchetti, Antonio Toral, and Piek Vossen. 2009. SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *SEW*, pages 123–128.

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 *ACL*, pages 2854–2864.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *ACL*, pages 1006–1017.

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*, pages 657–668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Zhendong Dong and Qiang Dong. 2006. *HowNet and the computation of meaning*. World Scientific.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *EMNLP-IJCNLP*, pages 5297–5306.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised Chinese word sense disambiguation method based on HowNet. In *COLING*, pages 1752–1757.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *EMNLP-IJCNLP*, pages 3509–3514.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample. In *SemEval*, pages 19–23.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *ACL*, pages 5670–5681.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *ACL*, pages 138–143.

Zi Lin and Yang Liu. 2019. Implanting rational knowledge into distributed representation at morpheme level. In *AAAI*, pages 2954–2961.

Yang Liu, Zi Lin, and Sichen Kang. 2018. Towards a description of chinese morpheme conceptions and semantic composition of word. *Journal of Chinese Information Processing*, 32(2):12–21.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *EMNLP*, pages 1402–1411.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. Incorporating glosses into neural word sense disambiguation. In *ACL*, pages 2473–2482.

Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2004. Optimizing feature set for Chinese word sense disambiguation. In *SENSEVAL-3*, pages 191–194.

Hyun-jung Park, Min-chae Song, and Kyung-Shik Shin. 2018. Sentiment analysis of korean reviews using cnn: Focusing on morpheme embedding. *Journal of Intelligence and Information Systems*, 24(2):59–83.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *EACL*, pages 99–110.

Shan Wang and Francis Bond. 2013. Building the Chinese open Wordnet (COW): Starting from core synsets. In *Workshop on Asian Language Resources*, pages 10–18.

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 *NAACL-HLT*, pages 5524–5531.

Hua Zheng, Yaqi Yan, Yue Wang, Damai Dai, and Yang Liu. 2021b. Chinese word-formation prediction based on representations of word-related features. In *CCL*, pages 386–397.

Dexi Zhu. 1982. *Yufa Jiangyi (Lectures on Grammar)*. The Commercial Press, China.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. In *NAACL-HLT*, pages 912–932.

A FiCLS Dataset Construction

A.1 Statistics

Table 6 shows the overall descriptions of 16 word-formations, including the explanation, example and instance percentage in FiCLS. The explanations function as an annotation guideline for the annotators. All annotators and reviewers are paid regarding the workload (0.1￥/annotation entry).

Word-Formation	**Explanation**	**Example**	**%**
联合 (Parallel)	morph$_1$ and morph$_2$ are similar, contrasting or complementary.	文体 (literary-physics)	34.40
定中 (Modifier-Head)	morph$_1$ modifies morph$_2$ (noun).	引文 (cited-paper)	18.72
述宾 (Verb-Object)	morph$_1$ operates on morph$_2$.	发文 (publish-paper)	14.66
单纯 (Single Morpheme)	The word is a single morpheme.	葡萄 (grape)	9.09
状中 (Adverb-Verb)	morph$_1$ modifies morph$_2$ (verb).	博引 (widely-cite)	5.81
连谓 (Verb-Consequence)	morph$_2$ is the consequence of morph$_1$.	休息 (stop-rest)	4.09
后缀 (Suffixation)	morph$_2$ is the suffix of morph$_1$.	花头 (trick-∅)	3.61
前缀 (Prefixation)	morph$_1$ is the prefix of morph$_2$.	老师 (∅-teacher)	3.47
述补 (Verb-Complement)	morph$_2$ is the action follows morph$_1$.	压低 (press-down)	2.50
重叠 (Overlapping)	morph$_1$ and morph$_2$ are the same.	白白 (vainly-vainly)	1.15
主谓 (Subject-Predicate)	morph$_1$ is the subject of morph$_2$.	眼花 (eyesight-dim)	1.13
介宾 (Preposition-Object)	morph$_1$ is a preposition, morph$_2$ is an object.	凭空 (from-nowhere)	0.49
方位 (Entity-Position)	morph$_1$ is an entity, morph$_2$ is a position.	期中 (semester-mid)	0.41
数量 (Number-Quantifier)	morph$_1$ is a number, morph$_2$ is a quantifier.	一点 (one-dot)	0.28
复量 (Quantifier-Quantifier)	Both morph$_1$ and morph$_2$ are quantifiers.	千米 (kilo-meter)	0.11
名量 (Noun-Quantifier)	morph$_2$ is the quantifier of morph$_1$.	花朵 (flower-bud)	0.07

Table 6: Descriptions of the total 16 word-formations. ∅ denotes the affix and % denotes the instance percentage. The first and the third columns are in the format of "Chinese characters (English translation)". We give a simple explanation in the second column to describe the relation between two characters, which functions as a guideline to the annotators.

A.2 Annotation Process

In the word-formation annotation process of FiCLS, our annotators include two professors and six postgraduates major in Chinese linguistics. For ease of annotation, we build an annotation interface, as shown in Figure 4.

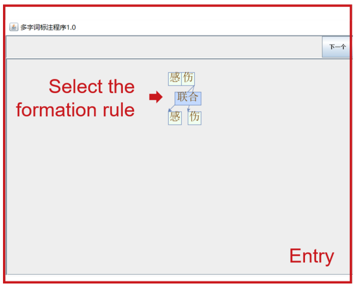

Figure 4: Human annotation interface.

The annotation process is as follows: (1) Equipped with the definition, annotators annotate each entry with the word-formation (selected from the total of 16 formation rules). Each entry is independently annotated by three annotators, who also note down a confidence score. If three annotations are the same, turn to (3); otherwise, turn to (2). (2) Another annotator reviews the conflicting annotations and confidence scores, and decides the final annotation. Turn to (3). (3) The annotation is collected into the final dataset.

It takes 20 seconds on average for each annotator to annotate an entry. Only 4,205 out of 20,382 entries enter Phase (2) in the annotation process.