Unsupervised Adverbial Identification in Modern Chinese Literature

Wenxiu Xie[1], **John S. Y. Lee**[2], **Fangqiong Zhan**[3], **Xiao Han**[2], **Chi-Yin Chow**[1]
[1]Department of Computer Science, City University of Hong Kong
[2]Department of Linguistics and Translation, City University of Hong Kong
[3]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
vasiliky@outlook.com, jsylee@cityu.edu.hk, zhjade2000@gmail.com
xhan42-c@my.cityu.edu.hk, tedchow@gmail.com

Abstract

In many languages, adverbials can be derived from words of various parts-of-speech. In Chinese, the derivation may be marked either with the standard adverbial marker DI, or the non-standard marker DE. Since DE also serves double duty as the attributive marker, accurate identification of adverbials requires disambiguation of its syntactic role. As parsers are trained predominantly on texts using the standard adverbial marker DI, they often fail to recognize adverbials suffixed with the non-standard DE. This paper addresses this problem with an unsupervised, rule-based approach for adverbial identification that utilizes dependency tree patterns. Experiment results show that this approach outperforms a masked language model baseline. We apply this approach to analyze standard and non-standard adverbial marker usage in modern Chinese literature.

1 Introduction

In many languages, adverbials can be derived from words of other parts-of-speech (POS). The adverbials are often morphologically marked in the derivation process. In English, the most common marker is the suffix “-ly”, for example in the derivation of “happily” from the adjective “happy”. In Chinese, the focus of our study, the standard adverbial marker is DI (地)[1], which can be suffixed to a wide variety of adverbs, adjectives and verbs to form adverbials. However, a non-standard adverbial marker, the suffix DE (的), is also used frequently, in some contexts on par with the standard marker (Zhang, 2012a).

Chinese parsers tend to be trained on standard texts that mostly employ the standard adverbial marker. Indeed, many Chinese treebank guidelines (Yu et al., 2001; Xia, 2000) mention only DI as the adverbial marker. As a result, even state-of-the-art parsers might not accurately detect adverbials formed with the non-standard marker. This affects not only parser accuracy but also downstream NLP tasks, as well as linguistic research on the DE vs. DI choice.

This paper investigates the task of adverbial identification, including adverbials marked both in the standard and non-standard manner. Due to the lack of annotated data for non-standard adverbials, our research focus is on unsupervised methods. We propose a simple yet effective approach for adverbial identification based on POS tag and dependency tree patterns. In an evaluation on texts drawn from modern Chinese literature, our approach achieved over 87% accuracy, outperforming a state-of-the-art masked language model.[2] To our knowledge, this is the first reported study on identifying DE- and DI-adverbials in Chinese text.

The rest of the paper is organized as follows. The next section provides the necessary linguistic background for adverbials in Chinese. Section 3 summarizes previous work. Section 4 describes our approach and the baselines. Section 5 presents our evaluation dataset. Section 6 discusses experimental results. Section 7 applies our research to the analysis of adverbial marker choice of a prominent Chinese author. Finally, Section 8 concludes.

2 Adverbials in Chinese

Every suffixed adverbial in Chinese can be rendered in one of two forms. The adjective *gaoxing* 高兴 ‘happy’, for example, can be transformed into an adverbial either with the standard DI suffix (*gaoxing-DI* 高兴地 ‘happily’) or the non-standard DE suffix (*gaoxing-DE* 高兴的 ‘happily’). We will

[1] Although the suffix is pronounced “de”, we use the shorthand DI based on the character’s pronunciation “di” in other contexts, in order to differentiate it from the non-standard marker, which has the same pronunciation “de”.

[2] All evaluation data is publicly released at https://github.com/wxx2021/Modern-Chinese-Literature-Adverbial-Marker-Datatset-LaTeCH-CLfL-2021.

Proceedings of LaTeCH-CLfL 2021, pages 91–95
Punta Cana, Dominican Republic (Online), November 11, 2021.

henceforth refer to the former as “DI-adverbials” and the latter as “DE-adverbials”.

Detecting the DE-adverbials can be challenging because of the dual role of the DE suffix: it serves both as the non-standard adverbial marker and as the (standard) attributive marker. A DE-suffixed word can therefore potentially be either an adverbial or an adjectival phrase. For example, *gaoxing-DE* is an adverbial in the expression *gaoxing-DE chang zhe ge* ‘happily sing a song’, but it is an adjectival phrase in the expression *gaoxing-DE rizi* ‘happy day’ (Table 1). The former is often parsed incorrectly as an adjective since training data tend to favor the standard marker.

In the rest of this paper, we will use the term *base word* to refer to the word from which the adverbial is derived, and *head word* to refer to the word modified by the adverbial. In Figure 1a, the adverbial *gaoxing-DE* ‘happily’ has *gaoxing* ‘happy’ as its base word and *chang* ‘sing’ as its head word.

3 Background

During the Vernacular Language Movement (Weng, 2020), Chinese intellectuals proposed to use DE as the standard attributive marker and DI as the standard adverbial marker (Table 1). Many Chinese treebank guidelines, such as those from Peking University (Yu et al., 2001) and the Chinese Treebank (Xia, 2000), also focus on the DI suffix in their treatment of adverbials. However, the division of labor between DE and DI has never been strictly observed, and the non-standard DE-adverbials have persisted.

Neglection of the DE-adverbials hampers both parsing accuracy and linguistics research. When parsers fail to recognize DE-adverbials and other adverbials that modify adjectives, adverbs or a sentence (Yang, 1999), accuracy in adverbial identification would be affected (Xing et al., 2020). Research on linguistic variation in Chinese often addressed DE as possessive and attributive marker, but not as adverbial marker (Zhang, 2012b). Because of the ambiguity with the DE suffix, most quantitative analyses on DI- vs. DE-adverbial usage required manual annotation or was restricted to relatively small sample sizes (Tan, 2004; Zhang, 2012a; Ho, 2015).

Previous research has utilized both rule-based and machine learning approaches to study Chinese numeral classifiers that form quantity noun phrases (Guo and Zhong, 2005; Peinelt et al., 2017).

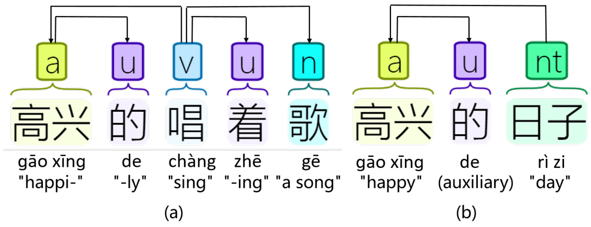

Figure 1: Use of the DE suffix as adverbial marker and attributive marker and its disambiguation through dependency relations: (a) DE marks the base word *gaoxing* ‘happy’ to form the adverbial *gaoxing-DE*, with *chang* ‘sing’ as its head word; (b) DE marks the base word *gaoxing*‘happy’ to form the adjectival phrase *gaoxing-DE*, with *rizi* ‘day’ as its head word

Similar to our study, it was found effective to leverage syntactic criteria and linguistic rules to analyze relations between numeral classifiers, nominal head words and nouns.

4 Approach

While the DI suffix is a reliable identifier for DI-adverbials, a naive search for the DE suffix would yield low precision in retrieving DE-adverbials since DE may also mark attributives. It is essential to determine the POS of the head word: DE marks an adverbial if the head word is an adjective or verb, but it marks an attributive if the head word is a noun.

4.1 Proposed method

We investigate a parser-based approach in the following three settings, using the HanLP Chinese parser (He, 2020) for word segmentation, POS tagging and dependency parsing.[3]

POS only This baseline predicts “adverbial” if the base word is tagged as “adverb” (d in HanLP, e.g., *gaoxing*/d DE/u). Note that the DE-adverbial in Figure 1a would be falsely rejected. If the word segmentor combines the base word and suffix as one word, the POS tag of the word is also required to be “adverb” (e.g. *gaoxing-DE*/d).

POS+base This second baseline aims to improve recall with more relaxed POS constraints. It predicts “adverbial” when the base word is tagged as “adjective” (a) or “verb” (v), in

[3] Full description of the POS tag set can be accessed at http://www.hankcs.com/nlp/part-of-speech-tagging.html#h2-8.

Suffix	Role	Status	Example
DI	adverbial marker	standard	*gaoxing-DI chang zhe ge* 高兴地唱着歌 'happily sing'
DE	adverbial marker	non-standard	*gaoxing-DE chang zhe ge* 高兴的唱着歌 'happily sing'
DE	attributive marker	standard	*gaoxing-DE rizi* 高兴的日子 'happy day'

Table 1: Examples illustrating the various syntactic roles of the DI suffix and DE suffix

Method	**DI**			**DE**			**All**
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Accuracy**
MLM	1	0.8344	0.9097	**1**	0.6600	0.7951	0.7833
POS only	1	0.7232	0.8393	0.9970	0.7628	0.8643	0.7843
POS+base	0.7666	**0.9462**	0.8470	0.7645	**0.9732**	0.8563	0.8382
POS+base+head	**1**	0.9356	**0.9667**	0.9725	0.7920	**0.8730**	**0.8743**

Table 2: Adverbial identification performance

addition to "adverb" (`d`) as above. If the word segmentor combines the base word and suffix as one word, the tags for "auxiliary" (`u`) and idiom (`i`) are also accepted. This approach would correctly recognize the DE-adverbial in Figure 1a, but also falsely identify Figure 1b as an adverbial.

POS+base+head Our proposed approach aims to balance precision and recall with an additional constraint on the head word, which is automatically identified via dependency relations in the parse tree (Figure 1). It predicts "adverbial" only when the head word is tagged as "verb" (`v`), "adjective" (`a`), "idiom" (`i`), "abbreviation" (`j`), or "preposition" (`p`), since Chinese prepositions often function like verbs. The POS requirement on the head word is waived if the head word has a child word in a "subject-predicate" dependency relation, since the relation strongly suggests the head word can be modified by an adverbial. This approach would correctly reject Figure 1b as an adverbial since the head word *rizi* 'day' is tagged as a noun.

4.2 Masked language model baseline

As an additional baseline, we evaluated BERT[4], a state-of-the-art masked language model (Devlin et al., 2018; Wolf et al., 2020). Similar to the parsers, BERT is trained mostly on standard Chinese text (e.g., Chinese Wikipedia) that skews towards the standard adverbial marker. We mask the characters DE and DI in the input text.[5] We predict "adverbial" if BERT ranks DI higher than DE as a candidate word for the masked position.

5 Data

There is no publicly available, large-scale corpus of Chinese text with annotations on adverbial markers. We focused on modern Chinese literary works written during and after the Vernacular Language Movement, when the division of labor between DE and DI was formalized. We created a corpus by extracting all works by four prominent Chinese authors (Guo Moruo, Lao She, Lu Xun, and Mao Duo) posted on the *Baiwan Shuku* website.[6] There are a total of 911 works containing over 6 million characters.

A native speaker of Chinese with formal training in linguistics examined three words in this corpus: *manman* 慢慢 'slow', *keqi* 客气 'courteous', and *gaoxing* 高兴 'happy'. Each of the 1,057 occurrences was labelled as a base word in a DE-adverbial (447 instances); a base word in a DI-adverbial (465 instances); or neither (145 instances).

6 Experimental Results

POS+base+head. As shown in Table 2, our proposed approach achieved 87.43% accuracy in adverbial identification. It outperforms POS+base, which has the second highest accuracy at 83.82%, indicating the effectiveness in examining the head word.[7]

[4] https://huggingface.co/bert-base-chinese

[5] For example, *gaoxing-DE chang zhe ge* (Figure 1a) would be masked as *gaoxing* [MASK] *chang zhe ge*

[6] http://www.millionbook.com/mj/index.html, accessed in July 2019.

[7] The improvement is statistically significant at $p < 0.0158$ by McNemar's Test. The improvement is also significant against MLM at $p < 0.0001$ and POS only at $p < 0.0001$

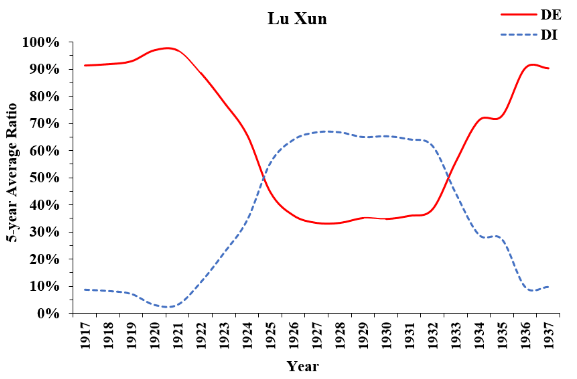

Figure 2: The proportion of DE-adverbials and DI-adverbials in the works of Lu Xun from 1917 to 1937.

The proposed approach also scored the highest F1 in identifying DI and DE adverbials. The relatively low recall resulted from the POS ambiguity for the head word. The parser sometimes erred on interpreting DE as marking attributives rather than adverbials, as many Chinese verbs share the same form as nouns. Recall was also affected by word segmentation errors, and parser errors where the base word is parsed as root.

POS only. The adverbials extracted by this approach were almost always true positives, with high precision for both DE- and DI-adverbials. The recall, however, was affected by noise in POS tagging, especially for DE-adverbials that do not modify verbs.

POS+base. This method outperformed the proposed approach in recall both for DI-adverbials (94.62%) and DE-adverbials (97.32%). However, it suffered from low precision (76.66% for DI-adverbials and 76.45% for DE-adverbials), partly due to false positives with incompatible head word POS.

Masked Language Model (MLM). This model achieved 100% precision but at the expense of low recall, especially for DE-adverbials (66.00%). It had difficulty recognizing a DE-adverbial especially when the verb is located at a long distance, in which case it tends to rank DE higher than DI.

7 Application

As an application of this research, we present a case study on the DE- vs. DI-adverbial usage of Lu Xun, arguably the most influential writer in modern Chinese literature. Literary works are ideal for studying adverbial markers since marker choices are more likely to be consciously rather than randomly made.

We used the POS+base+head method (Section 4.1) to automatically label *all* suffixed adverbials in his writings in our corpus. Over his career Lu Xun chose the DE suffix 56.9% of the time and the DI suffix in the rest. This overall figure, however, masks three distinct periods in which his marker usage varied.

Figure 2 shows a diachronic analysis, where each data point represents the average percentage of DE-adverbials and DI-adverbials within the 5-year window centered on the year on the horizontal axis. The initial period, from 1917 to 1924, was dominated by the non-standard marker DE. As he came under the influence of the Vernacular Movement, Lu Xun entered a second period around 1925. The DE percentage dropped dramatically from 65.7% in 1924 to 44.7% in 1925, reflecting his "innovative work in style" in several significant publications in 1925 (Gunn, 1991). In the 1930s, the call for a "mass language" gathered steam towards the end of the Vernacular Movement, with the goals of eradicating illiteracy and giving ordinary people access to writing. Subsequently the Latinized New Writing movement blurred the distinction between the two markers, leading to the re-emergence of the unified use of DE. An advocate of Latinization, Lu Xun reverted to non-standard markers, with the DE percentage approaching a similar level as the initial period.

8 Conclusions

This paper has presented the first study on identifying both standard and non-standard adverbial markers. We have proposed an unsupervised, rule-based algorithm based on dependency tree patterns. In an evaluation on modern Chinese literary text, our proposed method achieved an accuracy of over 87%, outperforming a state-of-the-art masked language model. Finally, we applied it on a diachronic analysis of adverbial marker choice in the works of a prominent author.

In future work, we aim to further improve performance in adverbial identification by incorporating supervised methods. As the DE vs. DI choice continues to evolve, it would also be interesting to examine contemporary authors and analyze their adverbial marker usage in comparison to those in the last century.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edward Gunn. 1991. *Rewriting Chinese: Style and innovation in twentieth-century Chinese prose*. Stanford university press.

Hui Guo and Huayan Zhong. 2005. Chinese classifier assignment using svms. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Han He. 2020. HanLP: Han Language Processing.

James Ho. 2015. From the use of three functional words “的, 地, 得” examining author’s unique writing style–and on dream of red chamber author issues. *BIBLID*, 120(1):119–150.

Nicole Peinelt, Maria Liakata, and Shu-Kai Hsieh. 2017. Classifierguesser: A context-based classifier prediction system for Chinese language learners. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 41–44.

Yongkang Tan. 2004. 状中结构助词混用的几个影响源 [in Chinese]. *Journal of Chongqing Radio Television University*, 3:42–43.

Jeffrey Weng. 2020. *Vernacular Language Movement*. Oxford University Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*, page 38.

Dan Xing, Endong Xun, Chengwen Wang, Gaoqi Rao, and Luyao Ma. 2020. Construction of adverbial-verb collocation database based on large-scale corpus. In *Workshop on Chinese Lexical Semantics*, pages 585–595. Springer.

Rongxiang Yang. 1999. On early modern Chinese adverbs. *Journal of Peking University (Humanities and social Sciences)*, 3.

Shiwen Yu, Jianming Lu, Xuefeng Zhu, Huiming Duan, Shiyong Kang, Honglin Sun, Hui Wang, Qiang Zhao, and Weidong Zhan. 2001. Processing norms of modern Chinese corpus. Technical report.

Yisheng Zhang. 2012a. On the selection of adverbial markers in contemporary Chinese [j]. *Chinese Linguistics*, 4.

Zheng-Sheng Zhang. 2012b. A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, 8(1):209–240.