Automatic Refinement of Syntactic Categories in Chinese Word Structures

Jianqiang Ma
University of Tübingen, Department of Linguistics
Wilhelmstr. 19, 72074 Tubingen, Germany
jianqiang.ma@uni-tuebingen.de

Abstract

Annotated word structures are useful for various Chinese NLP tasks, such as word segmentation, POS tagging and syntactic parsing. Chinese word structures are often represented by binary trees, the nodes of which are labeled with syntactic categories, due to the syntactic nature of Chinese word formation. It is desirable to refine the annotation by labeling nodes of word structure trees with more proper syntactic categories so that the combinatorial properties in the word formation process are better captured. This can lead to improved performances on the tasks that exploit word structure annotations. We propose syntactically inspired algorithms to automatically induce syntactic categories of word structure trees using POS tagged corpus and branching in existing Chinese word structure trees. We evaluate the quality of our annotation by comparing the performances of models based on our annotation and another publicly available annotation, respectively. The results on two variations of Chinese word segmentation task show that using our annotation can lead to significant performance improvements.

Keywords: Chinese word structure, automatic annotation, Chinese word segmentation

1. Introduction

1.1. Why Chinese word structure annotation

As an analytic language that lacks inflectional morphemes, the word formation in Chinese is mostly achieved by recursive compounding and derivation, which yield hierarchical word structures. Yet, in Chinese NLP, words are traditionally treated as structureless tokens that are recovered from consecutive written characters without delimiters in Chinese text, by Chinese word segmentation (CWS). Not until recently has Chinese word structure annotation (Li, 2011; Li and Zhou, 2012; Ma et al., 2012; Zhang et al., 2013; Zhao, 2009) been developed and shown to be highly effective for joint CWS, POS tagging and parsing. Such annotations are useful in at least two ways:

- the character-based parsing model trained upon it can use head characters and sub-words to offset data sparsity, which is particularly useful for parsing.
- the morphological rules learned from it can contribute to better models for word recognition, especially for out-of-vocabulary (OOV) words, an important source of errors in CWS and POS tagging.

Since sparsity and OOVs are common challenges in NLP, word structure annotation is potentially useful for other Chinese NLP applications as well.

1.2. Syntactic categories in word structure trees

Among the above-motioned annotations, only Zhang et al. (2013) is publicly available [1]. The hierarchical bracketing and head characters of 37,382 word types in CTB5, the version 5 of the Chinese Treebank (Xue et al., 2005) are manually annotated, which yielded word structure trees. Nevertheless, the syntactic categories of all the sub-words and characters in each word structure tree specified by the (word, POS tag) pair are uniformly assigned as the same category of that particular POS tag. While this treatment avoids complicated structure disambiguation in annotation, it offers relatively little information about the underlying compounding and derivation process.

For example, Figure 1 shows how the word 开幕 “inaugurate/inauguration” is annotated as two distinct structures with nominal and verbal function, respectively. In the case of 开幕 “inaugurate/inauguration”, its nominal and verbal entry are not explicitly related. Moreover, in their annotation scheme, the character 幕 “curtain” in Figure 1 (b) *must* be tagged as the V label because it appears in a verbal entry. But this assignment contradicts with the fact that this character does not have verbal function at all in modern Chinese. The assignment of the N tag to the 开 “open” in Figure 1 (a) is also questionable.

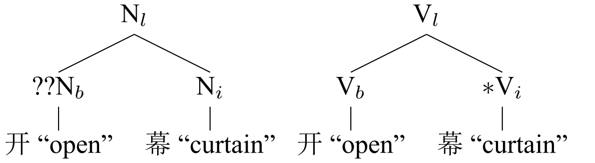


Figure 1: The word structure tree for 开幕 “inaugurate/inauguration” in Zhang et al. (2013) annotation.

Compounding, as the predominate word formation process in Chinese, is pervasive for and occurs between virtually all open-class lexical categories, unlike English or German, in which the majority of compounding occurs between nouns. Many linguists believe that Chinese compounding rules resemble syntactical rules (Zhu, 1985; Xue, 2001; Lu, 2005). Using a syntax metaphor, Zhang et al. (2013) annotation resemble a “partial-labeled” treebank, in which many bracketed constituents lack proper syntactic labels. Empirically, their annotation helps NLP tasks by reducing the data sparseness, but the incomplete annotation compromise the

[1] http://ir.hit.edu.cn/ mszhang/data.zip

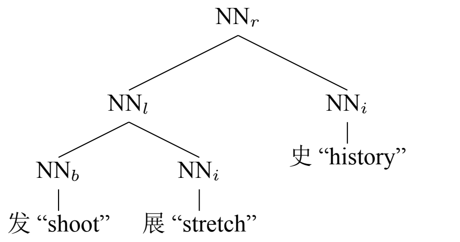


Figure 2: The word structure tree for word 发展史 "development history" in Zhang et al. (2013) annotation. The subscripts are head direction sub-tags $l/r/c$ and character position sub-tags b/i

word recognition capacity of models trained upon it.

1.3. The proposed automatic syntactic category annotation

To overcome these shortcomings, we propose a method that automatically annotates syntactic categories of word constituents, include word, sub-word and character types across word structure trees in Zhang et al. (2013) annotation. Based on work in Chinese morphology (Section 2.1; 2,2), we propose a hypothesis called "one syntactic category per source" as the foundation for the global induction of the syntactic identities of word constituents (Section 2.3). Algorithmically, our method projects the syntax label of the root node to its head components along the head-finding path of word structure trees to achieve such X-bar theory inspired (Section 2.4) syntactic category induction (Section 3, Algorithm 1). This is further complemented by similarity based re-propagation (Section 3, Algorithm 2) to reach a near perfect coverage.

Our annotation is publicly available [2] . The annotated word 发展史 "development history" in Zhang et al. (2013) and our annotation are shown in Figure 2 and Figure 3, respectively. We will show a summary of our annotation and present evaluations on the usefulness of our annotation on two NLP tasks (Section 4). It turns out that the models trained upon our annotation can achieve relative error reduction up to 46% , comparing with that of the Zhang et al. (2013) annotation, which suggests that our annotation is potentially useful for many Chinese NLP tasks.

2. Background and Annotation Hypothesis

2.1. Chinese morphology

Focusing on compounding, morphology of Chinese has been well studied in early works such as (Lü, 1979; Chao, 1968; Zhu, 1985). More recent work (Huang, 1984; Dai, 1992; Duanmu, 1998; Packard, 2000; Xue, 2001; Feng, 2009) are in the framework of generative linguistics. As a representative, Xue (2001) has proposed a system that derives virtually all the complex words with syntactic rules or with the morphology module after syntactic analysis. The boundary of syntax and morphology further blurs and the operation scope of syntax rules expands most part of the morphology. Our annotation largely follows the analyses in Xue (2001).

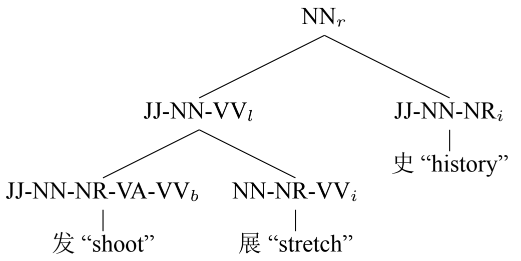


Figure 3: The word structure tree for word 发展史 "development history" in our annotation. It is tagged with automatically induced syntactic categories for word constituents. The category for the subword (also a word itself) 发展 "develop/development/developmental", for example, reflects its syntactic distribution in Chinese: it can occur as the nominal, verbal and adjective part in a phrase or a word. We use the same subtags as Zhang et al. (2013).

2.2. Poor inflection and distribution based lexical categorization

The fact that Chinese words are poor in inflections makes it harder to determine their lexical categories. Semantically related words that have the same root but distinct forms in English, such as "develop", "development" and "developmental" usually map to the same word form in Chinese, which is 发展 in this case. One could stick to the lexical category scheme in English and argue that there are actually three 发展 "develop/development/developmental", which correspond to the English words, respectively. Then use zero derivation to describe the derivation [3].

But since there is neither form difference nor significant meaning shift, it is more natural to treat 发展 "develop/development/developmental" as an unified word type. As many Chinese words can occupy different syntactic slots just like 发展 "develop/development/developmental", one should embrace multiple function of words and categorize words according to their syntactic distribution (Zhu, 1985; Lu, 2005) rather than spitting each multiple-function word into distinct single-function ones. For instance, 发展 "develop/development/developmental" belongs to a lexical category that can be described as $[+nominal, +verbal, +adjective, -other]$.

Zhang et al. (2013) uses the single-function, splitting scheme, following CTB treebank. Their word structure annotation is thus for each (word, POS tag) pair. In contrast, we follow the linguistics research to acknowledge multiple functions of words and categorize words based on their syntactic distributions. Our actual annotation is based on two approximations:

- we collapse (word, POS tag) pairs that share the same word form to a single word type. This may mix homographs together, but given that homographs only cover a small portion of the lexicon, the adverse effect is limited.

[2] http://www.sfs.uni-tuebingen.de/∼jma/word_str.txt

[3] Even for zero derived words in English, e.g. 'hammer (noun)' to 'hammer (verb)', there are (inflectional) form differences (e.g. 'hammered' indicates a verb) to distinguish two word types.

- For each word, we use the the POS tag distribution in the CTB, represented as a 0-1 vector, to approximate the syntactic distribution. For example the syntactic category of ' 发展' is denoted as NN-VV-JJ, which indicates this word has been tagged as Noun, Verb and Adjective, but not other tags in CTB.

To further demonstrate the impacts of the above two approximations, in Figure 4, we show our word structure annotation of the word 开幕 "inaugurate/inauguration", the Zhang et al. (2013) annotation of which is shown in Figure 1.

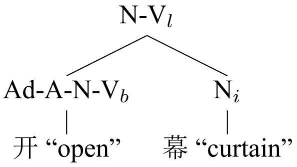


Figure 4: The word structure tree for 开幕 "inaugurate/inauguration" in our annotation. In contrast to the Zhang et al. (2013) annotation in Figure 1, this work annotates the same word 开幕 "inaugurate/inauguration" as one unified lexical entry. Moreover, each character (and word constitute in general) is annotated with its full set of syntactic functions (separated by "-"), which better reflects the combinatorial patterns among different categories.

2.3. One category per source

As the syntactic distribution based categorization is equally applicable to word constituents, characters and sub-words, the treatment of lexical category for word in the previous section can be generalized to word constituents as well. This leads to a hypothesis called "one category per source": for a constituent type, the syntactic category should be the same for each token (occurrence) of this type across word structure trees, which is likely to hold for the majority of constituents and is closely related to (Gale et al., 1992; Yarowsky, 1993) in semantics. An anaphora in syntax is that if a phrase is labeled as NP in a sentence, it is almost always a NP, when it occurs in other sentences.

Based on this hypothesis, we maintain a global *word constituent to syntactic category mapping*, $C2C$, during our automatic annotation. As one constituent type is supposed to have one category, we aggregate all the syntactic labels that have been assigned to its different tokens to get its final syntactic category.

2.4. X-bar theory and syntactic label propagation

Inspired by X-bar theory (Chomsky, 1970; Jackendoff, 1977), we hypotheses that the syntactic category of the word (XP for word structure tree), as well as the intermediate constituents that cover the head (X'), is projected from the syntactic category of the head character (X). The syntactic categories of the head character and head-containing sub-words are thus bounded to comply with the syntactic category of the word.

Our annotation uses a process called syntactic label propagation, which processes each word structure tree to propagate the syntactic label of each word, which is obtained by the process in Section 2.2., to these head constituents, utilizing the manual head annotations. Using the C2C mapping described in Section 2.3., we have an aggregated syntactic category of each constituent type that have occurred along head-finding paths. And the following factors further boost the coverage of the syntactic label propagation:

- Besides left-headed and right-headed structure, in Chinese word structure trees, there are substantial amounts of coordinate structures, in which both branches are head-containing.
- Most characters in Chinese are also single-character words, the syntactic categories of which can be directly passed to the characters.

For those constituents that are not covered by above process, we use a similarity-based re-propagation to induce their syntactic categories, which will be described in detail in Section 3..

3. The Annotation Algorithm

Given the annotation hypothesis in Section 2, we design the following algorithms that automatically refine the syntactic categories of annotated Chinese word structures in Zhang et al. (2013). Following the idea in Section 2.2., we first induce new syntactic category for each word type, which is simply the vector that represents the set of POS tags that has been co-occurred with the word. Then we apply **Algorithm 1**, an implementation of ideas in Section 2.4., to propagate the syntactic categories of each word in a top-down manner along the head-finding path of each word structure tree, using the map C2C as defined in Section 2.3..

For those word constituents that are not touched by Algorithm 1, we apply **Algorithm 2**, to propagate syntactic labels from similar tree node. A node A1 is said to be similar to A2, if both of them have occurred at the A position in following two rule templates (1) $X \to AB$ or $X \to BA$, in which the LHS and the sibling of A on the RHS are constant; and (2) $A \to LR$, in which the non-terminals L and R on the RHS are fixed. The rule templates describe similar rules that have occurred in the derivation of word structure trees in the annotation.

Algorithm 1 Syntactic label propagation

```
for Tree T in Forest do
    $tagset$ ← C2C[T.tags]
    S.push(T)                         ▷ Stack S keeps the all sub-trees
    while S is non-empty do
        C=S.pop()                     ▷ C is the current (sub)tree
        C2C[C.tags] = C2C[C.tags] +tagset
        if C is a non-terminal then
            head_children=C.get_head_children()
            for Child in head_child do
                S.push(Child)
            end for
        end if
    end while
end for
```

After running Algorithm 1 and 2, the syntactic categories of 99.9% constituents have been annotated, leaving only 177

Algorithm 2 Similarity-based re-propagation

```
for each T(ree) in Forest do
    for each subtree $C$ that $C.tags$ not in $C2C$ do
        collect rule $C.parent \rightarrow CC.rsibling$
        collect rule $C \rightarrow C.lchildC.rchild$
        FindSimTree($C.parent, C.rsibling,$
$C.lchild, C.rchild$)
    end for
end for
function FindSimTree(X, B, L, R)
    find set $S$ of nodes $A$ in Trees in Forest such that:
    $X \rightarrow AB$
    $A \rightarrow LR$
    return $S$
end function
```

constituents unannotated, for which we simply inherit the syntactic category of the word.

4. Annotation Result and Evaluation

The summary of annotated word structure are in Table 1. As analyzed in Section 2.4., the word constituent coverage of syntactic label propagation (Algorithm 1) is very high, and the similarity based re-propagation (Algorithm 2) covers almost all the remaining unannotated constituents. Table 2 shows the statistics of the binarized context-free grammars that are extracted from the original annotation and our annotation, respectively. While our annotation greatly enrich the grammar, a potential drawback is the increased size of the grammar, which we leave for future work.

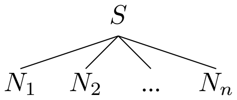
num. of words been annotated	36850
const. coverage Algorithm 1	94.3%
const. coverage Algorithm 2	99.9%

Table 1: Statistics of our word structure annotation

grammar from our annotation	
Num of rules	35415
Num of binary rules	**25989**
Num of non-terminals	1768
Num of pre-terminals	1282
grammar from original annotation	
Num of rules	20612
Num of binary rules	606
Num of non-terminals	177
Num of pre-terminals)	81

Table 2: Comparison of the Grammar size extracted from two annotations

In order to evaluate the effectiveness of our annotation, we have designed experiments on two word-structure parsing based NLP tasks, which are Chinese word segmentation (CWS) and joint Chinese word segmentation and POS tagging (joint seg-tag), respectively. In both tasks, a sentence is represented as a flat tree as the following one, in which S is the sentence node and each N_i is the word structure tree that covers i_{th} word W_i.



The difference of the tasks lies in the definition of subtree N_i, as shown in Table 3. Note that our annotation is for each word type W, while Zhang et al. (2013) annotation is for each word, POS tag pair (W, T) . The two tasks are similar to the standard ones in the literature such as (Ng and Low, 2004), except that they are based on word-structure parsing. Two word structure annotations are plugged in the training corpus according to scheme in Table 3, then parsers [4] are trained based on the PCFG grammar extracted from such flat-tree sentence level annotations and used to parse the testing corpus in the form of raw character sequence. We use the CTB corpus with standard splitting of data (Jiang et al., 2008). The parsing results are then converted into word-segmentation results by extracting characters covered by each node directly under S to a word.

grammar	CWS	joint seg-tag
original	tree of (W_i, T')	tree of (W_i, T_i)
our	tree of W_i	$[T_i$[tree of W_i]]

Table 3: Experiment settings matrix w.r.t tasks and grammars. Note T_i is the POS tag for the exact *token* of W_i in the sentence, while T' is the most frequent tag for the type of W_i,

	word segmentation		
	Precision	Recall	F-score
original grammar	0.884	0.884	0.884
our grammar	0.923	0.905	**0.914**

Table 4: Evaluations results on Chinese word segmentation

	joint segmentation and tagging		
	Precision	Recall	F-score
original grammar	0.878	0.898	0.888
our grammar	0.937	0.944	**0.940**

Table 5: Evaluations results on joint segmentation and POS tagging

We evaluate the performance on word segmentation for both tasks, using standard metrics of precision, recall, F-score and OOV-recall (Zhao and Liu, 2010), among which F-score measures overall accuracy. The evaluation results are shown in Table 4 and Table 5. The models trained on our grammar significantly outperform the model trained on the original grammar. In particular, the relative error reduction of F-score on joint word segmentation is **46%** (absolute improvement of 5.2% over 88.8%). These results demonstrate that our annotation is highly effective for NLP tasks.

[4] we use the Stanford Parser with the simple PCFG setting for experiments: http://nlp.stanford.edu/software/lex-parser.shtml

5. Related Work

Zhao (2009) has annotated Chinese word structures as in-word character dependencies, without any part-of-speech tags or dependency labels. Within a framework of discriminative dependency parsing for word segmentation, models using manually annotated character dependencies outperforms the ones with trivial character dependency schemes. This suggests that character dependency-based word structure annotations are useful.

Other works that develop word structure annotations in literature (Li, 2011; Li and Zhou, 2012; Ma et al., 2012; Zhang et al., 2013) all use PCFG style binary trees to represent word structures, as this work does. The first two works use the same annotation, in which words are classified into flat words and non-flat words. They have annotated the structures of non-flat words, which contain productive suffix and/or prefix and cover about 35% of word types in the CTB5 corpus. Ma et al. (2012) have proposed a semi-automatic approach to Chinese word structure annotation, covering more morphological phenomena, including compounding, which is a more popular word formation process. But their work conduct annotations for two-character words only.

The annotation in Zhang et al. (2013) is the only one publicly available so far. They have exploited the word structure annotation in their character-based parsing models, which lead to improvements in word segmentation, POS tagging and constituency parsing. But the drawback of their annotation, as discussed in Section 1 already, is that the categories of the nodes in their word structure trees reflect the word formation process in limited ways. Using syntactically inspired algorithms, our work automatically refines the syntactic categories in these trees. Our refinement can better capture the combinatorial tendencies between categories and leads to improvement in the practical tasks mentioned in Section 4..

6. Conclusion

We have proposed a syntactically-inspired method for automatically refinement of syntactic categories in Chinese word structure trees. The annotation follows Chinese word formation theory such as Xue (2001), which describes compounding and derivation with syntax-like rules. We have developed techniques of distribution based lexical categorization, head character-directed syntactic label propagation and similarity-like re-propagation to fully automate the annotation process. The refined annotation is made publicly available and is expected to benefit many Chinese NLP tasks, especially Chinese word segmentation.

7. Acknowledgments

The research described in this paper has received funding from the European Commission's 7th Framework Program under grant agreement num. 238405 (project CLARA, a Marie Curie ITN). The author would like to thank anonymous reviews, Dr. Dale Gerdemann, Prof. Chunyu Kit and Dr. Scott Martens for helpful comments and suggestions on drafts of this paper.

8. References

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chomsky, N. (1970). Remarks on nominalizationl. *Readings in English transformational grammar*.

Dai, X.-L. (1992). *Chinese morphology and its interface with syntax*. Phd thesis, Ohio State University.

Duanmu, S. (1998). Wordhood in chinese. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, pages 135–196.

Feng, S. (2009). *Interactions between morphology, syntax and prosody in Chinese (2nd Edition)*. Peking University Press.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

Huang, C.-T. J. (1984). Phrase structure, lexical integrity, and chinese compounds. 19(2):53–78.

Jackendoff, R. (1977). *X-bar syntax*. The MIT Press.

Jiang, W., Mi, H., and Liu, Q. (2008). Word lattice reranking for chinese word segmentation and part-of-speech tagging. In *Proc. 22nd Int. Conf. Comput. Linguist. - COLING '08*, pages 385–392. Association for Computational Linguistics.

Li, Z. and Zhou, G. (2012). Unified dependency parsing of chinese morphological and syntactic structures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1445–1454. Association for Computational Linguistics.

Li, Z. (2011). Parsing the internal structure of words: a new paradigm for chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1405–1414. Association for Computational Linguistics.

Lü, S. (1979). Hanyu yufa fenxi wenti, "problems in the analysis of chinese grammar". *Beijing: Shangwu Yinshuguan*.

Lu, J. (2005). *A Course on Modern Chinese Grammar Research (3rd Edition)*. Peking University Press.

Ma, J., Kit, C., and Gerdemann, D. (2012). Semi-automatic annotation of chinese word structure. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 9–17, Tianjin, China, December. Association for Computational Linguistics.

Ng, H. and Low, J. (2004). Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proc. EMNLP*, volume 4.

Packard, J. L. (2000). *The morphology of Chinese*. Cambridge University Press Cambridge.

Xue, N., Xia, F., Chiou, F.-d., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.*, 11(2):207–238, June.

Xue, N. (2001). *Defining and automatically identifying words in Chinese*. Phd thesis, University of Dalaware.

Yarowsky, D. (1993). One sense per collocation. In *Pro-*

ceedings of the workshop on Human Language Technology, pages 266–271. Association for Computational Linguistics.

Zhang, M., Zhang, Y., Che, W., and Liu, T. (2013). Chinese parsing exploiting characters. In *51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 125–134. Association for Computational Linguistics.

Zhao, H. and Liu, Q. (2010). The cips-sighan clp 2010 chinese word segmentation bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 199–209.

Zhao, H. (2009). Character-level dependencies in chinese: Usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 879–887, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhu, D. (1985). *Yu Fa Da Wen, "Questions and Answers on Grammar"*. Commercial Press, Beijing.