Character-level Dependency Annotation of Chinese

Li Yixuan
Université Paris3 - Sorbonne Nouvelle
LPP (CNRS)
yixuan.li@sorbonne-nouvelle.fr

Abstract

In this paper, we propose a new model for annotating dependency relations at the Mandarin character level with the aim of building treebanks to cope with the unsatisfactory performance of existing word segmentation and syntactic analysis models in specific scientific domains, such as Chinese patent texts. The result is a treebank of 100 sentences annotated according to our scheme, which also serves as a training corpus that facilitates the subsequent development of a joint word segmenter and dependency analyzer that enables downstream tasks in Chinese to be separated from the non-standardized pre-processing step of word segmentation.

1 Introduction

Word segmentation has long been a chicken-and-egg problem in Chinese. The notion of distinct words with spaces as natural boundaries in languages using the Latin alphabet has never been widely agreed upon in Chinese. In the absence of both natural delimiters and inflection marks, two main indicators of wordhood (Magistry et al., 2012), the distinction between words and larger lexical units in Chinese has been an unfamiliar and confusing concept since it was introduced by Zhang Shizhao in 1907. This has resulted in a low agreement rate of 76% on lexicality among native Chinese speakers (Sproat et al., 1997).

All existing Mandarin treebanks and syntactic annotation schemes for Mandarin Chinese employ word segmentation as the first step in the annotation. However, their segmentation criteria are far different without a clear unified standard. At the same time, dependency analyzers trained on these treebanks end up with inconsistent results with each other, especially on corpora containing a large number of domain-specific new terms, such as patent texts (Li et al.)

It is in this context that we decided to explore the idea of developing a character-based Chinese annotation schema. A treebank annotated with additional internal relations of words can be used as a resource to train a joint segmenter-parser, combining the two steps into one. Moreover, (Li et al., 2019) also showed that character-level annotations, even coarse ones, can help improve the results of dependency analysis for Chinese of different text types.

In the most widely accepted morphological theory of Chinese (Feng, 1997; Zhang, 2003; Dong, 2011), complex words are derivative words or compound words. The second group includes five types: modifier-head type, coordinative type, predicate-object type, predicate-complement type, and subject-predicate type. He (He et al., 2012) and Chi (Chi et al., 2019) suggest in their work that there is a parallelism between compound word structure and syntactic structure in Chinese, from which from which it is feasible to build a new dependency model that unifies the character level with the current word level relationship. Some previous works have also discussed the possibility of the joint dependency parsing and multi-word expression recognition on other languages (Candito et al., 2014; Nasr et al., 2015).

From this perspective, it is important to integrate the new word internal relations of the new words into dependency trees built on the basis of similar distributional criteria. This is why this work chose to base itself on a variant of UD (Gerdes et al., 2018), Surface-Syntactic Universal Dependencies (SUD), which is a near-isomorphic but more surface syntactic alternative schema of UD with a more classical word distribution-based dependency structure that favors functional heads. And to obtain the relationships between these roles, we applied syntactic tests that allowed us to

Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023), pages 42 - 53
March 9-12, 2023 ©2023 Association for Computational Linguistics

Word internal structure	Examples			SUD
Coordination compounds	价值 国家 查封 始终 明亮 高矮	*jià zhí* *guó jiā* *chá fēng* *shǐ zhōng* *míng liàng* *gāo ǎi*	‘price_N’ + ‘value_N’ = ‘value_N/V’ ‘country_N’ + ‘family_N’ = ‘country_N’ ‘examine_V’ + ‘close_V’ = ‘seize_V’ ‘begin_V’ + ‘finish_V’ = ‘all along_ADV’ ‘bright_ADJ’ + ‘bright_ADJ’ = ‘bright_ADJ’ ‘tall_ADJ’ + ‘short_ADJ’ = ‘height_N’	conj
Modifier compounds	蜂巢 汉字 飞机 火红 深蓝 滚烫 迟到 鼠窜 夜游	*fēng cháo* *hàn zì* *fēi jī* *huǒ hóng* *shēn lán* *gǔn tàng* *chí dào* *shǔ cuàn* *yè yóu*	‘bee_N’ + ‘nest_N’ = ‘beehive_N’ ‘Chinese_ADJ’ + ‘character_N’ = ‘Chinese character_N’ ‘fly_V’ + ‘machine_N’ = ‘plane_N’ ‘fire_N’ + ‘red_ADJ’ = ‘red as fire_ADJ’ ‘dark_ADJ’ + ‘blue_ADJ’ = ‘dark blue_ADJ’ ‘roll(ing)_V’ + ‘hot_ADJ’ = ‘boiling hot_ADJ’ ‘late_ADJ’ + ‘arrive_V’ = ‘be late_V’ ‘rat_N’ + ‘flee_V’ = ‘scamper off like a rat_V’ ‘night_N’ +’ tour_V’ = ‘noctivagation_N/V’	mod
Subject-predicate compounds	目睹 性急 海啸	*mù dǔ* *xìng jí* *hǎi xiào*	‘eye_N’ + ‘see_V’ = ‘witness_V’ ‘temper_N’ + ‘impatient_ADJ’ = ‘impatient_ADJ’ ‘sea_N’ + ‘howl_V’ = ‘tsunami_N/V’	subj
Predicate-object compounds	结果 睡觉 喝水	*jié guǒ* *shuì jiào* *hē shuǐ*	‘bear_V’ + ‘fruit_N’ = ‘bear fruit_V’/ ‘result_N’ ‘sleep_V’ + ‘sleep_N’ = ‘sleep_V’ ‘drink_V’ + ‘water_N’ = ‘drink water_V’	comp:obj
Predicate-complement compounds	请教 推动 说明 来自 可变 书本 雪花	*qǐng jiào* *tuī dòng* *shuō míng* *lái zì* *kě biàn* *shū běn* *xuě huā*	‘ask_V’ + ‘teach_V’ = ‘ask (obj) to teach/consult_V’ ‘push_V’ + ‘move_V’ = ‘push (obj) to move_V’ ‘speak_V’ + ‘clear_ADJ’ = ‘explain_V’ ‘come_V’ + ‘from_ADJ’ = ‘come from_V’ ‘can_AUX’ + ‘change_V’ = ‘changeable_ADJ’ ‘book_N’ + “Classifier” = ‘book_N’ ‘snow_N’ + ‘flower_N’ = ‘snowflakes_N’	comp:obl, comp:pred, comp:aux
Simple words - transliterated words - onomatopoeia - reduplicated words	车 咖啡 叮咚 侃侃	*chē* *kā fēi* *dīng dōng* *kǎn kǎn*	‘car_N’ ‘coffee_N’ ‘ding-dong_Onomatopoeia’ ‘eloquently_ADV’	flat

Table 1: List of Chinese word internal structures with examples and English translation.

identify the head and internal structure of the composite based on distributional criteria.

After discussing Chinese morphology and syntactic theory, the parallelism between Chinese compound word structure and syntactic structure is discussed especially in sections 2 and 3. The next two sections explain the complete annotation process, including two sub-steps. (1) automatic tokenization, POS tagging and dependency parsing using existing NLP pipelines (Section 4.1); and (2) manual correction and annotation follo-wing our SUD-based character-level annotation schema (Section 4.2). Then, Section 4.3 describes the conversion of the character-level treebank to a standard word-level UD treebank and the evaluation of the automatically converted treebank.

2 The Annotation Schema for Chinese Word Internal Relations

Instead of the conventional first step of word segmentation in Chinese treebank annotation, the annotation of character-based treebanks starts with the analysis of the relations between individual characters. Such relations can be typical syntactic relations, internal relations of words that do not conform to any syntactic relations in modern Chinese, or the third between the two kinds of relations mentioned above, which are more frozen than independent syntactic constituents but still largely corresponds to certain syntactic structures

In this section, we explain the criteria for wordhood and parts-of-speech. Our model annotates these relations between characters at all three levels of granularity simultaneously. Without the word segmentation process, all characters of a sentence are separated. Moreover, the word level is distinguished from the syntactic level by the sub-relation ":m", instead of the blank. The criteria for this distinction are described in Section 2.1, and the next Section 2.2 explains the choice of part-of-speech labels, especially at the character level.

2.1 Wordhood and word boundaries

One of the most widely used Chinese word segmentation standards is the Penn Chinese

Treebank (3.0) Segmentation Guidelines (Xia, 2000a). The guidelines introduced the concept of "word" based on the smallest syntactic unit, which was largely followed by the later UD Chinese Hong Kong Treebank (Poiret et al., 2021). The guidelines provided for Bakeoff 2005 is another applicable standard and includes a table summarizing decisions for a range of difficult cases. Magistry (2017) summarizes all the different segmentation guidelines to discuss wordhood in Chinese and defines a word as “the smallest sequence of autonomous characters to which we can attribute at least one word class”. Kratochvíl (1967) first proposed a more syntactic definition-based approach, which was later developed by (Huang, 1984; Duanmu, 1998; Packard, 2000; Nguyen, 2006). This approach proposes a set of widely applicable linguistic criteria to test whether a sequence of characters can be considered as a word: (1) Conjunction Reduction, (2) Freedom of Parts (3) Semantic Composition, (4) Exocentric Structure, (5) Adverbial Modification, (6) XP Substitution, (7) Productivity Criterion, (8) Syllable Count and (9) Insertion.

In this work, we mainly follow the provided test set, focusing on the independence criterion, the productivity criterion, and the presence of part-of-speech variation when the expression is used as a word (see Section 4.2 for a detailed analysis).

(1) 喝了。 给我水。
hē le *gěi wǒ shuǐ*
‘(I) drank.’ ‘Give me water.’

Taking the three predicate-object compounds in Table 1 as an example, both 结果 (*jié guǒ* ‘result’) and 睡觉 (*shuì jiào* ‘sleep’) are considered as words, while 喝水 (*hē shuǐ* ‘drink’) is a syntactic unit, since all characters in the latter can be used independently as a word, as follows. Therefore, in our work, structures considered as words are annotated as purely syntactic relations, such as A-not-A (e.g., 来不来 *lái bù lái* ‘come or not come’).

2.2 Choices for parts-of-speech tags

Whether the parts-of-speech are based on meaning or syntactic distribution has long been a central issue in POS tagging (Xia, 2000b). Since almost all Chinese characters have multiple parts of speech and have neither delimiters nor inflection marks, which are the two main indicators of languages using the Latin alphabet (Magistry et al., 2012), the distinction between different parts-of-speech is mainly indicated by the distribution position. Therefore, instead of considering semantics, we placed the choice of part-of-speech labels on distributional position in both word-level and character-level annotations.

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Table 3: List of UD POS tags.

Based on an automatic POS tagging described in Section 3.1, we manually correct the results referring to the Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) (Xia, 2000b) for the word-level and to Xinhua Dictionary for the character-level. Especially, as the choices for POS tags and for the word internal relation labels on characters are being made simultaneously, the relation type has a heavy influence on the POS choice, which is discussed in Section 3.

Based on the automatic POS tagging described in Section 3.1, we manually corrected the results on word-level by referring to the Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) (Xia, 2000b) and the character-level by referring to the Xinhua Dictionary. In particular, since the selection of POS tags and internal relation labels of words is going to be done simultaneously, the type of relationship has a great influence on the selection of POS, which will be discussed in Section 3.

And one method often used to identify its POS during annotation is to test whether a character can be combined with a functional character specifically reserved for a particular POS (see Table 2 for details).

This schema retains all 17 part-of-speech (UDPOS) tags of UD[1] (Nivre et al., 2016) in Table 3.

3 Correspondence between Chinese word internal structures and dependency relations in SUD

The annotation of syntactic relations is based on the Surface-Syntactic Universal Dependency (SUD) model proposed by (Gerdes et al., 2018). And based on this, we added our own character-level annotation labels by analogy with the surface-syntactic relations of SUD.

In this section, we introduce six categories of character-level relations in modern Chinese vocabulary. For each category, we describe the definition of a category, and its correspondence with the syntactic relations in SUD, and give some criteria to test whether a compound belongs to a certain category (see the full decision tree in Appendix A).

3.1 Coordination compounds

Coordinated compounds are composed of two or more morphemes that are usually synonymous, antonymic, or semantically related. The meaning of a compound can be a combination of its morphemes, completely independent of the meaning of its components, or inclined to one of its characters.

In terms of POS, a coordinating complex can consist of the following components:

(1) Two nouns characters: N1 + N2

In Table 1, 价值 (*jià zhí* 'value') and 国家 (*guó jiā* 'country') are two examples of this subcategory. In 价值 the two characters are synonyms and the meaning of the compound is the their synthesis, while in 国家 (*guó jiā* 'country') the meaning is inclined to 国 (*guó* 'country').

(2) Two verbs characters: V1 + V2

Among examples of this subcategory, 查封 (*chá fēng* 'to seize'), which consists of a sequence of two verbs is itself a verb and 始终 (*shǐ zhōng* 'all along'), which consists of a pair of antonyms is usually used as an adverb.

(3) Two adjectives characters: A1 + A2

Similar to subcategories (1) and (2), the external POS of the compound can be the same as (e.g. 明亮 *míng liàng* 'bright' is an adjective) or different to (e.g. 高矮 *gāo ǎi* 'height', the external POS of the word as a whole is a noun while both characters are adjectives).

All coordination structures are considered as "conj" relations in SUD, with the edges oriented from left to right. Which means the first character of a coordination compound is always the head in its internal relation.

We proposed a set of tests to decide whether a compound word "AB" can be assigned to each of the three subcategories of coordination: whether "AB" can be extended[2] into a "先 A 后 B (*xiān A hòu B* 'first A and then B')" structure or "边 A 边 B (*biān A biān B* 'A while B')" structure for subcategory (2), and into a "A 而不 B (*A ér bù B* 'is A but not B')" structure or "又 A 又 B (*yòu A yòu B* 'not only A but also B')" structure for subcategory (3). As for subcategory (1), the two noun characters can usually be extended into "A 和 B (*A hé B* 'A and B')".

3.2 Modifier compounds

A common modifier compound may consist of two or three characters. In the first case, a term AB, where A (or the modifier) modifies B (the head, which can be a noun, an adjective or a verb). In the example of Table 1:

- 蜂巢 (*fēng cháo* 'beehive'), 汉字 (*hàn zì* 'Chinese character') and 飞机 (*fēi jī* 'plane') all have a nominative center character. However, the modifier can also be a noun (as in the first word 蜂巢), an adjective (as in the second word 汉字) or a verb (as in the third word 飞机).

- Modifier compounds with an adjective center like the noun-centered compounds above. The modifier can be a noun (e.g. 火红 *huǒ hóng* 'red as fire'), an adjective (e.g. 深蓝

[1] In regard to the specificity of the patent writing style, INTJ dose not actually appear in the final treebank.

[2] The notion "expend" here allows to add extra characters if necessary (examples are given in following subsections) , considering the disyllabification in modern Chinese. In this case, the single character can be seen as a shortened form of the disyllabic term.

shēn lán 'dark blue') or a verb (e.g. 滚烫 *gŭn tàng* 'boiling hot').

- 迟到 (*chí dào* 'be late') is an example with a verbal center character and an adjective/adverb modifier, while noun characters can also be used as a modifier of a verbal character as shortened forms of oblique structures such as "V as N", "V with N", "V towards N", etc.

And in the second case, a term ABC, where AB together modifies C (the center character). In contrast to the variety of POS of bisyllabic modifier compounds, most trisyllabic modifier compounds are nouns, where the central character is also considered as a suffix in some works, according to its productivity.

The syntactic head (center character) of a modifier compound is always its last character, and the external POS of the entire term is always the same as the POS of its head character. Compound words in this group are annotated with a "mod" label and the direction of the edge runs from right to left..

For modifier compounds, the tests set includes a check for presence or absence of the following deformations (example (2)):

1. Possible expansion with 的/地/得 *de* [3], e.g. 蜂巢 (*fēng cháo* 'beehive') can be extend-ed into 蜜蜂的巢 (*mì fēng de cháo* 'hive of bee'), where 蜂 (*fēng* 'bee') stands for 蜜蜂 (*mì fēng* "honeybee", which is itself a modifier-head compound with 蜂 *fēng* 'bee' as its head).
2. Paradigm of the head character, such as the productive character 巢 (*cháo* 'nest') can combine with 蜂 and 鸟 (*niăo* 'bird').
3. Possible expansion into a corresponding phrase for those with a verbal center character, e.g. 鼠窜 (*shŭ cuàn* 'scamper off like a rat') is expended into 像鼠一样窜 (*xiàng shŭ yī yàng cuàn* 'scamper off like a rat'); 夜游 (*yè yóu* 'noctivagation') is expended into 在夜里游 (*zài yè lĭ yóu* 'tour in the night').

(2)	蜂巢	鸟巢
	fēng cháo	*niăo cháo*
	'beehive'	'bird nest'

3.3 Subject-predicate compounds

In subject-predicate compounds, similar to modifier compounds, the head character is always the last character, which is either a verb (e.g. 目睹 *mù dŭ* 'wittiness', 海啸 *hăi xiào* 'tsunami') or an adjective (e.g. 性急 *xìng jí* 'impatient'), while the first character is a noun, which serves as the subject of the head character. Unlike modifier compounds, the external POS of a subject-predicate compound does not always correspond to the POS of the head character.

The subject-predicate structure is annotated as "subj", with the marginal direction running from right to left.

Together with the predicate-object compounds and predicate-complement compounds, the test for the latter three types is that at least one character in the compound can have one of the aspect markers 了 (*le*) /着 (*zhe*) /过 (*guo*) without changing the meaning. This means that the character can only be a verbal character. The subject-predicate compounds differ from the other two in that they have only one verbal character in the second position, and their first character can be modified by the noun modifying particle '的 (*de*)' without a change of meaning, which means that this first character is a nominal character. In the example of 海啸 (*hăi xiào* 'tsunami'), it is possible to say 啸着 (*xiào zhe* 'is howling') and ADJ 的海 (*ADJ de hăi* 'ADJ sea').

3.4 Predicate-object compounds

In contrast to the subject-predicate structure, the first character in a predicate-object compound is the head character and the second character is the direct object of the verbal head character on the first position. This second character is usually a noun character (e.g. 结果 *jié guŏ* 'result', 睡觉 *shuì jiào* 'sleep', 喝水 *hē shuĭ* 'drink water').

The predicate-object structure is considered equivalent to the "comp:obj" relationship in SUD with a left-to-right edge.

[3] 的/地/得 *DE* are noun modifier particle, adjective modifier particle and verb modifier particle in Chinese.

Unlike the subject-predicate compounds, the predicate-object and predicate-complement compounds have a verbal head character in the first position. Although they are both annotated as "comp", the predicate-object compounds always have a noun character on the second position, while there is usually no nominal character in the second position of the predicate-complement compounds.

3.5 Predicate-complement compounds

There are two types of predicate-complement compounds. The first type can be compared to predicate-complement structure at the syntactic level: the first character of predicate-complement compounds is a verbal head character, which is similar to the predicate-complement structure, and its second character is a verbal or adjective character that acts as a resultative or directional complement of the head character in the first position.

A predicate-complement compound is always a verb and has a "comp"-like internal relationship marked as different types of sub-relations in SUD, such as "comp:obl" (for oblique arguments of verbs, adjectives, adverbs, nouns or pronouns, e.g. 来自 *lái zì* 'come from'), "comp:pred" (for predicative arguments of verbs, e.g. 请教 *qǐng jiào* 'consult', 推动 *tuī dòng* 'push (obj) to move') and "comp:aux" (for the argument of auxiliaries, e.g. 可变(*kě biàn* 'variable'), and corresponds to the "aux" relationship defined by UD). In this version of annotation, all sub-relations of predicate-complement are simply annotated as "comp".

The second type has a noun head character in its first position and in its second position a classifier (e.g. 书本 *shū běn* 'books') or a second noun character indicating the category or form of the first noun character (e.g. 雪花 *xuě huā* 'snowflakes'). The external POS of a compound of this type is always noun. This type can be easily identified by the presence of a classifier as the character in the second position.

3.6 Non-compound words and terms with unclear internal structures

In addition to the compound words in modern Chinese, there are also words that contain multiple characters but whose internal structure does not directly correspond to the syntactic relationships in modern Chinese, such as polysyllabic monograms, transliterated words, and onomatopoeic words. We borrowed the tag "flat"[4] from the UD/SUD schema, and established the corresponding character-level relationship "flat:m" for them.

Note that the subclass ":m" is specifically designed for relationships between Chinese characters. Thus, transliterated words using Chinese characters are marked as "flat:m", but foreign words are always marked as "flat".

Another point is that our annotation schema no longer contains the confusing label "compound". In the original UD schema, "compound" relations contained noun-noun compounds, verb and verb-object compounds (subdivided into "compound: dir", "compound:ext", "compound:vo" and "compound:vv"), and their boundaries with "nmod", "scomp", "xcomp" and their word segmentation boundaries are not very clear.

4 Construction and annotation of the Character-based Chinese Patent Tree-bank

To apply this annotation schema in a real corpus, we chose patents, one of the most challenging genres for syntactic parsing tasks due to their syntactic complexity and frequent use of uncommon domain-specific terms.

4.1 Collection of the data and automatic an-notation

We built the Chinese patent treebank by randomly selecting 100 sentences of patent claims from November 2017 to September 2018[5], which have been segmented to reduce the length of the sentences[6] In addition to line breaks, the ";" and ":" are also segmented. The shortened sentences were then split on individual characters as shown in Figure 1.

The obtained character-level treebanks were first automatically annotated with (1) word segmentation[7], (2) POS tags and (3) dependency analysis, based on votes from three state-of-the-

[4] The label "flat" is used to link names without internal structure in UD and SUD annotation.

[5] All patents are collected from the official site of CNIPA (China National Intellectual Property Administration, former SIPO): http://patdata1.cnipa.gov.cn/

[6] A Chinese patent claim sentence contain between 50 and 70 characters in average, which is extremely long compared to general texts, and even harder to parse.

[7] The results of word segmentation are present by ":m" on the relation labels.

art language processing pipelines. spaCy[8], Stanza[9] and Trankit[10].

Like the automatic annotation method for characters, the automatic POS tagging at the word level is based on a poll of three language processing pipelines. Unlike character-level annotation, the one single label we reserve for word-level annotation is the part of speech for each single character, which is saved as an external POS ("ExtPos") for the character combination.

These automatically annotated sub-relations ":m" and POS tags are later manually corrected according to the criteria described in Section 2.

4.2 Problematic cases

Cases of disagreement among annotators in annotated patent claim sentences can be divided into three main types: (1) compounds containing functional characters, (2) compounds involving resultative complements, and (3) compounds with obscure internal relation.

- Functional characters

Compound words containing functional characters or classic Chinese structures are usually highly frozen terms. However, many of them have a large paradigm.

(3) 之前 *zhī qián* 'before' 之后 *zhī hòu* ’after'

之间 *zhī jiān* 'between' 之内 *zhī nèi* 'within'

(4) 以前 *yǐ qián* 'before' 以后 *yǐ hòu* 'after'

以来 *yǐ lái* 'since' 以内 *yǐ nèi* 'within'

As a literal replacement for 的 (*de* 'PART indicating pre-modification') in Chinese, 之 (*zhī* 'PART') is usually combined with a positional word, such as in the example (3) below. These words containing 之 (*zhī* 'PART') are considered as a single word in the Penn Segmentation guidelines. Taking the change of part-of-speech we also annotated the relations in the terms 之前 (*zhī qián* 'before') and 之后 (*zhī hòu* 'after') as internal relations of words labeled as "mod:m", although each character is independent,. While 之间 (*zhī jiān* 'between') and 之内 (*zhī nèi* 'within') are annotated as the syntactic relation "mod".

(5) 其中 *qí zhōng* 'among (them)' 其间 *qí jiān* 'between (them)'

其实 *qí shí* 'in fact'

The other two problematic function words 以 (*yǐ* 'ADP') and 其 (*qí* 'PRON') can also be combined with positional words such as 之 (*zhī* 'PART'). Under the same criteria, all four expressions in example (4) are annotated as word internal relations "comp:m", in which 以 is the head. And 其中 (*qí zhōng* 'among (them)'), 其间 (*qí jiān* 'between (them)') and 其实 (*qí shí* 'in fact') in example (5) are labeled as "det:m", in which 其 (*qí* 'PRON') as the dependent character.

所 *(suo3* 'PART') is a extremely productive character used in the "所+VERB" structure and often found in patent claims, can 所 *(suo3* 'PART') can be seen as a function word capable of converting VERB into an ADJ-liked unit. Evolving from the ancient 所 structure, 所 *(suo3* 'PART') is sometimes omitted in modern Chinese (especially in

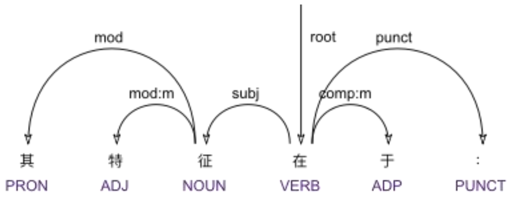

Figure 1: An example of character-based SUD Chinese Patent treebank.

[8] https://spacy.io/

[9] https://stanfordnlp.github.io/stanza/

[10] https://trankit.readthedocs.io/en/latest/

spoken language). In our schema, 所 *(suo3* ‘PART’) is considered as the head character of the structure and the relation is systematically annotated as “comp:m” relation in the Chinese patent treebank.

(6) 所述
suǒ shù
‘said’

所以
suǒ yǐ
‘because’

第一 章 中 (所) 描述 的 方法
dì yī zhāng zhōng (suǒ) miáo shù de fāng fǎ
‘The method described in chapter 1’

The last remarkable problematic function word is 自 (*zi4* ‘self’). As with 所 *(suo3* ‘PART’), 自 (*zi4* ‘self’) is always combined with VERB to form a self-reflexive verbal expression, in which the pronoun 自 (*zi4* ‘self’) acts as the subject and the object at the same time, e.g. 自测 (*zì cè* ‘self-evaluate’) in example (7) is equivalent to 自己测试自己 (*zì jǐ cè shì zì jǐ* ‘one evaluate himself’). This structure is annotated as “subj:m” with 自 as the dependent. A special case is the word 自由 (*zì yóu* ‘free; freedom’), which is too frozen that it is difficult to observe the syntactic-liked structure, and is thus annotated as “flat:m” like compounds with obscure internal relation.

(7) 自测
zì cè
‘self-evaluate’

自由
zì yóu
‘free; freedom’

- Resultative complements

The resultative complements can be seen as a single word by itself or as part of a VERB-complement compound (Xia, 2000a), depending on the segmentation criteria.

We adopt the test proposed by Xia based on syllable count in and segment it only if the verb or the complement have more than 2 syllables or the complement is the finished aspect mark 完 (*wán* ‘finish’). In (8) only 浸没 (*jìn mò* ‘submerged’) is remained unsegmented and is annotated as “comp:m” structure, while 连接至 (*lián jiē zhì* ‘connected to’) is segmented into 连接 (*lián jiē* ‘connect’) and the adposition 至 (*zhì* ‘ADP’), 配置有 (*pèi zhì yǒu* ‘configured with’) is segmented into 配置 (*pèi zhì* ‘configure; configuration’) and 有 (*yǒu* ‘have’) and 清干净 (*qīng gān jìng* ‘clean up’) into 清 (*qīng* ‘clean_V’) and 干净 (*gān jìng* ‘clean_ADJ’) with the syntactic label “comp”.

(8) 浸没
jìn mò
‘submerged’

连接 至
lián jiē zhì
‘connected to’

配置 有
pèi zhì yǒu
‘configured with’

清 干净
qīng gān jìng
‘clean up’

- Obscure internal relation

This type involves those compounds usually highly frozen and whose internal structure is not obvious anymore, just like the example of 自由 discussed above.

Other examples in (9) are 根据 (*gēn jù* ‘according to; proof’) and 作用 (*zuò yòng* ‘effect; function’). 根据 is hard to label due to the ambiguity: the structure can be interpreted as “root proof” or “root occupies”. According to the preference to distributional standards of the annotation schema, the first structure is chosen so that the external POS is same to that of the head character 据 (*ju4* ‘occupy; proof ’). As for 作用 (*zuo4 yong4* ‘effect; function’), the choice of relation label is between “comp:m” and “conj:m” as it dose not correspond to any of the tests of them. It is simply annotated “flat:m” instead to avoid a tedious study on the etymology.

(9) 根 据
gēn jù
root occupy; proof
‘according to; proof’

作 用
zuò yòng
compose use
‘effect; function’

4.3 Convertibility

The character-based Chinese treebank can be easily converted to a standard word-based treebank by simply combining all relations with the sub-relation ":m". The part-of-speech of the merged words is used as the external POS annotated on the head characters of the compound words.

The conversion from the SUD schema to the original UD schema is performed by Grew Match following the method proposed in (Gerdes et al., 2018).

The UD version of the treebank is released on https://github.com/UniversalDependencies/UD_Chinese-PatentChar.

5 Conclusion and Future Works

In this paper, we propose a new character-based Chinese annotation model. Instead of starting with non-standardized, information-wasting word segmentation, we analyze the word internal structures and distribute syntactic-liked labels based on the parallelism between compound word structure and syntactic structure in Chinese. Finally, we annotated the first character-level tree database consisting of 100 patent claim sentences.

Based on this character-level treebank, we have the possibility to train a character-based dependency analyzer by bootstrapping that can handle both word segmentation and syntactic analysis simultaneously.

In future work, we are also interested in developing a premoderne Chinese treebank containing a richer character-level structure[11].

Acknowledgments

I gratefully acknowledge the financial support for my doctoral study provided by the China Scholarship Council (csc).

References

Marie Candito and Matthieu Constant. 2014. Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference.

Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor Variety Criteria For Chinese Word Extraction. Computational Linguistics, 30:75–93.

Chi Changhai and Lin Zhiyong. A New Discussion on the Parallelism Between Compound Word Structure and Syntactic Structure in Chinese [J]. Journal Of Zhejiang University (Hu-Manties And Social Sciences), 2019, 49(5): 210-223

Chen Gong, Zhenghua Li, Min Zhang, and Xinzhou Jiang. 2017. Multi-grained Chinese Word Segmentation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 692–703, Copenhagen, Denmark. Association Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An Annotation Scheme Near-isomorphic to UD. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Kim Gerdes and Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 131–140, Berlin, Germany. Association for Computational Linguistics.

Chen Gong, Zhenghua Li, Min Zhang, and Xinzhou Jiang. 2017. Multi-grained Chinese Word Segmentation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 692–703, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Gong, Zhenghua Li, Bowei Zou, and Min Zhang. 2020. Multi-Grained Chinese Word Segmentation with Weakly Labeled Data. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2026–2036, Barcelona, Spain (Online). International Committee on Computational Linguistics

Xia Fei. 2000a. The Segmentation Guideliens for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive

[11] The word-level UD premoderne Chinese treebank is released on Removed for anonymous submission

Science Technical Report No. IRCS-00-06, http://repository.upenn.edu/ircs_reports/37/

Xia Fei. 2000b. The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-07, http://repository.upenn.edu/ircs_reports/38/

Harbin Institute of Technology Research Center for Social Computing and Information Retrieval (哈尔滨工业大学信息检索研究中心) [HIT-SCIR]. 2010. HIT-CIR Chinese Dependency Treebank Annotation Guideline (HITCIR 汉语依存树库标注规范).

Yang He and Yanlei Cui. The Similarities and Differences of Chinese Compound Word Structure and Syntactic Structure and Their Roots [J] (汉语复合词结构与句法结构的异同及其根源). Language Studies, 2012 (1): 6

Paul Kratochvíl. 1967. On The Phonology Of Peking Stress. Transactions of the Philological Society, 66(1):154–178.186

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12), pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Yixuan Li, Chuanming Dong, and Kim Gerdes. 2019. Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies. In Depling 2019 - International Conference on Dependency Linguistics, Paris, France.

Pierre Magistry. 2013. Unsupervised Word Segmentation and Wordhood Assessment.

Pierre Magistry and Benoît Sagot. 2012. Unsupervized word segmentation: the case for Mandarin Chinese. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 383–387, Jeju Island, Korea. Association for Computational Linguistics. 200

Alexis Nasr, Carlos Ramisch, José Deulofeu and André Valli. 2015. Joint Dependency Parsing and Multiword Expression Tokenisation.

Étienne Van Tien Nguyen. 2006. Unité lexicale et morphologie en chinois mandarin: vers l'élaboration d'un dictionnaire explicatif et combinatoire du chinois.

Jerome Lee Packard. 2000. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, United Kingdom.

Richard Sproat. 1990. A Statistical Method For Finding Word Boundaries in Chinese Text. International Journal of Computer Processing of Languages, 4:336–351.

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pages 133–143, Sapporo, Japan. Association for Computational Linguistics.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1994. A Stochastic Finite-State Word- Segmentation Algorithm for Chinese.

Andi Wu. 2003. Customizable Segmentation Of Morphologically Derived Words in Chinese. Int. J. Comput. Linguistics Chin. Lang. Process., 8.

Nainwen Xue, Xia Fei, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207–238.

Nianwen Xue, Xiuhong Zhang, Zixin Jiang, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2013. Chinese Treebank 8.0 LDC2013T21. Linguistic Data Consortium, Philadelphia, https://catalog.ldc.upenn.edu/ldc2013t21

A Decision Tree for Word Internal Rela-tion Labeling

Figure 2 shows the complete decision tree for word internal relation annotation. The criteria are mostly distributional with some semantic test in addition, such as whether the two character are synonym/antonyms. The synonym/antonyms here are strictly limited to polar antonyms (大 *da4* 'big' and 小 *xiao3* 'small') and coordinated structure like 春夏秋冬 (*shun1 xia4 qiu1 dong1* "four seasons").

B Comparison of the character-level Chinese treebank to the SUD and UD word-level treebank

And here is a comparison between the character-based treebank (Figure 3), the SUD word-based treebank (Figure 4) and the UD word-based tree-bank (Figure 5) of the same sentence in Chinese.

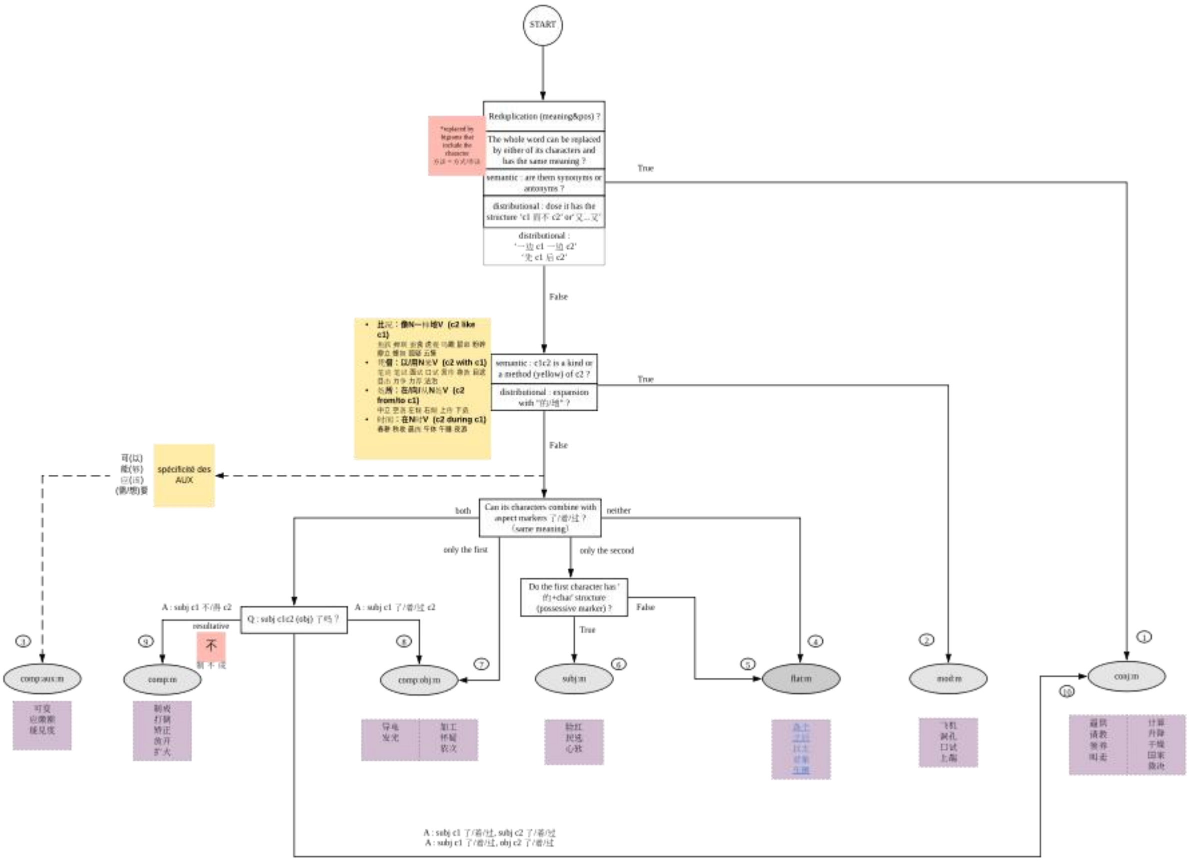
START
Reduplication (meaning&pos) ?
The whole word can be replaced by either of its characters and has the same meaning ?
semantic : are them synonyms or antonyms ?
distributional : dose it has the structure 'c1 而不 c2' or '又...又'
distributional : '一边 c1 一边 c2' '先 c1 后 c2'
True
False
semantic : c1c2 is a kind or a method (yellow) of c2 ?
distributional : expansion with "的/地" ?
True
False
像N一样地V (c2 like c1)
以/用N来V (c2 with c1)
在/向/从N出V (c2 from/to c1)
在N时V (c2 during c1)
spécificité des AUX
可(以)
能(够)
应(该)
(需/想)要
Can its characters combine with aspect markers 了/着/过 ? (same meaning)
both
neither
only the first
only the second
Do the first character has '的+char' structure (possessive marker) ?
True
False
Q : subj c1c2 (obj) 了吗 ?
A : subj c1 不/得 c2
resultative
不
A : subj c1 了/着/过 c2
A : subj c1 了/着/过, subj c2 了/着/过
A : subj c1 了/着/过, obj c2 了/着/过
comp:aux:m
comp:m
comp:obj:m
subj:m
flat:m
mod:m
conj:m
1
2
3
4
5
6
7
8
9
10
可变
打開
矫正
扩大
导电
发光
加工
怀疑
脸红
心软
飞机
口试
上海
请教
计算
升降
回家

Figure 2: Decision tree for word internal relation annotation.

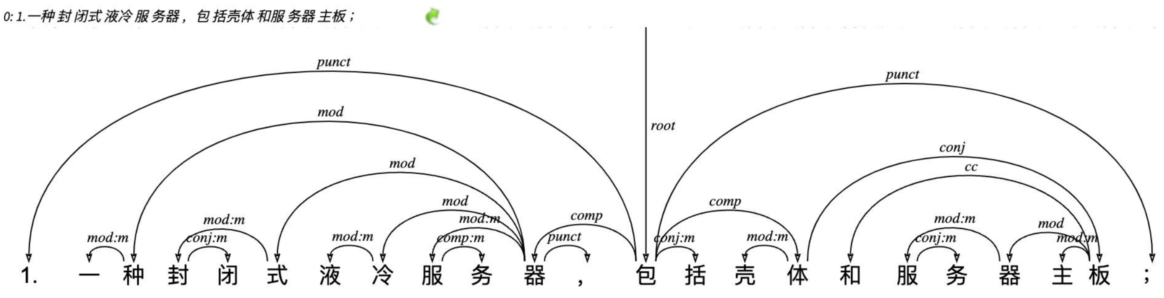
0: 1.一种 封 闭式 液冷 服 务器，包 括壳体 和服 务器 主板；
punct
mod
root
punct
mod
conj
cc
mod
comp
mod:m
mod:m
comp
mod:m
mod:m
conj:m
mod:m
comp:m
punct
conj:m
mod:m
conj:m
mod
mod:m
1. 一 种 封 闭 式 液 冷 服 务 器 ， 包 括 壳 体 和 服 务 器 主 板 ；

Figure 3: SUD character-based treebank.

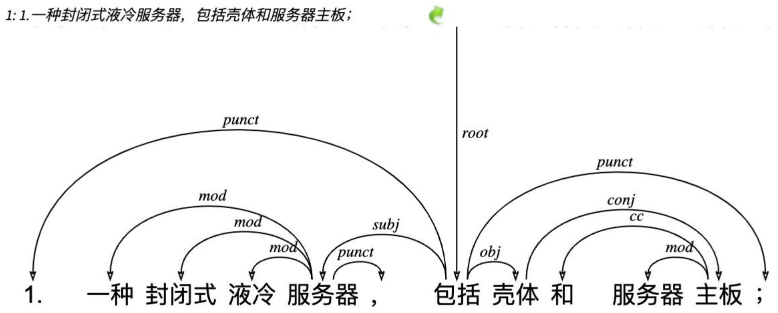
1: 1.一种封闭式液冷服务器，包括壳体和服务器主板；
root
punct
punct
mod
conj
mod
subj
cc
mod
punct
obj
mod
1. 一种 封闭式 液冷 服务器 ， 包括 壳体 和 服务器 主板 ；

Figure 4: SUD word-based treebank.

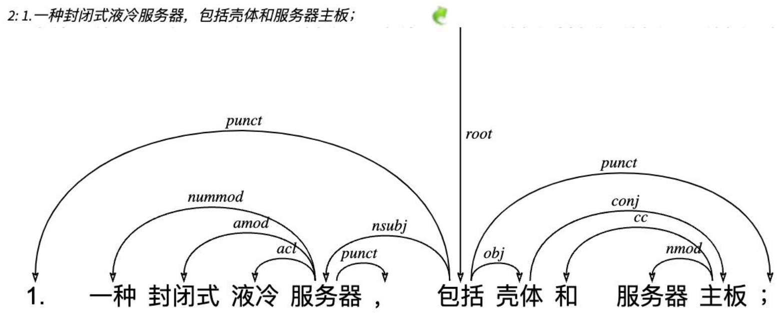
2: 1.一种封闭式液冷服务器，包括壳体和服务器主板；
root
punct
punct
nummod
conj
amod
nsubj
cc
acl
punct
obj
nmod
1. 一种 封闭式 液冷 服务器 ， 包括 壳体 和 服务器 主板 ；

Figure 5: UD word-based treebank.