Constructing the Corpus of Chinese Textual ‘Run-on’ Sentences (CCTRS): Discourse Corpus Benchmark with Multi-layer Annotations

Kun Sun
University of Tübingen, Germany
kun.sun@uni-tuebingen.de

Rong Wang
University of Stuttgart, Germany
rong4ivy@163.com

Abstract

Combining the annotation strengths of PDTB and RST, this study constructs a specialized Chinese discourse corpus on “run-on” sentences. “Run-on” sentences are a typical and prevalent form of discourse/text in Chinese. Despite their widespread use in Chinese, previous studies have only explored “run-on” sentences by using small-scale examples. In order to carry out computational tasks in realistic context and increase diversity of discourse corpora resources, we establish this discourse corpus. The present study selects 500 “run-on”sentences and annotates them on the levels of discourse, syntax and semantics. We mainly adopt an integrated annotation pipeline combining with RST and PDTB to process these sentences. After that, three state-of-the-art discourse parsers are employed to test the feasibility of this corpus, and the result shows that this corpus performs stably and can be used as a benchmark for evaluating discourse parsing.

1 Introduction

Discourse corpora annotated with discourse relations have become important in many down-stream NLP tasks including machine translation (Guzmán et al., 2014), machine reading comprehension (He et al., 2017) and automatic summarization (El-Kassas et al., 2021). Several discourse corpora have been proposed in previous work, grounded with various discourse theories. Currently there have been two influential discourse annotation systems: PDTB (Penn Discourse Treebank) and RST (rhetorical structure theory) (Mann and Thompson, 1988; Carlson et al., 2003; Webber, 2004; Webber et al., 2019). The two systems have their own strengths. However, few discourse corpora could have annotated using the strengths from the two systems. The two annotation systems have been adopted to annotate discourse structure in a few languages. However, for example, few Chinese corpora were both annotated for discourse properties and publicly available (Zhou and Xue, 2015; Jiang et al., 2018). The available annotated texts are primarily newspaper articles. The other problem is that these annotated Chinese discourse corpora seldom annotated other relevant information considering the characteristics of the Chinese language.

The Chinese language is known to be a discourse-oriented language (Tsao, 1979; Chu, 1998; Li, 2005). It is characterized by a very common but special linguistic phenomenon that is called “run-on” sentences. Native Chinese linguists usually refer to this type of sentence as *liu shui ju*, a “flowing-water sentence” (流水句), in which the metaphor “liu shui” (flowing water) vividly describes the physical feature and the logical relationships between the segments of such a sentence. Linguists working on the Chinese language often boast about the special characteristics of “run-on” sentence as such sentences may be seen as grammatically unacceptable in English but they are nonetheless widespread in Chinese (Sheng, 2016; Wang and Zhao, 2017). However, the term “run-on” does not describe the characteristics of these Chinese sentences in a precise way. Instead, this term simply helps those unfamiliar with the Chinese language to understand what they are. A “run-on” English sentence is like this: “*My cat meowed angrily, I knew she wanted food, I hurried to go to shop and make her food.*” By contrast, a Chinese “run-on” sentence would read as follows “*My cat meowed angrily, I hurried to go to shop and make her food.*” In other words, when the middle clause in the English sentence is missing, it feels that there is a semantic leap and it resembles a “run-on” sentence in Chinese.

Linguists working on Chinese have been aware of this phenomenon for a long time (Lu and Zhu, 1979; Hu and Jingsong, 1989; Chao, 1968; Shen, 2012). According to previous research, this means the general characteristics of “run-on” sentences can be summarized phonetically, syntactically, and semantically. By contrast, English, Japanese, and Korean do not have "run-on" sentences. The compound sentence structures are closed in these languages and usually consist of multiple clauses. The clauses are linked through logical relationships such as parallelism, causation, and concession. That means there are clear boundaries between single, complex, or compound sentences in these languages. This is not the case with Chinese run-on sentences. "Run-on" sentences seem to be "sentences". However, a "run-on" sentence is actually a discourse although it is composed of several segments enclosed by a full stop. The reason for this is that there is no clear boundary between the sentence and the text in many cases.

Further quantitative/computational research on "run-on" sentences and realistic Chinese discourse requires annotating a corpus and thus obtaining data. This study targets to construct the corpus of Chines textual "run-on" sentences (CCTRS). Clearly, discourse relations are the core characteristics of textual "run-on" sentences. As discussed at the outset, borrowing the annotation styles from the two mature discourse corpora (PDTB, RST), we annotated the discourse relations for "run-on" sentences. Although some discourse parallel corpora were created using PDTB and RST piplelines separately (Potsdam Commentary Corpus, Stede and Neumann, 2014; GUM corpora, Zeldes, 2017), no discourse corpora have merged the two pipelines into one integrated system to annotate discourse relations previously. The CCTRS is the first to do this.

This corpus (CCTRS) is to provide a benchmark of realistic datasets in Chinese computational discourse analysis. Compared with the past discourse corpora, the CCTRS accommodates an integrated method to annotate discourse relations and makes multilayer annotations regarding other semantic and discourse information. We believe that these annotation data are to promote further development on discourse relation recognition and discourse-level NLP tasks in realistic context. Further, the CCTRS can increase diversity of discourse corpora resources. The data from the CCTRS can help investigate linguistic problems quantitatively and promote the improvement of algorithms for zero anaphora resolution from the perspective of discourse relations.

2 Related Work

The question of how discourse units are effectively incorporated into a unified meaningful text has been addressed from a variety of perspectives, such as Hobbs's theory of coherence relations (Hobbs, 1979), the rhetorical structure theory (Mann and Thompson, 1988), the centering theory (Grosz et al., 1995), the discourse representation theory developed by (Asher et al., 2003), and the framework of lexicalized tree-adjoining grammar (L-TAG, Webber (2004)). Annotations on aspects of discourse structure have been made in text corpora on the basis of these theories.

L-TAG theory holds that discourse relations can be lexicalized, implying that two clauses linked by a connective contribute to two distinct arguments. Adopting this lexically-grounded predicate-argument approach, the Penn discourse treebank (PDTB3.0, Webber et al., 2019) provides annotations of discourse structures for English. However, Chinese discourse uses very few conjunctions and connectives (75% discourse connectives are implicit). Zhou and Xue (2015) followed the PDTB guidelines in establishing Chinese discourse treebank (CDTB). The other two Chinese discourse corpora were established following the PDTB styles (Zhou et al., 2014; Long et al., 2020). RST(rhetorical structure theory), another influential discourse theory, assumes there is a hierarchy of discourse segments that collectively span a full text. The RST discourse treebank (RST-DT, Carlson et al., 2003) has been adopted to annotate discourse in a variety of languages. Nevertheless, in attempting to apply RST to the annotation of Chinese discourse, we find that there is in many cases no way of distinguishing between a nucleus and its satellite.

We illustrate the characteristics of a "run-on" sentence in Chinese using an example, which consists of a sequence of clauses, with its English translation (see Example 1 in the **Appendix**). Example 1 helps us understand why an integrated annotation system combining PDTB and RST was taken in the CCTRS. The subsection 3.2.1 will give a detailed account of why and how the two systems were taken to integrate into an annotation pipeline. This Chinese example is a typical "run-on" sentence. First, the sentence has no connectives that clarify the temporal and logical relations between the clauses. With the semantic relationships between clauses left implicit, interpreting the sentence may sometimes require a considerable creative effort on the part of the reader. For example, after introducing the cry from a child, the sentence directly moves ahead to the clouds in the sky without mentioning the missing link of "*I felt bored and raised my head to find*", leaving a gap in the logical progression from cause to effect. The semantic leap between (3) and (4) leaves the reader/listener much more leeway in filling the gap. By contrast, an English translation usually provides the missing information for the semantic leap between (3) and (4), such as "*Strangely enough, when raising my head, I found that...*". Nevertheless, in English, we can find a similar phenomenon in discourse. It is possible that neither a discourse relation nor entity-based coherence can be inferred between the adjacent sentences. The phenomenon in English is actually the same as the semantic leap in Chinese "run-on" sentences. In the Penn Discourse Treebank (PDTB, Webber et al., 2019), such semantic leap in English discourse is annotated as "`NoRel`"(no relationship), which occurs with a very low frequency (0.67%).

Considering the uniqueness of "run-on" sentences, we also annotated the other types of information at grammatical, semantic and discourse levels. For example, topic chain plays a key role in discourse coherence. Semantic information, such as animacy also is very helpful in recognizing run-on sentences. It is the first time to annotate them in Chinese corpora. In the following, we will introduce them separately.

3 Annotation Scheme

3.1 Text and "run-on" sentence selection

The criterion of "run-on" sentences chosen for the corpus is easy to define and annotate while taking into account various research contributions. Take the following as an example, which is a direct translation from Chinese (in order to save the paper length, and the original example is in the Appendix).

[Example 2] *The river was full of people (1), four long vermilion boats were sliding in the pool(2), the water of the dragon boat had just risen(3), the water in the river was all bean green(4), the weather was so bright(5), the drums were sounding(6), [`semantic leap` (≈ `NoRel` in the PDTB)] Cuicui pursed her lips without saying a word(7), her heart was full of unspeakable joy(8).*

Clearly a semantic leap occurs from the sixth clause to the seventh clause, i.e., from the scene of dragon boat racing on the river to Cuicui directly and we do not know what relationship defines the discourse semantic relationship between (6) to (7). Although the definitions of "run-on" sentences are different, they all acknowledge the existence of semantic leaps between segments, which means this is an appropriate standard for selecting "run-on" sentences. *Semantic leap* is quite similar to `NoRel` in PDTB. We investigate these sentences both formally and semantically. Such a procedure is conducive to annotation and obtaining data. The second criterion is that all segments must be enclosed by a sentential-final period. In Example 2, clause (8) ends with a period. Although semantic leaps often occur between different sentences, these are not considered in this study. This is because we only interested in how semantic leaps occur within a block of clauses that native Chinese speakers/readers perceive as having a complete meaning. In short, we followed the two criteria to select "run-on" sentences. Note that the two criteria were selected from the characteristics described by previous related studies, and not from our own.

Further, in many cases, it is not easy to distinguish which clause is head or which clause is subordinate given that we attempt to establish head/subordinate between two clauses. For example, among the former six clauses in Example 2, it is hard to confirm which clause is a head. This indicates that the distinction between nucleus and satellite in the RST may not be applied in many cases, particularly in fiction genre.

As is well-known, multiple clauses can be joined by using commas without conjunctions in Chinese texts, with the period occurring at the end of the block of clauses and indicating the completeness of the meaning or idea therein rather than the completeness of a sentential structure (Lu and Zhu, 1979:322; Huang and Xiao, 2016; Xue and Yang, 2011, Sun and Lu, 2022). In this study, the selection of "run-on" sentences is limited to the fiction genre. 500 "run-on" sentences (equal to 500 texts) were carefully selected from ten well-known contemporary Chinese fiction (with 2.6 million Chinese characters in total).

3.2 Annotations at different levels

Combining the theoretical study of "run-on" sentences and the observation of examples, we venture the hypothesis that the factors contributed to the characteristics of run-on sentences involve verb valence, clause structure, discourse semantics, topic chain, and other kinds of semantic information surrounding the clause with a semantic leap. As mentioned above, "run-on" sentences are actually treated as text/discourse. The discourse structure should be highlighted. The semantic leap is one of the most important traits of "run-on" sentences, which means we need to pay particular attention to the semantic status of the clause containing the semantic leap. We took these aforementioned factors into account and annotated them. These annotations can be classified into three separate types that represent the core features of "run-on" sentences, namely, discourse, grammatical and semantic. The following details these annotations in this corpus.

Sense	Class
Temporal	Succession (Su) Precedence (Pr) Simultaneous (Si)
Contingency	Cause-effect (Ce) Conditional (Co) Purpose (Pu)
Comparison	Contrast (Cn) Concession (Cc) Conjunction (Cj) Leap (Le)
Expansion	Continuation (Cu) Progression (Pg) List (Lt)

Table 1: The hierarchy of discourse structure for annotation tags in the CCTRS. There are only two-class hierarchy tags: sense and class.

3.2.1 Discourse relations

The annotations of discourse structure and discourse relations for the current corpus combine two of the most successful annotation systems (PDTB and RST-DT) and they take into account the characteristics of Chinese discourse. The following provides a detailed account of them respectively.

The English PDTB annotation system is a shallow discourse annotation, which is reasonable, neat, and easy to operate. RST semantic labels are more numerous and repetitive, while PDTB semantic labels are hierarchical and consistent. However, the PDTB cannot capture the global coherence. The CCTRS uses the semantic tagging of the PDTB and CDTB. However, unlike PDTB and CDTB, we did not explicitly annotate the discourse connectors (e.g., "because of", "despite"). There are two reasons for this: first, the number of explicit discourse connectors in "run-on" sentences is quite small. Second, the discourse connectors are supposed not to influence discourse relations becuase of few discourse connectives used in Chinese discourse (75.7% discourse connectives are implicit in Chinese).

Specifically, given the unbalanced data on PDTB relations, the samples are sparse. For instance, in the PDTB, *temporal* has three classes, but *contingency* includes eight classes. The CCTRS annotation system therefore has at most four specific labels under each sense. The high-frequency semantic relations in previous Chinese discourse are selected as tags. Due to the semantic peculiarities between the segments of the "run-on" sentences, two new semantic tags such as *leap*, and *continuation*. were taken in the annotation system. This makes it easier to compare this with the CDTB and with the PDTB corpora of other languages.

The majority of specific semantic labels were adopted from the PDTB rather than RST-DT. We also added two tags which do not occur in PTDB or CDTB. Table 1 shows the tags for discourse relations in our corpus. For example, here *Le* as tag is used to represent the discourse semantic leap relationship. Once *Le* appears, this means there is a semantic leap, so we can posit that a flow sentence is formed. The semantic leap (*Leap*) in "run-on" sentence is similar to NoRel(no relationship) in the PDTB. When talking about semantic leaps, we mean that no proper discourse relation can be employed to describe the semantic relation between the two segments. An English example from the PDTB illustrates the similarity between these phenomena between English and Chinese (See Example 3 in the Appendix). In Example 3, there is a semantic leap in the place of NoRel. NoRel indicates that a semantic leap occurs between two discourse units. According to the PDTB, these are cases where no discourse relation or entity-based coherence relation can be inferred between adjacent sentences. As a matter of fact, NoRel indicates that a semantic leap occurs between two discourse units. However, there are some differences between a semantic leap in Chinese and a NoRel in English. The first difference is that a semantic leap occurs within a "sentence", at least in a unit enclosed by a full stop in Chinese. However, NoRel in English seldom occurs within a sentence. According to PDTB3.0 manual, there is not any case where NoRel occurs within a sentence. The other difference is that the frequency of NoRel use in English discourse is quite low. There are 254 `NoRel` cases in the PDTB, out of a total of annotated 40600 discourse relations. This suggests that `NoRel` cases only account for quite a small proportion of English discourse (about 0.63%). Although there is no direct statistical evidence concerning semantic leaps in Chinese, we believe that semantic leaps in discourse occur much more frequently than English according to our observation and the relevant studies. This claim is based on our observations and on a number of studies concerning Chinese "run-on" sentences.

The other new tag is *continuation*. It describes successive actions, or several events in succession. The tag of *continuation* was annotated when explicit time adverbials do not occur. Otherwise, the relation with explicit time adverbial was taken as *succession*. The relation of *continuation* is similar to *progression* in which one discourse unit represents a progression from the other, in extent, intensity, scale, etc. In contrast, there is no progression in extent, intensity etc. for *continuation*. In Chinese, they are defined by two different terms respectively, 承接 and 递进 .

We used RST constituent tree structure to annotate discourse structure for "run-on" sentences such that we can obtain the global coherence information. However, in view of the characteristics of Chinese discourse structure, we did not use the concepts of nucleus and satellite in the RST style annotation, because the relationship between nucleus and satellite is not very significant in some cases of fiction genre. Hence we did not annotate which is the nucleus and which is the satellite. For example, in Example 2, clauses 1-6 form a node ("1-6") and clauses 7-8 forms the other node ("7-8"). The first node ("1-6") has a "conjunction" relation with the other node ("7-8"). However, we almost cannot distinguish which node is the nucleus. More details can be seen in subsection 4.2.

According to Demberg et al. (2019), 76% of PDTB relations can be mapped with RST ones in the same English texts (53% directly mapped). This provides evidence that RST could be closely correlated with PDTB, that is, RST and PDTB could be merged together to make annotations for Chinese discourse. In short, our corpus used RST and PDTB strengths to annotate discourse structure and relations. We used RST tree to represent discourse structure but abandoned the concept of neclearity and satellite. We used two-class PDTB tags to annotate discourse relations but did not depend on discourse connectives.

3.2.2 Topic chain

When a clause does not have an argument before the predicate verb, it is recognized as co-referential zero amphora. The topic chain is considered as having a topic followed by several comment clauses, but usually the topic in the comment clauses is invisible, that is, is a *co-referential zero anaphora*. Within a topic chain, the topic controls and manages its comment clauses and the comment clauses are linked coherently, being dependent on some mechanism (Sun, 2019). Topic chain can help improve the performance of zero anaphora resolution in Chinese discourse (Kong et al., 2019). According to our observation, we feel that topic chain is closely related with discourse relations. Our follow-up study shows that topic chain is a good predictor for the occurrence of semantic leap in discourse. Here we annotated the number of clauses between the implicit topic and zero anaphor. This is called the *topic distance*, and it is mutually affected with discourse relations.

3.2.3 Grammatical status

The grammatical structure and the number of the *valence* of predicate verb in each clause was annotated. In English, a clause is the combination of a subject and a verb. There are two types of English clauses. Independent clauses consist of a subject and verb that make up a complete thought and they make sense on their own. In a dependent clause, the subject and verb don't make up a complete thought. Dependent clauses always need to be attached to an independent clause as they are too weak to stand alone. However, Chinese is very likely to use some VPs or NPs independently in discourse and these phrases can work independently as finite clauses in some cases. These independent VPs or NPs in Chinese are similar to English dependent clauses. The difference is that without the use of any grammatical device, these VPs and NPs can work independently. We use the following tags to annotate the grammatical status of each clause: *SVO (subject + predicative verb + object), SV (subject + predicative verb), SVC (subject + predicative verb + complement), VP (verb phrase), NP (noun phrase), and AP (adjective phrase).*

The *valence* indicates the number of arguments that are associated with a particular verb in a sentence. Most verbs are at least mono-valence. This means they have one argument, which is the subject of the sentence that performs the *action* stated by the verb. There are also divalent verbs, which require both a subject and a direct object upon which the action is performed, and trivalent verbs that also need an indirect object that is part of the action. Valence is related to transitivity of verbs, although these are not identical concepts as the transitivity is based purely on objects and not the subject (Gao et al., 2014). We used the numeric value to record the valence information, which is based on the number of arguments of the predicate verb in a clause.

3.2.4 Semantic information

The animacy information and action information were annotated. The term *animacy* was explicated by Comrie (1989), who listed the hierarchical sequence for nominative entities according to the degree of animacy. Comrie (1989) outlined how the grammatical or semantic characteristics of nouns are dependent on how sentient or 'alive' the referent of a noun is. Animacy can have different effects on the grammar of a language, such as the choice of pronoun (what/who), case endings, word order, or the form a verb takes when associated with a noun. These constitute the animacy degree. However, using the traits of a semantic leap between two clauses as our criteria, we use only three animate features in making the annotations: *human*, *nonhuman*, and *inanimate*.

4 Annotation Methods

4.1 Annotation procedures

We then divided a "run-on" sentence into multiple segments. A segment is discourse unit, that is, roughly a unit of discourse that makes sense in pairs and individually. Generally the first thing to look at is the unit separated by commas. If it contains a verb, then this unit can be treated as a sentence or clause segment. If the segment separated by commas does not contain a verb, it may be a noun phrase that must be based on the situation, that is, the noun phrase is part of the content of the small sentence behind or in front. If this is the case, it cannot be divided independently. Sentence segment here is similar to discourse units or elementrary discourse units (EDU) in RST or PDTB. Generally, segment division is not easy, so two of the three annotators first divide the segments.

In fact, there are three levels of sentence annotation: grammatical, semantic, and discourse. Among them, a discourse relation between segments is used to mark not only the semantic relationship between two adjacent segments, but also the relationship between cross-segments (i.e., constituency structure), as described above. Grammatical annotation focuses on the syntactic form of each segment, whether it is a subject-predicate structure, a noun phrase, and also on the valence of the verb from the number of elements in the argument. The semantic annotation mainly focuses on the animacy of the subject noun of the segment before and after the semantic leap relation and whether or not the verb is dynamic or static. It is explained below using sample examples.

The three annotators are native Chinese speakers with linguistics background (two are MA students majoring in linguistics and the third one has obtained PhD degree in linguistics). The three annotators received training from the authors and achieved over 80% agreement on six pilot passages. They then independently annotated all sentences and met to resolve all discrepancies. We used standard corpus annotation methods to check the reliability of our annotations, which will be presented in the following subsection.

4.2 Sample example

After a "run-on" sentence was segmented (into EDU), we annotated each segment (EDU). The following example illustrates how to annotate a "run-on" sentence. All 500 "run-on" sentences (texts) were annotated in the same manner. An example (Example 4) is used to illustrate annotations (here is the English translation, and the original example is in the Appendix).

[Example 4] *At this time an airplane flew over(1), and a white band trailed behind the airplane (2), and lasted for a long time(3), and the sky was cut open (4), or the sky cracked (5), and leaked(6), and the fish disappeared (7).*

This sentence in Example 4 is divided into seven sentence segments (EDU). The grammatical structure is used to mark the structure of each segment. For example, the first segment, "the plane flew over", is a "subject-predicate" structure, which is represented by "SV"; the third segment, "long-lasting", is a verb phrase, which is represented by "VP". The semantic sequence, which semantically relates two segments or groups of segments to each other, is indicated by "1*2" for the first segment and the second segment and “1-3*4” for the first to third segment and the fourth segment. All these segments can construe an RST tree structure where head or nucleus for each node is not distinguished, shown in Figure 1 in the Appendix. The semantic relation label adopted from the PDTB was introduced above. Here we need to introduce a new concept, that is, the *relation distance* is the linear distance between the segments forming a discourse semantic relationship (Sun and Xiong, 2019), e.g., “1*2” is “2-1=1”, and “1-3*4” is “4-1=3”. When a segment group is present, it is indicated by “1-3*4”. When a number of sentence segments are linked by “-”, such as “1-3”, it indicates that these sentence segments are formed into one integrated unit, which is similar to RST structure.

The verb valence is the number of valences of the verb in the segment, for example, the verb ”fly over” in the first segment has only one subject, so the number of valences is “1”. The verb "not loose" in the third segment has a valence of 0. The *topic distance* is the number of segments between the occurrence of a subject that is omitted but that may occur in subsequent segments. Animacy and action are specifically true of the state of the subject and verb in the two segments where a semantic leap occurs. Animacy refers to whether the subject is a living being or a human being, “ina” stands for “inanimate”, “nonhum” for “nonhuman”, and “hum” for “human”. Only three most distinctive features on animacy were chosen for simplicity and convenience. The action refers to the state in which the verb appears, whether it is dynamic or stative (Bach, 1986; Mcintosh, 1977), “dyn” is dynamic, and “sta” is stative. The main annotation information in Example 4 is shown in Table 2. [1]

4.3 Annotation reliability

A consistency assessment can be used to measure the objectivity of the corpus annotations. The assessment of the consistency of annotations provides a more objective picture of the quality of the annotation. The CDTB and the PDTB can assess the consistency of certain metrics such as discourse relations. To eliminate inconsistent annotations, we also used Kappa values (Siegel and Castellan, 1981) to assess the consistency of annotations evaluated using protocol rates. The final assessment of the degree of agreement between three annotators regarding the 500 sentences is shown in Table 3. The agreement ratio of the corpus is greater than 80% and the kappa value is greater than 0.6 for discourse structure, kernel, and relations. Krippendorff (1980) states that a k value greater than 0.6 for annotated data indicates good quality annotation. Table 3 shows that the annotations of the three annotators were very consistent as was the kappa perspective. This suffices to prove the reliability of our annotations.

5 Corpus Statistics

5.1 Frequency distribution

We look at simple statistics, mainly the frequency distribution of the various types of annotations in the CCTRS. With 500 sentences (texts) from the fiction genre, the CCTRS annotated 2286 discourse relations and 650 topic chains. Moreover, 954 clauses and 990 clauses were annotated by animacy and action information respectively. 2281 verbs were annotated by valency information. Figure 1 in the Appendix shows the distribution of frequencies for different discourse relations in the three discourse corpora. “Leap” occupies the largest proportion in the CCTRS, by contrast, “conjunction” takes the largest share in both the CDTB or the PDTB.We compare the distribution of frequencies among the three corpora. As shown in Figures 2 & 3 in the Appendix, the distribution of the three groups of data is basically similar. Generally speaking, such distribution of frequencies abides by the power law (Kello et al., 2010; Sun and Zhang, 2018).

5.2 Corpora comparisons

So far there have been four discourse corpora in Chinese, annotated in terms of PDTB and RST respectively. The four existing Chinese discourse corpora were annotated just with discourse relations. Compared with the four discourse corpora, the CCTRS annotations contain discourse relation, semantic information, grammatical structure and topic chain. This is the only multilayer annotation corpus. The differences among these corpora are seen in Table 4. A great deal of semantic and discourse information cannot be implemented through automatic annotation. In particular, there are very few corpus resources for the annotations of semantic and discourse information concerning aspects of Chinese language characteristics. For example, topic chain and animacy information annotations have never been annotated in previous Chinese cor-

[1] The corpus is available at: `https://github.com/fivehills/CCTRS-corpus-/tree/main`.

ID	EDU	grammatical structure	tree structure	discourse relation labels	relation distance	verb valency	topic distance	animacy
95-1	飞机飞过	SV	1*2	cu	1	1	NA	NA
95-2	飞机拖白带	SV	2*3	pr	1	2	2	NA
95-3	不敢	VP	1-3*4-7	ce	3	0	NA	NA
95-4	天被隔开	SV	4*5	cj	1	1	3	NA
95-5	天裂	SV	4-5*6	ce	2	1	NA	NA
95-6	漏水	VP	6*7	le	1	1	NA	ina
95-7	鱼不见了	NA	NA	NA	1	0	NA	nonhum

Table 2: Sample annotation sheet.

indicators	segments	RST spans	semantic relations	animacy	action
agreement	.92	.93	.92	.99	.98
Kappa	.82	.83	.81	.87	.86

Table 3: Consistency of annotation by the three annotators.

Corpus	Style	Genre	Multilayer annotations
CDTB (Zhou and Xue, 2015)	PDTB	newpaper	NO
CUHK (Zhou et al., 2014)	PDTB	newpaper	NO
TED-CDB (Long et al., 2020)	PDTB	TED talks	NO
MCDTB (Jiang et al., 2018)	RST	newspaper	NO
CCTRS (ours)	RST& PDTB	fiction	YES

Table 4: comparison of the CCTRS to related Chinese discourse datasets.

pora but they are closely related with discourse relations. Topic chain and animacy are closely related zero amphora (co-reference) (Kong et al., 2019). Moreover, although Jiang et al. (2018) annotated the macro discourse information in Chinese discourse using RST tree structure they adopted different tags from both RST tags and PDTB tags. In this way, it could be a little difficult to compare their tags with other similar discourse corpora (including discourse corpora in other languages). Our CCTRS used PDTB tags so that we can easily compare with either RST-style or PDTB-style corpora.

6 Benchmark for Discourse Parsers

We further apply the CCTRS as a benchmark for comparing and evaluating discourse parsers. For the 500 texts in the CCTRS, 71 are used for development set and 73 for test set, and the remaining 356 texts for training. We implement two parsing sub-tasks. RST parsing usually includes the following sub-tasks: span prediction, nuclearity indication, and relation classification. As mentioned above, there are two different kinds of nodes in the RST tree: nucleus and satellite. The nuclearity indication task aims to predict the nucleus or satellite given two EDUs or spans. However, our corpus does not contain the information on nuclearity. This way, the subtask of nuclearity indication was not be included in RST tree building. Additionally, we used PDTB tags to annoate discours relations. As a result, we can divide RST building into two tasks: span prediction and PDTB relation recognition.

6.1 Methods

We applied the standard micro averaged F1 scores on *Span (Sp.)* between EDUs, and *discourse relation (Rel.)*. The micro-averaged F-1 scores over labelled attachment decisions is applied to make valid comparison (Morey et al., 2017). Span describes the accuracy of RST tree structure construction, while discourse relation assesses the ability to categorize the discourse relations. A typical RST parser also needs to distinguish nuclearity and satellite. However, we did not require this task. Following previous PDTB relation recognition studies, we adopted the 13 relations defined in Table 2. We modified three typical RST parsers recently developed using three different methods: bottom-up (Feng and Hirst, 2014), top-down (Zhang et al., 2020), and LSTM (Koto et al., 2021). The three methods were employed to test the feasibility of our corpus.

6.2 Results

After using three RST parsers in the training data (356 texts), we required the three parsers to recognize span and discourse relations. Table 5 shows the average performance of the three RST parsers on development/test data. The human agreement from the annotators is presented for comparison. It seems that Koto et al. (2021) gets better performance than the other two systems. However, the three parsers perform quite similarly in span and relation recognition. According to the performance

Method	**Sp**	**Rel**	**F-1**
Feng and Hirst (2014)	63.2	43.6	42.9
Zhang et al. (2020)	62.8	44.2	43.1
Koto et al. (2021)	64.3	45.1	43.6
human	83	81	71.3

Table 5: Results over the test set calculated using micro-averaged F-1.

data, we can judge that the performance in the three parsers is very stable in analying the CCTRS. We can also see that human performance is still much higher than the three parsers, meaning there is large space for improvement in future work. Overall, due to the performance, we can conclude that the CCTRS is highly capable of using as a benchmark of discourse parsing.

7 Conclusion

“Run-on” sentences are both typical of and prevalent in Chinese discourse. This study collected 500 “run-on” sentences that were annotated at different levels. The main idea is to integrate the strengths of RST and PDTB with the Chinese discourse characteristics. This paper presented the annotation framework, construction workflow and statistics. Further, this multilayer discourse corpus can serve as an evaluating benchmark of realistic Chinese discourse parsing. Moreover, the CCTRS can provide us with valuable language sources to explore in computational linguistics and linguistics, such as co-referential zero anaphora resolution, and predicting semantic leaps in sentence.

Acknowledgments

This study was supported by *China Postdoctoral Science Foundation* (Outstanding Fund No. `2018T110581`) We appreciate helps provided by Prof. Haitao Liu at Zhejiang University, China. Many thanks go to Ms. Yiyun Zhou, and Ms. Mengjie Xu, and Dr. Yi Shan for their helps in text processing and annotations.

References

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Emmon Bach. 1986. The algebra of events. *Linguistics and Philosophy*, pages 5–16.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Yuen Chao. 1968. *A grammar of spoken Chinese*. Univ of California Press.

Chauncey Cheng-Hsi Chu. 1998. *A discourse grammar of Mandarin Chinese*. Peter Lang Pub Incorporated.

Bernard Comrie. 1989. *Language Universals and Typology (the 2nd edition)*. The University of Chicago Press.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue & Discourse*, 10(1):87–135.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521.

Song Gao, Hongxin Zhang, and Haitao Liu. 2014. Synergetic properties of chinese verb valency. *Journal of Quantitative Linguistics*, 21(1):1–21.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Jack Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Ming Hu and Jingsong. 1989. liú shuǐ jù chū tàn [a study of “run-on” sentence]. *Language Teaching and Research*, (4):42–54.

Bo Huang and Xu Xiao. 2016. *Xiandai Hanyu [Modern Chinese] (4th Edition)*. Higher Education Press.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. Mcdtb: a macro-level chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Christopher T Kello, Gordon DA Brown, Ramon Ferrer-i Cancho, John G Holden, Klaus Linkenkaer-Hansen, Theo Rhodes, and Guy C Van Orden. 2010. Scaling laws in cognitive sciences. *Trends in cognitive sciences*, 14(5):223–232.

Fang Kong, Min Zhang, and Guodong Zhou. 2019. Chinese zero pronoun resolution: A chain-to-chain approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–21.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726.

Klaus Krippendorff. 1980. Validity in content analysis.

Wendan Li. 2005. *Topic chains in Chinese: A discourse analysis and applications in language teaching*, volume 57. Lincom Europa.

Wanqiu Long, Bonnie Webber, and Deyi Xiong. 2020. Ted-cdb: A large-scale chinese discourse relation dataset on ted talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803.

Shu Lu and De Zhu. 1979. *Yufa xiuci jianghua [Lectures on grammar and rhetoric]*. Commercial Press.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Carey Mcintosh. 1977. A matter of style: stative and dynamic predicates. *PMLA*, 92(1):110–121.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on rst discourse parsing? a replication study of recent results on the rst-dt. In *Conference on Empirical Methods on Natural Language Processing (EMNLP 2017)*, pages pp–1330.

Jia Shen. 2012. "lingju" he "liushuiju"—wei chao yuanren xiansheng dancheng 120zhounian erzuo ["minor clauses" and "flowing sentence"—in memorial of 120 birthday of mr. chao yuanren]. *Studies of the Chinese Language*, (5):403–415.

Cai Sheng. 2016. *xiàn dài hàn yǔ liú shuǐ jù yán jiū [A Study of Modern Chinese "Run-on" Sentences]*. Ph.D. thesis, Jinlin University.

Sidney Siegel and N John Castellan. 1981. *JR.(1988): Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929.

Kun Sun. 2019. The integration functions of topic chains in chinese discourse. *Acta Linguistica Asiatica*, 9(1):29–57.

Kun Sun and Xiaofei Lu. 2022. Predicting native chinese readers' perception of sentence boundaries in written chinese texts. *Reading and Writing*, pages 1–22.

Kun Sun and Wenxin Xiong. 2019. A computational model for measuring discourse complexity. *Discourse Studies*, 21(6):690–712.

Kun Sun and Lili Zhang. 2018. Quantitative aspects of pdtb-style discourse relations across languages. *Journal of Quantitative Linguistics*, 25(4):342–371.

Feng-Fu Tsao. 1979. *A functional study of topic in Chinese: the first step towards discourse analysis*. Student Book.

Wen Wang and Yong Zhao. 2017. lùn hàn yǔ liú shuǐ jù de jù lèi shǔ xìng[on the sentence class properties of chinese "run-on" sentences. *Chinese Teaching in the World*, 31(2):171–180.

Bonnie Webber. 2004. D-ltag: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 631–635.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395.

Lanjun Zhou, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. The cuhk discourse treebank for chinese: Annotating explicit discourse connectives for the chinese treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 942–949.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Appendix

A. Examples

[Example 1]
(1) 夏天义是进了夏天智家的院子,
Xia Tianyi enter PFV Xia Tianzhi’s yard,
(2) 我没有进去,
I Not enter,
(3) 只听见白雪的孩子一声比一声尖着哭,
only hear Baixue’s kids one M louder than cry
(4) 原本天上还是铁锈红的云,
previously sky up already rust-red clouds,
(5) 一时间黑气就全罩了。
one time black air all over the cover PFV
[Free translation:] Xia Tianyi walked into Xia Tianzhi’s courtyard, but I did not go in, and then I only heard the sound of the kid crying, and crying louder and louder. Strangely enough, when raising my head, I found that the original sky is still rust-red clouds, a time when the black gas is all over.

[Example 2]
(1) 河边站满了人,
river stand fully people
(2) 四只朱色长船在潭中滑着,
four M red long boats in river floating
(3) 龙船水刚刚涨过,
dragon boats water just rise
(4) 河中水皆豆绿,
river middle water all green
(5) 天气又那么明朗,
weather so sunny
(6) 鼓声蓬蓬响着,
drums sounds PFV
(7) 翠翠抿着嘴一句话不说,
Cuicui purse PFV one word Not say
(8) 心中充满了不可言说的快乐。
in heart filled with indescribable happiness
[Free translation] The river was full of people (1), four long vermilion boats were sliding in the pool(2), the water of the dragon boat had just risen(3), the water in the river was all bean green(4), the weather was so bright(5), the drums were sounding(6), [semantic leap] Cuicui pursed her lips without saying a word(7), her heart was full of unspeakable joy(8).

[Example 3]
Jacobs Engineering Group Inc.’s Jacobs International unit was selected to design and build a microcomputer-systems manufacturing plant in County Kildare, Ireland, for Intel Corp. Jacobs is an international engineering and construction concern. [NoRel] Total capital investment at the site could be as much as $400 million, according to Intel. (WSJ_1081)

[Example 4]
(1) 这时候一架飞机飞过,
this time a M airplane flew over
(2) 飞机后拖了条白带,
airplane behind trailed PVF M white band
(3) 经久不散,

long-standing,
(4) 天就被割开了,
sky Bei cut open PFV
(5) 或者是天裂了,
or sky cracked PFV
(6) 漏了水,
leaked PFV water
(7) 鱼也不见了。
fish also disappeared.
[Free translation] At this time an airplane flew over, and a white band trailed behind the airplane for a long time, and the sky was cut open, or the sky cracked and leaked, [semantic leap] and the fish disappeared.

[Abbreviations of glosses]
PFV – perfective aspect
M – measure word

B. Figures

In this section there are three figures.

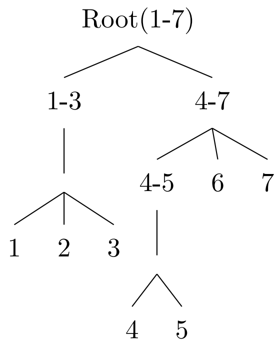

Figure 1: An RST constituent tree for Example 4 and Table 2 in the CCTRS

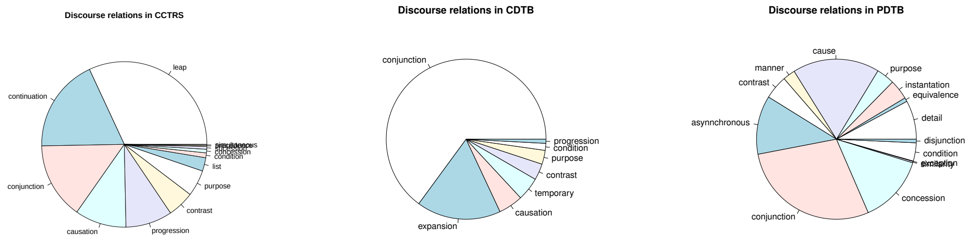

Figure 2: Proportions of discourse relations in the three discourse corpora

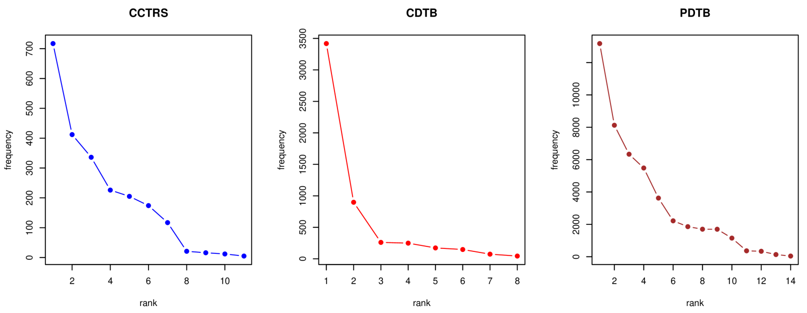

Figure 3: Frequency distribution of semantic relations in the three discourse corpora