TED-CDB: A Large-Scale Chinese Discourse Relation Dataset on TED Talks

Wanqiu Long [‡†], **Bonnie Webber**[†], and **Deyi Xiong**[‡]
[†] University of Edinburgh, Edinburgh, UK
[‡] College of Intelligence and Computing, Tianjin University, Tianjin, China
Wanqiu.long@ed.ac.uk, bonnie.webber@ed.ac.uk, dyxiong@tju.edu.cn

Abstract

As different genres are known to differ in their communicative properties and as previously, for Chinese, discourse relations have only been annotated over news text, we have created the TED-CDB dataset. TED-CDB comprises a large set of TED talks in Chinese that have been manually annotated according to the goals and principles of Penn Discourse Treebank, but adapted to features that are not present in English. It serves as a unique Chinese corpus of spoken discourse. Benchmark experiments show that TED-CDB poses a challenge for state-of-the-art discourse relation classifiers, whose F1 performance on 4-way classification is <60%. This is a dramatic drop of 35% from performance on the news text in the Chinese Discourse Treebank. Transfer learning experiments have been carried out with the TED-CDB for both same-language cross-domain transfer and same-domain cross-language transfer. Both demonstrate that the TED-CDB can improve the performance of systems being developed for languages other than Chinese and would be helpful for insufficient or unbalanced data in other corpora. The dataset and our Chinese annotation guidelines has been made freely available.[1]

1 Introduction

Recent years have witnessed increasing attention to the properties of discourse for a wide variety of natural language processing (NLP) tasks, e.g., machine translation (Ohtani et al., 2019; Voita et al., 2019), summarization (Isonuma et al., 2019; Xu et al., 2020), machine reading comprehension (Mihaylov and Frank, 2019). One of those interesting properties is the coherence between clauses and sentences arising from shallow discourse relations. As empirical approaches for modeling discourse relations usually require corpora annotated with such relations, Penn Discourse Treebank (PDTB) (Prasad et al., 2008b), based on the idea that the discourse relations are grounded in an identifiable set of discourse connectives or Altlex expressions, has been widely applied in the field of natural language processing. Largely because PDTB is effective to extract discourse semantic features, it serves as a useful substrate for the development and evaluation of neural models in many downstream NLP applications (Qin et al., 2017; Nie et al., 2019; Narasimhan and Barzilay, 2015).

Because for Chinese, discourse relations have only been annotated over news text and few of the resulting corpora are freely available, we have created the TED-CDB dataset. TED-CDB currently comprises 72 TED talks in Chinese (∼268.1K words), annotated with 15,540 discourse relations — almost 3 times as many as the CDTB (Zhou and Xue, 2015). Because Tonelli et al. (2010) have shown that discourse relations in spoken discourse are expressed differently than in written text, for scenarios involving Chinese spoken discourse (e.g., dialogue, spoken language translation), TED-CDB boasts unprecedented potential for exploitation and application.

Our contributions comprise:

- the largest PDTB-style Chinese discourse corpus over spoken monologues (Section 3.1). Table 1 compares the TED-CDB with other discourse-annotated Chinese corpora.
- new annotation elements to accommodate Chinese-specific discourse phenomena (Section 3.2).
- benchmark results on Level-2 discourse relation classification for future comparison with other models (Section 5).
- experiments with cross-domain and cross-lingual transfer learning that show that the TED-CDB can improve the performance of

[1] https://github.com/wanqiulong0923/TED-CDB

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2793–2803,
November 16–20, 2020.

Corpus	Domain	Total Relations	Availability
CDTB (Zhou and Xue, 2015)	News report	5,534	Through LDC
CUHK (Zhou et al., 2014)	News report	-	From owner
HIT-CTDB (Zhang et al., 2013)	Internet news	21,505	From owner
NTU (Huang and Chen, 2011a)	Sino and travel set	3,081	From owner
TED-CDB (ours)	TED Talks	15,540	Freely public available

Table 1: Comparison of our corpus to related data sets. "-" means the work do not mention the number.

systems being developed for languages other than Chinese and would be helpful for insufficient or unbalanced data in other corpora (Section 6).

2 Related Work

Most annotations in PDTB style are conducted on written texts originating from news reports. Before 2015, there has been just one corpus for spoken discourse (Tonelli et al., 2010), where PDTB annotations are constructed on Italian dialogues. Recently, researchers have realized that the PDTB annotation guidelines should be used more widely instead of just being confined to construct corpora of written texts. Zeyrek et al. (2018) annotate 6 TED talks for 7 languages. Scheffler et al. (2019) build a discourse corpus on Twitter Conversations.

Regarding Chinese discourse corpora for discourse relations, as illustrated in Table 1, there are mainly 4 Chinese discourse corpora based on the PDTB framework (Prasad et al., 2008a). Zhou and Xue (2012) present a PDTB-style discourse corpus for Chinese, which is further expanded to contain 164 documents, namely the Treebank (CDTB) (CDTB)(Zhou and Xue, 2015). Huang and Chen (2011b) construct a Chinese discourse corpus with 81 articles. They adopt the top-level senses from PDTB sense hierarchy and focus on the annotation of inter-sentential relations. Zhou et al. (2014) present the CUHK Discourse Treebank. They adapt the annotation scheme of Penn Discourse Treebank 2 (PDTB-2) to Chinese and re-annotate the documents of the Chinese Treebank and with only inter-sentence explicit discourse relations. The largest Chinese discourse relation corpus for written texts is HIT-CDTB (Zhang et al., 2013), which presents a new Chinese discourse relation hierarchy adapted from the PDTB system. Nevertheless, these four corpora can only be acquired by either purchasing or applying from the owners.

Therefore, the scarcity of Chinese datasets, especially the lack of corpora for spoken monologues have significantly inspired to build TED-CDB.

3 The TED-CDB Corpus

This section describes the annotation procedure for TED-CDB, including details on the data, annotation scheme, annotation process and agreement study among the annotators.

3.1 Data Description

TED talks (TED is short for technology, entertainment and design), as examples of planned spoken monologues delivered to a live audience (Greenbaum et al., 1996), are given by experts from different fields and different countries, most of which are translated to various languages.

Hai and Sandra (2020) indicate that Chinese translations as a whole can be reliably distinguished from texts originally written in Chinese, for texts translated into a target language possess linguistic properties that are very different from comparable texts originally written in this language. Hence, we collect two types of TED talks: (1) 26 TED talks originally presented in English and translated into Chinese, and (2) 56 TED talks originally presented in Chinese (in Taipei, Shanghai or Chengdu). Together, these 72 TED talks contain 268,099 words after preprocessing.

3.2 Annotation Scheme

Our annotation scheme has been adapted from the PDTB 3.0 relation hierarchy. In the PDTB 3.0 relation hierarchy, there are 4 top-level senses (Expansion, Temporal, Contingency, Contrast) and their second- or in some cases third- level senses, as shown in table 2. To this hierarchy, an additional second-level sense – Progression – has been added under Expansion, specifically for Chinese.

Discourse relations are taken to hold between two abstract object arguments, named Argument 1 and Argument 2. Generally, the arguments are clauses or sentences. Using the PDTB annotator tool, we annotated an explicit connective, identified its two arguments in which the connective occurs, and then labeled the sense for explicit relation. For implicit relations, when we inferred the type of

Temporal	Synchronous	–
	Asynchronous	Precedence
		Succession

Contingency	Cause	Reason
		Result
		Negative-result
	Condition	Arg1-as-cond
		Arg2-as-cond
	Negative condition	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal
Comparison	Contrast	–
	Similarity	–
	Concession	Arg1-as-denier
		Arg2-as-denier

Expansion	Conjunction	–
	Disjunction	–
	Equivalence	–
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Execption	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Table 2: PDTB-3 Sense Hierarchy (Webber et al., 2019). The Level-2 senses are used in assessing system performance (Section 5.1).

relation between two arguments, we tried to insert a connective for this relation. If a connective conveys more than one sense or more than one relation can be inferred, multiple senses would be assigned to the token. And we use a set of consistency rules due to specific linguistic properties in Chinese such as ellipsis of subject, pair connectives.

As some syntactic and textual contexts could not been annotated in our previous work (Long et al., 2020), we loosen the constraints on arguments, connectives, and distance of arguments. In this way, more relations are acquired effectively on the same texts, thus revealing the discourse coherence and structure more fully and clearly. The following are the main additions to our annotation scheme, which future efforts at Chinese discourse annotation might consider adopting as well. In the examples throughout the paper, explicit connectives are underlined, while connectives inserted for implicit relations are both underlined and parenthesized. Sense labels are indicated after the connectives.

Relations have been annotated across non-adjacent sentences

While relations between non-adjacent sentences have only been annotated in the PDTB if Arg1 of an explicit connective is not adjacent to Arg2, implicit relations between non-adjacent sentences were not annotated, except in a small-scale study by Prasad et al. (2017) of relations between paragraph-initial sentences and material in the previous text. In contrast, we annotate relations across non-adjacent sentences not only for explicit relations but also implicit relations. We believe that this would be useful for annotating spoken monologues in general, not just for Chinese. That is, in communicating with an audience, speakers often insert material meant to explain the details of the first argument to audience. Relations can be found across non-adjacent sentences in our annotations. The following are two examples – the first, of an explicit relation, and the second, of an implicit relation.

(1) [我们在空间很早的时候，是做了一个接宝藏的游戏]$_1$。[这种设计在现在看起来好像有点不可思议，但是当时确实有效。因为它帮我们留住了一些实在等不了的用户，也避免了用户流失。所以从早一开始，我们空间跟游戏就息息相关了]$_2$。Then$_{\text{ASYCHROUNOUS}}$[后来，我们的团队也参与去做了QQ农场的游戏]$_3$。
"[When we started to do Qzone, we designed a game about collecting treasures]$_1$. [This design may seem a bit weird now, but it worked at the time. Because it helps us retain some users who can't wait, and also avoids the loss of users. So from the early beginning, our space has been closely related to games]$_2$. [Then$_{\text{ASYNCHRONOUS}}$, we joined to make the game of QQ farm]$_3$. "

(2) [我的研究告诉我，意识不单单是智力的表现，而是更多的有关于我们的本性，作为活着、能呼吸的有机体]$_1$。[意识和智力差别是很大的。就算你不聪明你也会感到痛苦，但前提是你得活着]$_2$。(Therefore)$_{\text{RESULT+SPEECHACT}}$[(所以)在 接下来我要讲给你们的故事中，我们对周围世界的意识体验，以及我们自己的存在，都是 被控制的错觉，都源自我们的生命体]$_3$。
"[My research tells me that consciousness is not just a manifestation of intelligence, but more about our nature as a living, breathable organism]$_1$. [The difference between consciousness and intelligence is very large. You will feel pain even if you are not smart, but only if you have to live]$_2$. [(Therefore)$_{\text{RESULT+SPEECHACT}}$

stories I want to tell you, our conscious experience of the world around us and our own existence are all controlled illusions, all of which originate from our living bodies]$_3$."

In the above examples, there is an explicit Temporal discourse relation between sentences 1 and 3 in the first example and an implicit relation between sentences 1 and 3 in the second example, and there are several sentences being added between the two non-adjacent sentences, which give details for sentence 1. The intervening materials are annotated as "Arg2-as-detail" with respect to the given Arg1(sentence 1).

Verbs can serve as explicit connectives

We follow the practice in the PDTB-3 of using PropBank annotation of modifier relations (ARG-M) to seed intra-sentential discourse relations (Webber et al., 2019). For Chinese, we adopt conventions from Chinese PropBank Annotation (Xue and Palmer, 2009). This allows additional discourse relations to be included. It is the first work to explore Chinese verbs which can signal discourse relations.

(3) [他 失 误 了]$_1$,making$_{\text{RESULT}}$[使得我 们 比 赛 输 了]$_2$。
"[He made a mistake]$_1$, [making$_{\text{RESULT}}$ us lose the game]$_2$."

In this example, the verb phrase "使得" can be identified, while the discourse relations can be expressed through a combination of the adverbial of Arg2 and the anaphoric reference to Arg1 as the implicit subject. In terms of Chinese PropBank Annotation, "使得我们比赛输了 （made us lose the game)" is the ARGM-ADV, and there is a relation expressing Cause.result between between the two clauses.

(4) [我到柏林]$_1$to$_{\text{PURPOSE}}$[去参加一个16天的德语强化]$_2$。
"[I went to Berlin]$_1$[to$_{\text{PURPOSE}}$ attend a 16 days' German intensive course]$_2$."

In the Example (3),"参加一个16天的德语强化（attend a 16 days' German intensive course)" is the purpose and has been labelled as an ARGM-PRP adjunct in the Chinese PropBank (Xue and Palmer, 2003). There is a verb "去" ， which is translated to "to" in this English translation. While the verb "去" is a poly semantic word, and it often refers to "go" in English, it tends to act as a structural auxiliary word in this example. There are several Chinese verbs that have the same function like "来" ， "让"， "用". They always signal senses of relation like Condition, Purpose, Result and Manner.

Noun phrases can serve as arguments

Noun phrases have been annotated as arguments previously in Chinese discourse corpora like the CDTB (Zhou and Xue, 2015). While the Chinese NomBank (Xue, 2006) annotates the nominalized predicate, and also the Chinese Proposition Bank (Xue and Palmer, 2009) performs similar annotation of nominalized verbs. Accordingly, we do not annotate all noun phrases as arguments but those nouns which are nominalizations of their verbal form. Chinese verbs and their nominalizations share the same form, but we identify this kind of arguments, depending on whether the structure NP + 的 (of) + nominalizations of predicate appears. Moreover, in this structure, the NP can always be regarded as the object or subject of the nominalized predicate for the argument.

(5) [我 们 对 他 自 由 的 限 制]$_1$ made$_{\text{RESULT}}$[使他没有办法加入]$_2$。
"[Our towards him freedom limitation]$_1$ [made$_{\text{RESULT}}$ him cannot join]$_2$."

The noun phrase in this example is a typical NP + 的 (of) + nominalizations of predicate structure, the nominalized verb "限制" (limit) can be seen as the predicate of the object NP "自由" (freedom). And the rnoun phase can be paraphrased into "我们限制他的自由" (we limit his freedom).

Punctuation can serve as an AltLex

AltLex (Alternative Lexicalizations) are expressions that convey the SENSE of a discourse relation, without being explicit connectives. If the Altlex Expressions like "这 导 致 了" (this cause), "一个例子是" (one example is...), "原因是" (the reason is) appear, the insertion of connectives become redundant. Although this kind of expressions are usually referred to words before, we have actually found that punctuation like colon play the role of Altlex expressions. With it, the relation of details can usually be expressed without adding additional connectives. Colon as AltLex Expression is the first attempt for PDTB-style corpora among all languages.

(6) [我觉得应该为我的接下来其他去参加的伙伴提供以下的几个提示:$_{\text{RESULT}}$]$_1$ [请您带好你的头灯]$_2$。
"[I think I should provide the following tip for those who will attend it:$_{\text{RESULT}}$]$_1$ [please take your headlights]$_2$."

In Example (5), we can see that the colon is sufficient to display the relation between the clauses. Hence, we have reasons to regard it as a special kind of Altlex expressions.

	Agreement	Kappa
Relation type	0.96	0.94
Senses (Top level)	0.94	0.92
Senses (Second level)	0.85	0.83
Senses (Third level)	0.83	0.81

Table 3: Agreement study

3.3 Annotation Process

The annotator team comprised a professor as the supervisor, an experienced annotator and a researcher of PDTB as counselors, 6 master degree candidates as annotators. All annotators are engaged in research on Natural Language Processing and have a certain theoretical foundation of linguistics. With the professional guidance and rich annotation experience, the quality and the efficiency can be initially guaranteed. To ensure annotation quality, the entire annotation process has the following phrases:

- Training and discussion. The experienced annotator trained the six annotators through training meeting, based on the Chinese tutorial[2] we made on PDTB guidelines and our adaptation scheme;

- Self-pre-annotation. The annotators tried to independently annotate the same texts, finding samples for different senses of relations, exploring problems respectively and discussing issues together, and the experienced annotator checked their work and provided advice for each of them. This step repeated three times until the annotators were all well trained;

- Group pre-annotation. To ensure consistency between the annotators, they were divided into two groups to annotate the same texts and compare their annotation;

- Formal annotation. We annotated 10 TED talks per cycle. During each cycle, the annotated texts from the annotators would be handed in to the experienced annotator who gathered problems existing in their annotation and gave suggestions. Uncertain or new issues would be discussed in the weekly meeting. After each cycle, we exchanged the partner between different groups;

- Check and improve. This phase is very critical for minimizing errors.

3.4 Agreement Study

To ensure annotation consistency, we measured annotators' consistency in annotating specific types of relations which are explicit, implicit, Altlex, NoRel, EntRel, Hypophora, senses from the top level to the third level. Kappa is a quantitative measure of reliability for two raters that are rating the same thing, corrected for how often that the raters may agree by chance. The formulas are:

$$K = \frac{P_o - P_e}{1 - P_e}, P_o = P(consistent)/500; \quad (1)$$

$$P_e = P(correct) + P(incorrect); \quad (2)$$

$$P(correct) = (\frac{A+B}{A+B+C+D}) * (\frac{A+C}{A+B+C+D}); \quad (3)$$

$$P(incorrect) = (\frac{C+D}{A+B+C+D}) * (\frac{B+D}{A+B+C+D}); \quad (4)$$

A quantifies instances where both the annotators' annotations are correct; D does so where both annotators' annotations are incorrect. B quantifies instances where annotator 1 is incorrect while annotator 2 is correct, while C does the reverse. P_o refers to the agreement rate for 500 instances. P(consistent) quantifies instances where the annotators are consistent. We compute the Kappa value and agreement rate between two annotators and then get the average Kappa value and agreement rate among the six annotators. Our results of agreement study can be seen from Table 3.

As is indicated in the Table 3, we achieve relatively high agreement results and Kappa value for the discourse relation type and top-level senses ($\geqslant$ 0.9). Moreover, strong results on the second-level and third-level senses were also achieved, with an agreement rate of 0.85 and Kappa value of 0.83 for the second level senses and agreement rate of 0.83 and Kappa value of 0.81 for the third level.

4 Statistics on TED-CDB

Table 4 shows statistics on TED-CDB. The corpus contains 15,540 discourse relations, which is almost three times as large as the number of discourse relations in Chinese Discourse Treebank (Zhou and Xue, 2015). Of these, 5,531 are explicit relations, while 7,015 are implicit. This means that implicit relations are more frequent in Chinese spoken discourse, while approximately the same number of explicit and implicit relation are

[2]https://github.com/wanqiulong0923/TED-CDB

RelType	Explicit	Implicit	AltLex	EntRel	Hypophora	Norel	Total
Intra-Sentence	5,301	3,209	943	390	4	0	9,847
Inter-Sentence	230	3,806	91	614	355	597	5,693
Total	5,531	7,015	1,034	1,004	359	597	15,540

Table 4: Distribution of 6 kinds of relations annotated in the TED-CDB, within and across sentences.

Figure 1: Distribution of the first and second level senses in TED-CDB (a) and CDTB (b)

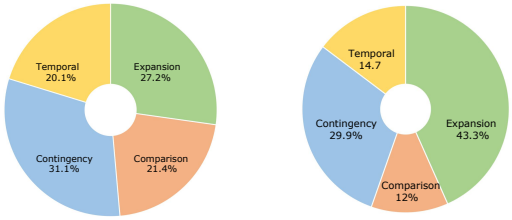

Figure 2: Distribution of the first level senses for explicit relations (a) and implicit relations (b) in TED-CDB.

found in the PDTB-3. There is also a large number of Altlex relations (1034). This type of relations is crucial for automatically identifying discourse relations under the circumstance of no explicit connectives. In our work, we try to detect all possible Altlex expressions that are capable of conveying the discourse relations.

The number of the intra-sentential relations and inter-sentential relations in PDTB-3 are almost the same, but clearly, we can see that the discourse relations in our corpus are more commonly annotated within the sentence, consisting of 9,847 intra-sentential relations and 5,693 inter-sentential relations. The reason perhaps lies in the use of punctuation, which is quite different in Chinese than in English. For example, a comma sometimes serves the same function as a full stop in English (Xue and Yang, 2011). Therefore, a long Chinese sentence may require the use of multiple English sentences to express the same content and preserve grammatically (Li et al., 2014). This may be why there are more intra-sentential relations in Chinese than in English.

We also compared the CDTB and our TED-CDB with respect to the sense distribution. This is displayed in Figure 1(a) and 1(b). CDTB uses an annotation style similar to the PDTB for the texts from the Chinese Treebank corpus. For a discourse relation, one of eight discourse relation senses is assigned. Although all senses in the CDTB are at the same level of the hierarchy, we can map them to the four top-level relation senses in the PDTB hierarchy according to their definitions: Alternative → Expansion; Causation → Contingency; Conditional → Contingency; Conjunction → Expansion; Contrast → Comparison; Expansion → Expansion; Purpose → Contingency; Temporal → Temporal, progression → Expansion; From Figure 1(b), most relations in CDTB are Expansion, constituting the largest percentage of 82%, while the percentage of other 3 types of relation is less than a quarter. On the contrary, Figure 1(a) clearly shows that TED-CDB sees a balanced and rich distribution over the senses. The percentage of Expansion is higher than other types of relations, but it just represents 38%, while contingency, temporal, and comparison can validate their existence, accounting for 29%, 18% and 15% respectively. Moreover, there are several different second-level senses under each of the four top-level senses, among which Cause is the most.

To explore the discourse differences between implicit and the explicit relations, we compare the distribution of top-level senses between the two corpora. Figure 2(a) and 2(b) show that there are more Contingency relations among the explicits, whereas there are more Expansion relations among the implicits. The statistics also tell us that in explicit relations, “因为” (because) and “所以” (so) are the top two most frequent connectives.

5 Experiments

This section describes benchmark experiments for discourse relation recognition on our dataset.

5.1 Methods

We used the state-of-the-art pretrained language models and fine-tuned them on our corpus to con-

	Expansion		Comparison		Contingency		Temporal		Total Macro F1		Total Acc	
	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit
ERNIE	91.75	**69.19**	97.05	45.68	94.12	50.36	91.89	**64.22**	93.70	57.36	93.65	60.14
BERT-wwm	**93.16**	66.32	**97.46**	43.09	**96.64**	53.18	**95.07**	58.37	**95.08**	55.24	**94.92**	58.00
BERT-wwm-ext	92.41	67.89	95.73	45.75	93.60	**55.80**	93.21	60.70	93.74	**57.53**	93.65	**60.57**
ROBERTa-wwm-ext	91.75	67.70	94.92	**47.56**	93.13	52.04	92.11	60.38	92.98	56.92	92.92	59.29

Table 5: Results on level-1 discourse relation classification; F1 score (%) of each level-1 relation on PDTB-3 setting for both explicit and implicit relation on TED-CDB. ; Total macro F1 and Total Accuracy are for all level-1 senses.

	Explicit				Implicit			
Relations	ERNIE	BERT-wwm	BERT-wwm-ext	ROBERTa-wwm-ext	ERNIE	BERT-wwm	BERT-wwm-ext	ROBERTa-wwm-ext
Conjunction	76.47	**77.67**	76.19	**77.67**	50.00	**51.15**	48.32	47.97
Concession	**93.13**	92.61	90.55	92.93	35.46	42.03	43.84	**47.95**
Cause	**89.95**	89.42	87.38	89.76	54.59	56.36	**58.60**	57.99
Contrast	**61.54**	16.67	33.33	36.36	0.00	35.29	36.36	**48.00**
Condition	**86.05**	81.40	78.57	81.48	35.29	37.04	31.58	**40.00**
Synchronous	84.85	86.27	**88.24**	87.62	00.00	14.29	0.00	**17.39**
Purpose	**81.08**	72.22	68.42	78.95	28.57	26.67	25.00	**34.78**
Asynchronous	89.98	89.06	90.32	**90.62**	**58.22**	58.07	59.69	56.68
Negative-condition	**72.73**	44.44	66.67	66.67	00.00	00.00	00.00	00.00
Progression	**78.05**	71.43	71.43	71.11	00.00	00.00	00.00	**9.10**
Substitution	84.85	70.97	**87.50**	84.85	41.67	26.67	**45.45**	40.00
Disjunction	**100.00**	95.65	**100.00**	**100.00**	0.00	44.44	44.44	**58.82**
Level-of-detail	74.41	**78.26**	75.56	72.34	52.26	**56.47**	51.28	51.72
Instantiation	**75.86**	**75.86**	73.33	73.33	12.12	30.43	31.57	**32.56**
Similarity	62.50	66.67	**71.43**	**71.43**	-	-	-	-
Manner	**66.67**	58.82	50.00	62.50	00.00	**26.67**	21.05	16.67
Exception	**100.00**	**100.00**	85.71	85.71	-	-	-	-
Equivalence	40.00	**50.00**	**50.00**	40.00	26.67	26.67	**37.50**	23.53
Total Macro F1	**78.78**	73.19	74.7	75.74	24.68	33.27	33.42	**36.45**
Total Acc	85.45	83.45	82.91	**85.55**	47.93	49.79	**49.93**	49.79

Table 6: Results on level-2 discourse relation classification; F1 score(%) for each level-2 relation in the PDTB-3 hierarchy plus the "Progression" sense relation for both explicit and implicit relation on TED-CDB; Total macro F1 and Total Accuracy are for all level-2 senses in the hierarchy; "-" means there is no the type of sense in the test set.

duct the benchmark test. Particularly, we used the following three baselines:

- BERT, a bidirectional encoder from transformers (Devlin et al., 2019) which is tuned towards two objectives: masked language modeling and next sentence prediction. We adopted two BERT systems: BERT-wwm and BERT-wwm-est (Cui et al., 2019). "-wwm" denotes whole word masking, which means that if a part of a complete word (i.e., word-piece) is replaced by [mask], the other parts of the same word will also be replaced by the mask. "-est" denotes the model trained on a larger data (5.4B).

- ERNIE (115M)[3], a.k.a Enhanced Representation through Knowledge Integration (Sun et al., 2019), which is trained with not only Wikipedia data but also community QA, Baike (similar to Wikipedia), etc.

- ROBERTa, a robust BERT (Liu et al., 2019). We used ROBERTa-wwm-est-large.[4]

For all models, we used the default hyper-parameters (batch=8, learning_rate=2e-5, epoch=10). BERT-wwm (110M) and BERT-wwm-ext have the same hidden size H=768 trained in different size of tokens (0.4B and 5.4B respectively). And ROBERTa-wwm-est (325M) has hidden size H=1024, which is trained in the same way as ROBERTa but without next sentence prediction, with more training steps.

We adopted the F1 and accuracy rate to evaluate both explicit and implicit relation recognition. Moreover, we evaluated the tasks on both the top level (4-way classification) and second level (18-way classification). We used 80% of the dataset as the training set, 10% as dev set and 10% as test set.

5.2 Results

As can be seen from Table 5 and 6, these pre-trained models perform differently on our dataset, but most of the differences are not large. With respect to the 4-way relation classification, all four models achieve high results for the explicit relations, with average accuracy and average F1 all above 92%. This may indicate good annotation consistency for the explicit relations in the corpus. On the other hand, implicit relation classification is much more difficult for the models, with an average accuracy of 60% and average F1 score of 57%. As for the second level (18-way classification), Table 6 shows

[3]https://github.com/PaddlePaddle/ERNIE
[4]https://github.com/ymcui/Chinese-BERT-wwm

	Acc	Macro F1
Expansion	98.92	96.59
Comparison	25.00	40.00
Contingency	11.11	16.67
Temporal	00.00	00.00
Total	93.45	38.31

Table 7: F1 score (%) and total accuracy (%) for level-1 implicit relation classification on CDTB

	Zero-shot for CDTB	TED-CDB
Comparison	50.70	38.16
Contingency	87.35	75.00
Temporal	44.90	70.19
Macro F1	60.98	61.11
Acc	77.00	66.75

Table 8: F1 score (%) and total accuracy (%) for comparison of 3-way implicit relation classification. The left is the result for zero-shot learning from TED-CDB to CDTB, while the right is for TED-CDB.

that the models still obtain quite good results for explicit relations. However, it becomes more challenging for them to classify the implicit relations for the second level. Even the best model among them, ROBERTa-wwm-ext just achieves an accuracy of 49.79% and F1 of 36.45%. In short, we can see how challenging it is for the state-of-the-art models to improve the performance of implicit relation classification on our TED-CDB corpus, which can be used as a testbed for future efforts devoted to spoken discourse relation recognition.

6 Transfer Learning via TED-CDB

We also conducted transfer learning experiments across discourse corpora in different domains and languages. In particular, we considered the following two tasks for transfer learning: (1) training on TED-CDB snf testing on CDTB and (2) training on TED-CDB and testing on TED-MDB. The former is for transfer learning across domains of the same language, while the latter for transfer learning across seven languages within the same domain. The goal of these transfer learning experiments is to investigate if TED-CDB would be helpful for improving the performance of systems being developed for other languages and for insufficient or unbalanced data in other corpora.

6.1 Same-Language Cross-Domain Learning

While the best pre-trained models just can achieve an accuracy of around 60% for 4-way classification and less than 50% for 18-way classification on implicit relations on our dataset, we have noticed that models in previous work (Rutherford et al., 2017) can achieve a significantly higher accuracy of more than 85% on CDTB. Therefore, we used the BERT-wmm model with the same parameters as in our baseline experiments to perform 4-way implicit relation classification on CDTB. Table 7 shows that, although the accuracy of the model is 93.45%, its F1 score is just 38.31%. A closer look at the model performance at each type of sense shows that this high accuracy can be attributed to the most common sense relation, Expansion, on which the accuracy is 98.92%. However, accuracy for the other, less-frequent senses is much lower. In particular, the relation of Temporal gets 0 for both Accuracy and Macro F1. The reason behind this is that the sense distribution of CDTB is quite unbalanced, and most of the annotated relations are Expansion as shown in 1(b), while the number of implicit relation of Temporal can be counted. In other words, the training data for other sense types are not sufficient. Therefore, we wonder whether it is useful that our dataset serves as training set to test all the three types of relations in CDTB, while the relations of Expansion category are removed from both datasets. The model we used here is the BERT-wwm, whose parameters are the same as before.

Table 8 shows the 3-way implicit relation classification results on TED-CDB and those of the zero-shot transfer learning from TED-CDB on CDTB. Compared with the model performance for 3-way implicit relation classification on TED-CDB, Contingency and Comparison get better scores when these three kinds of relations in CDTB are used as the test set for models fine-tuned on TED-CDB. However, for the type of Temporal, the model trained on TED-CDB does not perform well for the CDTB test set. We looked into the test set and discovered that there are only 7 implicit relations of Temporal and that the annotation for several is not consistent with what we tend to annotate, for example:

(6) [集体分东西]$_1$,[他分到的一份肯定最差]$_2$。
"[If the group distribute things]$_1$,[what he gets must be the worst]$_2$. "

	Expansion		Comparison		Contingency		Temporal		Total Macro F1	
	Cross validation	Zero-shot	Cross validation	Zero-shot	Cross validation	Zero-shot	Cross validation	Zero-shot	Cross validation	Zero-shot
English	55.66	67.30	00.00	45.45	34.27	44.64	38.14	61.54	32.02	54.73
German	31.60	62.30	00.00	29.17	32.55	28.57	37.66	46.67	25.45	41.67
lithuanian	75.58	65.47	00.00	23.81	03.05	36.36	11.90	22.22	22.63	36.97
Polish	13.58	61.92	00.00	22.22	25.46	26.51	20.05	19.51	14.77	32.54
Portuguese	67.16	66.89	00.00	36.36	14.29	30.51	30.67	28.57	28.93	40.58
Russian	55.43	58.09	00.00	27.27	20.75	18.35	20.39	23.53	24.14	31.81
Turkish	00.00	62.26	00.00	34.62	27.19	26.53	27.61	48.28	13.70	42.92

Table 9: F1 Score (%) for cross validation within TED-MDB and zero-shot transfer learning from TED-CDB to TED-MDB; The task is 4-way (level-1) implicit relation classification; Total Macro F1 are for all level-1 senses in each language.

For this example, we might annotate it as contingency. Condition, whereas in CDTB the sense of Temporal is assigned to the two arguments.

6.2 Same-Domain Cross-Language Learning

TED-MDB (Zeyrek et al., 2018) corpus annotation follows the PDTB 3.0 framework. It contains manual annotation of 6 TED talks in seven languages (English, Turkish, European Potuguese, Polish, German, Russian, and Lithuanian). The sub-corpus for each language is quite small, with about 200 implicit discourse relations each, compared with the ∼7.0 K implicit relations in the TED-CDB. Therefore, we can see whether the TED-CDB can help them. For this experiment, the multilingual BERT was used, which is as large as BERT-wwm but the training data is expanded to cover 104 languages. We used the multilingual BERT implementation from Huggingface.[5] The design for these experiments is making a comparison between a cross validation within the TED-MDB and a zero-shot transfer learning from TED-CDB to TED-MDB. Due to the unbalanced distribution of senses in TED-MDB, using the method of Easy Ensemble (Liu, 2009), we divided the Expansion data of every language in the TED-MDB into 4 parts and then each part was added into the data of other types to become the training set. Finally, we integrated these training sets from 6 language into one training set, and the left data for one language would be the test set. Therefore, what we used here is 4-fold cross validation where each fold is used as the test set exactly once. The average test set accuracy is then reported.

Table 9 shows the results for transfer learning from TED-CDB to TED-MDB and cross validation within TED-MDB for the task of 4-way implicit relation classification. Comparing the performances with and without our TED-CDB as training set suggests that using the model trained on TED-CDB leads to noticeably better performance for all 7 languages in TED-MDB. In addition, when TED-CDB is used for training, the performance for the 7 languages is close to that for TED-CDB data itself as test set. In particular, from the table, it is noteworthy that the performance on Comparison dramatically increases with the model trained on TED-CDB.

[5]https://github.com/huggingface/transformers

7 Conclusion

We have presented TED-CDB, a large-scale dataset for discourse relations on spoken monologues in Chinese. It is equipped with high-quality annotations and linguistic elements tailored for both Chinese and the genre of spoken monologue. The benchmark results of pretrained language models suggest that TED-CDB is a challenging dataset, which can be used to promote further development on discourse relation recognition and discourse-level NLP tasks. Moreover, we display the ability of TED-CDB to help address the issue of insufficient or unbalanced data on other corpora and improve the performance of models for other languages.

Acknowledgement

The present research was supported by the National Natural Science Foundation of China (Grant No. 61861130364), Natural Science Foundation of Tianjin (Grant No. 19JCZDJC31400) and a Newton International Fellowship from the Royal Society (London)(NAF\R1\180122). We would like to thank the annotators' efforts. Also we are grateful to the anonymous reviewers for their insightful comments. The corresponding author is Professor Deyi Xiong (dyxiong@tju.edu.cn).

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese BERT. *ArXiv*, abs/1906.08101.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

D.S.E.U.S. Greenbaum, S. Greenbaum, and Oxford University Press. 1996. *Comparing English Worldwide: The International Corpus of English*. Clarendon Press.

Hu Hai and Kübler Sandra. 2020. Investigating translated chinese and its variants using machine learning. *Journal of Natural Language Engineering*, page 1–34.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011a. Chinese discourse relation recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011b. Chinese discourse relation recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288, Baltimore, Maryland. Association for Computational Linguistics.

Tian-Yu Liu. 2009. Easyensemble and feature selection for imbalance data sets. *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pages 517–520.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 1907.11692.

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for chinese TED talks. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1018–1025, Marseille, France. European Language Resources Association.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Katherine Forbes Riley, and Alan Lee. 2017. Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16, Saarbrücken, Germany. Association for Computational Linguistics.

Rashmi Prasad, Samar Husain, Dipti Sharma, and Aravind Joshi. 2008b. Towards an annotated corpus of discourse relations in Hindi. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 974–984, Vancouver, Canada. Association for Computational Linguistics.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291, Valencia, Spain. Association for Computational Linguistics.

Tatjana Scheffler, Berfin Aktaş, Debopam Das, and Manfred Stede. 2019. Annotating shallow discourse relations in twitter conversations. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 50–55, Minneapolis, MN. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced representation through knowledge integration. *ArXiv*, abs/1904.09223.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Nianwen Xue. 2006. Semantic role labeling of nominalized predicates in Chinese. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 431–438, New York City, USA. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese treebank. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 47–54, Sapporo, Japan. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Nat. Lang. Eng.*, 15:143–172.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 125–134, Sofia, Bulgaria. Association for Computational Linguistics.

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, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.

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