CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing

Nianwen Xue, Brandeis University,
xuen@brandeis.edu
Hwee Tou Ng, National University of Singapore,
nght@comp.nus.edu.sg
Sameer Pradhan, cemantix.org and Boulder Learning, Inc.,
pradhan@cemantix.org
Attapol Rutherford, Brandeis University,
tet@brandeis.edu
Bonnie Webber, University of Edinburgh,
bonnie@info.ed.ac.uk
Chuan Wang, Brandeis University,
cwang24@brandeis.edu
Hongmin Wang, National University of Singapore,
wanghm@comp.nus.edu.sg

Abstract

The CoNLL-2016 Shared Task is the second edition of the CoNLL-2015 Shared Task, now on Multilingual Shallow discourse parsing. Similar to the 2015 task, the goal of the shared task is to identify individual discourse relations that are present in natural language text. Given a natural language text, participating teams are asked to locate the discourse connectives (explicit or implicit) and their arguments as well as predicting the sense of the discourse connectives. Based on the success of the previous year, we continued to ask participants to deploy their systems on TIRA, a web-based platform on which participants can run their systems on the test data for evaluation. This evaluation methodology preserves the integrity of the shared task. We have also made a few changes and additions in the 2016 shared task based on the feedback from 2015. The first is that teams could choose to carry out the task on Chinese texts, or English texts, or both. We have also allowed participants to focus on parts of the shared task (rather than the whole thing) as a typical system requires substantial investment of effort. Finally, we have modified the scorer so that it can report results based on partial matches of the arguments. 23 teams participated in this year's shared task, using a wide variety of approaches. In this overview paper, we present the task definition, the training and test sets, and the evaluation protocol and metric used during this shared task. We also summarize the different approaches adopted by the participating teams, and present the evaluation results. The evaluation data sets and the scorer will serve as a benchmark for future research on shallow discourse parsing.

1 Introduction

The shared task for the Twentieth Conference on Computational Natural Language Learning (CoNLL-2016) is a follow-on to the CoNLL-2015 shared task, and it is on *Multilingual Shallow Discourse Parsing (SDP)*. While the 2015 task focused on newswire text data in English, this year we added a new language, Chinese. Given a natural language text as input, the goal of an SDP system is to detect and categorize discourse relations between discourse segments in the text. The conceptual framework of the Shallow Discourse Parsing

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1–19,
Berlin, Germany, August 7-12, 2016.

task is that of the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008; Prasad et al., 2014), where a discourse relation is viewed as a predicate that takes two abstract objects as arguments. The two arguments may be realized as clauses or sentences, or occasionally phrases. It is "shallow" in that sense that the system is not required to output a tree or graph that covers the entire text, and the discourse relations are not hierarchically organized. As such, it differs from analyses according to either Rhetorical Structure (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003).

The rest of this overview paper is structured as follows. In Section 2, we provide a concise definition of the shared task. We describe how the training and test data are prepared in Section 3. In Section 4, we present the evaluation protocol, metric and scorer. The different approaches that participants took in the shared task are summarized in Section 5. In Section 6, we present the ranking of participating systems and analyze the evaluation results. We present our conclusions in Section 7.

2 Task Definition

The goal of the shared task on shallow discourse parsing is to detect and categorize individual discourse relations. Specifically, given a newswire article as input, a participating system is asked to return the set of discourse relations it can identify in the text. A discourse relation is defined as a relation taking two abstract objects (events, states, facts, or propositions) as arguments (Prasad et al., 2008; Prasad et al., 2014). Discourse relations may be expressed with explicit connectives like *because*, *however*, *but*, or implicitly inferred between two argument spans interpretable as abstract objects. In the current version of the PDTB, only adjacent spans are considered. Each discourse relation is labeled with a sense selected from a sense hierarchy. Its argument spans may be sentences, clauses, or in some rare cases, noun phrases. To detect a discourse relation, a participating system needs to:

1. Identify the text span of an explicit discourse connective, if present, or the position between adjacent sentences as the proxy site of an implicit discourse relation;
2. Identify the two text spans that serve as arguments to the relation;
3. Label the arguments as *Arg1* or *Arg2*, as appropriate;
4. Predict the sense of the discourse relation (e.g., "Cause", "Condition", "Contrast").

A full system that outputs all four components of the discourse relations usually comprises a long pipeline, and it is hard for teams that do not have a pre-existing system to put together a competitive full system. This year we therefore allowed participants to focus solely on predicting the sense of discourse relations, given gold-standard connectives and their arguments.

3 Data

3.1 Training and Development

The training and development sets for English remain exactly the same as those used in the CoNLL-2015 shared task. Details regarding how the data was adapted from the Penn Discourse TreeBank 2.0 (PDTB 2.0) are provided in the overview paper of the CoNLL 2015 shared task (Xue et al., 2015). The Chinese training and development sets are taken from the Chinese Discourse TreeBank (CDTB) 0.5 (Zhou and Xue, 2012; Zhou and Xue, 2015), available from the LDC (http://ldc.upenn.edu), supplemented with additional annotated data from the Chinese TreeBank (Xue et al., 2005).

The CDTB adopts the general annotation strategy of the PDTB, associating discourse relations with explicit or implicit discourse connectives and the two spans that serve as their arguments. In the case of explicit discourse relations (Example 1), there is an overt discourse connective, which may be realized syntactically as a subordinating or coordinating conjunction, or a discourse adverbial. Implicit discourse relations are cases where there is not an overt discourse connective (Example 2). Like PDTB, CDTB also annotates Alternative Lexicalizations (AltLex) and Entity Relations (EntRel) when no explicit or implicit discourse relations can be identified.

(1) [**Conn** 尽管] [**Arg1** 亚洲 一些
even though Asia some
国家 的 金融 动荡 会 使
country DE financial turmoil will make
这些 国家 的 经济 增长
these country DE economy growth
受到 严重 影响] , [**Conn** 但]
experience serious impact , but
[**Arg2** 就 整 个 世界 经济 而 言 ,
to whole CL world economy ,
其他 国家 的 强劲 增长 势头
other country DE strong growth momentum
会 弥补 这 一 损失] 。
will compensate this one loss .

"Even though the financial turmoil in some Asian countries will affect the economic growth of these countries, as far as the economy of the whole world is concerned, the strong economic growth of other countries will make up for this loss."

(2) 其中 [**Arg1** 出口 为
among them export be
一百七十八点三亿 美元 , 比
17.83 billion dollar , compared with
去年 同 期 下降
last year same period decrease
百分之一点三] ; [**Arg2** 进口
1.3 percent ; import
一百八十二点七亿 美元 , 增长
18.27 billion dollar , increase
百分之三十四点一] 。
34.1 percent .

"Among them, export is 17.83 billion dollars, an 1.3 percent increase over the same period last year. Meanwhile, import is 18.27 billion dollars, which is a 34.1 percent increase."

The CDTB also differs somewhat in its annotation practices. The first difference is in the way that implicit discourse relations are identified. PDTB uses sentence-final punctuation (periods, question or exclamation marks) to identify where implicit discourse relations might occur. However, since the concept of "sentence" is less formalized in Chinese, and since a comma may serve as a sentence-final marker (as well as sentence-internal punctuation), CDTB identifies implicit relations by examining commas in addition to periods, question and exclamation marks, and disambiguating them to identify those serving as sentence-final markers. Teams that exploited these language-specific characteristics did well on the Chinese task (Section 6). Table 1 shows that the distribution of explicit and implicit discourse relations also differs between Chinese and English: while there are about equal numbers of explicit and discourse relations in English, implicit discourse relations outnumber explicit discourse relations in Chinese. The second difference in annotation practices is how the arguments are labeled. In the PDTB, the argument that is introduced by a discourse connective (e.g., a subordinate conjunction) is labeled *Arg1* while the other argument is labeled *Arg2*. Since there are much fewer explicit discourse relations than implicit discourse relations, the argument labels are defined "semantically", meaning they are defined based on how arguments are interpreted. For example, for a Causation relation, *Arg1* is the cause while *Arg2* is the result. Since arguments are defined semantically, there is less of a need to have Level-3 subtypes as in the PDTB. For example, *Contingency:Cause:Reason* and *Contingency:Cause:Result* are essentially the same relation, just with the arguments reversed. For this reason, CDTB adopts a flat set of 10 relations (Table 2), which are used in this shared task without any modification.

The above discussion shows that PDTB-style discourse relations are substantially, but not fully language-independent due to different lexicalizations (e.g., explicit vs implicit discourse connectives) and grammaticalizations (the formalization of the concept of sentence). As we shall see in Section 5 where we discuss different approaches, teams that exploited these language-specific properties did well on the Chinese task. For example, the way in which implicit discourse relations are annotated impacts how the arguments for implicit discourse relations are identified. In addition, because the smaller number of explicit discourse relations, it makes less sense to train separate models for explicit relations alone because many of the discourse connectives in the training data will not repeat in the test data. In addition, the senses of discourse relations are less evenly distributed in Chinese than in English. For example, "Conjunction" is a very common category, presumably because without explicit discourse connectives, a discourse relation is harder to judge, leading annotators

	Train	Dev	Test
Implicit	6,706	251	281
Explicit	2,225	77	96
EntRel	1,098	50	71
AltLex	211	5	7
Total	10,240	328	455

Table 1: The distribution of discourse relation types in the Chinese data

to use "Conjunction" as a default category.

3.2 Test Data

We provide two test sets for each language: a test set from a publicly available annotated corpus, and a blind test set specifically prepared for this task. The official ranking of the systems is based on their performance on the *blind test set*. We reused the English test sets from the 2015 shared task, details of which can be found in (Xue et al., 2015). For Chinese, one test set is from the CDTB, and uses the same data source as the training data. The *blind test set* is from Chinese Wikinews.

3.2.1 Data Selection and Post-processing

For the blind test data, 29,892 words of Chinese newswire texts were selected from a dump of Chinese Wikinews[1] created on 23rd October 2015, and annotated in accordance with the CDTB-0.5 annotation guidelines.

The raw Wikinews data was pre-processed as follows:

- News articles were extracted from the Wikinews XML dump[2] using the publicly available WikiExtractor.py script.[3]
- Additional processing was done to remove any remaining XML annotations and produce a raw text version of each article (including its title and date).
- Articles written purely in simplified Chinese were identified using the Dragon Mapper[4] Python library, and segmented using the NUS Chinese word segmenter (Low et al., 2005).
- Sentences in each article were manually segmented such that adjacent sentences were separated by a carriage return, and one extra carriage return was added between two paragraphs to ease paragraph boundary identification.
- Each article was named according to its unique Wikinews ID, accessible online at `http://zh.wikinews.org/wiki?curid=`ID.

Since longer articles with many multi-sentence paragraphs are more consistent with the CDTB-0.5 texts, 64 articles were randomly selected among the articles with more than 400 characters. Word segmentation errors and some typos were manually corrected.

3.2.2 Annotations

The *blind test set* was annotated by two of the shared task organizers, one of whom (seventh author) was the main annotator (MA) while the other (first author) acted as the reviewing annotator (RA), reviewing each relation annotated by the MA and recording agreement or disagreement. Annotation involved marking the relation type (Explicit, Implicit, AltLex), sense (alternative, causation, conditional, conjunction, contrast, expansion, purpose, temporal, EntRel, NoRel), and arguments (Arg1 and Arg2), using the PDTB annotation tool.[5]

Before commencing official annotation, the MA was trained in CDTB-0.5 style annotation by the RA. After a review of the guidelines, the MA annotated some CDTB texts that were already annotated, and then compared his annotations with the standard annotations. Some differences were discussed between the MA and the RA to further strengthen MA's knowledge of the guidelines.

4 Evaluation

The scorer that computes all of the available evaluation metrics is open-source with some contribution from the participants during the task period[6].

4.1 Main evaluation metric: End-to-end discourse parsing

A shallow discourse parser (SDP) is evaluated based on the end-to-end F_1 score on a per-

[1] `https://zh.wikinews.org/`

[2] `https://dumps.wikimedia.org/zhwikinews/20151020/zhwikinews-20151020-pages-meta-current.xml.bz2`

[3] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

[4] `http://dragonmapper.readthedocs.io/en/latest/index.html`

[5] `https://www.seas.upenn.edu/~pdtb/tools.shtml#annotator`

[6] `http://www.github.com/attapol/conll16st`.

Sense	Definition
Alternative	Relation between two alternatives
Causation	Relation between cause and effect
Condition	Relation between a supposed condition and a supposed result
Conjunction	Relation between two equal-status statements serving a common communicative function
Contrast	Relation between two statements, co-occurrence of which seems contradictory, counter-intuitive, out-of-ordinary, etc.
Expansion	Relation in which one argument is an elaboration or restatement of another
Purpose	Relation between an action and the intention behind it
Temporal	Relation that is temporal in nature, expressing temporal precedence, etc.
Progression	Relation in which one argument represents a progression from the other, in extent, intensity, scale, etc.
EntRel	Relation between two statements that are connected only by the fact that they are about the same entity or entities.

Table 2: Definitions of relation senses in the Chinese data.

discourse relation basis for both languages. The input to an SDP consists of documents with gold-standard word tokens along with their automatic parses. We do not pre-identify discourse connectives or any other elements of the discourse annotation. The SDP must output a list of discourse relations comprising argument spans and their labels, explicit discourse connectives where applicable, and the senses. The F_1 score is computed based on the number of predicted relations that match a gold standard relation exactly. Like the 2015 edition of the task, a relation is correctly predicted if and only if the text spans of its two arguments are correctly predicted (Arg1 and Arg2), as is its sense. The results from this evaluation is shown in Table 5.

An argument is considered correctly identified if and only if it matches the corresponding gold standard argument span exactly, and is also correctly labeled (Arg1 or Arg2). In the main evaluation, partial matching is given no credit. Sense classification evaluation is less straightforward, since senses are sometimes annotated partially or annotated with two senses. To be considered correct, the predicted sense for a relation must match one of the two senses if there is more than one sense. If the gold standard is partially annotated, the sense must match with the partially annotated sense although the blind test set contains no partial annotation.

4.2 Supplementary Evaluation: Discourse relation sense classification

Although the submissions are ranked based on the end-to-end F_1 score, discourse relation sense classification subtask has gained much attention from the community within the past years including some participants from last year. We provide the data and evaluation setup for participants who are only interested in the discourse relation sense classification subtask and for those who want to evaluate their system without the error propagation from argument extraction.

In this supplementary evaluation, the input is gold-standard argument pairs and their corresponding explicit discourse connectives if applicable. The goal is to fill in the senses including EntRel. The results from this evaluation are shown in Table 9

4.3 Component-wise and partial evaluation

For analytical purposes, the scorer also provides component-wise evaluation with error propagation and a breakdown of the discourse parser performance for explicit and non-explicit discourse relations. The scorer computes the precision, recall, and F_1 for the following tasks:

- Explicit discourse connective identification.
- Arg1 identification.
- Arg2 identification.
- Arg1 and Arg2 identification.

- Sense classification with error propagation from discourse connective and argument identification.

For purposes of evaluation, an explicit discourse connective predicted by a parser is considered correct if and only if the predicted raw connective includes the gold raw connective head, while allowing for the tokens of the predicted connective to be a subset of the tokens in the gold raw connective. We provide a function that maps discourse connectives to their corresponding heads. The notion of discourse connective head is not the same as its syntactic head. Rather, it is thought of as the part of the connective conveying its core meaning. For example, the head of the discourse connective "At least not when" is "when", and the head of "five minutes before" is "before". The non-head part of the connective serves to semantically restrict the interpretation of the connective.

Although Implicit discourse relations are annotated with an implicit connective inserted between adjacent sentences, participants are not required to provide the inserted connective. They only need to output the sense of the discourse relation. Similarly, for AltLex relations, which are also annotated between adjacent sentences, participants are not required to output the text span of the AltLex expression, but only the sense. The EntRel relation is included as a sense in the shared task, and here, systems are required to correctly label the EntRel relation between adjacent sentence pairs.

We also provide partial evaluation to assess how well a system does when we relax the criteria. The official full evaluation metric produces low scores due to error propagation from argument extraction. Partial evaluation instead allows 'fuzzy matching' in arguments. The extracted Arg1 and Arg2 are correct if and only if the average of F_1 score of the extracted Arg1 and Arg2 is greater than 0.7. This allows us to evaluate the sense classification of that relation even if the argument extraction is not perfect. The evaluation is also done for both explicit and non-explicit relations separately (Table 8) and together (Table 6).

4.4 Closed and open tracks

In keeping with the CoNLL shared task tradition, participating systems were evaluated in two tracks, a *closed* track and an *open* track. A participating system in the closed track could only use the provided PDTB training set but was allowed to process the data using any publicly available (i.e., non-proprietary) natural language processing tools such as syntactic parsers and semantic role labelers. In contrast, in the open track, a participating system could not only use any publicly available NLP tools to process the data, but also any publicly available (i.e., non-proprietary) data for training. A participating team could choose to participate in the closed track or the open track, or both.

The motivation for having two tracks in CoNLL shared tasks was to isolate the contribution of algorithms and resources to a particular task. In the closed track, the resources are held constant so that the advantages of different algorithms and models can be more meaningfully compared. In the open track, the focus of the evaluation is on the overall performance and the use of all possible means to improve the performance of a task. This distinction was easier to maintain for early CoNLL tasks such as noun phrase chunking and named entity recognition, where competitive performance could be achieved without having to use resources other than the provided training set. However, this is no longer true for a high-level task like discourse parsing where external resources such as Brown clusters have proved to be useful (Rutherford and Xue, 2014). In addition, to be competitive in the discourse parsing task, one also has to process the data with syntactic and possibly semantic parsers, which may also be trained on data that is outside the training set. As a compromise, therefore, we allowed participants in the closed track to use the following linguistic resources, in addition to the training set:

For English,

- Brown clusters
- VerbNet
- Sentiment lexicon
- Word embeddings (word2vec)

For Chinese, the following resources are provided, both trained on Gigaword Simplified Chinese data:

- Brown clusters (implementation from (Liang, 2005))
- Word embeddings (word2vec)

To make the task more manageable for participants, we provided them with training and test data with the following layers of automatic linguistic annotation produced using state-of-the-art NLP tools:

For English,

- Phrase structure parses predicted using the Berkeley parser (Petrov and Klein, 2007);
- Dependency parses converted from phrase structure parses using the Stanford converter (Manning et al., 2014).

For Chinese,

- Phrase structure parses predicted with 10-fold cross validation on CTB8.0 using the transition-based Chinese parser (Wang and Xue, 2014);
- Dependency parses converted from phrase structure parses using the Penn2Malt converter.

4.5 Evaluation Platform: TIRA

We use a new web service called TIRA as the platform for system evaluation (Gollub et al., 2012; Potthast et al., 2014). Traditionally, participating teams have been asked to manually run their system on the blind test set without the gold standard labels, and submit the output for evaluation. Starting with the 2015 shared task, however, we shifted this evaluation paradigm, asking participants to deploy their systems on a remote virtual machine, and to use the TIRA web platform (tira.io) to run their systems on the test sets without actually seeing them. The organizers would then inspect the evaluation results, and verify that participating systems yielded acceptable output.

This evaluation protocol allowed us to maintain the integrity of the blind test set and reduce the organizational overhead. On TIRA, the blind test set can only be accessed in the evaluation environment, and the evaluation results are automatically collected. Participants cannot see any part of the test sets and hence cannot do iterative development based on the test set performance, which preserves the integrity of the evaluation. Most importantly, this evaluation platform promotes replicability, which is crucial for proper evaluation of scientific progress. Reproducing all of the results is just a matter of a button click on TIRA. All of the results presented in this paper, along with the trained models and the software, are archived and available for distribution upon request to the organizers and upon the permission of the participating team, who holds the copyrights to the software. Replicability also helps speed up the research and development in discourse parsing. Anyone wanting to extend or apply any of the approaches proposed by a shared task participant does not have to re-implement the model from scratch. They can request a clone of the virtual machine where the participating system is deployed, and then implement their extension based off the original source code. Any extension effort also benefits from the precise evaluation of the progress and improvement since the system is based off the exact same implementation.

5 Approaches

Teams could participate in either English or Chinese or both, and either submit an end-to-end system or just compete in the discourse relation sense prediction component. All end-to-end systems for English adopted some variation of the pipeline architecture proposed by Lin et al (2014) and perfected by Wang and Lan (2015), which has components for identifying discourse connectives and extracting their arguments, for determining the presence or absence of discourse relations in a particular context, and for predicting the senses of the discourse relations. Here we briefly summarize the approaches used in each subtask.

Connective identification The identification of discourse connectives is not a simple dictionary lookup as some discourse connective expressions are ambiguous and may function as discourse connectives in some context but not in others. Several approaches to this

ID	Institution	Learning methods	Resources used	Extra resources
steven	Aicyber.com	-	-	-
bit (Jian et al., 2016)	BIT	SVM (for English explicits, English and Chinese implicits), rule-based method for Chinese explicit	Word embeddings	General Inquirer lexicon, HowNet, Central News of Taiwan
ttr (Rutherford and Xue, 2016)	Brandeis University	Feedforword (implicit sense only, pooling before hidden layers)	word embeddings	no
clac (Laali et al., 2016)	Concordia	CRF, decision tree (C4.5), Convolutional Network (implicit discourse senses)	syntactic parses, word embeddings	no
devenshu (Jain and Majumder, 2016)	DA-IICT	Maxent (openNLP)	syntactic parses	no
ecnucs (Wang and Lan, 2016)	ECNU	Liblinear, convolutional network for implicit relation (for English implicit)	phrase structure parses	no
goethe (Schenk et al., 2016)	Goethe University Frankfurt	Feed-forward neural network, CRF (connective and argument extraction), SVM (explicit sense)	syntactic parses, Brown clusters	no
gtnlp	Georgia Tech	-	-	-
tbmihaylov (Mihaylov and Frank, 2016)	Heidelberg	Liblinear (scikit-learn) (for explicit sense), CNN (for implicit sense)	word embeddings	no
aarjay	IIT-Hyderabad	-	-	-
iitbhu (Kaur et al., 2016)	IITBHU	Naive Bayes, MaxEnt	syntactic parses, MPQA subjectivity, VerbNet, Word embeddings (word2vec)	no
cip2016 (Kang et al., 2016)	Institute of Automation, CAS	MaxEnt (Mallet)	syntactic parses, word embeddings	no

Table 3: Approaches of participating systems (Part I). Teams that have not submitted a system description paper are marked with $*$.

subtask are represented in this competition. One is to collect all candidate discourse connective by looking up a list of possible connectives compiled from the training data and train a classifier to disambiguate them. There are two variants in this approach: one strategy is to train a classifier for each individual discourse connective expression (Oepen et al., 2016), and the other is to train one classifier for all discourse connective expressions (Wang and Lan, 2016; Kong et al., 2015; Laali et al., 2016). Alternatively, connective identification is treated as a token-level sequence labeling task, solved with sequence labeling models like CRF (Stepanov and Riccardi, 2016).

Argument extraction Different strategies were used for extracting the arguments for explicit and for implicit discourse relations. Determining the arguments of implicit discourse relations is relatively straightforward. Most systems adopted a heuristics–based extraction strategy that parallels the PDTB annotation strategy for implicit discourse relations: for each pair of adjacent sentences that do not straddle a paragraph boundary, if an explicit discourse relation does not already exist, posit

ID	Institution	Learning methods	Resources used	Extra resources
nguyenlab (Nguyen, 2016)	JAIST	CRF (CRF++) for detecting connectives and arguments, SMO and Random Forest for classifying senses	phrase structure trees, MPQA Subjectivity lexicon, word embeddings	none
gw0 (Weiss and Bajec, 2016)	Univ. of Ljubljana	Focused RNN (sense only, for both explicit and implicit)	word embeddings	none
olslopots (Oepen et al., 2016)	Olso-Potsdam-Teesside	SVM (SVM^{light}), heuristic argument extraction	Brown clusters	none
purduenlp (Pacheco et al., 2016)	Purdue University	SVM (explicit sense), Feedforword (implicit sense)	word embeddings	Wikipedia (for training event embeddings)
stepanov (Stepanov and Riccardi, 2016)	University of Trento	CRF++, AdaBoost	Brown clusters, dependency/phrase structure parses, VerbNet, MPQA Lexicon	none
tao0920 (Qin et al., 2016)	SJTU	SVM (explicit sense), CNN (implicit word sense)	word embedings (implicit word sense)	none
Rival2710 (Li et al., 2016b)	SJTU	Maxent (OpenNLP)	syntactic parses	none
lib16b (Kong et al., 2016; Li et al., 2016a)	Soochow University	Maxent (OpenNLP), SVM (for Chinese)	syntactic parses, Brown clusters	none
Soochow (Fan et al., 2016)	Soochow	Averaged perceptron (for both sequence labeling and sense)	syntactic parses, Brown clusters	none
ykido (Kido and Aizawa, 2016)	University of Tokyo	SVM and Maxent (Scikit-learn)	Word embeddings, parse trees, MPQA subjectivity lexicon	none
VTNLPS16 (Chandrasekar et al., 2016)	Virginia Polytechnic and State University	Maxent (NLTK), Averaged Perceptron	syntactic parses (phrase/dependency), Brown clusters	none
nikko	University of Washington	-	-	-

Table 4: Approaches of participating systems (Part II). Teams that have not submitted a system description paper are marked with $*$.

an implicit discourse relation. It is possible that no discourse relation exists, but such cases are rare and most systems choose to ignore such a possibility (Oepen et al., 2016; Laali et al., 2016; Chandrasekar et al., 2016).

The extraction of the arguments for explicit discourse relations is more involved as their distribution is more diverse. The two arguments of an explicit discourse relations can be in either the same or different sentence. Identifying the argument spans of explicit discourse relations thus resembles finding the text span for discourse connectives, and there are two general approaches. One is to treat it a sequence labeling task and solve it with sequence labeling models like CRF (Fan et al., 2016; Stepanov and Riccardi, 2016), and the other is to identify candidate argument spans and train a binary classifier to determine if the candidate argument span is a true (fragment of) argument span. The difference is that the arguments are typically clauses or sentences while discourse connectives are typically single words (e.g., "as") or multi-word expressions

(e.g.,"as long as"). Candidate arguments are typically identified with the help of syntactic parse trees rather than dictionaries (Oepen et al., 2016; Wang and Lan, 2016; Kong et al., 2016). The argument spans do not align perfectly with constituents in a tree, and participating systems have adopted two strategies to cope with this. One is to first identify pieces of an argument and compose them (Wang and Lan, 2016; Kong et al., 2016), and the other is to identify whole arguments but then edit them based on linguistically motivated heuristics (Oepen et al., 2016) or the prediction of classifiers (Laali et al., 2016).

Relation sense classification All systems have separate classifiers for explicit and implicit discourse connectives. For explicit relations, the discourse connective itself is the best predictor of the discourse relation. Many discourse connectives are unambiguous, always mapping to one discourse relation sense. For ambiguous discourse connectives, discourse relation sense classification amounts to word sense disambiguation. For explicit discourse relation senses, participants have generally adopted "conventional" machine learning techniques such as SVM and MaxEnt models that rely on manually designed features. Explicit discourse relation senses can be predicted with high accuracy. The main challenge is predicting implicit discourse relation senses, which has received a considerable amount of attention in recent years (Pitler et al., 2009; Biran and McKeown, 2013; Rutherford and Xue, 2014). Determining implicit discourse relation senses relies on information from the two arguments of the relation. For this subtask, there is a good balance between "conventional" machine learning techniques such as Support Vector Machines and Maximum Entropy models that rely heavily on handcrafted features, and neural network based approaches. A wide variety of features have been used for this subtask, and they include features extracted from syntactic parses (Kang et al., 2016; Kong et al., 2016; Stepanov and Riccardi, 2016; Jain and Majumder, 2016; Wang and Lan, 2016; Fan et al., 2016), Brown clusters (Kong et al., 2016; Stepanov and Riccardi, 2016; Oepen et al., 2016; Laali et al., 2016; Chandrasekar et al., 2016; Pacheco et al., 2016), VerbNet classes (Stepanov and Riccardi, 2016; Kaur et al., 2016), and the MPQA lexicon (Stepanov and Riccardi, 2016; Kaur et al., 2016). However, features extracted from the two arguments for "conventional" machine learning methods are generally weak predictors of relation sense. Neural network based learning methods that are capable of learning representations for classification purposes seem to be particularly appealing in this learning scenario and many teams trained neural network models for the subtask of predicting the sense of implicit discourse relations. A variety of neural network architectures are represented. (Schenk et al., 2016) used a feedforward neural network, with dependency structures used to re-weight the word embeddings used as input to the network. (Wang and Lan, 2016; Qin et al., 2016) achieved competitive performance using a Convolutional Neural Network architecture for this subtask. Finally, (Weiss and Bajec, 2016) produced competitive results with a focused RNN. Word embeddings were typically used as input to the neural network models and different pooling methods have been used to derive the vectors for arguments. Rutherford and Xue (2016) used simple summation pooling in a feedforward network and achieved competitive performance in classifying implicit discourse relation senses.

Language (in-)dependence of the task To achieve competitive results, teams that participated in the Chinese task made significant changes to their systems, based on the linguistic characteristic and style of annotation for the Chinese data (Kang et al., 2016; Wang and Lan, 2016). The majority of Chinese discourse connectives are paired or discontinuous. When identifying discourse connectives, a system has to allow the possibility that different parts of the same connective may be separated from each other. The ECNU team devised a strategy that allowed their system to identify candidate discourse connectives that are discontinuous (Wang and Lan, 2016). Also, because different parts of a paired connective are text-bound to different arguments, it is no longer possible to follow the PDTB practice of labelling an argument based on whether it is bound to a connective or not (i.e, Arg2 is argument bound to the con-

nective, while Arg1 is the other argument). As a result, the argument labels in the CDTB are defined semantically. The CAS team made labeling the argument a separate task from identifying the text spans of the argument (Kang et al., 2016), and (Wang and Lan, 2016) use a combination of classifiers and rules to determine the argument labels. Finally, because implicit discourse relations in Chinese text are not restricted to adjacent sentences with unambiguous punctuation marks, competitive Chinese systems realized the importance of disambiguating mid-sentence punctuation marks as anchors for identifying the argument spans (Kang et al., 2016; Wang and Lan, 2016).

6 Results

We provide no separate rankings for the closed track and open track, even though there are a few teams that used external resources. Also, no overall ranking is provided based on both English and Chinese, due to imbalanced participation.

Table 5 shows the performance of end-to-end systems based on the strict match of argument spans. We present results on three data sets for each language. For English the three data sets are (1) the blind test set (official); (2) the standard WSJ test set; and (3) the standard WSJ development set. The three data for Chinese are (1) the blind test set; (2) the CDTB test set; and (3) the CDTB development set. The official rankings are based on the blind test sets annotated specifically for this shared task. The three data sets for English are exactly the same as those we used for the 2015 shared task (Xue et al., 2015) so we can measure progress from year to year. The top-ranked submission for English is by the Olso-Potsdam-Teesside team, and their overall score based on strict match is 27.77% F1 score, which represents an improvement of 3.77% over last year's winning system submitted by the East China Normal University (ECNU) (Wang and Lan, 2015). Four other teams also beat the score of last year's winning system. There is considerable fluctuation in the rankings across the three data sets, with the ECNU system receiving the highest score on both the WSJ development and test sets. The top ranked Chinese system was submitted by the Institute of Automation, Chinese Academy of Sciences, although the difference between the top two teams is only 0.3%. However, the rankings are very stable across data sets. Since there are many more teams that participated in the English task than the Chinese task, we decided not to provide an overall ranking based on the results of both languages. (In such a putative ranking, the ECNU system would be ranked top.)

Table 6 provides the ranking based on partial match of argument spans. The ranking remains largely unchanged when the scorer setting is changed from strict match to partial match for English. For the Chinese evaluation, the ranking is also to a large extent consistent with that based on strict match. For both English, the overall parser scores based on F1 score are considerably higher when the scorer shifts from a strict match setting to a partial match setting, indicating that error propagation is a serious issue when there is a long pipeline. Tables 7 and 8 present the accuracy of individual components for explicit and implicit discourse relations based on strict and partial match respectively. For English, the parser accuracy for explicit discourse relations is generally higher than that for implicit discourse relations, although the argument span extraction accuracy is higher for implicit discourse relations than for explicit discourse relations.

The overall parser accuracy for implicit relations is dragged down by the lower accuracy in predicting discourse relation sense, as is shown is Table 9, which compares the accuracy of classifying explicit and implicit discourse relation sense. This pattern does not consistently hold for results on Chinese across the three data sets. On the blind test set, the parser accuracy for some of the teams is actually higher for implicit discourse relations than for explicit discourse relations. Our hypothesis is that this is caused by the fact that there are much more instances for implicit discourse relations than explicit discourse relations. In this situation, the difference in discourse relation sense accuracy between explicit and implicit discourse relations is much smaller in Chinese than in English, an observation that is largely born

Language	Participant	Parser			Connective			Argument		
	ID	P	R	F	P	R	F	P	R	F
Blind Test										
English	oslopots	27.75	27.79	27.77	93.53	90.12	91.79	48.30	48.18	48.24
English	ecnucs	25.69	26.30	25.99	90.11	92.61	91.34	48.06	46.82	47.43
English	stepanov	26.22	24.07	25.10	84.89	92.55	88.56	48.55	52.84	50.60
English	tao0920	24.41	24.81	24.61	88.67	93.73	91.13	48.47	47.64	48.05
English	goethe	24.29	24.65	24.47	86.87	92.00	89.36	46.73	46.01	46.37
English	li16b	30.36	20.26	24.31	90.47	92.80	91.62	30.19	45.17	36.19
English	Soochow	24.49	18.94	21.36	89.57	92.57	91.04	33.66	43.34	37.90
English	clac	21.11	21.01	21.06	91.91	88.56	90.20	39.04	39.20	39.12
English	nguyenlab	20.31	20.43	20.37	79.50	91.51	85.08	39.04	38.78	38.91
English	VTNLPS16	19.51	21.09	20.27	88.13	90.41	89.25	39.45	36.47	37.90
English	rival2710	12.62	18.94	15.15	98.56	98.21	98.38	36.15	24.09	28.91
English	devanshu	12.69	9.18	10.65	77.70	93.71	84.96	14.89	20.55	17.27
English	nikko	7.35	10.34	8.59	67.27	87.38	76.02	17.45	12.40	14.50
English	iitbhu	4.60	6.87	5.51	86.87	91.30	89.03	24.98	16.74	20.05
Standard WSJ Test (Section 23)										
English	ecnucs	30.26	31.16	30.70	93.50	94.42	93.96	48.48	47.12	47.79
English	tao0920	29.90	30.65	30.27	92.42	94.88	93.63	49.10	47.96	48.52
English	goethe	29.54	30.03	29.78	89.82	93.57	91.65	49.97	49.14	49.55
English	li16b	33.07	25.48	28.78	94.04	95.38	94.71	36.87	47.92	41.68
English	oslopots	27.47	28.89	28.16	96.42	92.52	94.43	50.18	47.77	48.94
English	stepanov	27.64	27.96	27.80	90.57	94.36	92.43	48.68	48.19	48.44
English	Soochow	27.47	25.84	26.63	94.69	94.79	94.74	42.75	45.50	44.08
English	nguyenlab	25.52	23.88	24.67	83.42	92.22	87.60	40.59	43.46	41.97
English	clac	23.94	24.91	24.42	93.61	88.52	91.00	42.55	40.94	41.73
English	VTNLPS16	20.80	25.84	23.05	89.92	91.51	90.71	42.24	34.04	37.70
English	rival2710	20.13	22.33	21.17	99.67	98.19	98.92	40.33	36.39	38.26
English	devanshu	13.19	10.23	11.53	67.06	94.36	78.40	14.96	19.31	16.86
English	nikko	7.08	10.03	8.30	48.00	88.96	62.35	17.69	12.51	14.66
English	iitbhu	5.66	9.20	7.01	92.09	93.92	93.00	29.09	17.91	22.17
Standard WSJ Development (Section 22)										
English	ecnucs	39.90	40.98	40.43	95.29	95.15	95.22	57.24	56.46	56.85
English	goethe	39.87	40.55	40.21	92.79	94.32	93.55	57.66	57.26	57.46
English	tao0920	38.20	38.99	38.59	93.82	95.08	94.45	56.55	56.04	56.29
English	li16b	39.86	32.10	35.56	93.53	94.93	94.22	42.97	54.03	47.87
English	oslopots	34.13	35.30	34.71	96.32	92.25	94.24	55.43	54.26	54.84
English	clac	32.12	33.10	32.60	94.26	89.90	92.03	49.72	48.87	49.29
English	stepanov	32.46	32.46	32.46	91.47	95.25	93.32	54.39	55.04	54.71
English	Soochow	32.72	30.47	31.56	93.53	95.07	94.29	47.84	52.08	49.87
English	VTNLPS16	28.58	33.59	30.88	93.09	92.95	93.02	50.21	43.33	46.52
English	nguyenlab	30.59	28.76	29.65	85.00	91.60	88.18	45.82	49.25	47.47
English	rival2710	27.78	30.33	29.00	99.71	98.40	99.05	48.40	44.93	46.60
English	devanshu	17.74	13.85	15.56	69.71	94.05	80.07	18.18	23.60	20.54
English	nikko	10.68	15.41	12.62	56.91	90.21	69.79	24.44	17.25	20.22
English	iitbhu	5.76	9.23	7.10	91.47	95.25	93.32	31.96	20.31	24.84
Blind Test										
Chinese	cip2016	29.13	24.99	26.90	48.76	66.51	56.27	38.93	45.07	41.78
Chinese	ecnucs	26.74	26.46	26.60	60.95	65.34	63.07	41.79	42.19	41.99
Chinese	li16b	23.31	23.61	23.46	63.07	65.99	64.50	38.55	38.03	38.29
Chinese	goethe	16.12	10.76	12.90	45.23	46.97	46.08	21.94	32.13	26.07
Chinese	nikko	3.70	2.43	2.93	36.75	68.42	47.82	3.05	4.64	3.68
Standard Xinhua Test										
Chinese	cip2016	39.67	42.20	40.89	67.71	78.31	72.63	56.26	52.24	54.18
Chinese	ecnucs	37.60	43.30	40.25	65.63	80.77	72.41	54.73	47.43	50.82
Chinese	li16b	34.07	37.14	35.54	75.00	80.00	77.42	48.79	44.76	46.69
Chinese	goethe	30.16	20.22	24.21	66.67	74.42	70.33	30.11	44.05	35.77
Chinese	nikko	4.59	3.74	4.12	45.83	84.62	59.46	6.37	7.84	7.03
Standard Xinhua Development										
Chinese	ecnucs	38.32	47.52	42.42	85.71	86.84	86.27	58.75	47.37	52.45
Chinese	cip2016	39.47	43.08	41.20	79.22	88.41	83.56	55.87	50.71	53.17
Chinese	li16b	33.86	39.16	36.32	79.22	83.56	81.33	52.74	45.60	48.91
Chinese	goethe	24.68	20.10	22.16	61.04	65.28	63.09	31.59	38.66	34.77
Chinese	nikko	5.47	5.48	5.48	64.94	84.75	73.53	6.27	6.25	6.26

Table 5: Scoreboard for the CoNLL-2016 shared task showing performance (**strict** scoring) across the subtasks and the three data partitions—blind test, standard test and development set for both English and Chinese.

out in the results shown in Table 9.

7 Conclusions

Twenty three teams from three continents participated in the CoNLL-2016 Shared Task on multilingual shallow discourse parsing.

Language	Participant	Parser			Argument		
	ID	P	R	F	P	R	F
		Blind Test					
English	oslopots	44.14	44.25	44.20	80.41	80.64	80.53
English	ecnucs	40.13	41.19	40.65	77.57	79.94	78.74
English	stepanov	38.34	35.24	36.72	79.85	72.41	75.95
English	tao0920	39.51	40.20	39.85	77.66	79.22	78.43
English	goethe	39.66	40.28	39.97	76.64	78.02	77.32
English	li16b	49.01	32.75	39.27	75.93	48.32	59.06
English	Soochow	42.60	33.09	37.24	78.21	58.36	66.85
English	clac	37.29	37.14	37.22	77.55	77.19	77.37
English	nguyenlab	32.46	32.67	32.56	63.15	63.63	63.39
English	VTNLPS16	35.32	38.21	36.71	68.76	75.16	71.82
English	rival2710	22.60	33.91	27.13	51.30	80.80	62.75
English	devanshu	38.13	27.63	32.04	77.12	53.25	63.00
English	nikko	15.34	21.59	17.94	25.09	36.10	29.61
English	iitbhu	20.45	30.52	24.49	45.82	71.62	55.89
		Standard WSJ Test (Section 23)					
English	ecnucs	46.91	48.27	47.58	80.35	82.99	81.65
English	tao0920	45.84	46.93	46.37	79.51	81.66	80.57
English	goethe	45.48	46.25	45.86	79.14	80.66	79.90
English	li16b	50.13	38.60	43.62	75.82	56.94	65.03
English	oslopots	43.04	45.22	44.10	77.25	81.67	79.40
English	stepanov	41.94	42.38	42.16	78.21	79.12	78.66
English	Soochow	45.74	43.00	44.33	79.46	74.07	76.67
English	nguyenlab	38.78	36.28	37.49	66.09	61.26	63.58
English	clac	39.81	41.40	40.59	76.46	79.88	78.14
English	VTNLPS16	34.85	43.26	38.60	60.77	77.27	68.03
English	rival2710	35.62	39.48	37.45	72.15	81.06	76.34
English	devanshu	38.76	30.03	33.84	76.64	57.61	65.78
English	nikko	14.20	20.10	16.64	22.62	32.62	26.71
English	iitbhu						
		Standard WSJ Development (Section 22)					
English	ecnucs	54.12	55.04	54.58	83.29	84.63	83.96
English	goethe	53.09	53.62	53.36	82.21	82.87	82.54
English	tao0920	52.21	52.84	52.52	82.46	83.32	82.89
English	li16b	56.53	45.17	50.22	80.58	62.33	70.29
English	oslopots	48.02	49.22	48.61	81.41	83.41	82.40
English	clac	46.04	47.02	46.52	79.50	81.09	80.29
English	stepanov	45.77	45.38	45.58	81.05	79.95	80.50
English	Soochow	49.50	45.53	47.43	82.41	74.76	78.40
English	VTNLPS16	41.13	47.94	44.28	65.66	77.46	71.08
English	nguyenlab	42.40	39.63	40.97	69.67	64.30	66.88
English	rival2710	41.42	44.89	43.08	75.75	82.48	78.97
English	devanshu	44.63	34.52	38.93	80.12	59.80	68.49
English	nikko	18.02	25.92	21.26	27.12	39.17	32.05
English	iitbhu						
		Blind Test					
Chinese	cip2016	46.67	40.31	43.26	72.48	61.52	66.55
Chinese	ecnucs	42.10	41.69	41.89	68.76	68.02	68.39
Chinese	li16b	41.64	42.22	41.93	65.35	66.38	65.86
Chinese	goethe	32.40	22.13	26.30	64.50	42.88	51.51
Chinese	nikko	9.06	5.95	7.18	9.77	6.34	7.69
		Standard Xinhua Test					
Chinese	cip2016	51.43	55.38	53.33	71.05	76.96	73.89
Chinese	ecnucs	50.86	58.68	54.49	68.57	80.45	74.03
Chinese	li16b	53.02	57.80	55.31	69.54	76.46	72.83
Chinese	goethe	45.02	30.77	36.55	69.44	46.30	55.56
Chinese	nikko	11.08	9.01	9.94	14.36	11.63	12.86
		Standard Xinhua Development					
Chinese	ecnucs	50.95	63.19	56.41	65.07	82.37	72.70
Chinese	cip2016	53.79	59.27	56.40	68.85	76.68	72.55
Chinese	li16b	50.34	58.22	54.00	66.26	77.65	71.50
Chinese	goethe	39.62	32.38	35.63	58.95	47.32	52.50
Chinese	nikko	11.72	11.75	11.73	10.47	10.50	10.48

Table 6: Scoreboard for the CoNLL-2016 shared task showing performance (**partial** scoring) across the subtasks and the three data partitions—blind test, standard test and development set for both English and Chinese.

The shared task required the development of an end-to-end system, and the best system achieved an F1 score of 27.77% on the blind test set for English, and 26.90% for Chinese, reflecting the serious error propagation problem in such a system. The shared task exposed the most challenging aspect of shallow discourse parsing as a research problem, help-

Language	Participant	Explicit					Implicit			
	ID	Parser F	Connective F	A1 F	A2 F	A12 F	Parser F	A1 F	A2 F	A12 F
Blind Test										
English	oslopots	34.45	91.79	52.43	75.20	43.95	21.89	64.60	76.40	52.02
English	ecnucs	33.94	91.34	51.05	74.20	42.84	19.54	61.05	75.83	51.15
English	stepanov	31.74	88.56	50.28	72.05	41.84	19.46	66.83	79.11	58.05
English	tao0920	31.64	91.13	48.43	73.57	41.40	19.01	61.61	77.82	53.35
English	goethe	30.74	89.36	48.84	71.97	41.07	19.63	60.77	74.63	50.44
English	li16b	31.18	91.62	47.18	68.85	38.25	16.10	42.22	44.83	33.73
English	Soochow	27.47	91.04	41.86	69.84	33.46	15.06	51.80	59.96	42.50
English	clac	31.10	90.20	48.37	70.61	39.89	12.19	54.06	60.94	38.44
English	nguyenlab	30.83	85.08	52.17	70.07	41.39	12.55	42.97	55.08	37.06
English	VTNLPS16	29.33	89.25	46.08	67.94	38.62	13.56	49.47	59.06	37.35
English	rival2710	25.31	98.38	41.29	73.43	33.39	10.11	37.30	41.70	26.30
English	devanshu	20.67	84.96	37.17	51.13	28.52	1.12	17.23	25.09	6.55
English	nikko	20.93	76.02	35.16	48.37	26.63	2.28	17.65	14.43	8.31
English	iitbhu	7.93	89.03	28.94	25.81	9.03	4.15	37.03	41.49	26.24
Standard WSJ Test (Section 23)										
English	ecnucs	40.31	93.96	51.39	76.43	44.31	22.38	64.66	66.86	50.83
English	tao0920	40.53	93.63	50.38	76.73	44.90	21.36	65.18	67.84	51.67
English	goethe	40.44	91.65	50.41	75.95	45.22	20.60	67.17	68.32	53.28
English	li16b	36.57	94.71	47.14	71.14	40.81	19.82	51.56	51.56	42.68
English	oslopots	39.38	94.43	51.99	72.57	43.93	18.02	69.92	71.45	53.47
English	stepanov	39.60	92.43	49.64	76.51	44.56	17.56	65.58	67.78	51.80
English	Soochow	32.97	94.74	44.99	72.09	37.40	20.51	63.36	66.81	50.52
English	nguyenlab	39.39	87.60	53.81	71.79	45.28	11.67	46.59	50.90	39.06
English	clac	35.72	91.00	47.29	72.56	40.23	13.95	60.05	57.13	43.11
English	VTNLPS16	36.41	90.71	47.21	69.62	40.87	13.31	50.97	47.32	35.39
English	rival2710	32.51	98.92	42.15	75.48	36.24	11.70	56.28	53.14	39.95
English	devanshu	23.70	78.40	33.44	48.26	27.87	1.18	21.05	22.77	7.30
English	nikko	21.13	62.35	28.43	43.91	23.22	2.70	19.88	15.46	10.92
English	iitbhu	11.93	93.00	29.98	34.35	12.80	4.24	38.59	36.44	27.42
Standard WSJ Development (Section 22)										
English	ecnucs	51.13	95.22	62.01	81.26	55.11	31.10	68.84	73.81	58.39
English	goethe	50.87	93.55	61.97	78.87	54.41	30.99	70.45	74.10	60.14
English	tao0920	49.70	94.45	60.99	79.94	53.44	28.98	68.71	74.19	58.80
English	li16b	42.97	94.22	52.89	74.81	46.37	27.54	58.31	59.77	49.51
English	oslopots	46.44	94.24	60.72	75.83	51.37	24.09	71.25	77.46	58.03
English	clac	44.57	92.03	56.71	75.95	48.67	21.67	63.70	63.70	49.87
English	stepanov	45.89	93.32	55.66	79.07	49.36	20.89	69.51	74.51	59.40
English	Soochow	39.30	94.29	51.00	73.98	42.70	24.21	66.86	72.12	56.76
English	VTNLPS16	46.21	93.02	57.90	73.77	50.26	19.08	56.58	55.20	43.59
English	nguyenlab	46.68	88.18	60.72	72.31	52.33	14.61	49.14	56.40	43.12
English	rival2710	41.49	99.05	50.55	78.89	45.29	18.57	61.09	60.72	47.71
English	devanshu	31.31	80.07	41.89	54.39	34.97	2.07	20.47	23.27	7.81
English	nikko	30.36	69.79	40.76	52.30	32.82	4.42	21.59	19.81	14.31
English	iitbhu	10.69	93.32	31.21	27.91	11.70	5.10	41.22	40.96	32.25
Blind Test										
Chinese	cip2016	24.46	56.27	38.53	44.44	26.50	27.12	55.26	60.17	44.57
Chinese	ecnucs	28.88	63.07	41.13	47.53	31.81	24.74	54.21	54.99	42.36
Chinese	li16b	20.63	64.50	37.40	39.93	23.31	23.75	53.27	54.74	41.93
Chinese	goethe	18.56	46.08	32.76	34.92	20.70	10.80	40.91	35.88	27.55
Chinese	nikko	6.21	47.82	13.10	23.22	7.13	1.69	12.87	8.12	2.37
Standard Xinhua Test										
Chinese	cip2016	48.59	72.63	55.87	68.16	49.16	38.69	62.66	67.62	53.79
Chinese	ecnucs	45.09	72.41	59.77	62.07	47.13	38.21	59.55	65.26	50.12
Chinese	li16b	26.88	77.42	41.94	54.84	29.03	36.34	59.08	65.62	49.15
Chinese	goethe	28.73	70.33	39.56	57.14	28.57	22.26	42.81	45.55	36.30
Chinese	nikko	12.16	59.46	21.62	32.43	12.16	2.36	12.70	11.23	5.61
Standard Xinhua Development										
Chinese	ecnucs	53.59	86.27	67.97	70.59	56.21	39.43	59.86	65.25	50.50
Chinese	cip2016	45.21	83.56	54.79	68.49	45.21	39.82	62.82	67.98	53.41
Chinese	li16b	34.67	81.33	46.67	60.00	34.67	36.09	59.47	65.68	50.30
Chinese	goethe	17.45	63.09	40.27	44.30	20.13	22.67	45.70	45.34	36.93
Chinese	nikko	11.76	73.53	19.12	30.88	11.76	4.12	20.60	12.36	5.07

Table 7: F-score (**strict** scoring) of all subtasks separated by *Explicit* and *Implicit* discourse relations across the three data partitions—blind test, standard test and development set for both English and Chinese.

ing future research better calibrate their efforts. The evaluation data sets and the scorer we prepared for the shared task will be a useful benchmark for future research on shallow discourse parsing.

Language	Participant	Explicit				Implicit			
	ID	Parser F	A1 F	A2 F	A12 F	Parser F	A1 F	A2 F	A12 F
Blind Test									
English	oslopots	56.66	71.96	81.73	71.74	33.23	84.47	88.98	86.31
English	ecnucs	57.25	70.19	79.67	71.69	26.90	79.53	84.11	82.73
English	stepanov	51.03	65.71	78.05	63.83	24.24	80.54	85.49	84.43
English	tao0920	55.82	70.67	80.04	70.73	26.68	79.59	84.01	82.59
English	goethe	54.02	68.12	78.26	68.11	28.32	79.79	84.37	82.23
English	li16b	54.46	69.67	79.78	67.37	21.11	49.40	52.67	49.47
English	Soochow	54.11	64.79	79.52	66.08	19.73	64.14	69.07	66.11
English	clac	52.43	67.54	78.38	65.23	23.59	82.34	87.03	84.78
English	nguyenlab	52.74	66.52	75.65	67.13	17.16	56.09	60.85	56.31
English	VTNLPS16	54.46	68.47	78.14	68.70	22.83	70.89	75.26	72.02
English	rival2710	45.24	63.64	85.10	59.71	15.30	57.20	60.76	56.00
English	devanshu	47.79	62.33	73.16	62.86	17.04	60.11	63.30	60.84
English	nikko	40.65	56.08	56.50	50.23	5.19	23.16	21.81	16.52
English	iitbhu	47.19	62.48	64.70	54.53	11.51	56.74	60.37	55.63
Standard WSJ Test (Section 23)									
English	ecnucs	69.21	72.16	88.62	74.89	28.60	82.78	85.65	86.55
English	tao0920	67.80	71.38	88.25	73.29	27.77	82.30	85.16	86.15
English	goethe	66.41	69.37	86.35	70.92	28.05	83.63	85.44	86.88
English	li16b	62.37	66.67	87.94	67.28	22.08	62.45	62.45	62.20
English	oslopots	65.96	70.75	86.90	71.27	24.36	84.74	86.47	85.85
English	stepanov	63.31	68.11	86.24	68.76	23.72	82.53	85.02	86.22
English	Soochow	64.13	67.17	86.50	69.50	25.12	79.54	81.94	82.55
English	nguyenlab	61.81	69.71	82.48	68.60	15.39	56.02	58.53	56.66
English	clac	62.49	66.58	84.78	67.77	20.11	82.34	84.96	85.99
English	VTNLPS16	64.55	70.39	84.26	70.39	19.37	64.65	65.45	64.25
English	rival2710	56.51	63.32	91.29	61.35	20.52	77.83	79.71	80.39
English	devanshu	51.36	56.71	70.68	56.92	19.49	66.27	65.63	65.38
English	nikko	42.54	46.26	51.51	43.83	4.79	25.04	20.87	17.24
English	iitbhu								
Standard WSJ Development (Section 22)									
English	ecnucs	75.04	77.29	87.74	79.80	36.63	80.60	86.22	86.47
English	goethe	72.52	76.32	85.40	76.88	36.84	81.15	85.71	86.47
English	tao0920	73.48	76.20	86.79	78.16	34.32	79.92	85.79	85.84
English	li16b	67.38	71.88	85.04	71.28	31.69	66.61	68.73	68.10
English	oslopots	69.97	77.04	85.61	76.31	29.22	82.75	88.30	86.92
English	clac	66.67	70.94	83.56	71.51	28.03	80.32	86.04	85.74
English	stepanov	68.26	71.48	85.82	72.14	26.11	80.42	86.07	86.46
English	Soochow	67.79	70.58	84.34	72.65	28.14	79.09	82.50	82.32
English	VTNLPS16	70.45	74.75	83.88	75.48	24.22	65.67	68.32	66.41
English	nguyenlab	66.61	73.55	82.23	73.60	17.96	57.08	62.29	59.23
English	rival2710	63.32	70.30	89.85	67.64	25.11	77.20	82.78	81.81
English	devanshu	57.76	65.74	71.45	62.41	22.73	64.36	67.75	65.58
English	nikko	50.74	57.46	58.07	52.50	6.98	26.25	24.39	20.07
English	iitbhu								
Blind Test									
Chinese	cip2016	48.11	53.21	53.21	48.77	40.82	71.33	70.39	67.02
Chinese	ecnucs	49.36	55.39	55.76	51.02	36.66	67.54	66.56	62.38
Chinese	li16b	43.36	53.12	55.65	47.17	38.09	67.77	64.60	61.57
Chinese	goethe	43.02	53.29	54.37	48.17	18.56	53.62	52.21	48.17
Chinese	nikko	18.16	22.53	31.72	16.22	3.14	19.92	9.50	4.44
Standard Xinhua Test									
Chinese	cip2016	64.80	64.80	73.74	63.47	49.87	74.15	75.46	73.39
Chinese	ecnucs	66.67	67.82	72.41	65.82	49.63	72.21	72.46	71.04
Chinese	li16b	63.44	65.59	65.59	61.90	50.46	71.11	73.46	69.79
Chinese	goethe	62.64	62.64	68.13	60.49	27.74	53.77	54.79	50.36
Chinese	nikko	35.14	40.54	39.19	33.33	3.84	18.61	11.82	7.77
Standard Xinhua Development									
Chinese	ecnucs	78.43	73.20	79.74	79.39	49.93	76.02	71.57	68.09
Chinese	cip2016	75.34	64.38	71.23	72.41	50.68	77.78	73.88	70.94
Chinese	li16b	65.33	56.00	74.67	62.90	49.41	75.96	72.61	68.77
Chinese	goethe	57.72	59.06	65.77	56.91	27.06	61.71	54.92	47.58
Chinese	nikko	29.41	33.82	39.71	26.98	6.97	23.45	14.58	6.30

Table 8: F-score (**partial** scoring) of all subtasks separated by *Explicit* and *Implicit*discourse relations across the three data partitions—blind test, standard test and development set for both English and Chinese.

Acknowledgments

We would like to thank the Penn Discourse TreeBank team and the Chinese Discourse TreeBank Team, for allowing us to use the PDTB corpus and the CDTB corpus for the shared task. Thanks also go the LDC (Linguistic Data Consortium), who helped distribute

Language	Participant	All Senses			Explicit Senses			Implicit Senses		
	ID	P	R	F	P	R	F	P	R	F
Blind Test										
English	aarjay	41.51	41.44	41.47	78.70	78.42	78.56	9.95	9.95	9.95
English	BIT	18.69	18.69	18.69	17.99	17.99	17.99	19.30	19.30	19.30
English	clac	50.04	49.96	50.00	76.35	76.08	76.22	27.72	27.72	27.72
English	ecnucs	54.10	54.01	54.06	77.48	77.34	77.41	34.20	34.15	34.18
English	goethe	52.36	52.27	52.32	76.53	76.26	76.40	31.85	31.85	31.85
English	gtnlp	54.35	54.26	54.30	75.09	74.82	74.95	36.75	36.75	36.75
English	gw0	52.48	52.44	52.46	75.32	75.18	75.25	33.08	33.08	33.08
English	gw0	19.93	19.93	19.93	18.35	18.35	18.35	21.29	21.29	21.29
English	nguyenlab	51.37	51.28	51.32	74.91	74.64	74.77	31.44	31.39	31.42
English	oslopots	53.60	53.52	53.56	77.74	76.62	77.17	33.84	33.84	33.84
English	PurdueNLP	23.82	23.82	23.82	22.02	17.63	19.58	29.10	29.10	29.10
English	steven	41.44	41.44	41.44	65.42	62.95	64.16	24.04	23.12	23.58
English	tao0920	53.94	53.85	53.89	75.95	75.54	75.74	35.38	35.38	35.38
English	tbmihaylov	54.69	54.51	54.60	78.34	78.06	78.20	34.56	34.46	34.51
English	ykido	51.90	51.86	51.88	76.05	74.82	75.43	32.31	32.31	32.31
English	ttr	-	-	-	-	-	-	37.67	37.67	37.67
Standard WSJ Test (Section 23)										
English	aarjay	50.90	50.90	50.90	89.70	89.70	89.70	15.60	15.60	15.60
English	BIT	20.41	20.41	20.41	24.62	24.62	24.62	16.58	16.58	16.58
English	clac	57.36	57.36	57.36	89.48	89.48	89.48	28.13	28.13	28.13
English	ecnucs	64.34	64.34	64.34	90.13	90.13	90.13	40.95	40.87	40.91
English	goethe	62.64	62.64	62.64	90.13	90.13	90.13	37.61	37.61	37.61
English	gtnlp	60.93	60.93	60.93	89.48	89.48	89.48	34.95	34.95	34.95
English	gw0	58.45	58.45	58.45	89.48	89.48	89.48	30.21	30.21	30.21
English	gw0	17.11	17.11	17.11	15.51	15.51	15.51	18.56	18.56	18.56
English	nguyenlab	57.36	57.36	57.36	88.72	88.72	88.72	28.83	28.83	28.83
English	oslopots	60.62	60.62	60.62	90.13	90.13	90.13	33.76	33.76	33.76
English	PurdueNLP	59.95	59.95	59.95	87.96	87.96	87.96	34.45	34.45	34.45
English	steven	44.70	44.70	44.70	73.88	71.48	72.66	20.83	20.34	20.58
English	tao0920	62.69	62.69	62.69	89.59	89.59	89.59	38.20	38.20	38.20
English	tbmihaylov	63.31	63.31	63.31	89.80	89.80	89.80	39.19	39.19	39.19
English	ykido	54.73	54.73	54.73	90.41	90.02	90.22	22.61	22.61	22.61
English	ttr	-	-	-	-	-	-	36.13	36.13	36.13
Standard WSJ Development (Section 22)										
English	aarjay	62.43	62.43	62.43	91.50	91.50	91.50	36.85	36.85	36.85
English	BIT	20.10	20.10	20.10	23.22	23.22	23.22	17.36	17.36	17.36
English	clac	62.22	62.22	62.22	90.74	90.74	90.74	37.12	37.12	37.12
English	ecnucs	67.97	67.97	67.97	92.56	92.56	92.56	46.51	46.33	46.42
English	goethe	66.90	66.90	66.90	91.35	91.35	91.35	45.45	45.39	45.42
English	gtnlp	63.92	63.92	63.92	90.29	90.29	90.29	40.72	40.72	40.72
English	gw0	61.36	61.36	61.36	91.81	91.81	91.81	34.58	34.58	34.58
English	gw0	60.65	60.65	60.65	89.68	89.68	89.68	35.11	35.11	35.11
English	nguyenlab	60.51	60.51	60.51	90.29	90.29	90.29	34.31	34.31	34.31
English	oslopots	65.70	65.70	65.70	91.35	91.35	91.35	43.12	43.12	43.12
English	PurdueNLP	62.22	62.22	62.22	89.68	89.68	89.68	38.05	38.05	38.05
English	steven	46.88	46.88	46.88	72.30	70.11	71.19	26.94	26.44	26.68
English	tao0920	67.83	67.83	67.83	92.26	92.26	92.26	46.33	46.33	46.33
English	tbmihaylov	64.13	64.13	64.13	91.20	91.20	91.20	40.32	40.32	40.32
English	ykido	57.74	57.74	57.74	90.29	90.29	90.29	29.11	29.11	29.11
English	ttr	-	-	-	-	-	-	40.32	40.32	40.32
Blind Test										
Chinese	BIT	33.51	33.51	33.51	75.27	75.27	75.27	18.11	18.11	18.11
Chinese	ecnucs	64.73	64.73	64.73	77.24	76.15	76.69	60.52	60.52	60.52
Chinese	goethe	63.73	63.73	63.73	80.39	80.39	80.39	57.59	57.59	57.59
Chinese	gw0	72.92	72.92	72.92	78.98	78.98	78.98	70.68	70.68	70.68
Chinese	gw0	57.97	57.97	57.97	29.15	29.15	29.15	68.60	68.60	68.60
Chinese	tao0920	61.02	61.02	61.02	75.82	73.67	74.73	56.35	56.35	56.35
Chinese	ttr	-	-	-	-	-	-	63.38	63.38	63.38
Standard Xinhua Test										
Chinese	BIT	37.00	36.92	36.96	94.74	93.75	94.24	21.73	21.73	21.73
Chinese	ecnucs	77.09	76.92	77.01	94.74	93.75	94.24	72.42	72.42	72.42
Chinese	goethe	77.09	76.92	77.01	96.84	95.83	96.34	71.87	71.87	71.87
Chinese	gw0	70.11	70.11	70.11	92.71	92.71	92.71	64.07	64.07	64.07
Chinese	gw0	50.77	50.77	50.77	3.13	3.13	3.13	63.51	63.51	63.51
Chinese	tao0920	72.91	72.75	72.83	93.68	92.71	93.19	67.41	67.41	67.41
Chinese	ttr	-	-	-	-	-	-	70.47	70.47	70.47
Standard Xinhua Development										
Chinese	BIT	36.03	36.03	36.03	92.21	92.21	92.21	21.90	21.90	21.90
Chinese	ecnucs	78.07	78.07	78.07	96.10	96.10	96.10	73.53	73.53	73.53
Chinese	goethe	75.72	75.72	75.72	96.10	96.10	96.10	70.59	70.59	70.59
Chinese	gw0	72.06	72.06	72.06	93.51	93.51	93.51	66.67	66.67	66.67
Chinese	gw0	68.15	68.15	68.15	94.81	94.81	94.81	61.44	61.44	61.44
Chinese	tao0920	76.76	76.76	76.76	97.40	97.40	97.40	71.57	71.57	71.57
Chinese	ttr	-	-	-	-	-	-	63.38	63.38	63.38

Table 9: Discourse relation sense classification evaluation results (Supplementary evaluation). All participants are given gold standard discourse connectives and argument pairs.

the training and development data to participating teams. We are also very grateful to the TIRA team, who provided their evaluation platform, and especially to Martin Potthast for his technical assistance in using the TIRA platform and countless hours of troubleshooting.

This work was partially supported by the National Science Foundation via Grant Nos. 0910532 and IIS-1421067 and by the Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2013-T2-1-150.

References

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

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Prashant Chandrasekar, Xuan Zhang, Saurabh Chakravarty, Arijit Ray, John Krulick, and Alla Rozovskaya. 2016. The virginia tech system at conll-2016 shared task on shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Ziwei Fan, Zhenghua Li, and Min Zhang. 2016. Finding arguments as sequence labeling in discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Tim Gollub, Benno Stein, and Steven Burrows. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, August.

Devanshu Jain and Prasenjit Majumder. 2016. Da-iict submission for pdtb-styled discourse parser. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Ping Jian, Xiaohan She, Chenwei Zhang, Pengcheng Zhang, and Jian Feng. 2016. Discourse relation sense classification systems for conll-2016 shared task. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Xiaomian Kang, Haoran Li, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2016. An end-to-end chinese discourse parser with adaptation to explicit and non-explicit relation recognition. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Manpreet Kaur, Nishu Kumari, Anil Kumar Singh, and Rajeev Sangal. 2016. Iit (bhu) submission on the conll-2016 shared task: Shallow discourse parsing using semantic lexicons. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Yusuke Kido and Akiko Aizawa. 2016. Discourse relation sense classification with two-step classifiers. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Fang Kong, Sheng Li, and Guodong Zhou. 2015. The SoNLP-DP system in the CoNLL-2015 shared task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Fang Kong, Sheng Li, Junhui Li, Muhua Zhu, and Guodong Zhou. 2016. Sonlp-dp system for conll-2016 english shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Majid Laali, Andre Cianflone, and Leila Kosseim. 2016. The clac discourse parser at conll-2016. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Junhui Li, Fang Kong, Sheng Li, Muhua Zhu, and Guodong Zhou. 2016a. Sonlp-dp system for conll-2016 chinese shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Zhongyi Li, Hai Zhao, Chenxi Pang, Lili Wang, and Huan Wang. 2016b. A constituent syntactic parse tree based discourse parser. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*

Todor Mihaylov and Annette Frank. 2016. Discourse relation sense classification using cross-argument semantic similarity based on word embeddings. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Minh Nguyen. 2016. Sdp-jaist: A shallow discourse parsing system conll 2016 shared task. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Eric Velldal, and Lilja Ovrelid. 2016. Opt: Oslo–potsdam–teesside. pipelining rules, rankers, and classifier ensembles for shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Maria Leonor Pacheco, I-Ta Lee Lee, Xiao Zhang, Abdullah Khan Zehady, Pranjal Daga, Di Jin, Ayush Parolia, and Dan Goldwasser. 2016. Adapting event embedding for implicit discourse relation recognition. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inferencing for unlexicalized parsing. In *Proceedings of the Human Language Technology/North American Chapter of the Association for Computational Linguistics.*

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through Brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.*

Attapol T. Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in english and chinese. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Do we really need all those rich linguistic features?

a neural network-based approach to implicit sense labeling. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Evgeny Stepanov and Giuseppe Riccardi. 2016. Unitn end-to-end discourse parser for conll 2016 shared task. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Zhiguo Wang and Nianwen Xue. 2014. Joint pos tagging and transition-based constituent parsing in chinese with non-local features. In *ACL (1)*, pages 733–742.

Gregor Weiss and Marco Bajec. 2016. Discourse sense classification from scratch using focused rnns. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the 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(2):397–431.