# The CoNLL-2015 Shared Task on Shallow Discourse Parsing

Nianwen Xue* Hwee Tou Ng† Sameer Pradhan‡
Rashmi Prasad◇ Christopher Bryant† Attapol T. Rutherford*
* Brandeis University
xuen,tet@brandeis.edu
† National University of Singapore
nght,bryant@comp.nus.edu.sg
‡ Boulder Language Technologies
pradhan@bltek.com
◇ University of Wisconsin-Milwaukee
prasadr@uwm.edu

## Abstract

The CoNLL-2015 Shared Task is on Shallow Discourse Parsing, a task focusing on identifying individual discourse relations that are present in a natural language text. A discourse relation can be expressed explicitly or implicitly, and takes two arguments realized as sentences, clauses, or in some rare cases, phrases. Sixteen teams from three continents participated in this task. For the first time in the history of the CoNLL shared tasks, participating teams, instead of running their systems on the test set and submitting the output, were asked to deploy their systems on a remote virtual machine and use a web-based evaluation platform to run their systems on the test set. This meant they were unable to actually see the data set, thus preserving its integrity and ensuring its replicability. In this 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 Nineteenth Conference on Computational Natural Language Learning (CoNLL-2015) is on *Shallow Discourse Parsing* (SDP). In the course of the sixteen CoNLL shared tasks organized over the past two decades, progressing gradually to tackle phenomena at the word and phrase level phenomena and then the sentence and extra-sentential level, it was only very recently that discourse level processing has been addressed, with coreference resolution (Pradhan et al., 2011; Pradhan et al., 2012). The 2015 shared task takes the community a step further in that direction, with the potential to impact scores of richer language applications (Webber et al., 2012).

Given an English newswire text as input, the goal of the shared task is to detect and categorize discourse relations between discourse segments in the text. Just as there are different grammatical formalisms and representation frameworks in syntactic parsing, there are also different conceptions of the discourse structure of a text, and data sets annotated following these different theoretical frameworks (Stede, 2012; Webber et al., 2012; Prasad and Bunt, 2015). For example, the RST-DT Corpus (Carlson et al., 2003) is based on the Rhetorical Structure Theory of Mann and Thompson (1988) and produces a complete tree-structured RST analysis of a text, whereas the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008; Prasad et al., 2014) provides a shallow representation of discourse structure, in that each discourse relation is annotated independently of other discourse relations, leaving room for a high-level analysis that may attempt to connect them. For the CoNLL-2015 shared task, we chose to use the PDTB, as it is currently the largest data set annotated with discourse relations.[1]

[1] http://www.seas.upenn.edu/~pdtb



*Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–16,
Beijing, China, July 26-31, 2015. 

The necessary conditions are also in place for such a task. The release of the RST-DT and PDTB has attracted a significant amount of research on discourse parsing (Pitler et al., 2008; Duverle and Prendinger, 2009; Lin et al., 2009; Pitler et al., 2009; Subba and Di Eugenio, 2009; Zhou et al., 2010; Feng and Hirst, 2012; Ghosh et al., 2012; Park and Cardie, 2012; Wang et al., 2012; Biran and McKeown, 2013; Lan et al., 2013; Feng and Hirst, 2014; Ji and Eisenstein, 2014; Li and Nenkova, 2014; Li et al., 2014; Lin et al., 2014; Rutherford and Xue, 2014), and the momentum is building. Almost all of these recent attempts at discourse parsing use machine learning techniques, which is consistent with the theme of the CoNLL conference. The resurgence of deep learning techniques opens the door for innovative approaches to this problem. A shared task on shallow discourse parsing provides an ideal platform for the community to gain crucial insights on the relative strengths and weaknesses of "standard" feature-based learning techniques and "deep" representation learning techniques.

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 a set of discourse relations contained in the text. A discourse relation, as defined in the PDTB, from which the training data for the shared task is drawn, is a relation taking two abstract objects (events, states, facts, or propositions) as arguments. Discourse relations may be expressed with explicit connectives like *because*, *however*, *but*, or implicitly inferred between abstract object units. In the current version of the PDTB, non-explicit relations are inferred only between adjacent units. Each discourse relation is labeled with a sense selected from a sense hierarchy, and its arguments are generally in the form of 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;
2. Identify the spans of text that serve as the two arguments for each relation;
3. Label the arguments as (*Arg1* or *Arg2*) to indicate the order of the arguments;
4. Predict the sense of the discourse relation (e.g., "Cause", "Condition", "Contrast").

## 3 Data

### 3.1 Training and Development

The training data for the CoNLL-2015 Shared Task was adapted from the Penn Discourse TreeBank 2.0. (PDTB-2.0.) (Prasad et al., 2008; Prasad et al., 2014), annotated over the one million word Wall Street Journal (WSJ) corpus that has also been annotated with syntactic structures (the Penn TreeBank) (Marcus et al., 1993) and propositions (the Proposition Bank) (Palmer et al., 2005). The PDTB annotates discourse relations that hold between eventualities and propositions mentioned in text. Following a lexically grounded approach to annotation, the PDTB annotates relations realized explicitly by discourse connectives drawn from syntactically well-defined classes, as well as implicit relations between adjacent sentences when no explicit connective exists to relate the two. A limited but well-defined set of implicit relations are also annotated within sentences. Arguments of relations are annotated in each case, following the *minimality principle* for selecting all and only the material needed to interpret the relation. For explicit connectives, Arg2, which is defined as the argument with which the connective is syntactically associated, is in the same sentence as the connective (though not necessarily string adjacent), but Arg1, defined simply as the other argument, is unconstrained in terms of its distance from the connective and can be found anywhere in the text (Exs. 1-3). (All the following PDTB examples shown highlight Arg1 (in italics), Arg2 (in boldface), expressions realizing the relation (underlined), sense (in parentheses), and the WSJ file number for the text with the example (in square brackets)).

(1) GM officials want *to get their strategy to reduce capacity and the work force in place* before **those**

**talks begin**. (Temporal.Asynchronous.Precedence) [wsj_2338]

(2) But that ghost wouldn't settle for words, *he wanted money and people – lots*. So **Mr. Carter formed three new Army divisions and gave them to a new bureaucracy in Tampa called the Rapid Deployment Force**. (Contingency.Cause.Result) [wsj_2112]

(3) Big buyers like Procter & Gamble say *there are other spots on the globe, and in India, where the seed could be grown*. "It's not a crop that can't be doubled or tripled," says Mr. Krishnamurthy. But **no one has made a serious effort to transplant the crop**. (Comparison.Concession.Contra-expectation) [wsj_0515]

Between adjacent sentences unrelated by any explicit connective, four scenarios hold: (a) the sentences may be related by a discourse relation that has no lexical realization, in which case a connective (called an *Implicit* connective) is inserted to express the inferred relation (Ex. 4), (b) the sentences may be related by a discourse relation that is realized by some alternative non-connective expression (called *AltLex*), in which case these alternative lexicalizations are annotated as the carriers of the relation (Ex. 5), (c) the sentences may be related not by a discourse relation realizable by a connective or AltLex, but by an entity-based coherence relation, in which case the presence of such a relation is labeled *EntRel* (Ex 6), and (d) the sentences may not be related at all, in which case they are labeled *NoRel*. Relations annotated in these four scenarios are collectively referred to as *Non-Explicit* relations in this paper.

(4) *The Arabs had merely oil*. Implicit=while **These farmers may have a grip on the world's very heart**. (Comparison.Contrast) [wsj_0515]

(5) *Now, GM appears to be stepping up the pace of its factory consolidation to get in shape for the 1990s*. **One reason is mounting competition from new Japanese car plants in the U.S.** that are pouring out more than one million vehicles a year at costs lower than GM can match. (Contingency.Cause.Reason) [wsj_2338]

(6) *Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29*. EntRel **Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group**. [wsj_0001]

In addition to the argument structure of relations, the PDTB provides sense annotation for each discourse relation, capturing the polysemy of connectives. Senses are organized in a three-level hierarchy, with 4 top-level semantic *classes*. For each class, a second level of *types* is defined, and there are 16 such types. There is a third level of *subtype* which provides further refinement to the second level *types*. In the PDTB annotation, annotators are allowed back off to a higher level in the sense hierarchy if they are not certain about a lower level sense. That is, if they cannot distinguish between the subtypes under a type sense, they can just annotate the type level sense, and if there is further uncertainty in choosing among the types under a class sense, they can just annotate the class level sense. Most of the discourse relation instances in the PDTB are annotated with at least a type level sense, but there are also a small number annotated with only a class level sense.

The PDTB also provides annotations of attribution over all discourse relations and each of their arguments, as well as of text spans considered as supplementary to arguments of relations. However, both of these annotation types are excluded from the shared task.

PDTB-2.0. contains annotations of 40,600 discourse relations, distributed into the following five types: 18,459 Explicit relations, 16,053 Implicit relations, 624 AltLex relations, 5,210 EntRel relations, and 254 NoRel relations. We provide Sections 2–21 of the PDTB 2.0 release as the training set, and Section 22 as the development set.

### 3.2 Test Data

We provide two test sets for the shared task: Section 23 of the PDTB, and a blind test set we prepared especially for the shared task. The official ranking of the systems is based on their performance on the *blind test set*. In this section, we provide a detailed description of how the blind test set was prepared.

#### 3.2.1 Data Selection and Post-processing

For the blind test data, 30,158 words of untokenized English newswire texts were selected from a dump of English Wikinews[2], accessed 22nd October 2014, and annotated in accordance with PDTB 2.0 guidelines.

The raw Wikinews data was pre-processed as follows:

- News articles were extracted from the Wikinews XML dump[3] using the publicly available WikiExtractor.py script.[4]

[2] https://en.wikinews.org/

[3] https://dumps.wikimedia.org/enwikinews/20141119/enwikinews-20141119-pages-articles.xml.bz2

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

- Additional processing was done to remove any remaining XML information and produce a raw text version of each article (including its title).
- All paragraphs were double spaced to ease paragraph boundary identification.
- Each article was named according to its unique Wikinews ID such that it is accessible online at `http://en.wikinews.org/wiki?curid=`ID.

Initially, 30k words of text were selected from this processed data at random. However, it soon became apparent that some texts were too short for PDTB-style annotation or otherwise still contained remnant XML errors. Another issue was that since Wikinews texts are written by members of the public, rather than professionally trained journalists, some articles were considered as not up to the same standards of spelling and grammar as the WSJ texts in the PDTB.

For these reasons, despite making the decision to allow the correction of extremely minor errors (such as obvious typos and occasional article or preposition errors), just under half of the original 30k word random selection was ultimately deemed unsuitable for annotation. Consequently, the remaining texts were selected manually from Wikinews, with a slight preference for longer articles with many multi-sentence paragraphs that are more consistent with WSJ-style texts.

#### 3.2.2 Annotations

Annotation of the blind test set was carried out by two of the shared task organizers, one of whom (fifth author) was the main annotator (MA) while the other (fourth author), a lead developer of the PDTB, 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, EntRel, NoRel), relation realization (explicit connective, implicit connective, AltLex expression), arguments (Arg1 and Arg2), and sense of a discourse relation, using the PDTB annotation tool.[5] Unlike the PDTB guidelines, we did not allow back-off to the top class level during annotation. Every relation was annotated with a sense chosen from at least the second type level. Also different from the PDTB, attribution spans or attribution features were not annotated.

Before commencing official annotation, MA was trained in PDTB-2.0. style annotation by RA. A review of the guidelines was followed by double blind annotation (by MA and RA) of a small number of WSJ texts not previously annotated in the PDTB, and differences were then compared and discussed. MA then also underwent self-training by first annotating some WSJ texts that were already annotated in the PDTB, and then comparing these annotations, to further strengthen knowledge of the guidelines.

After the training period, the entire blind test data was annotated by MA over a period of a few weeks, and then reviewed by RA. Disagreements during the review were manually recorded using a formal scheme addressing all aspects of the annotation, including relation type, explicit connective identification, senses, and each of the arguments. This was done to verify the integrity of the blind test data and keep a record of any confusion or difficulty encountered during annotation. Manual entry of disagreements was done within the tool interface, through its commenting feature. A recorded comment in the tool is unique to a relation token and is recorded in a stand-off style. Disagreements were later resolved by consensus between MA and RA.

#### 3.2.3 Inter-annotator Agreement

The record of disagreements was utilized to compute inter-annotator agreement between MA and RA. The overall agreement was 76.5%, which represents the percentage of relations on which there was complete agreement. Agreement on explicit connective identification was 96.0%, representing the percentage of explicit connectives that both MA and RA identified as discourse connectives. We note here that if a connective was identified in the blind test data, but was not annotated in the PDTB despite its occurrence in the WSJ (e.g.,"after which time", "despite"), we did not consider it a potential connective and hence did not include it in the agreement calculation. When the textual context allowed it, such expressions were instead marked as AltLex.

We also did a more fine-grained assessment to determine agreement on Arg1, Arg2, Arg1+Arg2 (i.e., the number of relations on which the annotators agreed on both Arg1 and Arg2), and senses. This was done for all the relation types considered

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

together, as well as for Explicit and Non-Explicit relation types separately. Sense disagreement was computed using the CoNLL sense classification scheme (see Section 3.3), even though the annotation was done using the full PDTB sense classification scheme (see Table 2). The agreement percentages are shown in Table 1. When multiple senses were provided for a relation, a disagreement on any of the senses was counted as disagreement for the relation; disagreement on more than one of the senses was counted only once. Absence of a second sense by one annotator when the other did provide one was also counted as disagreement.

As the table shows, agreement on senses was reasonably high overall (85.5%), with agreement for Explicit relations expectedly higher (91.0%) than for Non-Explicit relations (80.9%). Overall agreement on arguments was also high, but in contrast to the senses, agreement was generally higher for the Non-Explicit than for Explicit relations. Agreement on the Arg1 of Explicit relations (89.6%) is, not surprisingly, lower than for Arg2 (98.7%), because the Arg1 of Explicit relations can be non-adjacent to the connective's sentence or clause, and thus, harder to identify. For the Non-Explicit relations, in contrast, but again to be expected, because of the argument adjacency constraint for such relations, agreement on Arg1 (95.0%) and Arg2 (96.4%) shows minimal difference. Table 1 also provides the percentage of relations with agreement on both Arg1 and Arg2, showing this to be higher for Non-Explicit relations (92.4%) than for Explicit relations (88.7%).

Compared to the agreement reported for the PDTB (Prasad et al., 2008; Miltsakaki et al., 2004), the results obtained here (See Table 1) are slightly better. PDTB agreement on Arg1 and Arg2 of Explicit relations is reported to be 86.3% and 94.1%, respectively, whereas overall agreement on arguments of Non-Explicit relations is 85.1%. For the senses, although the CoNLL senses do not exactly align with the PDTB senses, a rough correspondence can be assumed between the CoNLL classification as a whole and the type and subtype levels of the PDTB classification, for which PDTB reports 84% and 80%, respectively.

### 3.3 Adapting the PDTB Annotation for the shared task

The discourse relations annotated in the PDTB have many different elements, and it is impractical to predict all of them in the context of a shared task where participants have a relatively short time frame in which to complete the task. As a result, we had to make a number of exclusions and simplifications, which we describe below.

The core elements of a discourse relation are the two abstract objects as its arguments. In addition to this, some discourse relations include supplementary information that is relevant but not necessary (as per the minimality principle) to the interpretation of a discourse relation. Supplementary information is associated with arguments, and optionally marked with the labels "Sup1", for material supplementary to Arg1, and "Sup2", for material supplementary to Arg2. An example of a Sup1 annotation is shown in (7). In the shared task, supplementary information is excluded from evaluation when computing argument spans.

(7) ($_{Sup1}$ Average maturity was as short as 29 days at the start of this year), *when short-term interest rates were moving steadily upward.* Implicit=for example **The average seven-day compound yield of the funds reached 9.62% in late April** . (Expansion.Instantiation) [wsj_0982]

Also excluded from evaluation, to make the shared task manageable, are attribution relations annotated in PDTB. An example of an explicit attribution is "he says" in (8), marked over Arg1.

(8) When **Mr. Green won a $240,000 verdict in a land condemnation case against the state in June 1983** , he says *Judge O'Kicki unexpectedly awarded him an additional $100,000* (Temporal.Synchrony) [wsj_0267]

The PDTB senses form a hierarchical system of three levels, consisting of 4 *classes*, 16 *types*, and 23 *subtypes*. While all classes are divided into multiple types, some types do not have subtypes. Previous work on PDTB sense classification has mostly focused on classes (Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Biran and McKeown, 2013; Li and Nenkova, 2014; Rutherford and Xue, 2014). The senses that are the target of prediction in the CoNLL-2015 shared task are primarily based on the second-level types and a selected number of third-level subtypes. We made a few modifications to make the distinctions clearer and their distributions more balanced, and these changes are presented in Table 2. First, senses in the PDTB that have distinctions that are too subtle and thus too difficult to predict are collapsed.

| | Arg1 agr | Arg2 agr | Arg1+Arg2 agr | Sense agr |
|---|---|---|---|---|
| **Explicit** | 89.6% | 98.7% | 88.7% | 91.0% |
| **Non-Explicit** | 95.0% | 96.4% | 92.4% | 80.9% |
| **Total** | 92.5% | 97.4% | 90.7% | 85.5% |

Table 1: Inter-annotator agreement on blind test data annotation in various conditions.

| **CoNLL senses** | **PDTB senses** |
|---|---|
| Temporal.Synchronous | same |
| Temporal.Asynchronous.Precedence | same |
| Temporal.Asynchronous.Succession | same |
| ∗ Contingency.Cause.Reason | Contingency.Cause.Reason + Contingency.Pragmatic cause |
| Contingency.Cause.Result | same |
| ∗ Contingency.Condition | Contingency.Condition + Contingency.Pragmatic condition + Subtypes of Contingency.Condition + Subtypes of Contingency.Pragmatic Condition |
| ∗ Comparison.Contrast | Comparison.Contrast + Comparison.Pragmatic contrast + Subtypes of Comparison.Contrast |
| ∗ Comparison.Concession | Comparison.Concession + Comparison.Pragmatic concession + Subtypes of Comparison.Concession |
| ∗ Expansion.Conjunction | Expansion.Conjunction + Expansion.List |
| Expansion.Instantiation | same |
| ∗Expansion.Restatement | Expansion.Restatement + Subtypes of Expansion.Restatement |
| ∗ Expansion.Alternative | Expansion.Alternative.Conjunctive + Expansion.Alternative.Disjunctive |
| Expansion.Alternative.Chosen alternative | same |
| Expansion.Exception | same |
| EntRel | same |

Table 2: Flat list of 15 sense categories used in CoNLL-2015, with correspondences to PDTB senses. Senses that involve a change from the PDTB senses are marked ∗.

For example, "Contingency.Pragmatic cause" is merged into "Contingency.Cause.Reason", and "Contingency.Pragmatic condition" is merged into "Contingency.Condition". Second, the distinction between "Expansion.Conjunction" and "Expansion.List" is not clear in the PDTB and in fact, they seem very similar for the most part, so the latter is merged into the former. Third, while "Expansion.Alternative.Conjunctive" and "Expansion.Alternative.Disjunctive" are merged into "Expansion.Alternative", a third subtype of "Expansion.Alternative", "Expansion.Alternative.Chosen Alternative" is kept as a separate category as its meaning involves more than presentation of alternatives. Finally, while "EntRel" relations are not treated as discourse relations in the PDTB, we have included this category as a sense for sense classification since they are a kind of coherence relation and we require systems to label these relations in the shared task. In contrast, instances annotated with "NoRel" are not treated as discourse relations and are excluded from the training, development and test data sets. This means that a system needs to treat them as negative samples and *not* identify them as discourse relations. These changes have resulted in a *flat* list of 15 sense categories that need to be predicted in the shared task. A comparison of the PDTB senses and the senses used in the CoNLL shared task is presented in Table 2.

| Relation | Sense | WSJ-Train | WSJ-Dev | WSJ-Test |
|---|---|---|---|---|
| **Explicit** | **Overall** | 14722 | 680 | 923 |
| | Expansion.Conjunction | 4323 | 185 | 242 |
| | Comparison.Contrast | 2956 | 160 | 271 |
| | Contingency.Condition | 1148 | 50 | 63 |
| | Temporal.Synchrony | 1133 | 68 | 71 |
| | Comparison.Concession | 1080 | 12 | 27 |
| | Contingency.Cause.Reason | 943 | 38 | 74 |
| | Temporal.Asynchronous.Succession | 842 | 51 | 64 |
| | Temporal.Asynchronous.Precedence | 770 | 49 | 36 |
| | Contingency.Cause.Result | 487 | 19 | 38 |
| | Comparison | 347 | 20 | 1 |
| | Expansion.Instantiation | 236 | 9 | 21 |
| | Expansion.Alternative | 195 | 6 | 5 |
| | Expansion.Restatement | 121 | 6 | 7 |
| | Expansion.Alternative.Chosen-alternative | 96 | 6 | 3 |
| | Expansion | 24 | 0 | 0 |
| | Expansion.Exception | 13 | 0 | 0 |
| | Temporal | 4 | 1 | 0 |
| | Temporal.Asynchronous | 3 | 0 | 0 |
| | Contingency | 1 | 0 | 0 |
| **Implicit** | **Overall** | 13156 | 522 | 769 |
| | Expansion.Conjunction | 3227 | 120 | 141 |
| | Expansion.Restatement | 2486 | 101 | 190 |
| | Contingency.Cause.Reason | 2059 | 73 | 113 |
| | Comparison.Contrast | 1614 | 82 | 127 |
| | Contingency.Cause.Result | 1372 | 49 | 89 |
| | Expansion.Instantiation | 1132 | 47 | 69 |
| | Temporal.Asynchronous.Precedence | 418 | 25 | 7 |
| | Comparison.Concession | 193 | 5 | 5 |
| | Temporal.Synchrony | 153 | 8 | 5 |
| | Comparison | 145 | 1 | 0 |
| | Expansion.Alternative.Chosen-alternative | 142 | 2 | 15 |
| | Temporal.Asynchronous.Succession | 125 | 3 | 5 |
| | Expansion | 73 | 6 | 3 |
| | Expansion.Alternative | 11 | 0 | 0 |
| | Contingency.Condition | 2 | 0 | 0 |
| | Temporal | 1 | 0 | 0 |
| | Expansion.Exception | 1 | 0 | 0 |
| | Contingency.Cause | 1 | 0 | 0 |
| | Contingency | 1 | 0 | 0 |
| **EntRel** | **Overall** | 4133 | 215 | 217 |
| | EntRel | 4133 | 215 | 217 |
| **AltLex** | **Overall** | 524 | 19 | 19 |
| | Contingency.Cause.Result | 147 | 4 | 8 |
| | Expansion.Conjunction | 94 | 3 | 8 |
| | Contingency.Cause.Reason | 76 | 5 | 8 |
| | Expansion.Restatement | 57 | 0 | 1 |
| | Temporal.Asynchronous.Precedence | 42 | 2 | 2 |
| | Expansion.Instantiation | 33 | 1 | 1 |
| | Comparison.Contrast | 32 | 2 | 1 |
| | Temporal.Asynchronous.Succession | 18 | 0 | 0 |
| | Temporal.Synchrony | 16 | 1 | 0 |
| | Comparison.Concession | 4 | 0 | 1 |
| | Expansion | 2 | 0 | 0 |
| | Contingency.Condition | 2 | 0 | 0 |
| | Expansion.Exception | 1 | 0 | 0 |
| | Expansion.Restatement | 0 | 1 | 0 |
| | Expansion.Alternative | 0 | 0 | 0 |

Table 3: Distribution of senses across the four relation types in the WSJ PDTB data used for the shared task. The total numbers of the relations here are less than in the complete PDTB release because some sections (00, 01, and 24) are excluded for the shared task, following standard split of WSJ data in the evaluation community. We are intentionally withholding distribution over the blind test set in case there is a repeat of the SDP shared task using the same test set.

Table 3 shows the distribution of the senses across the four discourse relations within the WSJ PDTB data[6]. We are intentionally withholding the sense distribution across the blind test set in case there is a repeat of the SDP shared task using the same test set.

## 4 Evaluation

### 4.1 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 to use the following linguistic resources in the closed track, other than the training set:

- Brown clusters
- VerbNet
- Sentiment lexicon
- 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 processed with state-of-the-art NLP tools:

- 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))

As it turned out, all of the teams this year chose to participate in the closed track.

### 4.2 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 were asked to manually run their system on the blind test set without the gold standard labels, and submit the output for evaluation. This year, 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 the test sets. 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 very 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,

[6]There is a small number of instances in the PDTB training set that are only annotated with the class level sense. We did not take them out of the training set for the sake of completeness.

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.

### 4.3 Evaluation metrics and scorer

A shallow discourse parser is evaluated based on the end-to-end $F_1$ score on a per-discourse relation basis. The input to the system consists of documents with gold-standard word tokens along with their automatic parses. We do not pre-identify the discourse connectives or any other elements of the discourse annotation. The shallow discourse parser must output a list of discourse relations that consist of the 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. A relation is correctly predicted if (a) the discourse connective is correctly detected (for Explicit discourse relations), (b) the sense of the discourse connective is correctly predicted, and (c) the text spans of its two arguments are correctly predicted (Arg1 and Arg2).

Although the submissions are ranked based on the relation $F_1$ score, the scorer also provides component-wise evaluation with error propagation. The scorer computes the precision, recall, and $F_1$ for the following[7]:

- 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 the 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.

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). Systems are not given any credit for partial match on argument spans.

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.

Additionally, the scorer provides a breakdown of the discourse parser performance for Explicit and Non-Explicit discourse relations.

## 5 Approaches

The Shallow Discourse Parsing (SDP) task this year requires the development of an end-to-end system that potentially involves many components. All participating systems adopt some variation of the pipeline architecture proposed by Lin et al (2014), which has components for identify-

[7] Available at: `http://www.github.com/attapol/conll15st`

| System | learning methods | resources used | extra resources |
|---|---|---|---|
| ECNU | Naive Bayes, maxent | Brown clusters, MPQA subjectivity lexicon | no |
| Trento | CRF++, AdaBoost | Brown clusters, dependency/phrase structure parses | no |
| Soochow | Maxent in Open NLP | VerbNet, MPQA subjectivity lexicon, Brown clusters | no |
| JAIST | CRF++, LibSVM (SMO) | syntactic parses, Brown clusters | no |
| UIUC | Liblinear | Brown clusters, MPQA lexicon | no |
| Concordia | C4.5 (Weka) | ClearTK, syntactic parse | no |
| ∗UT Dallas | - | - | - |
| NTT | Rule-based argument extraction and SVM based sense classification | Brown clusters, dependency trees | no |
| AU KBC | CRF++ for both arguments and sense, and rules | MPQA, VerbNet, Brown clusters | no |
| CAS | OpenNLP maxent | phrase structure trees | no |
| Dublin 1 | RNN (Theano) for argument extraction, Maxent for others | syntactic features, skip-gram word embeddings | no |
| Dublin 2 | LibSVM, Theano, word2vec | Brown clusters | no |
| Goethe University Frankfurt | SVM, rule-based | Brown clusters, word embeddings | no |
| IIT | Naive Bayes, Maxent | syntactic parses, Boxer | no |
| SJTU | Maxent | no external resource used | no |
| ∗PKU | - | - | - |

Table 4: Approaches of participating systems. Teams that have not submitted a system description paper are marked with ∗.

ing 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. Most participating systems cast discourse connective identification and argument extraction as token-level sequence labeling tasks, while a few systems use rule-based approaches to extract the arguments. Sense determination is cast as a straightforward multi-category classification task. Most systems use machine learning techniques to determine the senses, but there are also systems that, due to lack of time, adopt a simple baseline approach that detects the most frequent sense based on the training data.

In terms of learning techniques, all participating systems except the two systems submitted by the Dublin team use standard "shallow" learning models that take binary features as input. For sequence labeling subtasks such as discourse connective identification and argument extraction, the preferred learning method is Conditional Random Fields (CRF). For sense determination, a variety of learning methods have been used, including Maximum Entropy, Support Vector Machines, and decision trees. In the last couple of years, neural networks have experienced a resurgence and have been shown to be effective in many natural language processing tasks. Neural network based models on discourse parsing have also started to appear (Ji and Eisenstein, 2014). The use of neural networks for the SDP task this year represents a minority, presumably because researchers are still less familiar with neural network based techniques, compared with standard "shallow" learning techniques, and it is difficult to use a new

learning technique to good effect within a short time window. In this shared task, only the Dublin University team attempted to use neural networks as a learning approach in their system components. In their first submission (Dublin I), Recurrent Neural Networks (RNN) are used for token level sequence labeling in the argument extraction task. In their second submission, paragraph embeddings are used in a neural network model to determine the senses of discourse relations.

The discussion of learning techniques cannot be entirely separated from the use of features and the linguistic resources that are used to extract them. Standard "shallow" architectures typically make use of discrete features while neural networks generally use continuous real-valued features such as word and paragraph embeddings. For discourse connective and argument extraction, token level features extracted from a fixed window centered on the target word token are generally used, and so are features extracted from syntactic parses. Distributional representations such as Brown clusters have generally been used to determine the senses (Chiarcos and Schenk, 2015; Devi et al., 2015; Kong et al., 2015; Song et al., 2015; Stepanov et al., 2015; Wang and Lan, 2015; Wang et al., 2015; Yoshida et al., 2015), although one team also used them in the sequence labeling task for argument extraction (Nguyen et al., 2015). Additional resources used by some systems for sense determination include word embeddings (Chiarcos and Schenk, 2015; Wang et al., 2015), VerbNet classes (Devi et al., 2015; Kong et al., 2015), and the MPQA polarity lexicon (Devi et al., 2015; Kong et al., 2015; Wang and Lan, 2015). Table 4 provides a summary of the different approaches.

## 6 Results

Table 5 shows the performance of all participating systems across the three test evaluation sets: i) (Official) Blind test set; ii) Standard WSJ test set; iii) Standard WSJ development set. The official rankings are based on the blind test set annotated specifically for this shared task. The top-ranked system is the submission by East China Normal University (Wang and Lan, 2015). As discussed in Section 4, the evaluation metric is very strict, and is based on exact match for the extraction of argument spans. For the detection of discourse connectives, only the head of a discourse connective has to be correctly detected. Errors in the beginning of the pipeline will propagate to the end, and other than word tokenization, all input to the participating systems is automatically generated, so the overall accuracy reflects results in realistic situations. The scores are very low, with the top system achieving an overall parsing score of 24.00% (F1) on the blind test set and 29.69% (F1) on the Wall Street Journal (WSJ) test set. For comparison purposes, the National University of Singapore team re-implemented the state-of-the-art end-to-end parser described in (Lin et al., 2014), and this system achieves an F1 of 19.98% on the WSJ test set. This shows that a fair amount of progress has been made against the Lin et al baseline.

The rankings are generally consistent across the two test sets, with the largest change in ranking from the NTT team and the Goethe University team. This is perhaps not a coincidence: both teams used rule-based approaches to extract arguments. The rules worked well on the WSJ test set which draws from the same source as the development set, but might not adapt well to the blind test set, which is drawn from a different source. Machine-learning based approaches generally can better adapt to new data sets.

Due to the short time frame participants had to complete an end-to-end task, teams chose to focus on either argument extraction components or the sense classification components, or in the case of sense classification, either focus on the classification of senses for Explicit relations or senses for Non-Explicit relations. A detailed breakdown of the performance for Explicit versus Non-Explicit discourse relations is presented in Table 6. In general, parser performance for Explicit discourse relations is much higher than that of Non-Explicit discourse relations. The difficulty for Non-Explicit discourse relations mostly stems from Non-Explicit sense classification. This is evidenced by the fact that even for systems that achieve higher argument extraction accuracy for Non-Explicit discourse relations than Explicit discourse relations, the overall parser accuracy is still lower for Non-Explicit relations. The lower accuracy in sense classification thus drags down the overall parser accuracy for Non-Explicit discourse relations.

## 7 Conclusions

Sixteen teams from three continents participated in the CoNLL-2015 Shared Task on shallow dis-

| Rank | | Participant | | Argument | | | Connective | | | Parser | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **O** | **L** | **Organization** | **ID** | **F** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** |
| | | | | **Blind Test** | | | | | | | | |
| 1 | 1 | East China Normal University | wangj | 46.37 | 45.77 | 46.98 | 91.86 | 93.48 | 90.29 | 24.00 | 23.69 | 24.32 |
| 2 | 2 | University of Trento | stepanov | 38.86 | 37.25 | 40.61 | 89.92 | 92.57 | 87.41 | 21.84 | 20.94 | 22.83 |
| 3 | 3 | Soochow University | kong | 33.23 | 35.57 | 31.18 | 91.62 | 92.80 | 90.47 | 18.51 | 19.81 | 17.37 |
| 4 | 4 | Japan Advanced Institute of Science and Tech. | nguyen | 32.11 | 42.72 | 25.72 | 61.66 | 88.55 | 47.30 | 18.28 | 24.31 | 14.64 |
| 5 | 5 | UIUC Cognitive Computing Group | song | 41.31 | 40.48 | 42.18 | 87.98 | 89.11 | 86.87 | 17.98 | 17.62 | 18.36 |
| 6 | 6 | Concordia University | laali | 23.29 | 35.67 | 17.29 | 90.19 | 87.88 | 92.63 | 17.38 | 26.62 | 12.90 |
| 7 | 7 | University of Texas Dallas | xue | 30.22 | 31.70 | 28.87 | 89.90 | 92.73 | 87.23 | 17.06 | 17.89 | 16.29 |
| 8 | 8 | Nippon Telegraph and Telephone Lab Japan | yoshida | 35.55 | 52.16 | 26.96 | 51.04 | 92.45 | 35.25 | 15.70 | 23.04 | 11.91 |
| 9 | 9 | AU KBC Research Center | devi | 33.17 | 35.12 | 31.43 | 84.49 | 92.32 | 77.88 | 15.02 | 15.90 | 14.23 |
| 10 | 10 | Chinese Academy of Sciences | xu15 | 21.95 | 28.88 | 17.70 | 82.60 | 93.02 | 74.28 | 12.62 | 16.60 | 10.17 |
| 11 | 11 | Dublin City University 1 | wangl | 22.09 | 19.26 | 25.89 | 79.43 | 84.87 | 74.64 | 11.15 | 9.72 | 13.07 |
| 12 | 12 | Dublin City University 2 | okita | 21.52 | 18.77 | 25.23 | 79.43 | 84.87 | 74.64 | 10.66 | 9.29 | 12.49 |
| 13 | 13 | Goethe University Frankfurt | chiarcos | 29.21 | 26.00 | 33.33 | 51.18 | 59.38 | 44.96 | 9.13 | 8.13 | 10.42 |
| 14 | 14 | India Institute of Tech. | mukherjee | 21.71 | 18.14 | 27.05 | 89.30 | 91.67 | 87.05 | 7.64 | 6.38 | 9.51 |
| 15 | 15 | Shanghai Jiao Tong University 1 | chen | 4.70 | 4.53 | 4.88 | 81.68 | 81.17 | 82.19 | 3.58 | 3.46 | 3.72 |
| 16 | 16 | Peking University | xu15b | 12.70 | 10.54 | 15.96 | 59.11 | 58.69 | 59.53 | 0.92 | 0.76 | 1.16 |
| | | | | **Standard WSJ Test (Section 23)** | | | | | | | | |
| 1 | 1 | East China Normal University | wangj | 49.42 | 48.72 | 50.13 | 94.21 | 94.94 | 93.50 | 29.69 | 29.27 | 30.12 |
| 2 | 2 | University of Trento | stepanov | 40.71 | 39.71 | 41.77 | 92.77 | 93.80 | 91.77 | 25.33 | 24.71 | 25.99 |
| 8 | 3 | Nippon Telegraph and Telephone Lab Japan | yoshida | 43.77 | 48.83 | 39.66 | 89.12 | 91.84 | 86.57 | 24.99 | 27.87 | 22.64 |
| 7 | 4 | University of Texas Dallas | xue | 30.26 | 31.78 | 28.88 | 89.33 | 91.20 | 87.54 | 21.72 | 22.81 | 20.73 |
| 6 | 5 | Concordia University | laali | 24.81 | 36.98 | 18.67 | 91.38 | 88.76 | 94.15 | 21.25 | 31.66 | 15.99 |
| 3 | 6 | Soochow University | kong | 37.01 | 34.69 | 39.66 | 94.77 | 95.39 | 94.15 | 20.64 | 19.35 | 22.12 |
| 5 | 7 | UIUC Cognitive Computing Group | song | 38.18 | 35.73 | 41.00 | 91.83 | 92.33 | 91.33 | 20.27 | 18.97 | 21.76 |
| 4 | 8 | Japan Advanced Institute of Science and Tech. | nguyen | 35.43 | 52.98 | 26.61 | 63.89 | 91.87 | 48.97 | 20.25 | 30.29 | 15.21 |
| 13 | 9 | Goethe University Frankfurt | chiarcos | 36.78 | 36.58 | 36.98 | 68.19 | 71.96 | 64.79 | 15.23 | 15.15 | 15.32 |
| 10 | 10 | Chinese Academy of Sciences | xu15 | 23.36 | 28.05 | 20.01 | 90.64 | 95.12 | 86.57 | 15.05 | 18.08 | 12.89 |
| 9 | 11 | AU KBC Research Center | devi | 31.26 | 31.76 | 30.79 | 86.44 | 94.36 | 79.74 | 14.61 | 14.84 | 14.39 |
| 11 | 12 | Dublin City University 1 | wangl | 25.46 | 21.74 | 30.74 | 87.99 | 90.40 | 85.70 | 12.73 | 10.87 | 15.37 |
| 12 | 13 | Dublin City University 2 | okita | 24.55 | 20.95 | 29.65 | 88.06 | 90.32 | 85.92 | 12.30 | 10.49 | 14.85 |
| 14 | 14 | India Institute of Tech. | mukherjee | 22.52 | 18.19 | 29.55 | 93.06 | 93.93 | 92.20 | 7.15 | 5.78 | 9.39 |
| 15 | 15 | Shanghai Jiao Tong University 1 | chen | 4.57 | 4.24 | 4.95 | 78.67 | 77.84 | 79.52 | 4.43 | 4.11 | 4.80 |
| 16 | 16 | Peking University | xu15b | 13.24 | 10.65 | 17.48 | 58.04 | 57.28 | 58.83 | 2.11 | 1.70 | 2.78 |
| | | | | **Development** | | | | | | | | |
| 1 | 1 | East China Normal University | wangj | 57.21 | 56.84 | 57.59 | 95.14 | 95.28 | 95.00 | 37.84 | 37.59 | 38.09 |
| 8 | 2 | Nippon Telegraph and Telephone Lab Japan | yoshida | 51.42 | 56.56 | 47.14 | 88.94 | 92.39 | 85.74 | 31.60 | 34.75 | 28.97 |
| 2 | 3 | University of Trento | stepanov | 45.34 | 44.99 | 45.68 | 93.79 | 94.35 | 93.24 | 30.27 | 30.04 | 30.50 |
| 3 | 4 | Soochow University | kong | 43.12 | 41.06 | 45.40 | 94.22 | 94.93 | 93.53 | 26.32 | 25.06 | 27.72 |
| 4 | 5 | Japan Advanced Institute of Science and Tech. | nguyen | 40.07 | 58.92 | 30.36 | 65.53 | 91.56 | 51.03 | 26.10 | 38.38 | 19.78 |
| 9 | 6 | AU KBC Research Center | devi | 42.96 | 42.28 | 43.66 | 92.63 | 98.03 | 87.79 | 25.76 | 25.35 | 26.18 |
| 6 | 7 | Concordia University | laali | 29.87 | 44.43 | 22.49 | 92.25 | 89.27 | 95.44 | 25.71 | 38.24 | 19.36 |
| 5 | 8 | UIUC Cognitive Computing Group | song | 43.44 | 41.24 | 45.89 | 91.45 | 93.27 | 89.71 | 25.12 | 23.84 | 26.53 |
| 7 | 9 | University of Texas Dallas | xue | 35.78 | 37.77 | 33.98 | 93.43 | 94.85 | 92.06 | 24.19 | 25.54 | 22.98 |
| 10 | 10 | Chinese Academy of Sciences | xu15 | 26.68 | 32.06 | 22.84 | 91.52 | 95.23 | 88.09 | 18.14 | 21.80 | 15.53 |
| 13 | 11 | Goethe University Frankfurt | chiarcos | 41.58 | 42.08 | 41.09 | 63.17 | 67.45 | 59.41 | 17.12 | 17.33 | 16.92 |
| 11 | 12 | Dublin City University 1 | wangl | 29.75 | 25.59 | 35.52 | 85.65 | 90.10 | 81.62 | 16.51 | 14.20 | 19.71 |
| 12 | 13 | Dublin City University 2 | okita | 29.09 | 24.98 | 34.82 | 86.33 | 90.35 | 82.65 | 15.36 | 13.19 | 18.38 |
| 14 | 14 | India Institute of Technology | mukherjee | 26.78 | 21.89 | 34.47 | 93.55 | 95.41 | 91.76 | 8.82 | 7.21 | 11.35 |
| 15 | 15 | Shanghai Jiao Tong University 1 | chen | 6.81 | 6.43 | 7.24 | 86.09 | 85.28 | 86.91 | 6.55 | 6.18 | 6.96 |
| 16 | 16 | Peking University | xu15b | 12.64 | 9.00 | 21.24 | 51.54 | 42.64 | 65.15 | 1.49 | 1.06 | 2.51 |

Table 5: Scoreboard for the CoNLL-2015 shared task showing performance across the tasks and the three data partitions—blind test, standard test (WSJ-23) and development. The Column **O** and **L** refer to official and local ranks. The red highlighted rows indicate a system (JAIST) that performed poorly on the WSJ test set, but did much better on the blind test set. The blue highlighted rows indicate the opposite phenomena for a system (NTT) that ranked higher on the WSJ development and test partitions, but dropped in rank on the blind test set.

| Rank | | | Participant | | Explicit | | | | | Non-Explicit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **O** | **E** | **I** | **Organization** | **ID** | **A12** | **A1** | **A2** | **Conn.** | **Parser** | **A12** | **A1** | **A2** | **Parser** |
| | | | | **Blind Test** | | | | | | | | | |
| 7 | 1 | 11 | University of Texas Dallas | xue | 40.04 | 49.68 | 70.06 | 89.90 | 30.58 | 21.61 | 25.02 | 34.77 | 5.20 |
| 1 | 2 | 1 | East China Normal University | wangj | 41.35 | 48.31 | 74.29 | 91.86 | 30.38 | 50.41 | 60.87 | 74.58 | 18.87 |
| 2 | 3 | 2 | University of Trento | stepanov | 39.59 | 49.03 | 70.68 | 89.92 | 29.97 | 38.31 | 43.29 | 56.57 | 15.77 |
| 6 | 4 | 15 | Concordia University | laali | 36.60 | 45.18 | 69.18 | 90.19 | 27.32 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4 | 5 | 8 | Japan Advanced Institute of Science and Tech. | nguyen | 34.23 | 44.08 | 51.35 | 61.66 | 27.20 | 30.44 | 36.90 | 46.13 | 11.25 |
| 9 | 6 | 10 | AU KBC Research Center | devi | 34.73 | 44.49 | 64.20 | 84.49 | 26.73 | 31.91 | 35.70 | 46.60 | 5.53 |
| 5 | 7 | 5 | UIUC Cognitive Computing Group | song | 30.05 | 37.89 | 60.11 | 87.98 | 23.32 | 50.18 | 59.52 | 74.40 | 13.57 |
| 3 | 8 | 4 | Soochow University | kong | 30.42 | 36.43 | 73.04 | 91.62 | 22.95 | 35.87 | 49.87 | 51.07 | 14.35 |
| 10 | 9 | 13 | Chinese Academy of Sciences | xu15 | 27.20 | 36.40 | 61.00 | 82.60 | 22.20 | 16.42 | 19.79 | 27.16 | 2.53 |
| 8 | 10 | 3 | Nippon Telegraph and Telephone Lab Japan | yoshida | 21.61 | 28.13 | 38.02 | 51.04 | 16.93 | 45.59 | 53.66 | 62.29 | 14.82 |
| 13 | 11 | 9 | Goethe University Frankfurt | chiarcos | 19.04 | 26.41 | 36.85 | 51.18 | 13.51 | 34.79 | 44.33 | 53.54 | 6.73 |
| 14 | 12 | 12 | India Institute of Technology | mukherjee | 13.65 | 22.32 | 61.99 | 89.30 | 12.36 | 26.24 | 37.03 | 41.49 | 4.98 |
| 11 | 13 | 6 | Dublin City University 1 | wangl | 12.47 | 18.05 | 36.65 | 87.81 | 9.12 | 27.84 | 39.46 | 44.27 | 12.74 |
| 15 | 14 | 16 | Shanghai Jiao Tong University 1 | chen | 10.55 | 13.94 | 48.97 | 81.68 | 8.04 | | | | 0.00 |
| 12 | 15 | 7 | Dublin City University 2 | okita | 11.10 | 16.65 | 28.13 | 79.43 | 7.85 | 27.61 | 39.24 | 44.05 | 12.30 |
| 16 | 16 | 14 | Peking University | xu15b | 3.57 | 6.07 | 20.89 | 59.11 | 2.32 | 18.02 | 26.46 | 28.85 | 0.10 |
| | | | | **Standard WSJ Test (Section 23)** | | | | | | | | | |
| 1 | 1 | 1 | East China Normal University | wangj | 45.20 | 50.66 | 77.40 | 94.21 | 39.96 | 53.09 | 67.17 | 68.41 | 20.74 |
| 2 | 2 | 5 | University of Trento | stepanov | 44.58 | 50.05 | 76.23 | 92.77 | 39.54 | 37.44 | 44.50 | 47.56 | 13.28 |
| 7 | 3 | 10 | University of Texas Dallas | xue | 41.57 | 49.75 | 68.55 | 89.33 | 37.59 | 19.45 | 24.74 | 25.37 | 6.55 |
| 8 | 4 | 3 | Nippon Telegraph and Telephone Lab Japan | yoshida | 38.82 | 46.07 | 68.38 | 89.12 | 34.47 | 48.81 | 57.99 | 60.08 | 15.11 |
| 4 | 5 | 9 | Japan Advanced Institute of Science and Tech. | nguyen | 38.16 | 43.82 | 56.25 | 63.89 | 33.22 | 32.44 | 38.85 | 38.85 | 8.01 |
| 6 | 6 | 15 | Concordia University | laali | 38.07 | 44.69 | 72.34 | 91.38 | 32.60 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5 | 7 | 4 | UIUC Cognitive Computing Group | song | 30.39 | 37.25 | 66.67 | 91.83 | 27.02 | 44.33 | 57.13 | 60.14 | 14.95 |
| 9 | 8 | 11 | AU KBC Research Center | devi | 30.77 | 36.64 | 49.68 | 86.44 | 26.78 | 31.66 | 38.28 | 43.29 | 4.82 |
| 10 | 9 | 13 | Chinese Academy of Sciences | xu15 | 28.70 | 36.07 | 63.53 | 90.64 | 25.75 | 17.32 | 23.35 | 23.48 | 2.95 |
| 3 | 10 | 2 | Soochow University | kong | 30.21 | 34.02 | 74.48 | 94.77 | 25.30 | 42.38 | 57.71 | 54.95 | 16.97 |
| 13 | 11 | 8 | Goethe University Frankfurt | chiarcos | 25.20 | 30.79 | 50.74 | 68.19 | 21.89 | 46.25 | 62.84 | 63.50 | 9.79 |
| 11 | 12 | 7 | Dublin City University 1 | wangl | 19.36 | 24.42 | 46.20 | 93.18 | 17.38 | 30.70 | 43.04 | 40.75 | 11.50 |
| 12 | 13 | 6 | Dublin City University 2 | okita | 14.66 | 21.10 | 38.20 | 88.06 | 13.21 | 30.73 | 43.01 | 40.72 | 11.72 |
| 14 | 14 | 12 | India Institute of Technology | mukherjee | 13.78 | 20.34 | 59.38 | 93.06 | 12.90 | 27.42 | 38.47 | 36.44 | 3.93 |
| 15 | 15 | 16 | Shanghai Jiao Tong University 1 | chen | 10.29 | 14.68 | 48.77 | 78.67 | 9.97 | | 0.09 | | 0.00 |
| 16 | 16 | 14 | Peking University | xu15b | 4.28 | 6.31 | 24.05 | 58.04 | 3.53 | 18.40 | 25.60 | 24.25 | 1.29 |
| | | | | **Development** | | | | | | | | | |
| 9 | 1 | 10 | AU KBC Research Center | devi | 54.69 | 62.90 | 75.91 | 92.80 | 49.11 | 35.03 | 40.89 | 45.10 | 7.64 |
| 1 | 2 | 1 | East China Normal University | wangj | 54.05 | 61.56 | 80.56 | 95.14 | 48.16 | 60.01 | 70.32 | 74.23 | 28.70 |
| 2 | 3 | 6 | University of Trento | stepanov | 51.33 | 57.10 | 78.70 | 93.79 | 46.89 | 40.08 | 45.91 | 49.42 | 15.69 |
| 8 | 4 | 3 | Nippon Telegraph and Telephone Lab Japan | yoshida | 47.90 | 55.68 | 72.16 | 88.94 | 43.02 | 54.92 | 62.48 | 67.47 | 20.27 |
| 7 | 5 | 11 | University of Texas Dallas | xue | 48.51 | 57.46 | 72.24 | 93.43 | 41.49 | 23.49 | 27.67 | 29.83 | 7.49 |
| 4 | 6 | 8 | Japan Advanced Institute of Science and Tech. | nguyen | 45.14 | 51.56 | 57.79 | 65.53 | 41.17 | 35.09 | 40.29 | 40.29 | 11.82 |
| 6 | 7 | 15 | Concordia University | laali | 45.91 | 53.16 | 75.34 | 92.25 | 39.52 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5 | 8 | 4 | UIUC Cognitive Computing Group | song | 34.78 | 43.18 | 65.97 | 91.45 | 31.18 | 49.88 | 60.59 | 64.47 | 20.00 |
| 10 | 9 | 13 | Chinese Academy of Sciences | xu15 | 33.16 | 41.71 | 67.99 | 91.52 | 30.25 | 19.30 | 23.13 | 23.83 | 4.35 |
| 3 | 10 | 2 | Soochow University | kong | 34.67 | 38.67 | 74.37 | 94.22 | 29.78 | 49.94 | 62.13 | 62.37 | 23.54 |
| 13 | 11 | 9 | Goethe University Frankfurt | chiarcos | 27.37 | 33.93 | 48.32 | 63.17 | 23.77 | 53.24 | 66.71 | 70.69 | 11.67 |
| 11 | 12 | 5 | Dublin City University 1 | wangl | 20.52 | 28.55 | 41.78 | 93.23 | 17.70 | 35.49 | 45.26 | 45.16 | 15.96 |
| 12 | 13 | 7 | Dublin City University 2 | okita | 18.59 | 26.27 | 37.33 | 86.33 | 15.82 | 35.49 | 45.32 | 45.13 | 15.07 |
| 14 | 14 | 12 | India Institute of Technology | mukherjee | 17.09 | 25.94 | 65.52 | 93.55 | 15.59 | 32.25 | 41.22 | 40.96 | 4.99 |
| 15 | 15 | 16 | Shanghai Jiao Tong University 1 | chen | 15.15 | 18.35 | 58.27 | 86.09 | 14.57 | | | 0.36 | 0.00 |
| 16 | 16 | 14 | Peking University | xu15b | 3.14 | 4.77 | 19.08 | 51.54 | 2.79 | 17.90 | 22.92 | 23.76 | 0.77 |

Table 6: Scoreboard for the CoNLL-2015 shared task showing performance split across Explicit and Non-Explicit subtasks on the three data partitions—blind test, standard test (WSJ-23) and development. The rows are sorted by the parser performance of the participating systems on the Explicit task. The Column **O**, **E**, **I** refer to official, Explicit and Non-Explicit task ranks respectively. The blue highlighted rows indicate participants that did not attempt the Non-Explicit relation subtask. The green highlighted row shows a team that probably overfitted the development set. Finally, the red highlighted row indicates a team that possibly focused on the Explicit relations task and even though their overall rank was lower, they did very well on the Explicit relations subtask. This is also the system that did not submit a paper, so we do not know more details.

course parsing. The shared task required the development of an end-to-end system, and the best system achieved an F1 score of 24.0% on the blind test set, 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, helping 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.

## Acknowledgments

We would like to thank the Penn Discourse TreeBank team, in particular Aravind Joshi and Bonnie Webber, for allowing us to use the PDTB corpus for the shared task. Thanks also go the LDC (Linguistic Data Consortium), who helped distribute 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

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*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*.

Christian Chiarcos and Niko Schenk. 2015. A minimalist approach to shallow discourse parsing and implicit relation recognition. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S, Pattabhi RK Rao, Vijay Sundar Ram R., and Malarkodi C.S. 2015. A hybrid discourse relation parser in CoNLL 2015. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

David A Duverle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association of Computational Lingustics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Sucheta Ghosh, Giuseppe Riccardi, and Richard Johansson. 2012. Global features for shallow discourse parsing. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

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.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the 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*.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Junyi Jessy Li and Ani Nenkova. 2014. Addressing class imbalance for improved recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

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.

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.*

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the Human Language Technology/North American Chapter of the Association for Computational Linguistics Workshop on Frontiers in Corpus Annotation*.

Truong Son Nguyen, Bao Quoc Ho, and Le Minh Nguyen. 2015. JAIST: A two-phase machine learning approach for identifying discourse relations in newswire texts. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

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, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. Technical report, University of Pennsylvania.

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.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computation Natural Language Learining: Shared Task*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computation Natural Language Learining: Shared Task*.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*.

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.

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*.

Yangqiu Song, Haoruo Peng, Parisa Kordjamshidi, Mark Sammons, and Dan Roth. 2015. Improving a pipeline architecture for shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Manfred Stede. 2012. *Discourse Processing*. Morgan & Claypool Publishers.

Evgeny Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2015. The UniTN discourse parser in CoNLL 2015 shared task: Token-level sequence labeling with argument-specific models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the 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*.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Longyue Wang, Chris Hokamp, Tsuyoshi Okita, Xiaojun Zhang, and Qun Liu. 2015. The DCU discourse parser for connective, argument identification and explicit sense classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Yasuhisa Yoshida, Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2015. Hybrid approach to PDTB-styled discourse parsing for CoNLL-2015. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics*.