CoNLL 2015 Shared Task on Shallow Discourse Parsing (SDP):

Information about the training and development data:

This release contains the training (Sections 2-21) and development (Section 22) sets for the CoNLL 2015 Shared Task on Shallow Discourse Parsing. The training and development data in this release is provided in JSON format as well as the traditional CoNLL tabular format, and the package can be obtained by contacting Ilya Ahtaridis (email address: ldc@ldc.upenn.edu) at the LDC after signing a data use agreement prepared by the LDC.

The training and development data for the shared task is derived from the Penn Discourse TreeBank (PDTB) 2.0, a 1-million-word Wall Street Journal corpus. 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 appears to relate the two. Arguments of relations are annotated in each case. For explicit connectives, arguments are unconstrained in terms of their distance from the connective and can be found anywhere in the text. Between adjacent sentences where no explicit connective appears, 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 provided to express the inferred relation, (b) the sentences may be related by a discourse relation that is realized by some alternative non-connective expression, in which case these alternative lexicalizations are annotated as the carriers of the relation, (c) the sentences may be related not by a discourse relation, but merely by an entity-based coherence relation, in which case the presence of such a relation is labeled, and (d) the sentences may not be related at all, in which case they are labeled as such. In addition to the argument structure of relations, the PDTB provides sense annotations for each discourse relation, capturing the polysemy of connectives. The senses for the discourse relations 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 “subtypes” which specify the semantic contribution of each argument. In this year’s shared task, the second-level “types” as well as a selected number of third-level “subtypes” are used as the target of prediction for the sense of the discourse connective.

PDTB, version 2.0. annotates 40,600 discourse relations, distributed into the following five types: 18,459 Explicit Relations, 16,053 Implicit Relations, 624 Alternative Lexicalizations, 5210 Entity Relations.

Evaluation

Evaluations Tracks:

There are two evaluation tracks. In the closed track, a participating system can only be trained on the provided PDTB training set and linguistic resources. To reduce the data preparation burden on participants, we provide the training, development and test data with the following layers of automatic linguistic annotation:

1. Phrase structure parses (predicted by the Berkeley Parser)
2. Dependency parses (Converted from the Berkeley Parser output, using the Stanford dependency converter)

The phrase structure and dependency parses are included in this package together with the PDTB data, in JSON and CoNLL tabular format. We will also provide the following publicly available linguistic resources that have been found to be useful in shallow discourse parsing literature, and they can be downloaded from the CoNLL 2015 Shared Task website (http://www.cs.brandeis.edu/~clp/conll15st):

1. Brown Clusters
2. VerbNet
3. Sentiment lexicon
4. Word embeddings (produced using Word2Vec)

To ensure meaningful comparisons among participating systems, we ask closed track participants to use only the version of linguistic resources that we provide. Participants can, however, re-process the training set themselves. If they use a third-party tool, we require that they use one that is publicly available so that the results can be replicated by other researchers who are interesting in working on this problem after the shared task is over.

In the open track, a participating system may use any publicly available NLP tools to process the data AND any publicly available (i.e., non-proprietary) data for training.

A participating team can choose to participate in the closed track or the open track or both.

Evaluation Metric:

The evaluation metric will be based on the F measure, the harmonic mean of precision and recall. The precision will be computed as the number of discourse relations correctly predicted by a participant system divided by the total number of discourse relations in the system output. The recall will be computed as the number of correctly predicted discourse relations divided by the total number of discourse relations in the gold standard test set. The scorer can be downloaded from the shared task website.

The evaluation will be done on a per-discourse relation basis. A relation is correctly predicted if (1) the discourse connective is correctly detected (for explicit discourse relations), (2) the sense of a discourse relation is correctly predicted, and (3) the text spans of the two arguments as well as their labels (Arg1 and Arg2) are correctly predicted. An example (explicit) discourse relation is given below:

[$_{\text{Arg1}}$ Use of dispersants was approved] when [$_{\text{Arg2}}$ a test on the third day showed some positive results], officials said. (CONTINGENCY:Cause)

Evaluation Data:

The blind test set will consist of 20,000 to 30,000 words of newswire text annotated following the PDTB annotation guidelines. In addition, we will also use Section 23 of the PDTB as an additional test set. Results on this additional test set will NOT be used to rank the systems, but they can be used as a reference for comparison with previously reported results. Section 22 will be used as the development set. Participants are asked NOT to use this additional test set to train or tune their systems, but they can use the development set to tune (but not train) their systems.

Evaluation Format:

This year we will experiment with a new evaluation format in which participants are asked to submit their systems to a server for automatic evaluation instead of submitting system output for the test set. The participants will develop their systems as usual and produce their output in JSON format. More details will be provided via discussion forums on the shared task website.

Organizers:

Nianwen Xue (chair), Brandeis University
Hwee Tou Ng, National University of Singapore
Sameer Pradhan, Harvard University
Rashmi Prasad, University of Wisconsin at Milwaukee
Christopher Bryant, National University of Singapore
Attapol Rutherford, Brandeis University