Large Discourse Treebanks from Scalable Distant Supervision

Patrick Huber and Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
{huberpat, carenini}@cs.ubc.ca

1 Introduction

Discourse parsing is an important upstream task in Natural Language Processing (NLP) with strong implications for many real-world applications, for example sentiment analysis (Bhatia et al., 2015; Nejat et al., 2017; Hogenboom et al., 2015), text classification (Ji and Smith, 2017) and summarization (Gerani et al., 2014), just to name a few. Despite this widely recognized role of discourse parsing in NLP, most recent discourse parsers (and consequently downstream tasks) still rely on small scale human annotated discourse treebanks (such as RST-DT (Carlson et al., 2003), Instr-DT (Subba and Di Eugenio, 2009) or PDTB (Prasad et al., 2008)), trying to infer general-purpose discourse structures from very limited data in a few narrow domains. While in principle diverse and large discourse treebanks could be annotated, the process is expensive, tedious and does not scale.

To overcome this dire situation and allow discourse parsers to be trained on larger, more diverse and domain independent datasets, we propose a framework to generate “silver-standard” discourse trees from distant supervision on the auxiliary task of sentiment analysis\(^1\). Our approach explicitly models discourse structures, distinguishing it from previous work on implicit discourse modelling, such as Liu et al. (2019), previously criticized in Ferracane et al. (2019).

2 Distant Supervision Approach

The overview of our approach, separated in two distinct phases, is shown in Figures 1 and 2. In the first phase (Figure 1), we train the neural Multiple-Instance Learning (MIL) model by Angelidis and Lapata (2018) to infer EDU-level sentiment and attention scores from the document-level gold-label sentiment of the Yelp’13 corpus (Tang et al., 2015). Subsequently (see Fig. 2), we use the fine-grained EDU-level information to run a CKY-style dynamic programming model, inferring the discourse structure of the document by optimizing the distance between the document-level gold-label sentiment and the discourse-aggregated sentiment prediction. We call the discourse treebank generated by this approach Yelp13-DT.

\(^1\)This work has been partly published and is partly under submission.
complexity of the computation to $O(n^3)$, the space-
complexity grows according to the Catalan num-
ber, making the prediction of discourse structures
for long documents intractable. We overcome this
limitation with a heuristic beam search approach,
empirically setting the beam-size to 10 trees per
CKY cell. For subtrees on low levels of the dis-
course tree the overall document sentiment is not
necessarily a good pruning criteria. We therefore
allow deep subtrees to be more diverse by adding
an exploration/exploitation trade-off, as commonly
used in Reinforcement Learning (RL). With these
two extensions of the standard CKY approach, we
can explore additional discourse attributes during
the CKY process, such as the ternary nuclearity
attribute. Furthermore, nearly arbitrary long docu-
ments can be effectively and efficiently processed,
significantly expanding the applicability of the ap-
proach to new domains. We call a second discourse
treebank, generated with the additional heuristic
and stochastic components and extended with nu-
clearity computations MEGA-DT.

### 3 Results

We evaluate our “silver-standard” discourse tree-
banks on the inter-domain discourse parsing task.
We therefore train the top-performing TwoStage
parser (Wang et al., 2017) on the Yelp13-DT and
MEGA-DT discourse-annotated review corpora and
evaluate the performance on the commonly used
RST-DT and Instr-DT treebanks in the news- and
instructions-domain (see bottom rows in the inter-
domain evaluation sub-table of Table 1). We com-
pare the performance against models trained and
evaluated on different domain gold-standard tree-

| Approach          | Structure RST-DT | Structure Instr-DT | Nucleary RST-DT | Nucleary Instr-DT |
|-------------------|------------------|--------------------|-----------------|-------------------|
| **Intra-Domain**  | Evaluation       |                    |                 |                   |
| CODRA(2015)       | 83.84            | 82.88              | 68.90           | 64.13             |
| Two-Stage(2017)   | 86.00            | 77.28              | 72.40           | 60.01             |
| **Inter-Domain**  | Evaluation       |                    |                 |                   |
| Two-Stage_{RST-DT}| ×                 | 73.57              | ×               | 49.78             |
| Two-Stage_{Instr-DT}| 74.32          | ×                  | 44.68           | ×                 |
| Two-Stage_{Yelp13-DT}| 76.41         | ×                  | –               | –                 |
| Two-Stage_{MEGA-DT}| **77.82**       | **75.18**          | **44.88**       | **54.87**         |

Table 1: Results of the average micro precision measure for structure- and nuclearity-prediction, evaluated on the RST-DT and Instr-DT corpora. Inter-domain subscripts identify the training set. Inter-domain results averaged over 10 independent runs. Best performance per sub-table is **bold**. († statistically significant with p-value ≤ .05 to the best inter-domain baseline (Bonferroni adjusted), × not feasible combinations, – Not pursued)

In general it can be observed, that the MEGA-
DT corpus dominates the comparison on the chal-
lenging but arguably useful inter-domain discourse
parsing task. This suggests that the approach taken
for the MEGA-DT treebank improves the structure
prediction through more diversity in low-level trees,
and properly addresses the nuclearity-prediction
task. To give further insight into the generation ap-
proach, we show a qualitative result of the “silver-
standard” annotation in Figure 3. The shown tree-
structure indicates that generated trees are non-
trivial, reasonably balanced and strongly linked
to the EDU-level sentiment.

In conclusion, our distant supervision approach
enables the NLP community to extend any exist-
ing sentiment-annotated dataset with discourse-
trees, allowing the automated creation of large-
scale domain/genre-specific discourse treebanks.

![Figure 3: Discourse tree generated with our approach, containing 72 EDUs. Node color indicates inferred EDU-level sentiment.](image-url)
References

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218.

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

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. *arXiv preprint arXiv:1906.01472*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.

Alexander Hogenboom, Flavius Frasincar, Franciska De Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Commun. ACM*, 58(7):69–77.

Yangfeng Ji and Noah A Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3).

Yang Liu, Ivan Titov, and Mirella Lapata. 2019. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Mitsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. *LREC*.

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*, pages 566–574. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.