MEGA RST Discourse Treebanks with Structure and Nuclearity from Scalable Distant Sentiment Supervision

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

Abstract

The lack of large and diverse discourse treebanks hinders the application of data-driven approaches, such as deep-learning, to RST-style discourse parsing. In this work, we present a novel scalable methodology to automatically generate discourse treebanks using distant supervision from sentiment-annotated datasets, creating and publishing MEGA-DT, a new large-scale discourse-annotated corpus. Our approach generates discourse trees incorporating structure and nuclearity for documents of arbitrary length by relying on an efficient heuristic beam-search strategy, extended with a stochastic component. Experiments on multiple datasets indicate that a discourse parser trained on our MEGA-DT treebank delivers promising inter-domain performance gains when compared to parsers trained on human-annotated discourse corpora.

1 Introduction

Discourse parsing is an important Natural Language Processing (NLP) task, aiming to uncover the hidden structure underlying coherent documents, as described by theories of discourse like Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) or PDTB (Prasad et al., 2008). Not only has discourse parsing been shown to enhance key downstream tasks, such as text classification (Ji and Smith, 2017), summarization (Gerani et al., 2014) and sentiment analysis (Bhatia et al., 2015; Nejat et al., 2017; Hogenboom et al., 2015), but it also appears to complement contextual embeddings, like BERT (Devlin et al., 2018), in tasks where discourse information is critical, such as argumentation analysis (Chakrabarty et al., 2019).

Traditionally, RST-style discourse parsing builds a complete, hierarchical constituency tree for a document (Morey et al., 2018), where leaf nodes are clause-like sentence fragments, called elementary-discourse-units (EDUs), while internal tree nodes are labelled with discourse relations (e.g., Evidence, Contrast). In addition, each node is given a nuclearity attribute, which encodes the importance of the node in its local context.

A key limitation for further research in RST-style discourse parsing is the scarcity of training data. Only a few human annotated discourse treebanks exist, each only containing a few hundred documents. Although our recent efforts using distant supervision from sentiment to generate large-scale discourse treebanks have already partly addressed this dire situation (Huber and Carenini, 2019), the previously proposed solution is still limited in: (i) Scope, by only building the RST constituency structure without nuclearity and relation labels; and (ii) Applicability, by relying on a non-scalable CKY solution, which cannot be applied to many real-world datasets with especially long documents.

In this work, we propose a significant extension to this line of research by introducing a scalable solution for documents of arbitrary length and further moving beyond just predicting the tree-structure by incorporating the nuclearity attribute, oftentimes critical in informing downstream tasks (Marcu, 2000; Ji and Smith, 2017; Shiv and Quirk, 2019). Inspired by the recent success of heuristic search in NLP tasks involving trees (e.g., Fried et al. (2017); Mabona et al. (2019)), we develop a beam-search strategy implementing an exploration-exploitation trade-off, as commonly used in reinforcement-learning (RL) (Poole and Mackworth, 2010).

Remarkably, by following this heuristic approach, any large corpus annotated with sentiment can be turned into a discourse treebank on which a domain/genre specific discourse parser can be trained. As a case study for this process, we annotate, evaluate and publicly release a new discourse-augmented Yelp '13 corpus (Tang et al., 2015)

called MEGA-DT[1] (comprising ≈250,000 documents) with nuclearity attributed "silver-standard" discourse trees, solely leveraging the corpus' document-level sentiment annotation.

To evaluate the quality of our newly proposed MEGA-DT corpus, we conduct a series of experiments. We train the top-performing discourse parser by Wang et al. (2017) on MEGA-DT and compare its performance with the same parser trained on previously proposed treebanks. Specifically, we compare our discourse-annotated dataset against a smaller "silver-standard" treebank (Huber and Carenini, 2019) containing around ≈100,000 documents with ≤20 EDUs and two standard human annotated corpora in the news domain (RST-DT) (Carlson et al., 2002) and in the instructional domain (Subba and Di Eugenio, 2009).

Results indicate that while training a parser on MEGA-DT does not yet match the performance of training and testing on the same treebank (intra-domain), it does push the boundaries of what is possible with distant supervision. In most cases, training on MEGA-DT delivers statistically significant improvements on the arguably more difficult and useful task of inter-domain discourse prediction, where a parser is trained on one domain and tested/applied to another one.

Overall, this suggests that our new approach to distant supervision from sentiment can generate large-scale, high-quality treebanks, with MEGA-DT being the best publicly available resource for training a discourse parser in domains where no gold-standard discourse annotation is available.

2 Related Work

The most closely related line of work is RST-style discourse parsing, with the goal to obtain a complete discourse tree, including structure, nuclearity and relations. Based on the observation that these three aspects are correlated, most previous work has explored models to learn them jointly (e.g., Joty et al. (2015); Ji and Eisenstein (2014); Yu et al. (2018)). However, while this strategy seems intuitive, the state-of-the-art (SOTA) system on structure-prediction by Wang et al. (2017) applies a rather different strategy, first jointly predicting structure and nuclearity and then subsequently predicting relations. The main motivation behind this separation is the large number of possible output classes when predicting these three aspects together. The success of the system by Wang et al. (2017) on the widely used RST-DT corpus inspires us to also learn structure and nuclearity jointly, rather than combining all three aspects.

The second line of related work infers fine-grained information from coarse-grained supervision signals using machine learning. Due to the lack of annotated data in many domains and for many real-world tasks, methods to automatically generate reliable, fine-grained data-labels have been explored for many years. One promising approach in this area is Multiple Instance Learning (MIL) (Keeler et al., 1991). The general task of MIL is to retrieve fine-grained information (called instance-labels) from high-level supervision (called bag-labels), using correlations of discriminative features within and between bags to predict labels for instances. With the recent rise of deep-learning, neural MIL approaches have also been proposed (Angelidis and Lapata, 2018).

We previously combined the two lines of related work described above to create discourse structures from large-scale datasets, solely using document-level supervision (Huber and Carenini, 2019). When applied to the auxiliary task of sentiment analysis, we generated a "silver-standard" discourse structure treebank using the neural MIL model by Angelidis and Lapata (2018) in combination with a sentiment-guided CKY-style tree-construction algorithm, generating near-optimal discourse trees in bottom-up fashion (Jurafsky and Martin, 2014). Although our approach has shown clear benefits, it is inapplicable to many real-world datasets, as it does not scale to long documents and cannot predict nuclearity- and relation-labels. Addressing these limitations is a major motivation of this paper.

Further efforts to automatically generate discourse trees from auxiliary tasks have been mostly focused on latent tree induction, generating trees from text classification (Karimi and Tang, 2019) or summarization tasks (Liu et al., 2019). For both approaches, domain dependent discourse trees are induced during the neural training process. While either method has shown to improve the performance on the downstream task itself, subsequent research by Ferracane et al. (2019) indicates that the induced trees are often trivial and shallow, and

---

[1]Our new Discourse Treebank and the code to generate further "silver-standard" discourse treebanks can be found at: `https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/`

do not represent valid discourse structures.

The third stream of related work is on leveraging heuristic search algorithms in NLP tasks involving trees. For syntactic parsing, Vinyals et al. (2015) and Fried et al. (2017) show that a static, small beam size (e.g. 10) already achieves good performance, with Dyer et al. (2016) delivering promising results by using greedy decoding. As a recent example for discourse parsing, Mabona et al. (2019) successfully combine standard beam-search with shift-reduce parsing using two parallel beams for shift and reduce actions. Overall, recent work shows that beam-search approaches and their possible extensions can effectively address scalability issues in multiple parsing scenarios. In this paper, we extend the standard beam-search approach with a stochastic exploration-exploitation trade-off, as used in Reinforcement Learning, where signals also tend to be sparse and noisy.

3   Predicting Discourse from Sentiment

Previous work has shown that incorporating RST-style discourse trees can help to predict document-level sentiment (Bhatia et al., 2015; Nejat et al., 2017; Hogenboom et al., 2015). These findings give rise to the assumption that the sentiment of a document can also provide important information on its discourse structure. In the following sub-section, we shortly revisit our previous approach to exploit this assumption by solely relying on document-level sentiment annotations (Huber and Carenini, 2019). Afterwards, we will present our new approach to overcome scalability issues and jointly predict structure and nuclearity.

3.1   Predicting Discourse Structure for Short Documents

The discourse-structure tree of a document can be predicted from its global sentiment by combining Multiple Instance Learning (MIL) and the CKY algorithm. We will illustrate the process on the following negative sentiment example with a polarity of -0.5, pre-segmented into five EDUs: *[Bad poutine.]₁, [Fries were nice and thin,]₂, [had a good taste,]₃, [however, the gravy was cold]₄, [and the cheese was not melted.]₅*. The first step in generating the discourse structure for this example consist of assigning each EDU a sentiment polarity p_{EDU} within the interval of $[-1, 1]$ and an attention score a_{EDU} between $[0, 1]$, both learned through MIL from the overall document sentiment polarity.

To obtain the tuple $\{p_{EDU}, a_{EDU}\}$ for each EDU in a document, the neural MIL model (Angelidis and Lapata, 2018) is trained on a document-level sentiment dataset, with the goal to predict sentiment-labels for EDU-level instances. The model therefore generates a mapping from inputs (EDUs) to the respective outputs (sentiment-classes) by exploiting correlations between the appearance of EDUs in documents and the respective document gold-labels across a corpus. For example, the EDU *[had a good taste,]₃* will most likely appear predominantly in positive documents, allowing the MIL model to infer a positive EDU-level sentiment polarity p_{EDU} for this input. When applying the neural model by Angelidis and Lapata (2018), an attention mechanism is internally used to weight the importance of EDUs for the overall document sentiment. An attention-weight a_{EDU} is also extracted for each EDU and subsequently used as an importance score when aggregating subtrees using the CKY approach.

From those tuples $\{p_{EDU}, a_{EDU}\}$ assigned to leaf-nodes, the sentiment polarity p and attention score a for any internal node in an arbitrary constituency tree can be computed bottom-up by aggregating its two child nodes c_l, c_r. Out of the set of potential aggregation functions proposed in Huber and Carenini (2019), the best performing approach has shown to be:

$$p = \frac{p_{c_l} * a_{c_l} + p_{c_r} * a_{c_r}}{a_{c_l} + a_{c_r}} \quad a = \frac{a_{c_l} + a_{c_r}}{2} \quad (1)$$

By recursively applying this function from the leaf-nodes, we can compute the sentiment and attention of the root node, representing the full document.

The process of selecting the best discourse tree for a given document can be framed as finding the tree for which the sentiment of the root node (spanning the whole document) is the closest to the gold-standard sentiment annotation. A brute-force solution to this problem is to generate all possible discourse trees using the general CKY algorithm and selecting the best tree amongst all candidates. However, the computational complexity of this approach quickly explodes, as shown for our running example with 5 EDU leaf-nodes in Figure 1. From the decision-space of possible tree-structures, the tree with the shortest sentiment-distance from the gold-standard, computed at the root node, is selected.

Although this method has been shown to provide reasonably good trees when leveraged for
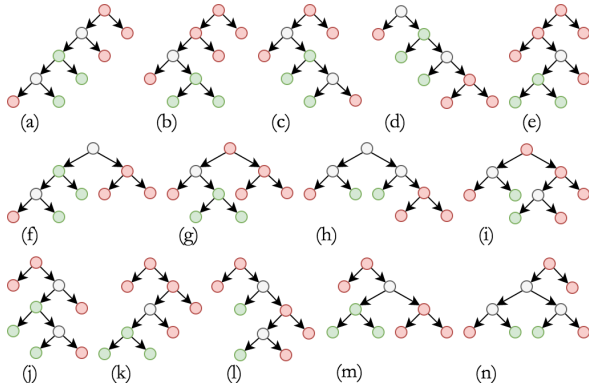
Figure 1: All 14 projective discourse trees annotated with sentiment for a 5 EDU document (using a simplified color-scheme, green = positive, red = negative, grey = neutral, omitting the attention attribute)

discourse-structure parsing, it is limited in two fundamental ways: (1) The approach is not scalable. Since its space complexity grows with the Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$ for trees with $n + 1$ EDUs, it can only be applied to short ($\approx\leq 20$ EDUs) documents (see bottom row in Table 1), making it impractical for many real-world datasets containing longer documents, such as the Yelp '13 (Tang et al., 2015), IMDB (Diao et al., 2014) or Amazon Review dataset (Zhang et al., 2015). (2) Due to the high computational complexity of the structure prediction itself, the inference of further RST-tree properties, such as nuclearity and relations, often critical for downstream tasks, are not feasible with this unconstrained CKY approach.

3.2 Predicting Discourse Structure and Nuclearity from Arbitrary Documents

Inspired by the recent success in applying beam-search to enhance the scalability of multiple NLP parsing tasks (Mabona et al., 2019; Fried et al., 2017; Dyer et al., 2016; Vinyals et al., 2015), we propose a novel heuristic beam-search approach that can automatically generate discourse trees containing structure- and nuclearity-attributes for documents of arbitrary length.

Stochastic Beam-Search In essence, the general CKY dynamic programming algorithm creates all possible binary trees covering the n EDUs by internally filling an $(n \times n)$ matrix, where each $cell(i, j)$ contains the information on all subtrees covering the text spans from EDU_i to EDU_j. Our heuristic beam-search solution limits the computational complexity of this process by reducing the number of subtrees stored in each cell to a constant beam-size B. This naturally raises the question on how

to select the B subtrees to preserve in each cell. We follow the intuitive assumption that subtrees for which the sentiment diverges most from the overall document sentiment (the only available supervision for this task) can be safely discarded. Out of the set of possible subtrees T for a given cell, only the subset T' with $|T'| = B$ is kept, containing the B subtrees with the closest sentiment polarity p_{t_i} to the gold-label (gl) sentiment of the document. Formally:

$$T' = \underset{t_i \in T, |T'|=B}{\arg\min} |p_{t_i} - gl| \quad (2)$$

However, one limitation of this heuristic rule is that it strictly prefers subtrees with sentiment closer to the overall document sentiment, independent of their distance from the root node. This can be problematic when applied in early stages of the tree-generation process, where only a few EDUs are combined. For instance, a mostly positive document might still contain certain negative subtrees at its lowest levels, which also need to be aggregated appropriately.

Ideally, we would like to support a high degree of exploration on low levels of the tree, only loosely forcing the sentiment of subtrees in the beam to align with the overall document gold-label sentiment; while on higher levels of the tree, the requirement of closely reflecting the document's gold-standard sentiment should be strictly enforced (i.e., exploiting the distant supervision).

We implement this strategy through a stochastic beam-search approach, which relies on a softmax selection using the Boltzmann–Gibbs distribution (Poole and Mackworth, 2010). The temperature coefficient τ thereby modulates the exploration-exploitation trade-off (similar to previous work in RL), by influencing the divergence of the softmax outputs. We then sample from the resulting, categorical probability distribution $P = ((Prob(t_1), ..., Prob(t_n))$, computed for every local subtree $t_i \in T$ to obtain a subset T' of size B (as shown in equation 3).

$$Prob(t_i) = \frac{e^{\frac{1}{|p_{t_i} - gl|}/\tau}}{\sum\limits_{t_j \in T} e^{\frac{1}{|p_{t_j} - gl|}/\tau}} \quad (3)$$

$$\tau = f(n, c) = (n - c) + 1 \quad (4)$$

In this work, the parameter τ is defined as a linear function $f(n, c)$ parameterized by the number
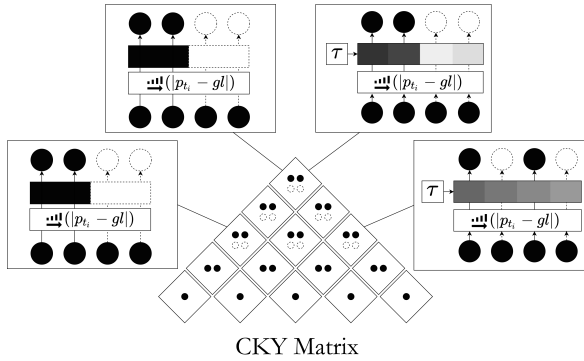
Figure 2: Standard beam-search approach (left) picking the top $B = 2$ tree-candidates with the smallest distance $|p_{t_i} - gl|$ in every CKY cell. Stochastic beam-search approach (right) calculating the Boltzmann–Gibbs distribution with the tree-coverage dependent temperature τ, modulating the subtree sampling process of the tree-candidates. (For readability, we only show a maximum of 4 subtrees per CKY cell)

Beam	20 EDUs	30 EDUs	100 EDUs
1	1.6KB	3.7KB	40KB
10	24KB	48KB	440KB
100	920KB	1.5MB	7.9MB
∞	3.6GB	1.9PB	400SB

Table 1: Upper-bounds for growth of spatial complexity using different beam sizes and unconstrained CKY (∞), assuming 1Byte per unit in memory. KB = 10^3, MB = 10^6, GB = 10^9, PB = 10^{15}, SB = 10^{54}

of EDUs c covered under the subtree t_i as well as the total number of EDUs n (see equation 4). This way, τ influences equation 3 such that for larger values of c (at the top of the tree), τ gets close to 1 and the sampling is likely to select subtrees with low distance $|p_{t_i} - gl|$. For subtrees with a small coverage c (at the bottom of the tree), τ becomes $>> 1$ and $Prob(t_i)$ resembles the uniform distribution, allowing for a high degree of exploration. For illustration, Figure 2 highlights the differences between the standard and the stochastic beam-search approach.

Analysis of Spacial Complexity The described system significantly reduces the spatial complexity, independent from whether a stochastic component is used. The complexity reduction can be easily observed by comparing the theoretical upper-bounds for the space consumption of the unrestricted CKY approach (eq. 5) against the upper-bounds for the heuristically constrained CKY method (eq. 6).

$$\sum_{i=1}^{n-1} \frac{4(n-i)}{i} \binom{2i-2}{i-1} \tag{5}$$

$$4n^2B + 4(n-1)B^2 \tag{6}$$

In both equations, n represents the number of leaf-nodes (EDUs) in the discourse tree. In eq. 5, the number of generated trees at every level of the tree is bound by the Catalan number, while in eq. 6 the bound has a quadratic dependency on the input-size and the beam-size. For the equations shown,

we assume that on every level of the tree, each of the possible subtrees is represented by 2 pointers to the child-nodes as well as a sentiment and attention value for the subtree itself. Table 1 compares the space capacities required with increasing document length, indicating that with a proper beam size, our heuristic strategy can deal with the tree structures for very long documents.

Integration of Nuclearity With this scalable solution, it is now possible to also take additional properties, like nuclearity, into account. The inherent advantage of generating nuclearity-attributed discourse trees becomes obvious when revisiting the definition in RST (Mann and Thompson, 1988), where the nuclearity-attribute encodes a notion of "importance" in the local context, with *Nucleus-Statellite (N-S)* and *Satellite-Nucleus (S-N)* attributions defining the directionality between two nodes, while the *Nucleus-Nucleus (N-N)* attribution implies equal importance (Morey et al., 2018). Expressing this notion of importance, it is not surprising that nuclearity-attribution is frequently critical in informing many downstream tasks like summarization and text categorization (e.g., Marcu (2000); Ji and Smith (2017); Shiv and Quirk (2019)).

Technically, we integrate the nuclearity attribute into the tree-generation process by assigning each subtree one of the three nuclearity classes *N-S*, *S-N* or *N-N*, following the assumption that the attention values a_{c_l}, a_{c_r} capture the nodes' relative importance in the tree. Starting from the leaf-node attention, extracted from MILNet, we propagate the attention values through the tree structure according to equation (1). More specifically, for a subtree where the attention value a_{c_l} is greater than the attention a_{c_r}, we will assign the *N-S* label, while *S-N* is assigned if the opposite is true. However, in this way, only two of the three possible nuclearity classes can be represented (namely *N-S* and *S-N*), as the attention values are distinct. To further account for the third class of *N-N*, we include

Treebank	#Documents	#EDUs*		
		min	avg	max
Instr-DT	176	2	33	248
RST-DT	385	2	56	240
Yelp13-DT	100,000	2	10	20
MEGA-DT	250,000	2	19	150

Table 2: Treebank size and distribution (*calculated on the training set)

an additional subtree at every merge in the CKY procedure, which averages not only the two attention values a_{c_l}, a_{c_r} (as shown in eq. 1) but also the child polarity scores p_{c_l}, p_{c_r}. This reflects the definition of the *N-N* nuclearity class according to RST, where an even importance for all child nodes is assumed in the multi-nucleus case. The additional complexity of doubling the number of trees in each cell is only manageable due to the use of our heuristic approach.

4 Evaluation

In this section, we evaluate our proposed method to generate the MEGA-DT discourse treebank by assessing the performance of a discourse parser when trained on MEGA-DT against our previously proposed "silver-standard" treebank (Huber and Carenini, 2019) as well as two commonly used, human-annotated, discourse corpora.

4.1 Treebanks

The two human-annotated treebanks are:
Instructional-DT (from here on called *Instr-DT*) by Subba and Di Eugenio (2009), which comprises of documents on home-repair instructions annotated with full RST-style discourse trees, separated into training- and test-set with a 90-10 split.
RST-DT by Carlson et al. (2002), containing news articles alongside with full RST-style discourse trees, in the standard 90-10 train-test split.

The two automatically annotated treebanks are:
Yelp13-DT, generated according to our previously proposed unconstrained CKY approach as described in Huber and Carenini (2019). We use the pre-segmented version of the Yelp'13 customer review dataset by Angelidis and Lapata (2018), separated into EDUs by applying the publicly available discourse segmenter proposed in Feng and Hirst (2014). Yelp13-DT contains short documents with ≤ 20 EDUs, only considering two nuclearity classes (namely N-S and S-N).
MEGA-DT, our novel treebank, is also generated

from the original Yelp'13 corpus, akin to Yelp13-DT. However, due to our newly proposed, scalable solution, MEGA-DT is much larger and more comprehensive, integrating all three nuclearity classes. A comparison of the key dimensions of all treebanks used in this work is shown in Table 2.

4.2 Discourse Parsers

To interpret our results in the context of existing work, we consider a diverse set of top-performing discourse parsers. Previous work by Morey et al. (2017) compares a set of competitive parsers, including **DPLP** (Ji and Eisenstein, 2014), **gCRF** (Feng and Hirst, 2014), **CODRA** (Joty et al., 2015) and Li et al. (2016). We further add the **Two-Stage** discourse parser by Wang et al. (2017) and the neural approach by Yu et al. (2018) into our final evaluation. Due to the top performance of the parser by Wang et al. (2017) on the structure-prediction of the widely used RST-DT corpus, and even more importantly, due to the separation of the relation computation from the structure/nuclearity prediction, we use the parser by Wang et al. (2017) in our inter-domain experiments.

4.3 Preliminary Evaluation

We run a set of preliminary evaluations on a randomly selected subset containing 10,000 documents from the Yelp'13 dataset. In general, the preliminary evaluation suggests that (1) A beam-size of 10 delivers the best trade-off between computational complexity and performance (out of {1, 5, 10, 50, 100}), when tested according to the distance between gold-label sentiment and model prediction. (2) We employ a sentence-first aggregation strategy, using sentence-boundary predictions from the NLTK toolkit[2]. By not allowing inter-sentence connections, unless the complete sentence is already represented by a subtree, we reach superior results in the preliminary evaluation compared to exploring the complete CKY space. This is consistent with previous findings showing that sentence boundaries are key signals for tree aggregations (Joty et al., 2015).

4.4 Experiments and Results

We train the discourse parser by Wang et al. (2017)[3] on our newly generated MEGA-DT corpus as well as the Yelp13-DT and the original RST-DT and

---

[2]www.nltk.org/api/nltk.tokenize.html
[3]www.github.com/yizhongw/StageDP/

	Structure				Nuclearity			
	RST-DT		Instr-DT		RST-DT		Instr-DT	
Approach	Par.	R-Par.	Par.	R-Par.	Par.	R-Par.	Par.	R-Par.
Right Branching	9.27	54.64	25.45	62.72	×	×	×	×
Left Branching	7.45	53.73	4.32	52.16	×	×	×	×
Hier. Right Branching	**48.74**	**74.37**	**50.68**	**75.34**	×	×	×	×
Hier. Left Branching	41.16	70.58	27.50	63.75	×	×	×	×
Majority Class	×	×	×	×	$^{(N\text{-}S)}$61.28	$^{(N)}$61.33	$^{(N\text{-}N)}$52.33	$^{(N)}$76.48
Intra-Domain Evaluation								
DPLP(2014)*	64.10	82.00	–	–	54.20	68.20	–	–
gCRF(2014)*	68.60	84.30	–	–	55.90	69.40	–	–
CODRA(2015)*	65.10	82.60	–	82.88	55.50	68.30	–	64.13
Li(2016)*	64.50	82.20	–	–	54.00	66.50	–	–
Two-Stage(2017)	**70.97**	**86.00**	**58.86**	79.43	**57.97**	72.40	**40.00**	62.39
Yu(2018)	–	85.50	–	–	–	73.10	–	–
Inter-Domain Evaluation								
Two-Stage$_{\text{RST-DT}}$	×	×	45.95	73.57	×	×	**27.18**	49.78
Two-Stage$_{\text{Instr-DT}}$	46.01	74.32	×	×	**22.22**	44.68	×	×
Two-Stage$_{\text{Yelp13-DT(2019)}}$	52.95	76.41	46.59	74.14	15.51	35.72	7.27	33.35
Two-Stage$_{\text{MEGA-DT}}$	†**55.76**	†**77.82**	†**50.23**	†**75.18**	15.86	**44.88**	20.31	†**54.87**
Human (2017)	78.70	88.30	–	–	66.80	77.30	–	–

Table 3: Results of the micro-averaged precision measure using the original Parseval method (Par.) and RST Parseval (R-Par.). Inter-domain subscripts identify the training set. Inter-domain results averaged over 10 independent runs. Models with stochastic components are averaged over 3 distinct generation processes. The best performance per sub-table is **bold**. * Results taken from Morey et al. (2017), † statistically significant with p-value $\leq .05$ to the best inter-domain baseline (Bonferroni adjusted), – non-published values, × not feasible combinations

Instr-DT corpora[4]. To verify the ability of the training treebanks to support the discourse parser in extracting domain-independent features of general discourse, we evaluate the performance on the inter-domain discourse parsing task, training the Two-Stage discourse parser on one domain (e.g., Yelp user reviews in MEGA-DT) and evaluating it on documents in a different domain (e.g., news articles in RST-DT). We compare the obtained performances against the classic and arguably easier intra-domain measure (training and testing on documents within the same domain).

The results of the final evaluation are summarized and aggregated in three sets of experiments in Table 3. In the first set (on top of Table 3), we show the micro-averaged original Parseval performance (Par.) (Morey et al., 2017) as well as the RST-Parseval measures (R-Par.) of standard linguistic baselines for the structure- and nuclearity-prediction task. Regarding the structure prediction (left), we compare the performance when applying a strictly right- or left-branching tree to the data, as well as hierarchical versions of those (right-/left-branching trees on sentence-level combined by right-/left-branching trees on document level). The

results indicate that the hierarchical right-branching tree resembles the original tree structure the closest on both metrics and either evaluation treebank[5]. As a baseline for the nuclearity prediction task, we compute the majority class on the training corpora. It is important to note that while the linguistic baselines for structure do not require available training data, the majority class measure depends on access to an annotated corpus in the target domain.

The second set of results shows the intra-domain performance of top performing discourse parsers, frequently evaluated against in the past. While all parsers except CODRA (Joty et al., 2015) have been only evaluated on RST-DT, we additionally train and evaluate the Two-Stage parser on the Instr-DT corpus. When comparing the intra-domain discourse parsing performance, the Two-Stage parser reaches the consistently best performance on RST-DT structure prediction, while the discourse parser by Yu et al. (2018) achieves the best results on the RST-DT nuclearity prediction using RST-Parseval. CODRA reaches the best performance on the Instr-

---

[4]Trained on an Intel Core i9 (10 Cores, 3.30 GHz) CPU

[5]Note that the performance of the Hierarchical Right-Branching baseline is higher than reported in Huber and Carenini (2019), because of an additional clean-up step required during data preprocessing. The competitive performance of this baseline is most likely attributed to the highly structured nature of the target domains.

| Approach | Structure | | | | Nuclearity | | | |
| | RST-DT | | Instr-DT | | RST-DT | | Instr-DT | |
	Par.	R-Par.	Par.	R-Par.	Par.	R-Par.	Par.	R-Par.
Two-Stage$_{\text{MEGA-DT-Base}}$	51.85	75.92	47.05	73.87	17.22	36.46	8.64	34.48
Two-Stage$_{\text{MEGA-DT +Stoch}}$	55.05	77.58	43.75	73.76	**17.76**	37.43	8.22	35.89
Two-Stage$_{\text{MEGA-DT +Nuc}}$	54.55	76.76	50.01	74.35	13.82	44.22	19.82	54.10
Two-Stage$_{\text{MEGA-DT}}$	**55.76**	**77.82**	**50.23**	**75.18**	15.86	**44.88**	**20.31**	**54.87**

Table 4: Ablation study showing the influence of nuclearity and stochasticity on the overall performance, measured as the micro-average precision using original Parseval (Par.) and RST Parseval (R-Par.). Results averaged over 10 runs (using 3 distinct generation processes if a stochastic components is included). The best performance is **bold**.
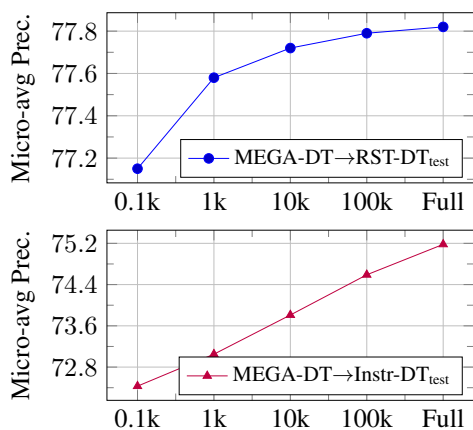

Figure 3: Performance-trend over increasingly large subsets tested on RST-DT (top) and Instr-DT (bottom). Each sample is generated as the average performance over 10 random subsets, drawn from 3 independently created treebanks.

DT corpus when evaluated with RST-Parseval.

The main contribution of this work is placed in the third set of results, where the Two-Stage discourse parser is trained and tested on different, non-overlapping domains (i.e., inter-domain). This task is arguably more useful and significantly more difficult than the task evaluated in the second set, which is reflected in the performance decrease for structure and nuclearity in the first two rows of the sub-table, confirming that the transfer of discourse-structures and -nuclearity between domains is a challenging task. The results presented in the third row of the sub-table show the performance of the Two-Stage parser when trained on Yelp13-DT, containing short documents with limited nuclearity annotations. The approach achieves consistently better performance compared to the first two rows on the inter-domain structure prediction task (For both, original Parseval and RST-Parseval), as we have previously shown in Huber and Carenini (2019). However, only considering two out of three nuclearity classes (N-S and S-N), the system performs rather poorly on the nuclearity classification task. The bottom row of the third sub-table dis-

plays the performance of the Two-Stage discourse parser when trained on our new MEGA-DT corpus. Training on MEGA-DT delivers statistically significant improvements over the best inter-domain baseline in all structure prediction tasks. Furthermore, our new system also achieves statistically significant gains on the Instr-DT nuclearity prediction, when evaluated according to the RST-Parseval metric. The nuclearity measure on RST-DT using RST-Parseval is statistically equivalent to the best baseline system. Overall, our MEGA-DT corpus appears to outperform previously published treebanks for inter-domain discourse parsing on every sub-task on at least one competitive metric.

In order to gain deeper insights into the effectiveness of our proposed treebank generation approach, we run a set of four additional evaluations. First, we evaluate the individual components of our system by showing an ablation study in Table 4, starting with the performance of the discourse parser trained with MEGA-DT-Base, a treebank generated with the standard beam-search approach and without integrating nuclearity. Adding each feature separately (+Stoch, +Nuc) we observe improvements on at least one of the sub-tasks; however, the combination of the two components produces the best performing MEGA-DT corpus. Second, we show the performance-trend over increasingly large subsets of MEGA-DT in Figure 3, tested on RST-DT (top) and Instr-DT (bottom). The two trends highlight consistent improvements with increasingly large dataset sizes, suggesting further possible gains with even larger treebanks. Third, we further analyze the nuclearity classification performance in Table 5, which presents four confusion matrices for the discourse parsing output of our MEGA-DT treebank, evaluated according to the original Parseval and RST-Parseval metrics on RST-DT and Instr-DT. The matrices show a potential explanation for the performance-gap between the original Parseval and the RST-Parseval metrics,

		Predicted					Predicted	
		N-N	N-S	S-N			N	S
Gold	N-N	243	11	40		N	1947	186
	N-S	570	28	96		S	1321	141
	S-N	244	5	47				
		N-N	N-S	S-N			N	S
Gold	N-N	85	1	9		N	470	30
	N-S	77	2	16		S	149	7
	S-N	23	1	3				

Table 5: Confusion Matrices for the model trained on MEGA-DT, evaluated on RST-DT (top) and Instr-DT (bottom). Left: Original Parseval, Right: RST-Parseval

identifying the over-prediction of the N-N class, especially for gold-label N-S nuclearities. Further, we frequently misclassify the gold-label N-S nuclearity class as S-N. Lastly, we present an additional qualitative analysis in Appendix A to investigate the strength and potential weaknesses of trees in MEGA-DT. We therefore show three randomly selected trees that closely/poorly reflect the authors gold-label sentiment respectively (see Figure 4 for a teaser). In general, the qualitative analysis shows that trees in MEGA-DT are non-trivial, reasonably balanced, strongly linked to the EDU-level sentiment and mostly well-aligned with meaningful discourse-structures.

5   Conclusions and Future Work

In this work, we present a novel distant supervision approach to predict the discourse-structure and -nuclearity for documents of arbitrary length solely using document-level sentiment information. To deal with the increasing spatial complexity, we apply and compare heuristic beam-search strategies, including a stochastic variant inspired by RL techniques. Our results on the challenging inter-domain discourse-structure and -nuclearity prediction task strongly suggests that the heuristic approach taken (1) enhances the structure prediction task through more diversity in the early-stage tree selection, (2) allows us to effectively predict nuclearity and (3) helps to significantly reduce the complexity of the unrestricted CKY approach to scale for arbitrary length documents.

In conclusion, our new approach allows the NLP community to augment any existing sentiment-annotated dataset with discourse trees, enabling the automated generation of large-scale domain/genre-specific discourse treebanks. As a case study for the effectiveness of the approach, we annotate and publish our MEGA-DT corpus as a high quality RST-style discourse treebank, which has been shown to outperform previously proposed discourse tree-banks (namely Yelp13-DT, RST-DT and Instr-DT) on most tasks of inter-domain discourse parsing. This suggests that parsers trained on our MEGA-DT corpus (or further domain-specific treebanks generated according to our approach) should be used to derive discourse trees in target domains where no gold-labeled data is available.

This work can be extended in several ways: (i) We plan to investigate into further functions for τ to enhance the exploration-exploitation trade-off. (ii) Additional strategies to assign nuclearity should be explored, considering the excessive N-N-classification shown in our evaluation. (iii) We plan to apply our approach to more sentiment datasets (e.g., Diao et al. (2014)), creating even larger tree-banks. (iv) Our new and scalable solution can be extended to also predict discourse relations besides structure and nuclearity. (v) We also plan to use a neural discourse parser (e.g. Yu et al. (2018)) in combination with our large-scale treebank to fully leverage the potential of data-driven discourse parsing approaches. (vi) Taking advantage of the new MEGA-DT corpus, we want to revisit the potential of discourse-guided sentiment analysis, to enhance current systems, especially for long documents. (vii) Finally, more long term, we intend to explore other auxiliary tasks for distant supervision of discourse, like summarization, question answering and machine translation, for which plenty of annotated data exists (e.g., Nallapati et al. (2016); Cohan et al. (2018); Rajpurkar et al. (2016, 2018)).
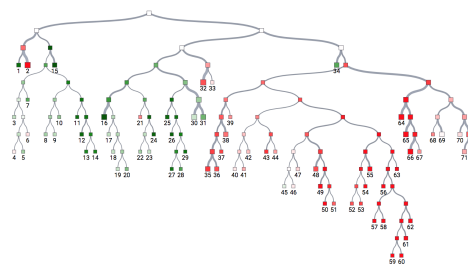

Figure 4: Teaser for a tree analyzed in Appendix A containing 72 EDUs and neutral document-level sentiment.

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, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2926–2936.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. *arXiv preprint arXiv:1602.07776*.

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 (Volume 1: Long Papers)*, pages 511–521.

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

Daniel Fried, Mitchell Stern, and Dan Klein. 2017. Improving neural parsing by disentangling model combination and reranking effects. *arXiv preprint arXiv:1707.03058*.

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.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2306–2316.

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 (Volume 1: Long Papers)*, volume 1, pages 13–24.

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).

Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson London.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. *arXiv preprint arXiv:1903.07389*.

James D Keeler, David E Rumelhart, and Wee Kheng Leow. 1991. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371.

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.

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. *arXiv preprint arXiv:1909.11049*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on rst discourse parsing? a replication study of recent results on the rst-dt. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Guİ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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.

David L Poole and Alan K Mackworth. 2010. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Vighnesh Shiv and Chris Quirk. 2019. Novel positional encodings to enable tree-based transformers. In *Advances in Neural Information Processing Systems*, pages 12058–12068.

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.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in neural information processing systems*, pages 2773–2781.

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.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural rst parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

A Qualitative Analysis of Generated Discourse Trees

The following examples are automatically generated trees from our MEGA-DT corpus. EDU leaf-nodes are enumerated and can be referenced with the discourse units in the description. The colour-saturation and -hue values represents the sentiment of the nodes, with a dark red (high saturation) representing a strongly negative subtree, white (low saturation) representing a neutral sentiment sub-tree and a dark green (high saturation) represents a strongly positive subtree. The thickness of edges and the size of nodes represent the attention of the subtree, which is strongly correlated with the subtree nuclearity.
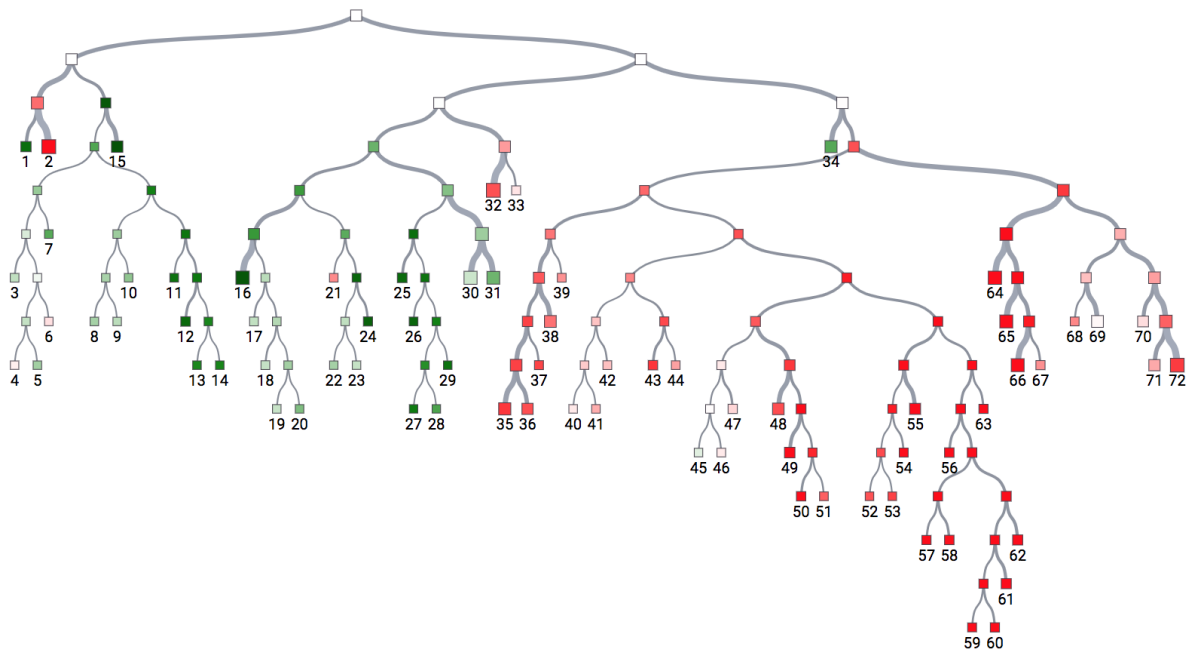
Figure 5: Accurately predicted example, Gold-label polarity: 0, Predicted polarity: 0.03,
Discourse: [amazing food.]₁, [awful, awful service.]₂, [the garlic bread. very good.]₃, [softer than i expected,]₄, [which was nice.]₅, [i also just wasn't expecting garlic bread.]₆, [so it was a nice surprise.]₇, [escargot -]₈, [i was the only one at the table (of 10)]₉, [to eat it.]₁₀, [they were great!]₁₁, [served bubbling hot, not rubbery at all, delicious sauce.]₁₂, [i kept the dish]₁₃, [to dip bread into just because of the sauce.]₁₄, [veal - amazing.]₁₅, [everything tasted fantastic.]₁₆, [ok, the carrots]₁₇, [that were on the side were a bit plain]₁₈, [and could have been softer, but the veal itself and the sauce]₁₉, [it was in, and the mushrooms and pasta.]₂₀, [i left nothing on my plate.]₂₁, [my husband got the same]₂₂, [and also had the same impression.]₂₃, [creme brulee - fantastic.]₂₄, [tasted great, good texture.]₂₅, [pleasantly surprised.]₂₆, [my husband got the tiramisu]₂₇, [and said]₂₈, [it was great.]₂₉, [so why the 3 stars]₃₀, [when the food was so amazing?]₃₁, [because of the terrible service. 1 -]₃₂, [we got water.]₃₃, [great.]₃₄, [but our server * never * asked us]₃₅, [if we wanted anything else.]₃₆, [when my husband finally stopped him to ask for a glass for my father in law, a coke for]₃₇, [and other drinks, our server looked very inconvenienced by it. 2 -]₃₈, [didn't get to order appetizers.]₃₉, [you see]₄₀, [i got escargot?]₄₁, [i ordered that with my meal.]₄₂, [our server never asked about appetizers]₄₃, [and went straight to meals.]₄₄, [also, my husband was walking with our daughter]₄₅, [when the ordering was starting]₄₆, [and needed an extra minute.]₄₇, [our server wanted to start with him.]₄₈, [when asked if he could start with someone else's order,]₄₉, [our server protested,]₅₀, [but eventually did move on to the next person.]₅₁, [you'd think]₅₂, [starting at the next person was]₅₃, [asking him to cut off his hand. 3 - empty glasses everywhere!]₅₄, [never got or was offered a refill on my drink.]₅₅, [or anyone else's.]₅₆, [when my father stopped our server well]₅₇, [after our meal was over]₅₈, [and asked]₅₉, [if i could get a coke,]₆₀, [our server said]₆₁, [i had never ordered one.]₆₂, [well of course i hadn't.]₆₃, [i never had a chance to! 4 -]₆₄, [offering dessert seemed a complete afterthought.]₆₅, [will i recommend this place to anyone else?]₆₆, [conditionally.]₆₇, [i'll make sure to tell them]₆₈, [that the food was very good, but not to go]₆₉, [if they want attentive service,]₇₀, [are on any kind of time constraint, expect refills on their drinks,]₇₁, [or are at all shy about getting a server's attention.]₇₂
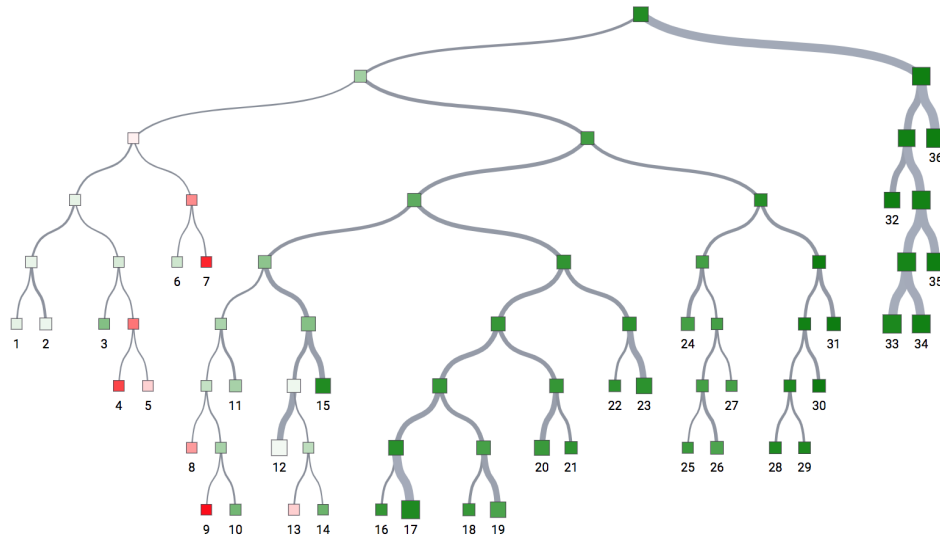
Figure 6: Accurately predicted example, Gold-label polarity: 0.5, Predicted polarity: 0.345,
Discourse: [it has been a hell of a work week]₁, [and friday could not have come any sooner.]₂, [this week was rough especially with the announcement]₃, [that we are officially in a recession]₄, [and the ambiguity was hitting me from all sides.]₅, [man, did]₆, [i need some distraction from all my worries.]₇, [so the bf (bacon) and i decided on thai]₈, [but wanted to venture out from the norm]₉, [and we are very glad]₁₀, [we did.]₁₁, [the thai hut exceeded our expectations.]₁₂, [we were a little skeptical at first with so many lackluster reviews]₁₃, [but we hit the jackpot on our night.]₁₄, [this place has a great vibe !]₁₅, [we were seated immediately]₁₆, [and staff was beyond courteous and attentive.]₁₇, [we were approached by a few staff members]₁₈, [which gave us the feeling of true teamwork.]₁₉, [our server was attentive]₂₀, [and even sparked up some conversation throughout our meal.]₂₁, [we started with a hot pot of chicken tom kha kai]₂₂, [and this soup hit the spot.]₂₃, [my worries were vanishing with every spoonful.]₂₄, [they say]₂₅, [chicken soup will cure a cold]₂₆, [and that menudo will feed a hangover.]₂₇, [well, i]₂₈, [now believe]₂₉, [tom kha kai is the cure for the blues]₃₀, [because it sure made me happy !]₃₁, [for our main dish we shared the red chicken curry.]₃₂, [a little heavy on the red and green peppers but very tasty and was the perfect match with the soup]₃₃, [so we will definitely be back]₃₄, [and will be sharing this place with some of our closest friends.]₃₅, [i've already made lunch plans for next wednesday.]₃₆
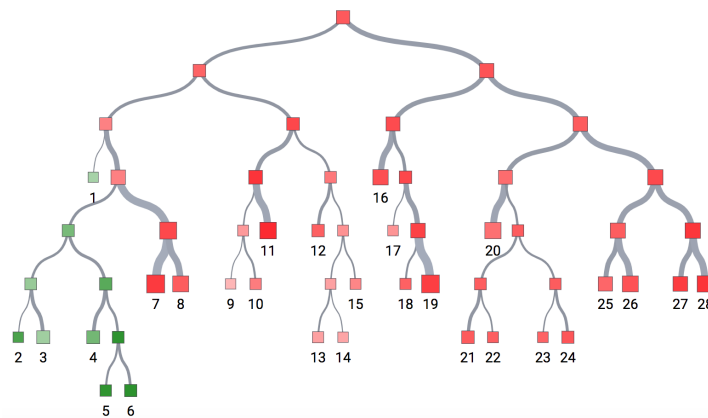
Figure 7: Accurately predicted example, Gold-label polarity: −0.5, Predicted polarity: −0.391,
Discourse: [stopped in here for a friday happy hour with co-workers.]₁, [the beer was decently]₂, [priced for happy hour.]₃, [the appetizers were decently priced,]₄, [which would be awesome]₅, [if they were good.]₆, [the chicken strips were terrible.]₇, [i have never eaten something so greasy and yet dry all at once.]₈, [they are beer battered (like fish)]₉, [which could be good,]₁₀, [but the execution on this was terrible.]₁₁, [the outside was really greasy]₁₂, [which took away all of the crispy goodness]₁₃, [that usually happens]₁₄, [when things are battered and deep fried.]₁₅, [the chicken itself was dry as a bone.]₁₆, [we also got an order of fries]₁₇, [that came out cold]₁₈, [and were just below mediocre.]₁₉, [the place was really warm,]₂₀, [which could be attributed to the summer heat,]₂₁, [but we were sitting inside,]₂₂, [so there is a fair assumption]₂₃, [that air conditioning would be involved.]₂₄, [i'll pass next time]₂₅, [my coworkers are planning a trip here.]₂₆, [i'd be better off]₂₇, [eating at mcdonald's.]₂₈

Figure 8: Inaccurately predicted example, Gold-label polarity: 1, Predicted polarity: −0.098,
Discourse: [upon first moving here 2 years ago,]₁, [i had the worse experience]₂, [attempting to get an airbrush spray tan at this salon.]₃, [they had only 2 people at specific times]₄, [that could spray you custom.]₅, [no problem showed up]₆, [and the tech could not figure out how to use the gun.]₇, [so awkward enough]₈, [him being a male]₉, [and standing there naked, i to get my money back]₁₀, [after waiting 20 min.]₁₁, [well a couple months back]₁₂, [they ran a deal for versa,]₁₃, [which is a booth spray tan.]₁₄, [i love this booth.]₁₅, [it is like airbrushing but private,]₁₆, [and this spray tan absolutely does not smell or stain your sheets!]₁₇, [i found this]₁₈, [upon leaving denver, co. and just]₁₉, [until i saw it online on living deals for amazon. :)]₂₀, [one down two]₂₁, [to go.]₂₂, [all for $ 29 :) love !]₂₃, [as far as the gym goes,]₂₄, [never used it !]₂₅

Figure 9: Inaccurately predicted example, Gold-label polarity: −0.5, Predicted polarity: 0.085,
Discourse: [i hate having]₁, [to write a poor review for this joint!]₂, [the owner is a really great guy]₃, [and the service was excellent.]₄, [the place is decorated well]₅, [and has a clean finished look.]₆, [i really wanted to love the pudding]₇, [but it really did n't work out for my wife and i. from first glance]₈, [the pudding was all very soupy]₉, [and while it tasted]₁₀, [okay, was not anything to write home about.]₁₁, [the shop is trying too hard]₁₂, [to be an ice cream or gelato setup.]₁₃, [i think all the flavors and take away from their core business model.]₁₄, [i think]₁₅, [they should focus on making the rice pudding more solid]₁₆, [and have a couple]₁₇, [warm pudding options.]₁₈, [i can envision a warm rice pudding with some nuts and raisins with some brown sugar or cinnamon on top]₁₉, [yum !]₂₀, [shoot for rich, creamy and full of flavor.]₂₁, [my best wishes go out to them]₂₂, [and hope]₂₃, [that the masses will enjoy it more than we did.]₂₄, [they are good folks]₂₅, [and deserve to be successful.]₂₆
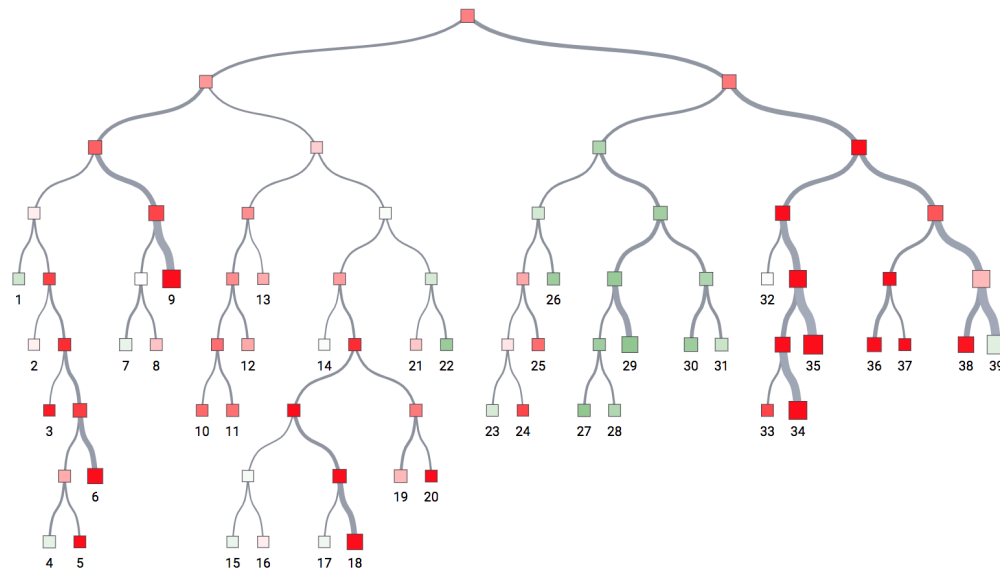
7456

Figure 10: Inaccurately predicted example, Gold-label polarity: -0.5, Predicted polarity: -0.006, Discourse: [i've been here a couple of times in the past.]$_1$, [usually at someone else's suggestion.]$_2$, [i ca n't say]$_3$, [that i recommend this place,]$_4$, [unless you like]$_5$, [your lunch served up with a lot of attitude.]$_6$, [the lady]$_7$, [that takes the orders at the counter]$_8$, [is usually abrasive and rude.]$_9$, [i am the type of person]$_{10}$, [who will kill the meanest person with kindness,]$_{11}$, [but there are places]$_{12}$, [where i draw the line.]$_{13}$, [so, i have drawn the line with rome's pizza.]$_{14}$, [the funny part about it all is]$_{15}$, [that my line is often zig-zag and curvy,]$_{16}$, [so i still go here]$_{17}$, [when someone else wants to go.]$_{18}$, [hehe.]$_{19}$, [my friends like the abuse i guess.]$_{20}$, [one friend says]$_{21}$, [the lady is nice to him.]$_{22}$, [the plus side]$_{23}$, [of going here is the fact]$_{24}$, [that they serve an average pizza by the slice with your custom toppings.]$_{25}$, [they also make hoagies and some other dishes.]$_{26}$, [they have a nice lunch special]$_{27}$, [that includes soda for a few bucks.]$_{28}$, [they also serve some typical american favorites like hot wings.]$_{29}$, [i usually order the of pizza lunch special]$_{30}$, [and get the unsweetened tea.]$_{31}$, [i'm not sure]$_{32}$, [why i forget,]$_{33}$, [but their tea tastes horrible]$_{34}$, [because the water from the fountain tastes terrible !]$_{35}$, [but it never fails, i forget]$_{36}$, [that i need to of water with me.]$_{37}$, [all in all, this place is a dive.]$_{38}$, [give it a try.]$_{39}$

7457