From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation

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

Abstract

Sentiment analysis, especially for long documents, plausibly requires methods capturing complex linguistics structures. To accommodate this, we propose a novel framework to exploit task-related discourse for the task of sentiment analysis. More specifically, we are combining the large-scale, sentiment-dependent MEGA-DT treebank with a novel neural architecture for sentiment prediction, based on a hybrid TreeLSTM hierarchical attention model. Experiments show that our framework using sentiment-related discourse augmentations for sentiment prediction enhances the overall performance for long documents, even beyond previous approaches using well-established discourse parsers trained on human annotated data. We show that a simple ensemble approach can further enhance performance by selectively using discourse, depending on the document length.

1 Introduction

Predicting whether a given word, sentence or document expresses a positive, neutral or negative sentiment is a fundamental task in Natural Language Processing (NLP). For instance, a recent survey of text mining papers from 1992-2017 has found that out of $4,346$ papers, 467 had a sentiment analysis component (Liu et al., 2019a). While early "bag-of-word" sentiment prediction models (Taboada et al., 2011) and their extensions (Wilson et al., 2009) already show promising results on the task, they all share one inherit limitation: Due to the absence of temporal information, they are not able to fully capture the semantics (and therefore the sentiment) of long texts, where different meanings oftentimes directly emerge from the word order, underlying syntax and discourse structures.

Recent models for sentiment analysis address this limitation by leveraging sequential paradigms (Dos Santos and Gatti, 2014; Kim, 2014; Tai et al., 2015; Adhikari et al., 2019b), simple hierarchical information (Yang et al., 2016), complex syntactic structures on sentence level (Socher et al., 2013) or discourse structures of multi-sentential text (Ji and Smith, 2017).

This paper follows the last line of aforementioned research, by developing a framework to exploit automatically generated, large-scale, domain-related discourse structures for sentiment prediction. Arguably, such framework can be especially beneficial for long documents that examine positive and negative aspects of a subject matter in complex rhetorical structures, like the ones shown in Figure 1.

More specifically, in this work, we generate complete and hierarchical RST-style discourse trees (Mann and Thompson, 1988) with leaf nodes representing clause-like document fragments, called elementary discourse units (EDUs) and internal tree nodes labelled with a nuclearity assignment (Nucleus, Satellite), encoding the importance of a node in its local subtree[1]. To incorporate these RST-style discourse structures, we employ a hybrid approach inspired by Bowman et al. (2016) and Choi et al. (2018), integrating a TreeLSTM (Tai et al., 2015) with the well-established Hierarchical Attention Network model (HAN) (Yang et al., 2016). From Ji and Smith (2017), we further adopt a non-competitive tree attention mechanism that is shown to be more appropriate in this context[2].

[1] Discourse relation are not considered in this work.

[2] We did not apply tree-transformers to the task, as in spite of recent proposals (e.g. Shiv and Quirk (2019), Nguyen et al. (2020)), no standard method has been widely agreed upon yet and results are still rather preliminary.

Proceedings of the 28th International Conference on Computational Linguistics, pages 185–197
Barcelona, Spain (Online), December 8-13, 2020

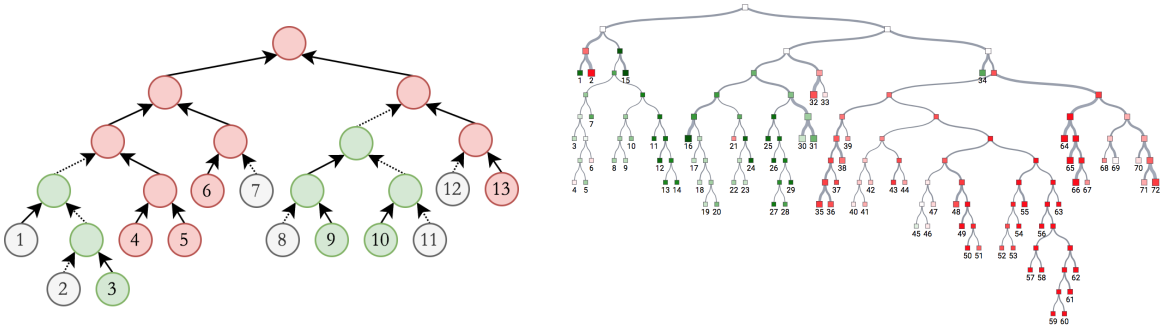

Figure 1: Sentiment annotated discourse trees for non-trivial documents containing 13 (left) and 72 (right) clause-like components with positive and negative constituents. Gold-label sentiment is negative (left) and neutral (right). Dashed/Thin lines indicate supplementary information, solid/thick lines indicate primary importance. Full text for left example is: *[I've been a member for a month now,]$_1$, [and I guess]$_2$, [I 'm able to get my workout done.]$_3$, [I do find myself annoyed]$_4$, [how cramped it is at the weights.]$_5$, [The equipment is older,]$_6$, [but it suffices.]$_7$, [I worked out at another studio on 3rd]$_8$, [and it was amazing!]$_9$, [It was so clean, nice, and new -]$_{10}$, [TV 's on every cardio machine.]$_{11}$, [When i came back to this location,]$_{12}$, [I felt bad.]$_{13}$* Due to space limitations a larger version as well as the corresponding full text for the right example is shown in Appendix A.

Aiming to enhance the task of sentiment analysis by using discourse, it seems intuitive to employ domain-related discourse structures. Therefore, instead of using the standard RST-DT discourse treebank in the news domain (Carlson et al., 2002), we decide to infer discourse structures automatically learned from sentiment annotations (Huber and Carenini, 2019) on our discourse-augmented Yelp'13 treebank called MEGA-DT (Huber and Carenini, 2020). This way, our framework goes from sentiment to sentiment, in the sense that the discourse structures used to improve the sentiment predictions are generated through distant supervision from sentiment itself. Our hypothesis is that a parser trained on a large "silver-standard" discourse treebank automatically generated from sentiment will generate more useful discourse trees for sentiment prediction than one trained on a small and generic treebank, even if such treebank is human-annotated for RST discourse structures.

In a series of experiments we show that while our novel approach to discourse-based sentiment prediction is statistically equivalent to the performance of sequential models, it does deliver substantial performance gains for long documents, where discourse plays a crucial role to reveal the sentiment of a complete document. Furthermore, our experiments indicate that the performance of discourse-based sentiment prediction is significantly improved when using discourse trees generated by distant supervision on sentiment, compared to the traditionally acquired RST-DT discourse corpus. Using an additional ensemble method, we can further improve the performance and, even if only by a small margin, significantly outperform individual models.

2 Related Work

This work is located at the intersection of recent approaches on discourse parsing and sentiment analysis and mostly influenced by four lines of research:

(1) RST-style Discourse Parsing is a valuable upstream task for many downstream models (e.g. Ji and Smith (2017), Gerani et al. (2014)). Different approaches either separate discourse parsing "vertically" into sub-tasks on sentence-level, paragraph-level and document-level (Joty et al., 2015; Ji and Eisenstein, 2014), or "horizontally", separating the prediction of structure and nuclearity from the relation computation (Wang et al., 2017). Furthermore, approaches have been explored to aggregate documents bottom-up using CKY (Joty et al., 2015) or employing local shift-reduce strategies, predicting the tree-structure through a sequence of actions based on linguistic features (Ji and Eisenstein, 2014; Subba and Di Eugenio, 2009; Wang et al., 2017) or dense representations (Yu et al., 2018). Empirically, Wang et al. (2017) show that the combination of horizontal separation with a shift-reduce parsing framework achieves

competitive performance, reaching state-of-the-art results on the structure-prediction task. In this work, we demonstrate the potential of this discourse parser trained on a large-scale sentiment-dependent treebank (MEGA-DT) to generate discourse trees for sentiment prediction, enhancing the performance on long and diverse documents.

(2) Neural Sentiment Analysis is a common sub-task in many real world systems with Kim (2014) being the first to show the effectiveness of convolutional neural networks for the task. Yang et al. (2016) followed shortly after with their Hierachical Attention Network model (HAN), proposing one of the first hierarchical models for text classification. HAN separates the task at the sentence-level and builds a model comprising of two hierarchical components, each with an additional attention mechanism. Further successful approaches to predict sentiment have been explored recently by Adhikari et al. (2019a), proposing a model based on BERT, and Adhikari et al. (2019b), applying a simple but more regularized BiLSTM to the task. In this fast moving area, our goal is to investigate the influence of discourse information on the task of sentiment analysis. We therefore decide to build our framework on the HAN model (Yang et al., 2016), which is the most established, yet recent approach in the field, previously re-implemented and tested in many studies. We inject discourse information using TreeLSTMs (Tai et al., 2015), which are also well-established compared to tree-transformers, for which architectural variants and results are still preliminary (e.g. Shiv and Quirk (2019), Nguyen et al. (2020)).

(3) Combining Discourse Parsing and Sentiment Analysis has been previously explored in multiple lines of work (Bhatia et al., 2015; Hogenboom et al., 2015; Nejat et al., 2017; Ji and Smith, 2017). Architecture-wise, the most closely related approach to our new model has been proposed by Ji and Smith (2017), where discourse trees generated by the DPLP parser (Ji and Eisenstein, 2014) trained on RST-DT are used in a recursive neural network to predict sentiment for multiple corpora. In their evaluation, the authors show slight improvements compared to the sequential HAN model. These initial positive results are a key motivation for our work, in which we aim to further improve the performance, especially on long documents, by not only training the discourse parser on a larger and more appropriate treebank (i.e. MEGA-DT), but also by improving the sentiment prediction, replacing recursive neural networks with superior TreeLSTMs, tightly integrated with HAN.

(4) (Discourse) Tree Learning tries to automatically infer discourse trees from large amounts of data. In popular approaches, trees are inferred directly while learning a neural model for a downstream task, such as text classification (Karimi and Tang, 2019) or extractive summarization (Liu et al., 2019b). Along this line of research, we previously proposed a similar objective in Huber and Carenini (2019), automatically generating discourse trees from distant supervision of a downstream task (sentiment analysis). However, we employed a rather different approach. Instead of trying to induce discourse trees directly during training of a neural network, we propose a dedicated system, comprising of well-established methods, to directly generate discourse trees. With the resulting large-scale, sentiment influenced discourse treebank called MEGA-DT, we reported promising results on the task of discourse parsing itself in Huber and Carenini (2020). Showing the potential of applying MEGA-DT to the task of sentiment prediction is a goal of this work.

3 Sentiment to Sentiment Framework

Our sentiment to sentiment framework involves three phases: A phase of discourse augmentation (Figure 2 (a)), in which we follow our previous approach described in Huber and Carenini (2019) and Huber and Carenini (2020). For each document in a corpus containing document-level sentiment annotation, we generate corresponding, task-dependent discourse trees. Then, this discourse augmented sentiment treebank is used to train a discourse parser. In the second phase (Figure 2 (b)), the trained discourse parser is applied to the original corpus, using the predicted trees to train our new discourse-based sentiment predictor. Finally, in the third phase (Figure 2 (c)), the trained framework is applied to any new document. First, the trained discourse parser generates the discourse tree for the document. Subsequently, this tree (along with the document itself) is fed to our sentiment predictor, which returns the most likely sentiment. In essence, we go from sentiment annotations to sentiment predictions through discourse augmentation.

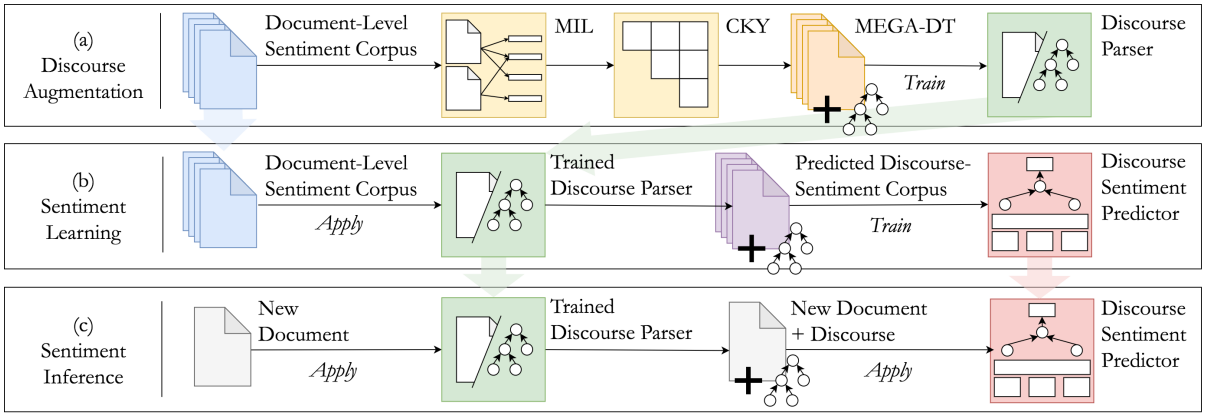

Figure 2: Our proposed sentiment analysis framework, containing, three phases of training/inference to integrate discourse parsing and sentiment analysis.

For the first phase, we briefly describe the discourse augmentation step adopted from our previous work (Huber and Carenini, 2019; Huber and Carenini, 2020) in section 3.1. For phase two, we focus on our novel sentiment predictor in section 3.2. The inference phase is straightforward and will be limited to the description in Figure 2 (c) for brevity.

3.1 Sentiment Inspired Discourse Trees

The approach to generate "silver-standard" partial discourse trees (incorporating structure and nuclearity) from distant sentiment supervision (Huber and Carenini, 2019; Huber and Carenini, 2020) comprises two major components. First, documents are annotated for sentiment and importance at the EDU-level using a neural Multiple-Instance Learning (MIL) method (Angelidis and Lapata, 2018), solely utilizing document-level supervision signals given in the original corpus. In particular, MIL infers a sentiment polarity label p_x within the interval of $[-1, 1]$ for each EDU x, depending on the distribution of words/EDUs within and between documents. Using the neural model by Angelidis and Lapata (2018), an additional attention mechanism is internally used to weight the importance of EDUs for the overall document sentiment. The attention-weight a_x in the interval $[0, 1]$ of EDU x is also extracted from the model and subsequently used as an importance score when aggregating sub-trees. Next, the tuples (p_x, a_x) are combined in a binary, bottom-up approach using dynamic programming, inspired by CKY (Jurafsky and Martin, 2014). With a multitude of possible discourse trees generated in this way, the tree-structure minimizing the divergence between the document sentiment gold-label and the predicted sentiment, obtained by combining the tuples (p_x, a_x) according to equation 1, is deemed to represent the document discourse-structure.

$$p = \frac{p_{c_l} * a_{c_l} + p_{c_r} * a_{c_r}}{a_{c_l} + a_{c_r}} \qquad a = \frac{a_{c_l} + a_{c_r}}{2} \tag{1}$$

p_{c_l} and p_{c_r} represent the sentiment polarity labels of the left and right sub-tree respectively. a_{c_l} and a_{c_r} represent the importance scores, retrieved from the internal MIL attentions. p and a are the respective labels for the parent sentiment polarity and importance score (Huber and Carenini, 2019).

As extensively described in Huber and Carenini (2020), the unconstrained CKY approach is not directly applicable for long documents (considered especially important in this work), since the spatial complexity of the CKY approach grows according to the Catalan number, with respect to the number of EDUs in a document. This effectively renders the unconstrained CKY approach insufficient for processing documents with over ≈ 20 EDUs, even on modern infrastructures[3]. To overcome this problem, we apply the augmentations proposed in Huber and Carenini (2020), reducing the spatial complexity through the application of a beam-search approach, improving the diversity in low-level trees through a stochastic extension. Further, we compute the additional nuclearity attribute, which has previously shown to be an

[3]We used an Intel Core i9-9820X (10 Cores, 3.30 GHz) with a RTX 2080 Ti (128 GB RAM) for our experiment.

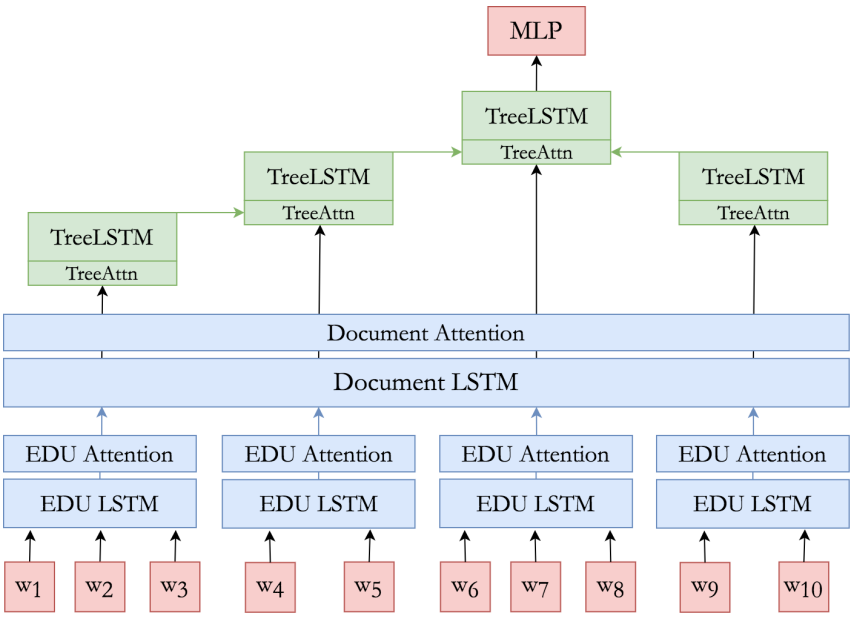

Figure 3: Topology of our hybrid approach using sequential HAN components (blue) in combination with an attention-extended discourse-inspired TreeLSTM (green) aggregation on the dependency discourse tree. Inputs and outputs are red.

important cue for a variety of downstream tasks (Marcu, 2000; Ji and Smith, 2017; Shiv and Quirk, 2019). With these extensions, the discourse-tree generation process can be effectively applied to documents of arbitrary length.

3.2 From Discourse to Sentiment

Discourse structure can be beneficial and complementary to sequential information for sentiment prediction, especially for long, complicated and nuanced documents (see Figure 1). We therefore take a balanced approach in this work, combining a sequential and tree-structured component to predict sentiment. Following the intuition by Bowman et al. (2016) and Choi et al. (2018), we encode low-level representations in a sequential manner and use the inferred trees on higher levels to guide the prediction of the document-level sentiment.

Sequential Model Component With the HAN model being a strong baseline for many tasks, despite its simple architecture, we decide to take advantage of this contextualization for individual EDUs, as well as for the document-level contextualization (see bottom in Figure 3). In the standard HAN model the first-level outputs (originally being sentence representations) are used as inputs to a document-level LSTM, augmented with an attention module, to generate the final hidden representation of a document. (see eq. 2 to 4).

$$u_i = tanh(Wh_i + b) \tag{2}$$

$$\alpha_i = \frac{exp(u_i^\top c)}{\sum_{j \in d} exp(u_j^\top c)} \tag{3}$$

$$h_d = \sum_{i \in d} \alpha_i h_i \tag{4}$$

With h_i as the hidden-state of EDU i, obtained from the document-level LSTM, c as the attention context-vector and d representing the set of all sentences/EDUs in the document. We inject discourse information by replacing the computation of the attention weighted sum of the EDU embeddings (equation 4) with a hierarchical TreeLSTM aggregation of the attention-weighted hidden states.

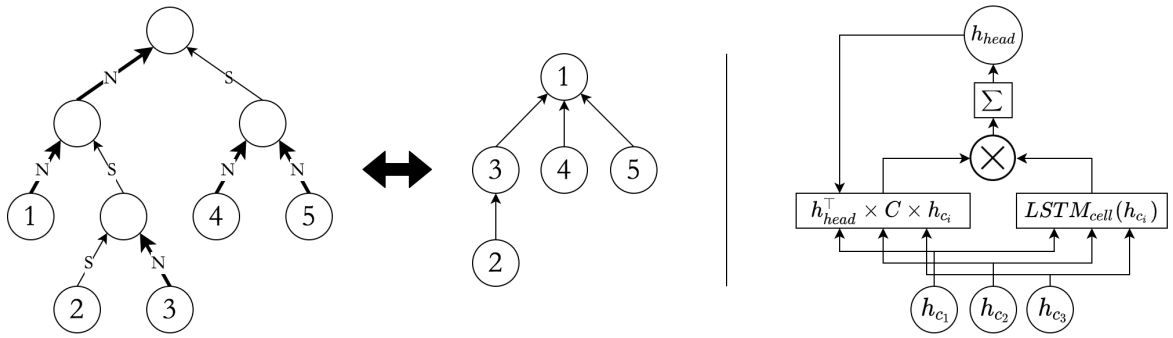

Figure 4: Left: Example transformation of a constituency tree to the respective dependency tree.
Right: Conditional attention module, weighting the importance of child node LSTM encodings $LSTM_{cell}(h_{c_i})$, given the initial head node hidden-state h_{head}

$$h_d = TreeLSTM(\forall_{i \in d} \alpha_i h_i) \tag{5}$$

We omit the description of the sentence-/EDU-level computations for brevity, as they are unchanged from the original HAN model.

Hierarchical Model Component Using a tree-guided hierarchical aggregation of EDU-level hidden-states to generate a discourse-level hidden representation of the document, we allow more important information according to the discourse tree to be more influential in the computation of the final document representation, as motivated by the examples in Figure 1. There are two crucial decisions on how to incorporate the discourse-guided tree aggregation:
(1) The tree representation. Although discourse parsing typically processes constituency tree-structures, most successful downstream applications of discourse parsing benefit from dependency discourse trees (e.g., Marcu (2000), Ji and Smith (2017), Shiv and Quirk (2019)). Even though both tree representations are conveying the same information and near-isomorphic conversions are available (Morey et al., 2018), we believe that this is because of the different role that nuclearity plays in the tree-representations. In particular, while in constituency trees nuclearity is an attribute of internal tree-nodes, head-dependent relations in the dependency tree are fundamentally shaped by the nuclearity attribution. This more explicit representation of nuclearity can benefit downstream applications. For this reason, we are converting the RST constituency trees into dependency representations (see left of Figure 4).
(2) The aggregation approach has a significant impact on the performance of the model. In this work, we choose the TreeLSTM model by Tai et al. (2015), an evolution of the recursive neural network used in Ji and Smith (2017). Following the intuition for tree-attention given by Ji and Smith (2017), we add a conditional, non-competitive attention module to the child-sum TreeLSTM, augmenting the aggregation of text-spans according to their position in the dependency discourse tree (see eq. 6 to 7). This extension has not been proposed as part of the TreeLSTM by Tai et al. (2015), however showed improved performance when used in combination with a recursive neural network for the task of discourse parsing (Ji and Smith, 2017), which lets us to believe it can also enhance the TreeLSTM for our problem at hand.

$$\alpha_i = \sigma(h_{head}^{\top} \times C \times h_{c_i}) \tag{6}$$

$$h_{head} = LSTM_{cell}(\sum_{i \in dep(h_{head})} \alpha_i h_{c_i}) \tag{7}$$

With C as the attention matrix of dimension $(|h_{head}| \times |h_{c_i}|)$, h_{head} representing the hidden-state of the head node and $dep(h_{head})$ returning the indices of the dependent child nodes of h_{head}. Please note that the hidden representation of every node in the dependency discourse tree is initialized with the attention-weighted EDU representation obtained from the sequential component and is updated by

the TreeLSTM function shown in equation 7. We combine the head-node EDU representation with the dependants' sub-tree encoding during the bottom-up tree aggregation process (see top of Figure 3 and right of Figure 4). We name our new model DAH (**D**iscourse **A**ugmented **H**AN).

4 Evaluation

In this section, we define the experimental setup and show empirical results of our novel approach, predicting sentiment using sentiment-inspired discourse parsing in the context of previous work. We present the datasets used in this work in section 4.1. Afterwards, the evaluation metrics and their intuitive justifications are mentioned in section 4.2, followed by a short description of the baselines (section 4.3). We finish the evaluation section by giving insights into our preliminary evaluations determining the system's hyper-parameters in section 4.4 and describe the final experiments and results in section 4.5.

4.1 Datasets

As shown in Figure 2, our proposed methodology requires two sets of corpora. In the first step, as described in section 3.1, we train a top-performing discourse parser (Wang et al., 2017) on a discourse corpus containing RST-style trees. In this step, we use two treebanks:
RST-DT: As a human-annotated gold-standard discourse treebank most widely used for discourse related research following the RST theory (Mann and Thompson, 1988). The dataset contains 385 discourse-annotated news articles from the Wallstreet Journal.
MEGA-DT: Our recently proposed "silver-standard" discourse corpus (Huber and Carenini, 2020), generated in an effort to provide an automatically annotated, large-scale discourse treebank. The corpus is based on the publicly available Yelp'13 sentiment dataset and contains around 250,000 documents annotated with full RST-style discourse trees containing structure and nuclearity attributes. The treebank has shown superior performance to small human-annotated datasets (including RST-DT) on the discourse domain-transfer task, reaching the best performance when evaluated on news/instruction treebanks.

To evaluate the potential of the discourse treebanks to predict sentiment in combination with our novel model architecture, we annotate a large-scale sentiment dataset with discourse trees generated by the discourse parser (Wang et al., 2017), trained on the corpora described above. The publicly available dataset used in this work is the **Yelp'13 dataset**, published by Tang et al. (2015) , containing customer reviews annotated with gold-label sentiment on a 5-point scale. For models incorporating discourse, the previously discourse segmented dataset published by Angelidis and Lapata (2018) is used with an 80%/10%/10% train/dev/test-split.

Please note that since we use the same base-corpus for training the discourse parser (MEGA-DT) and predicting sentiment for the final evaluation (Yelp'13), we restrict the data used to train the discourse parser to the training-portion of the corpus. This way we ensure that development- and test-documents are unseen during the whole training process.

4.2 Metrics

Previous models tackle the task of sentiment analysis by interpreting it as a classification problem. While this problem definition is valid for many text categorization tasks, we believe that sentiment analysis should be additionally evaluated as a regression task, taking the ordinal nature of the output into account. To more rigorously evaluate the models in our evaluation, we show four metrics for each system, including the commonly used accuracy and F1-score, as well as the Mean-Squared-Error (MSE) and Mean-Absolute-Error (MAE) metrics.

4.3 Baselines

We compare our new model against two closely related models, namely the Hierarchical Attention Network (HAN) by Yang et al. (2016) and the MILNet model (Angelidis and Lapata, 2018), which is used as part of the discourse-augmentation process itself in Huber and Carenini (2019) and Huber and Carenini (2020). With those two closely related baselines we ensure that possible confounding factors in the comparison are minimized, allowing for a clear picture on the effectiveness of incorporating discourse structures into the task of sentiment analysis.

Model	Yelp'13			
	Acc	F1	MSE	MAE
HAN	66.20	64.26	0.486	0.379
MILNet	64.19	61.93	0.584	0.417
$\text{DAH}_{\text{RST-DT}}$	65.71	63.49	0.496	0.384
$\text{DAH}_{\text{MEGA-DT}}$	$\simeq$ †66.07	$\simeq$ †64.09	$\simeq$ ‡0.491	$\simeq$ ‡0.381
Ensemble(HAN+$\text{DAH}_{\text{MEGA-DT}}$)	***66.27**	***64.30**	***0.483**	***0.377**

Table 1: Final evaluation on the Yelp'13 datasets, subscripts in model names indicate discourse-augmentation treebanks used to generate discourse trees. Best model for each metric is **bold**. $\simeq$Performance statistically equivalent to HAN model, †Discourse-augmentation treebank significantly better than RST-DT with p-value .05. ‡Discourse-augmentation treebank marginally significantly better than RST-DT with p-value .05-.1, *Statistically significant to best model on metric. All significance computations are Bonferroni adjusted.

4.4 Encodings and Hyper-Parameters

To support a fair comparison, we use the same encodings and model-dependent hyper-parameters in all systems. We replace the domain-depended pre-trained word2vec encodings (Mikolov et al., 2013) used in the original HAN model, with standard GloVe embeddings (Pennington et al., 2014). We add MSE and MAE evaluation metrics to the publicly available open-source deep learning toolkit for the original HAN model[4]. For the MILNet baseline, we align with our previous approach in Huber and Carenini (2019), which is also consistent with the adapted HAN model. Regarding our novel approach, we convert the constituency tree output of the discourse parser into a dependency tree according to Hayashi et al. (2016). We run preliminary evaluations on the development-set, comparing a set of loss-function (namely **Cross-Entropy**, MSE, MAE)[5] and interpreting the task as either, a classification- or a regression-problem. However, without any further fine-tuning and adaptations, using a regression-based loss is not advisable. In accordance with the intuition described above, we execute further hyper-parameter search on the main properties of the model itself, exploring a set of 5 learning rates ($\{0.1, 0.05, \mathbf{0.01}, 0.05, 0.001\}$) along with three optimization strategies (Adam (Kingma and Ba, 2014), AdaGrad (Duchi et al., 2011), **SGD** (Robbins and Monro, 1951)). We follow the original HAN implementation using 100 neurons per layer for the bi-directional word and sentence/EDU encodings. The TreeLSTM module contains 512 neurons. The mini-batch size used in all models is set to 64, as suggested in Yang et al. (2016). Dropout is set to 50% for all models.

4.5 Experiments and Results

We compare our novel model using multiple discourse representations obtained from sentiment-inspired discourse structures and standard treebanks against discourse-agnostic systems, solely based on sequential representations on word- and sentence-level. As motivated in Figure 1, we believe that discourse information is especially useful for long documents, where sentiment is generally expressed in a more diverse or subtle way as compared to short reviews with mostly a clear positive or negative sentiment. We align our evaluation with this intuition by comparing the systems' overall performance in Table 1 and further showing insights into the performance based on the document length in Figure 5.

The final comparison in Table 1 reports the performance of two baseline systems, not taking discourse information into account, along with two versions of our novel approach, incorporating discourse, and an ensemble method. The performance of all models is averaged over 5 independent runs with different random initializations. All models using discourse ($\text{DAH}_{\text{RST-DT}}$, $\text{DAH}_{\text{MEGA-DT}}$ and the ensemble of HAN and $\text{DAH}_{\text{MEGA-DT}}$) are trained with the top-performing discourse parser by Wang et al. (2017). All discourse-inspired models further employ an identical neural network architecture, allowing us to directly

[4] `https://github.com/castorini/hedwig`

[5] Selected hyper-parameter is **bold**

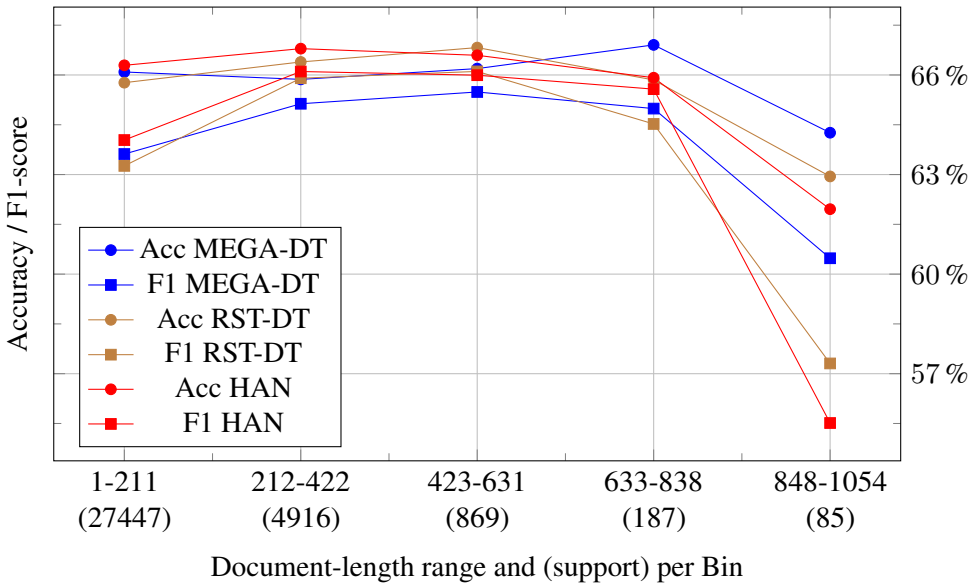

Figure 5: Accuracy and F1-score over document-lengths aggregated into 5 bins on the Yelp'13 dataset

evaluate the impact of different types of discourse trees on the task of sentiment analysis.

The best average performance of an individual model (not using the ensemble method) is achieved by the sequential HAN model shown in the first row in Table 1. Even though the average result over 5 independent runs for the $\text{DAH}_{\text{MEGA-DT}}$ system is below the HAN performance, they are statistically equivalent. When compared to the discourse-inspired $\text{DAH}_{\text{RST-DT}}$ model, the performance increase of $\text{DAH}_{\text{MEGA-DT}}$ is statistically significant on the accuracy and F1-score measures and marginally significant for the MSE and MAE. Interestingly, the MILNet model, which is used as an early part of the pipeline to generate the MEGA-DT discourse treebank, does perform substantially worse than the $\text{DAH}_{\text{MEGA-DT}}$ model, which leads us to believe that the combination of the CKY tree aggregation and the DAH sentiment neural-network are able to extract valid and important sentiment information and improve the performance despite the potential propagation of error from the early stage MILNet component. Besides the individual models, we also employ an additional experiment with a model-ensemble combining the two top performing models (HAN and $\text{DAH}_{\text{MEGA-DT}}$), taking their respective strength in different document-length-ranges (as revealed in Figure 5) into account. The model will be explained in more detail below.

The results shown in Table 1 indicate equal performance of our new $\text{DAH}_{\text{MEGA-DT}}$ methodology when compared to the original HAN model. However, discourse should arguably be more useful for long documents. Therefore, we further investigate into the document-length dependent performance of the models by splitting the test-set into 5 test-document length-depended bins to show the performance across different document sizes (measures by the number of words). We exclude the MILNet baseline in this evaluation due to its clearly inferior performance compared with the sequential HAN model as shown in Table 1.

The results shown in Figure 5 confirm our initial intuition on the usefulness of discourse structures for long documents, showing strong improvements for our discourse-dependent system in the two rightmost bins. While the performance generally drops for longer documents, the performance decrease is more severe for the sequential HAN model. Generally, we believe that the task of sentiment prediction is harder on longer and more diverse documents, however, we also partly account the performance decrease to the small number of long documents in the Yelp'13 corpus, as shown in the support for each of the bins on the horizontal axis of Figure 5. While the support shown here is on the test-portion, the general length-distribution on the training- and development-set are similarly skewed towards short documents.

It can further be seen that the significant performance increase on the overall dataset achieved by the $\text{DAH}_{\text{MEGA-DT}}$ over the $\text{DAH}_{\text{RST-DT}}$ can be mostly attributed to the performance increase in the two

right-most bins, containing documents with more than 632 words.

With this confirmation of our initial intuition, we generate a document-length-dependent ensemble of the two top-performing models (HAN and $\text{DAH}_{\text{MEGA-DT}}$) as mentioned above, to take advantage of the strength of both systems by selecting the appropriate classifier with a simple threshold – the document length. To determine the threshold, we evaluate both models on the development-set and select the average of the optimal threshold over 3 runs independently for each metric of interest. We then combine the results of the two top performing models on the test-set according to the determined threshold. As shown in Table 1, our ensemble approach significantly outperforms all the individual models, but admittedly only by a narrow margin. Nevertheless, overall the results indicate potential for further improvements in discourse-inspired sentiment analysis for long documents as well as in using ensembles of sequential and tree-driven models to effectively process documents with different levels of complexity.

5 Conclusion and Future work

In this work, we explore the next step along the recent line of research on discourse-inspired sentiment analysis, going from sentiment annotations to sentiment prediction through discourse augmentation. We integrate modern discourse parsing approaches into existing, sequential sentiment analysis frameworks, enhancing the model performance through the use of the large-scale MEGA-DT discourse dataset and a hybrid approach based on sequential and tree-based components (HAN combined with TreeLSTM). Our proposed approach shows to be especially beneficial when predicting sentiment for long documents containing mixed aspects, combined with complex rhetorical structures. Generating a model-ensemble with a simple threshold, based on the document length, improves the overall performance, showing statistically significant results.

We compare our newly developed model with the well-established HAN model. In future work, we plan to compare the standard DocBERT model (Adhikari et al., 2019a) and discourse-inspired versions of it, to further solidify the findings in this work. We also plan to generate other large-scale datasets according to Huber and Carenini (2020) and evaluate our model on further "silver-standard" discourse treebanks. Using a neural discourse parser, such as Yu et al. (2018) or Guz et al. (2020) to train on MEGA-DT is another extension of this work. Besides the task of sentiment analysis, extractive summarization has recently been shown to align well with discourse structures in a transformer framework (Xiao et al., 2020), giving rise to potential improvements using the DAH model on this task. As another extension, we intend to look into more sophisticated ways to ensemble the sequential- and discourse tree-based models.

Acknowledgments

We thank the anonymous reviewers and the UBC-NLP group for their insightful comments and suggestions. This research was supported by the Language & Speech Innovation Lab of Cloud BU, Huawei Technologies Co., Ltd.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019a. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019b. Rethinking complex neural network architectures for document classification. 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 4046–4051.

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.

Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*.

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

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Learning to compose task-specific tree structures. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159.

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.

Grigorii Guz, Patrick Huber, and Giuseppe CArenini. 2020. Unleashing the power of neural discourse parsers – a context and structure aware approach using large scale pretraining. In *Proceedings of COLING 2020, the 28th International Conference on Computational Linguistics*.

Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for rst discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136.

Alexander Hogenboom, Flavius Frasincar, Franciska De Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Communications of the 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.

Patrick Huber and Giuseppe Carenini. 2020. Mega rst discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

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*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

S. Liu, X. Wang, C. Collins, W. Dou, F. Ouyang, M. El-Assady, L. Jiang, and D. A. Keim. 2019a. Bridging text visualization and mining: A task-driven survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2482–2504.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019b. 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.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

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.

Xuan-Phi Nguyen, Shafiq Joty, Steven CH Hoi, and Richard Socher. 2020. Tree-structured attention with hierarchical accumulation. *arXiv preprint arXiv:2002.08046*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Vighnesh Shiv and Chris Quirk. 2019. Novel positional encodings to enable tree-based transformers. In *Advances in Neural Information Processing Systems*, pages 12081–12091.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

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.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

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.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Articles: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. Do we really need that many parameters in transformer forextractive summarization? discourse can help ! In *Proceedings of the 1st Workshop on Computational Approaches to Discourse (CODI)*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

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.

A Sentiment Annotated Discourse Trees for long Documents

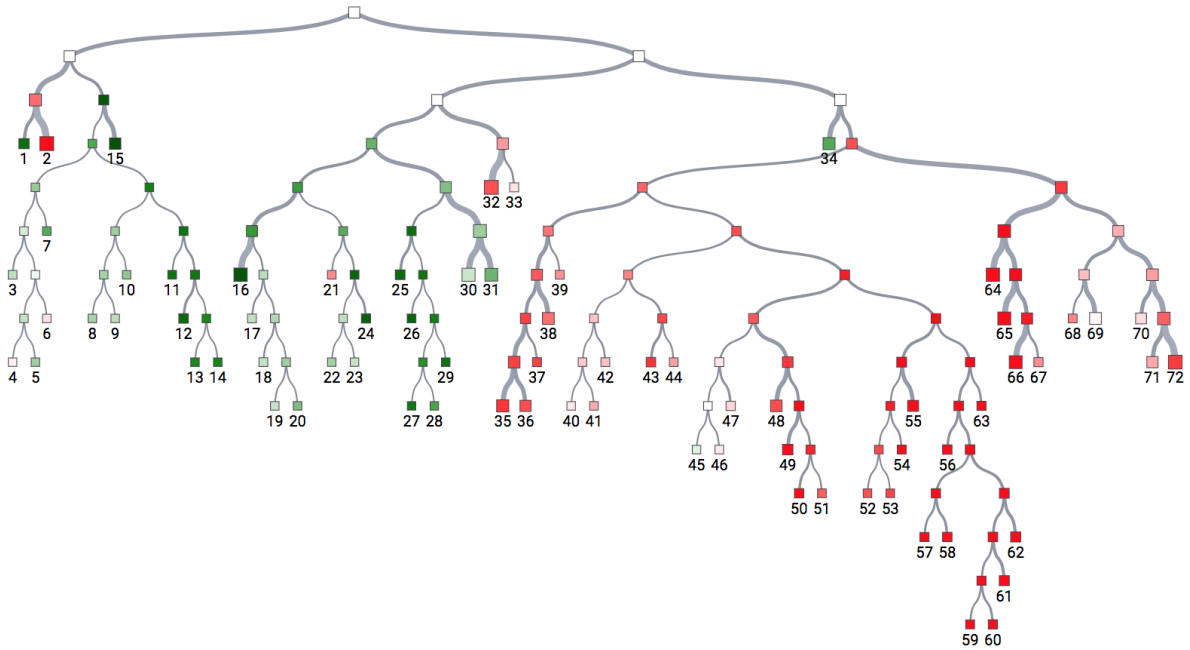

Figure 6: Discourse: [amazing food.]$_1$, [awful, awful service.]$_2$, [the garlic bread. very good.]$_3$, [softer than i expected,]$_4$, [which was nice.]$_5$, [i also just wasn't expecting garlic bread.]$_6$, [so it was a nice surprise.]$_7$, [escargot -]$_8$, [i was the only one at the table (of 10)]$_9$, [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.]$_{72}$