The Report Committee for Elisa Ferracane
certifies that this is the approved version of the following report:

Testing the usefulness of RST and more general representations for discourse analysis across domains and applications

APPROVED BY

SUPERVISING COMMITTEE:

Katrin Erk, Supervisor

Junyi Jessy Li, Co-Supervisor

Testing the usefulness of RST and more general representations for discourse analysis across domains and applications

by

Elisa Ferracane

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2018

Acknowledgments

I stand on the shoulders of giants, or in fact a *community* of giants. I here make an attempt to recognize them all.

I thank my supervisor Katrin Erk who bravely and enthusiastically embarked on this journey with me into unchartered waters well outside her areas of research. I thank my co-supervisor Jessy Li who has been helping me dive deeper into discourse even before she became a faculty member.

I am grateful to my incredible team of mentors: Greg Durrett, Byron Wallace, Rajka Smiljanic, and Colin Bannard. They have each helped guide my research and stretch my research beyond discourse processing.

I recognize my fellow graduate students for providing technical and emotional support, and a shared experience where we have learned and grown together: Stephen Roller, Su Wang, Eric Holgate, Alex Rosenfeld, Pengxiang Cheng, Cindy Blanco and Rachael Gilbert.

I am indebted to the National Science Foundation for awarding me a fellowship in the Graduate Research Fellowship Program.

Finally, I thank my family for their unwavering support and encouragement.

Elisa Ferracane
Austin, Texas
May 2018

Testing the usefulness of RST and more general representations for discourse analysis across domains and applications

Elisa Ferracane, M.A.
The University of Texas at Austin, 2018

Supervisors: Katrin Erk and Junyi Jessy Li

Discourse analysis is a task with enormous potential but is often met with lukewarm results. This report explores how well Rhetorical Structure Theory (RST) and more general representations of discourse can generalize across domains and tasks, and the validity of their underlying assumptions. Our first study attempts to uncover issues in Rhetorical Structure Theory (RST) discourse parsing by starting at the first step of discourse segmentation, and evaluate in the medical domain. Errors on our novel, small-scale medical corpus reveal differences at lower linguistic levels that affect the discourse segmenter, and point to problem areas in the way RST was operationalized.

Our second study focuses on more general representations of discourse that are learned by the model, and that have only a simple constraint of forming a dependency tree. We find these latent trees in fact do not represent discourse and focus instead on lexical cues. We propose a variant of this model that is able to learn deeper structures, but conclude that a different task which makes more use of discourse may be needed in order to produce more discourse-like structures.

Table of Contents

Chapter 1 Introduction 1

1.1 Report Outline 1

1.2 List of Report Contributions 2

Chapter 2 Background 3

2.1 Chapter Overview 3

2.2 Rhetorical Structure Theory (RST) 3

2.3 Structured Prediction and Attention 5

2.4 Chapter Summary 5

Chapter 3 Cross-Domain Discourse Segmentation 6

3.1 Chapter Overview 6

3.2 Motivation 7

3.3 Related Work 7

3.4 Experiment 8

3.4.1 The Corpora 9

3.4.2 Experimental Setup 9

3.5 Results 10

3.5.1 Error Analysis 10

3.6 Chapter Summary 12

Chapter 4 Latent Discourse Structures 14

4.1 Chapter Overview 14

4.2 Related Work 15

4.3 Models 17

4.3.1 Model Choice: Latent or Explicit? 17

4.3.2 Model Variants 18

4.4 Experiments 20

4.4.1 The Corpora 20

4.4.2 Experimental Setup 21

4.5 What is learned, and is it discourse? 21
4.6 Can we inject explicit discourse?... 29
4.7 Can we learn deeper structures? 30
4.8 Chapter Summary 31

Chapter 5 Conclusion... 32
5.1 Future Work 32

References.. 33

List of Figures

2.1 RST trees illustrating hypotactic and paratactic syntactic relations..... 4

List of Tables

3.1 Corpus statistics.......... 9

3.2 Performance of discourse segmentation using two different discourse segmenters and evaluated on two different domains. 10

3.3 Types of segmentation errors in the *Medical* domain with examples of predicted and gold EDU boundaries marked in square brackets (*the square brackets for the citation were changed to curly brackets to avoid confusion with EDU boundaries). 11

4.1 Statistics for the datasets used in the text classification tasks. The vocabulary size is calculated after preprocessing.......... 20

4.2 Accuracy on the three classification tasks using the baseline Liu and Lapata (2018). 22

4.3 Example of attention scores, derived dependency tree and text for a test sentence in Yelp 2017 (gold label=5). Root sentences are underlined. In this example the very light and very dark shadings suggest the model is very certain about the structure. 24

4.4 Example of attention scores, derived dependency tree and text for a test sentence in the Congressional Votes dataset (gold label=N). Root sentences are underlined. In this example, the mixed shading in the first column suggests the model is less certain about the root, but more certain that all sentences should be children of sentence 7... 25

4.5 Example of attention scores, derived dependency tree and text for a test sentence in the Writing Quality dataset (gold label=very good). Root sentences are underlined. In this example, the beginning of the document is likely to be the root.......... 26

4.6 Average tree height and percentage of nodes at each depth of the tree for the derived dependency trees. For the GUM corpus, the dependency trees are derived from the gold-labeled RST trees.......... 27

4.7 Top 10 words most associated with the root sentence (measured with PPMI). 28

4.8 Accuracy on sentiment analysis using the baseline Liu and Lapata (2018) and its variants. 29

4.9 Average tree height and percentage of nodes at each depth of the tree for the baseline model and its variants. 30

Chapter 1
Introduction

Discourse analysis is a concept that encompasses all other levels of Linguistics including syntax and semantics. It lives at all different levels of a text, from word to clause, sentence, paragraph, document, and in between. Despite these challenges, robust theories on discourse have been formulated and implemented for application to NLP tasks. While discourse influences most NLP tasks, it is less clear to what degree. Automated discourse parsers have been developed in an attempt to answer these questions, as well as computation models that learn discourse representations with much fewer constraints. In this report, we examine how well these representations for discourse, be it RST or learned, perform across different domains and tasks.

We first investigate RST segmentation, the first step in an RST-style discourse analysis, and how well it performs on out-of-domain documents. Errors on this data point at domain differences in the linguistic layers underpinning discourse, including syntax. Other errors highlight the ambiguity of discourse.

In our second study, we analyze a model that learns latent “discourse” structures, i.e. representations of documents using structured attention that only assumes these must be dependency trees. We find that these structures are not capturing discourse and instead focus on lexical cues. While our proposed variant is able to induce deeper structures, we still make no claims this is discourse.

1.1 Report Outline

The remainder of this report is structured as follows:

In Chapter 2, we discuss Rhetorical Structure Theory (RST) and structured prediction.

In Chapter 3, we present our first study on discourse segmentation for medical data. Our research question is to understand the magnitude and nature of segmentations errors in the medical domain. We present a quantitative and qualitative analysis showing the gap to be substantial, rooted in domain differences in the underlying linguistic levels and in ambiguities that arose when operationalizing RST.

In Chapter 4, we present our second study analyzing the latent structures learned by a structured attention that parses dependency trees. Our research questions are to understand what the model is learning, whether it is discourse, and whether we can induce more discourse-like structures. We find the model is learning lexical cues, but not discourse. Further, our proposed model induces deeper structures but are still very unlike RST trees.

I conclude my report in Chapter 5 and propose future avenues of research.

1.2 List of Report Contributions

In this report, I make the following contributions:

- We introduce a first, small-scale corpus of medical articles annotated with RST discourse segments.

- We present a quantitative and qualitative analysis of errors in the medical domain that provide insight into the shortcomings of discourse parsing.

- We perform a comprehensive analysis of the document-level latent structures in Liu and Lapata (2018), presenting ample evidence to refute the claim these are discourse. We instead show the model is learning mostly at the lexical level.

- We propose a (variant) model that induces deeper structures on the task of sentiment analysis, highlighting how design decisions of the neural network affect the learned representations.

Chapter 2
Background

2.1 Chapter Overview

In this chapter, we review some of the background critical to this report.

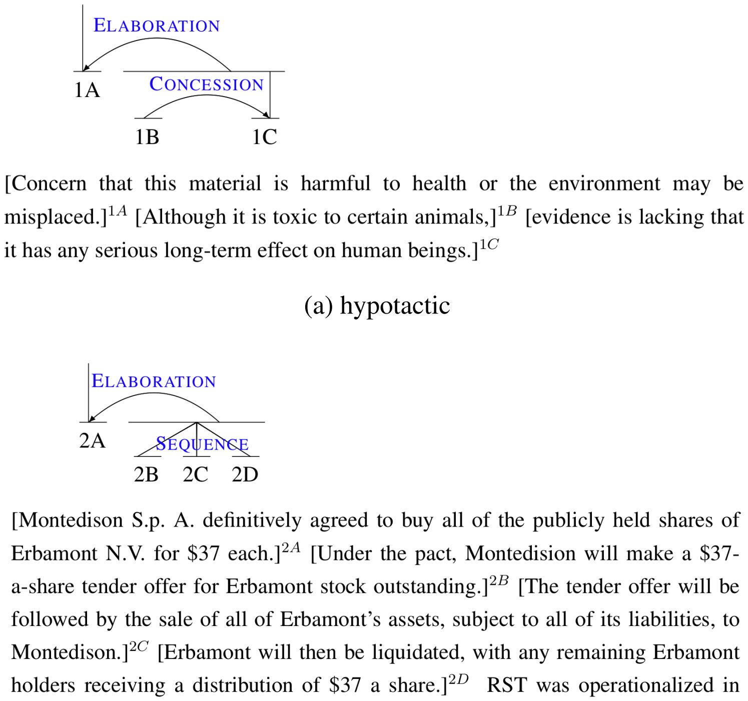
2.2 Rhetorical Structure Theory (RST)

The underlying principal of RST is that coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations (Mann and Thompson, 1988). Thus, the goal of RST is to describe the rhetorical organization of a text by using a hierarchical structure that captures the communicative intent of the writer. The first step in RST is to divide the text into elementary discourse units (EDUs), which generally correspond to clauses [1]. Two adjacent EDUs are related to each other by a discourse relation. This relation is characterized as either paratactic or hypotactic. In the more common hypotactic relation, typically identified with subordination, the EDU that is more central to the text's purpose is labelled as the *nucleus*, and the other (usually subordinating) EDU as the *satellite*. In the paratactic relation, typically identified with coordination, all EDUs are labelled as *nucleus*. These relations are then incrementally grouped together with other relations until forming a tree that spans the entire document. Figure 2.1 illustrates examples of hypotactic relations on the left, and paratactic relations on the right.

RST was operationalized in Carlson et al. (2001) with the RST Discourse Treebank (RST-DT).[2] This corpus consists of 385 Wall Street Journal articles from the Penn Treebank. Further research then proposed a set of rules to convert an RST tree into a dependency tree, where generally heads correspond to nuclear EDUs

[1] Clauses that are subjects, objects, or complements of a main verb are not treated as EDUs.

[2] 1https://catalog.ldc.upenn.edu/LDC2002T07



[Concern that this material is harmful to health or the environment may be misplaced.]^{1A} [Although it is toxic to certain animals,]^{1B} [evidence is lacking that it has any serious long-term effect on human beings.]^{1C}

(a) hypotactic

[Montedison S.p. A. definitively agreed to buy all of the publicly held shares of Erbamont N.V. for $37 each.]^{2A} [Under the pact, Montedision will make a $37-a-share tender offer for Erbamont stock outstanding.]^{2B} [The tender offer will be followed by the sale of all of Erbamont's assets, subject to all of its liabilities, to Montedison.]^{2C} [Erbamont will then be liquidated, with any remaining Erbamont holders receiving a distribution of $37 a share.]^{2D} RST was operationalized in

(b) paratactic

Figure 2.1: RST trees illustrating hypotactic and paratactic syntactic relations.

(Hirao et al., 2013).

With dependency trees being an appealing structure for computational models and the availability of the annotated corpus, RST has been widely adopted by the research community. In fact, RST has been shown to help with NLP tasks ranging from sentiment analysis to summarization, spanning domains such as online reviews and medical papers (Ji and Smith (2017), Li et al. (2016), Da Cunha et al. (2007), *interalia*). However, this same widespread use allowed the community to analyze its shortcomings, both in the theory and how it was operationalized. In our first study on discourse segmentation in the medical domain, we find evidence of problem areas for the way RST was operationalized.

2.3 Structured Prediction and Attention

Statistical models can learn very well to detect patterns in large amounts of data. However, we also need to incorporate biases such as linguistic theories that help the model learn a task better. In this setting, the model predicts structured objects such as parse trees instead of scalars.

With the shift to deep learning, research has focused more on models that learn latent representations of sentences or documents, constrained only by the end task. In our second study, we analyze one such model and attempt to incorporate some level of discourse theory.

2.4 Chapter Summary

In this chapter, we described the two major concepts underpinning this report: RST and structured prediction.

Chapter 3
Cross-Domain Discourse Segmentation

This chapter describes a short study on differences in automated RST discourse segmentation across two different domains (news and medical) and using two different RST segmenters. The work in this chapter was done in collaboration with Titan Page.

Code and data for all experiments in this chapter are available here: `https://github.com/elisaF/cross_domain_segmentation`.

3.1 Chapter Overview

Dividing a text into units is the first step in analyzing a discourse. This first critical step is often overlooked by researchers, despite several studies showing this to be a significant source of error that propagates to the downstream tasks of span, nuclearity and relation labeling (Feng, 2015). Furthermore, segmenting in a non-news domain, such as medical, has not been well studied and could present additional challenges.

In this study, we compare segmentation errors in the news versus medical domain and aim to understand the differences. Because segmentation has not been studied in the medical domain, we annotate a novel small-scale corpus. We use two different RST parsers to further understand whether the quality of the parser itself affects the error differences. Not surprisingly, the news domain outperforms the medical domain when using both segmenters. When using the higher-quality RST segmenter, we find it outperforms the poorer segmenter by a considerable margin in the news domain, but by a much smaller amount in the medical domain. Thus, we conclude that improving an RST segmenter in the news domain provides little benefit to medical. A qualitative analysis finds that the higher error rate in the medical domain is often due to idiosyncrasies found in both the news domain that the segmenter has learned, and in the medical domain that the segmenter has never seen before.

Our contributions in this work are two-fold: a small-scale corpus of medical documents annotated with RST-style discourse; a quantitative and qualitative analysis of the discourse segmentation errors in the medical domain that lay the groundwork for understanding both the limits of existing RST segmenters and problematic areas stemming from the way the RST theory was first operationalized.

3.2 Motivation

On the one hand, the segmentation task appears to be well-defined, as suggested by a high inter-annotator agreement (kappa=0.92), although this number is based on only 9 documents[1] (Carlson et al., 2001). The best discourse segmentation results are quite high at 92.6 F1 (Feng, 2015). Perhaps because of these strong results, segmentation is often dismissed as a solved task. Many RST parsers evaluate only on gold EDUs and do not include a segmenter. Nevertheless, previous studies (Soricut and Marcu, 2003; Fisher and Roark, 2007; Joty et al., 2015; Feng, 2015) identify this first step as a primary bottleneck for accurate discourse parsing. Critically, even using the best-performing segmenter degrades results by 10% on the downstream tasks of span, nuclearity and relation labeling when using predicted instead of gold EDUs (Feng, 2015).

3.3 Related Work

The challenge of evaluating a model on a domain different from where it was trained is a well-studied area of research. In RST, much work already recognizes differences when moving away from the news domain. Rimrott (2007) finds that certain relation types never occur in a small-scale analysis of scientific texts. Bachand et al. (2014) notes systematic differences in the distribution of discourse relations between news and online reviews.

[1] Although 53 documents were doubly-annotated in the RST-DT corpus, all except 9 were *pre-segmented* before being given to the annotators.

Despite these differences, most research using RST makes no attempt to modify the news-trained parsers (or segmenters), even though they are evaluating on considerably different domains. For example, the text classification model in Ji and Smith (2017) evaluates on online reviews using a news-trained RST discourse parser. We performed a small-scale study on Yelp reviews using their news-trained segmenter, and found the segmenter performance very poor. The EDU boundaries often occurred in the middle of a clause, and the EDUs themselves were often nonsensical: 2.2% of the EDUs were a punctuation mark, 6.4% were 1-word EDUs (such as 'I', 'like', 'really'), and 7.9% were 2-word EDUs (such as 'pleasantly surprised', 'completely unethical'). In fact, when corresponding with the paper authors, it was found that they did not use their own segmenter because of these issues.

Some studies do make an attempt to accommodate the different domain. For example, Da Cunha et al. (2007) used RST to summarize medical articles, but only used a subset of the relation types that were observed to occur in that domain. However, in both cases, differences in discourse *segmentation* are largely ignored.

The only study we are aware of that specifically focuses on discourse segmentation in different domains is that of Braud et al. (2017). The overall aim, though, is slightly different in that their goal is to design a segmenter for under-resourced languages by relying only on part-of-speech tags. The proposed neural network model that utilizes a multi-task learning setting to train on different RST corpora is able to outperform simple baselines. However, the study makes the simplifying assumption that each RST corpus represents a single domain, which is not always accurate (the GUM corpus(Zeldes, 2017) spans 4 different genres– interviews, news, travel guides, how-tos).

3.4 Experiment

We automatically segment the 11 Wall Street Journal articles and 11 sections of medical reports using the DPLP segmenter (Ji and Eisenstein, 2014) and the

Corpus	#documents	#tokens	#sentences	#EDUs
RST-DT SMALL	11	4031	166	403
MEDICAL	11	3324	166	385

Table 3.1: Corpus statistics.

segmenter proposed in Feng and Hirst (2014), which we refer to as FENG.[2] We evaluate the segmenter's ability to detect EDU boundaries present in the gold data (RST-DT and the novel medical corpus) using the metrics of precision, recall and F1.

3.4.1 The Corpora

The corpora statistics are summarized in Table 3.1. The Medical corpus consists of 2 clinical trial reports from PubMed Central which were divided into their sections. Each section was manually annotated, resulting in 11 labeled documents. The News corpus was created by sampling the same number of Wall Street Journal articles from the "Test" portion of the RST-DT, that were similar in length to the medical documents.

3.4.2 Experimental Setup

For the medical documents, XML formatting was stripped, and figures and tables were removed. The sections for *Acknowledgements*, *Competing Interests*, and *Pre-publication History* were not included. For the news documents, no preprocessing was performed.

The RST parsers, both of which employ the Stanford Core NLP pipeline (Manning et al., 2014) for preprocessing, were updated to use the same version of this software (2017-06-09). For the Feng parser, we enabled the option to perform a second pass of EDU segmentation using global features.

[2] We choose these two segmenters as they are the most widely-used and publicly available (most RST parsers do not include a segmenter).

RST SEGMENTER	DOMAIN	F1	P	R
DPLP	*News*	82.56	81.75	83.37
	Medical	75.20	77.26	73.25
FENG	*News*	**95.72**	**97.19**	**94.29**
	Medical	78.95	80.00	77.92

Table 3.2: Performance of discourse segmentation using two different discourse segmenters and evaluated on two different domains.

3.5 Results

Table 3.2 lists our results. As expected, the *News* domain outperforms the *Medical* domain, regardless of which segmenter is used. In the case of the DPLP segmenter, the gap between the two domains is about 7.4 F1 points. Note that the performance of this segmenter on *News* lags considerably behind the state of the art (-10 F1 points). When switching to the FENG segmenter, the performance on *News* increases dramatically (+13 F1 points). However, the performance on *Medical* increases by a mere 3.75 F1 points. Thus, large gains in *News* translate into only a small gain in *Medical*, underscoring the need to handle out-of-domain data.

3.5.1 Error Analysis

We perform an error analysis to understand the types of errors present in both domains, and contrast with those seen only in the *Medical* domain. We analyze errors only in the better-performing FENG segmenter.

Across both domains, the segmenter has trouble with infinitival clauses which are generally not treated as separate EDUs but the guidelines contain several nuanced exceptions. In the following example, the first infinitival clause is not separated because it is a complement of the verb, while the second is segmented as a modifier to the preceding noun phrase ‘its attempt’:

> [U.S. Memories is seeking major investors ***to** back its attempt*][***to** crack the $10 billion market . . .*]

ERROR TYPE	PREDICTED	GOLD
‘in’ modifier	[compare the safety of hydrochloride *in* patients][diagnosed with]	[compare the safety of hydrochloride][*in* patients][diagnosed with]
past verb	[Ten patients][*dropped* out of]	[Ten patients *dropped* out of]
title	[*Conclusion* The offer of a prize draw incentive]	[*Conclusion*][The offer of a prize draw incentive]
dash	[(95 % CI 0.96][– 1.13)]	[(95 % CI 0.96 – 1.13)]
citation*	[Studies have shown either small increase in response][{*6*} or more rapid]	[Studies have shown either small increase in response {*6*}][or more rapid]

Table 3.3: Types of segmentation errors in the *Medical* domain with examples of predicted and gold EDU boundaries marked in square brackets (*the square brackets for the citation were changed to curly brackets to avoid confusion with EDU boundaries).

Interestingly, Braud et al. (2017) also found this to be a large source of errors for their segmentations.

In the *Medical* domain, Table 3.3 summarizes and gives example of the most frequent errors. The first group of errors can be attributed to differences in syntactical constructions. The segmenter has trouble identifying embedded EDUs, in particular modifiers starting with ‘in’ describing the circumstance or background. Although ‘in’ phrases occur often in *News*, they usually refer to a location or time (e.g., ‘offices *in* New Orleans’). For many clauses with past tense verbs, an EDU boundary is surprisingly inserted between the subject and verb. The segmenter is likely confusing these cases for participial phrases that modify the noun, which would be treated as embedded EDUs (e.g., ‘[the charge][*related* to the action]’). Because this analysis partly relies on syntax, a source of this error could be traced back to the Stanford Core NLP parser, which is also trained on news articles.

The next group of errors relate more to differences in formatting and markup of the two domains. The section titles in the *Medical* domain are invariably never detected by the segmenter. Although *News* contains headers, they are always separated by a colon. Punctuation is a strong and easy signal for EDU boundaries. Although even punctuation can be ambiguous and conflicting across domains. While the em dash (‘–’) is typically used in the same context as parentheses or commas, we find cases in the *Medical* domain where it is used as a hyphen or en dash (‘-’), for example to specify a numerical range. Finally, the last error is one that would be common in many scientific domains. Citations do not occur in *News*, and the square brackets around citations used by PubMed Central articles are confusing to the segmenter.

3.6 Chapter Summary

As a first step in understanding discourse differences between domains, we analyze the performance of two discourse segmenters on *News* and *Medical*. For this purpose, we create a first, small-scale corpus of annotated medical documents. We find that both discourse segmenters perform better on *News*, as expected. However, we also find that large improvements in the *News* domain gained by using a better segmenter do not translate into substantial improvements in the *Medical* domain. An error analysis reveals difficulty in both domains for cases requiring a fine-grained syntactic analysis, as dictated by the RST-DT annotation guidelines. This finding suggests a need for either a clearer distinction in the guidelines, or more training examples for a model to learn to distinguish them. In the *Medical* domain, we find that differences in syntactic construction and formatting, including use of punctuation, account for most of the segmentation errors. We hypothesize syntactic errors can be partly traced back to syntactic parsers used in the preprocessing steps that are also trained on news.

Addressing even one of these issues may yield a multiplied effect on segmentation improvements as this domain is by nature highly repetitive and formulaic. However, a future avenue of research would be to first understand what impact these segmentation errors have on downstream tasks. On the one hand, using RST trees generated by the lower-performing DPLP parser nevertheless provides small gains to text categorization tasks such as sentiment analysis (Ji and Smith, 2017). On the other hand, understanding the verb form, which proved to be difficult in the *Medical* domain, has been shown to be useful in distinguishing text on experimental results from text describing more abstract concepts (such as background and introductory information) (de Waard and Maat, 2012).

Chapter 4
Latent Discourse Structures

This chapter focuses on latent discourse structures that are learned by a neural network, in contrast to the first chapter that examined explicit structures as dictated by a linguistic theory (RST).

Code and data for all experiments in this chapter are available here: `https://github.com/elisaF/structured`.

4.1 Chapter Overview

Neural networks have seen considerable success in learning latent representations for linguistic constructs ranging from syntax to discourse (Choi et al., 2017; Yogatama et al., 2017; Liu and Lapata, 2018). However, more recent work calls into question what the models are actually learning, finding they are not consistent with any linguistic theory (Williams et al., 2017). At the same time, these questions are spurred by the research community's desire to understand neural networks at a deeper level, instead of treating them as black boxes.

We find the Liu and Lapata (2018) model to be an excellent starting point for addressing these questions in regards to discourse and latent representations of documents. This model employs structured attention that is parsed into a non-projective dependency tree. Analyzing the attention weights and resulting tree allows us a window into the network. We utilize this model to perform an in-depth analysis to understand what the model is learning and whether the structured attention can be interpreted as discourse.

Our analysis and experiments on different inputs to the model show that the structured attention is mostly attending to a single sentence, the root of the dependency tree, which contains lexical cues useful for the task. The model is not learning discourse and the derived trees do not resemble those of any discourse theory. We propose a variant of the model that induces deeper latent structures.

Our contributions in this work are two-fold: we present a comprehensive analysis of the Liu and Lapata (2018) model that reveals what the model is learning and refute the claim that these latent structures represent discourse. Secondly, we propose a variant of the model that is able to learn deeper structures.

4.2 Related Work

We first describe the Liu and Lapata (2018) model, which we analyze and propose a variant of. We next discuss the growing area of research into understanding representations induced by neural networks.

Structured Attention Attention has gained widespread adoption as a way to selectively focus on parts of the input, which is particularly useful in models such as recurrent neural networks that do not handle long inputs well (Luong et al., 2015). Structured attention is then introduced as a way to learn unsupervised latent representations. Kim et al. (2017) use graphical models to induce hidden representations that are able to outperform models using simple attention on a variety of tasks. Importantly, they find the attention distributions of the learned structures to be interpretable and intuitive to the given task. In a similar vein, Liu and Lapata (2018) learns a structured attention that can be parsed as a non-projective dependency tree. We describe this model in more detail as we choose to analyze this model in our study.

The Liu and Lapata (2018) model consists of both sentence-level and document-level attention. The underlying attention is the same, but operates on different input. At a high level, the sentence-level model composes words into a single representation of a sentence. These sentence representations are then composed into a single representation of a document in the document-level model. In the sentence-level model, a bidirectional LSTM is run over a sentence, and these hidden representations $[h_1, h_2, \ldots, h_n]$ are used as the representations for the words in the sentence. They further decompose this output into a semantic part (e_t) and a structure part (d_t):

$$[e_t, d_t] = h_t \tag{4.1}$$

They then use the Matrix Tree algorithm to calculate the marginal probabilities, or *attention scores*, a_{ij} using the structure vectors $\boldsymbol{d}$. The attention is parsed into probabilities over a non-projective dependency tree. The semantic part of a word is then updated with attention by first calculating a context for all the possible parents of that word as follows:

$$p_i = \sum_{k=1}^{n} a_{ki} e_k + a_i^r e_{root} \tag{4.2}$$

where a_{ki} is the probability that k is the parent of i, e_k is the semantic vector of the parent, e_{root} is a special embedding for the root node, and a_i^r is the attention score for the i-th word being the root.

The parent vectors are then concatenated to the semantic vectors and passed through a non-linear function, resulting in the *updated* semantic vector r_i. While Liu and Lapata (2018) describes additionally concatenating a vector for all possible children c_i as illustrated in 4.3, their code only makes use of the parent vectors as in 4.4.

$$*r_i = tanh(\boldsymbol{W_r}[e_i, c_i, p_i])* \tag{4.3}$$

$$r_i = tanh(\boldsymbol{W_r}[e_i, p_i]) \tag{4.4}$$

where $\boldsymbol{W_r}$ are weights learned by the network.

Finally, a max pooling layer produces the final sentence representation v_i. Note that this pooling strategy chooses just *one* representation of a word (or *one* sentence for the document-level model) . This could prove detrimental to capturing a representation of the entire sentence (or entire document). The document-level model employs the same attention mechanism, but takes as input the outputs of the sentence-level model. That is, a biLSTM is run over the sentence representations $[v_i, \ldots, v_n]$ to yield a document representation q_i. At test time, the Chu-Liu-Edmonds algorithm is used to derive the maximum spanning tree over these attention scores.[1]

[1] We use the implementation in `https://github.com/atofigh/edmonds-alg.git`.

Interpreting latent structures Upon the heels of a surge in neural networks and attention, a large part of the community is now striving to understand what these structures represent. For example, Williams et al. (2017) examines latent parse trees learned by two neural network models. They find that while the models perform very well, the learned trees are shallower than their explicitly parsed counterparts (PTB parses) and do not resemble any linguistic theory. Our study finds similar results when interpreting document-level trees and comparing them to discourse parses.

Several papers have also proposed toy datasets of mathematical operations to understand how models are composing functions and gage their ability to produce a plausible tree (Hupkes et al., 2017; Nangia and Bowman, 2018). Hupkes et al. (2017) in particular propose 'diagnostic classifiers' to test hypotheses on what the hidden representations are encoding. While we find this research worthwhile and engaging, we feel that in order to meet our goal of understanding discourse and its latent representations, we need real-world data.

4.3 Models

In this chapter, we first explain our choice of model, and then describe our proposed variants of the Liu and Lapata (2018) model.

4.3.1 Model Choice: Latent or Explicit?

Models such as Liu and Lapata (2018) attempt to learn a *latent* structure for a document, while models such as Ji and Smith (2017) dictate an *explicit* structure, namely RST discourse dependency trees. While it would be desirable to compare and contrast both models, we were not able to replicate the performance of Ji and Smith (2017). Despite using their publicly available code on the same dataset (Yelp 2015), our accuracy on their *full* model (that uses RST trees) never exceeded 65.1, a considerable gap from the reported 71.8. Upon inspection of the parsed discourse trees, we noted very poor EDU segmentation. After corresponding with the authors, we learned they in fact did not use their own discourse segmenter, as reported in the paper. We thus attempted to reproduce their results first using sentences (parsed by Stanford Core NLP (Manning et al., 2014)) and then using the best-performing RST segmenter (Feng and Hirst, 2014). In both cases, the results were equally low. Suspecting that bad RST trees were the culprit, we tried to reproduce the results of their baseline *additive* model, which effectively runs a bidirectional LSTM over just the EDUs, without taking into account the tree structure. Accuracy approached 64.0, still a far cry from the reported 68.5. Finally, because tuning the model was very slow on dyNet (Neubig et al., 2017), we reimplemented the *additive* baseline model in PyTorch (Paszke et al., 2017), and the results were still the same. Because of the consistently negative results, we decided not to pursue this model and instead focus on latent structures by exploring the model in Liu and Lapata (2018).

4.3.2 Model Variants

We propose several variants of the Liu and Lapata (2018) model first by modifying its inputs, and finally by modifying its architecture.

baseline The baseline model uses the publicly available code for Liu and Lapata (2018) with both sentence- and document-level attention.

root-only The aim of this model is to understand how much the original model relies only on root sentences in the learned dependency tree. For all documents in the corpus, we remove all sentences that are not attached to the root. We then create new embeddings and retrain the **baseline** model on the pruned corpus.

edu This model attempts to inject explicit structure into the network. Specifically, we feed in EDUs instead of sentences into the **baseline** model.[2] We hypothesize this modification will encourage deeper structures as composition might be required to reconcile conflicting sentiments that exist at a finer-grained level than sentences as noted in Socher et al. (2013) (e.g., '[The potato and mushrooms were good,][but the green beans were way to "Al Dente"]'). For segmenting the text, we use Feng and Hirst (2014) with global features, which obtains the best segmentation on RST-DT.

[2] We do not take the route of modifying the word inputs into EDUs, as RST structures do not always result in trees that can be isolated into a single sentence. That is, it is possible for only one segment of a sentence to attach to another sentence (van der Vliet and Redeker, 2011)

deeper This model attempts to encourage deeper tree structures by making three changes to the original model: incorporating the updated semantic vectors from *subtrees* into the final tree representation, removing the document-level biLSTM, and using a different function at the pooling layer.

First, recall in 4.3 that the semantic vector r_i is updated with the probabilities of possible children c_i. However, this update only takes into account direct children, not the integrated information from a whole *subtree*. To further percolate information from subtrees, we propose an additional update to the semantic vector. Starting from the original updated semantic vector:

$$r_i = tanh(\mathbf{W}_r[e_i, c_i]) \tag{4.5}$$

we then propose a new representation for the children that incorporates the above semantic vector r_i:

$$c_i^{(2)} = \sum_{k=1}^{n} a_{ik} r_i \tag{4.6}$$

This new child representation is then incorporated into a new semantic vector:

$$r_i^{(2)} = tanh(\mathbf{W}_r[e_i, c_i^{(2)}]) \tag{4.7}$$

The second modification is to remove the document-level bidirectional LSTM. We believe the learned structures are shallow because context is already incorporated by the biLSTM without any need for deeper structures. As we discuss in section 4.5, this hypothesis is further confirmed by results on our earlier proposed model *root-only*.

Finally, we believe the max-pooling layer encourages the model to just choose one sentence, ignoring the learned structure. Instead, we propose a sum that is weighted by the probability of a given sentence being the root, that is using the learned attention score a_i^r:

$$q_i = \sum_{i=1}^{n} a_i^r r_i^{(2)} \tag{4.8}$$

		Number of documents				
Dataset	Classes	Total	Train	Dev	Test	Vocab.
Yelp 2017	5	1.1M	914,525	101,615	100K	131K
Congressional votes	2	1.7K	1,175	113	411	7.4K
Writing quality	2	8.5K	6,806	850	850	48.6K

Table 4.1: Statistics for the datasets used in the text classification tasks. The vocabulary size is calculated after preprocessing.

4.4 Experiments

We use the model with both sentence- and document-level structured attention as described in Liu and Lapata (2018) on a subset[3] of their same document classification tasks: sentiment analysis and Congressional vote prediction. We additionally evaluate on the task of identifying high-quality writing.

4.4.1 The Corpora

Table 4.1 summarizes the statistics for the three datasets we use on the document classification tasks. We describe each in more detail below.

Yelp 2017 We construct a dataset of 1.1M Yelp reviews from the Yelp 2017 Dataset Challenge, consisting of business reviews with a rating on a scale of 1 (negative) to 5 (positive) that we use as labels for our sentiment classification task. The data is split into roughly 80/10/10 for training, development and test. The reviews are balanced by rating in each partition. The review text is split into sentences using the Stanford Core NLP.

Congressional Votes This corpus consists of speech segments from Congressional floor debates that are labeled with how the speaker voted on a given bill ('yea' or 'nay'). We use the same dataset as Liu and Lapata (2018), which is the preprocessed version from Yogatama and Smith (2014).

Writing Quality This corpus was created by Louis and Nenkova (2013) and consists

[3] We do not evaluate on IMDB reviews or Czech reviews because the task of sentiment analysis is already analyzed in Yelp, and we are not focused on the non-projectivity of trees.

of science articles from the New York Times that are labeled as either ‘very good’ or ‘typical’. The ‘very good’ class was created by using as a seed the 63 articles in the New York Times corpus (Sandhaus, 2008) deemed to be high-quality writing by a team of expert journalists. The class was then expanded by adding all other science articles in the NYT corpus that were written by the seed authors (4,253 articles). For the ‘typical’ class, science articles by all other authors were included (19,520). Because the data is very imbalanced, we undersample the ‘typical’ class to be the same size as the ‘very good’. We split this data into 80/10/10 for training, development and test, with both classes equally represented in each partition. Sentences are split using Stanford Core NLP.

4.4.2 Experimental Setup

We follow the same preprocessing steps as in Liu and Lapata (2018) except for the *Writing Quality* dataset, we change the default maximum sentence length from 60 to 100 so that we discard 9% instead of 35% of the data. For training, we follow the same procedure with two modifications. First, we use different dimensions for the semantic and structure vectors which were found to perform slightly better (100 and 50, respectively, instead of the reported 75 and 25). Second, we critically do not batch at test time in order to accurately predict the maximum spanning tree. The code does not account for masking and therefore would not create a correctly sized tree for the shorter documents in a batch (during training, this problem is mitigated by batching with size 32 and grouping sentences of similar length into the same batch).

4.5 What is learned, and is it discourse?

In this part of our study, we aim to understand what the **baseline** model is learning about a given task by analyzing the latent structures, and whether these can be interpreted as discourse. We evaluate on the three described text classification tasks. We first analyze the attention scores to understand the distribution of probabilities for children and parents in the dependency tree (e.g., does the model

Model	Yelp 2013	Yelp 2017	Congressional votes	Writing quality
baseline	68.1	71.24	77.27	83.46

Table 4.2: Accuracy on the three classification tasks using the baseline Liu and Lapata (2018).

definitively choose one child, or are probabilities spread out across all children?). We then derive the dependency tree from the learned attention scores (using the Chiu-Liu-Edmonds algorithm) that we liken to an RST discourse dependency tree. Next, we calculate statistics on these trees and perform a qualitative analysis to conclude these structures in fact do not represent discourse. Instead, the model seems to focus mainly on lexical cues.

Table 4.2 summarizes the accuracy results on the **baseline** model. We first evaluate on Yelp 2013 (Tang et al., 2015) to confirm we are able to reproduce, within reason, the performance of Liu and Lapata (2018) (they report an accuracy of 68.6). While there is no prior work on Yelp 2017, the performance is consistent with previous year datasets, though slightly higher as can be expected with a larger dataset. Our performance on the Congressional Votes marginally exceeds that reported in (Liu and Lapata, 2018). For the Writing Quality, while we cannot directly compare our numbers with those reported in Louis and Nenkova (2013),[4] we confirm our performance is very similar.

Now that we have ensured the models are indeed learning the task well, we can confidently claim the latent structures paint an accurate picture of how a document is represented. We first examine the attention weights to understand how certain the model is about this representation. Recall the structured attention is parsed as a dependency tree. The attention weights are thus a square $n \times n$ matrix where n is the number of sentences, and an extra column is added at the beginning for the root symbol. For a given cell a_{ij}, its value represents the probability that i is a child of j. For example, a_{11} gives the probability of the first sentence attaching to the root.

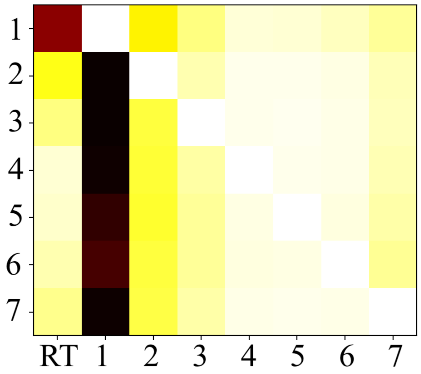
[4] They used a different sampling technique and performed cross-fold validation that would not work in this setting.

Attention Scores Here we describe patterns observed for the attention scores on each of the tasks. Tables 4.3, 4.4 and 4.5 illustrate the attention scores as a heat map for a test sentence. The sentences were chosen to be representative of patterns seen in each dataset.

For Yelp, the model often chooses a sentence with strong sentiment-bearing words as the root, and then the remaining sentences attach to that sentence. The probabilities are peaked, indicating the model is considerably certain about the structure. The trees decidedly do not represent discourse, and the model instead seems to focus on lexical cues.

In the case of Congressional Votes, the structures were similarly shallow, choosing the root to be the sentence where a speaker declares their position for or against the bill. However, the probabilities were more distributed across different sentences. This could be due to the nature of the data, or more likely due to the much smaller size. In this case as well, we do not deem these structures to be representing discourse and instead the model again seems to be learning lexical items.

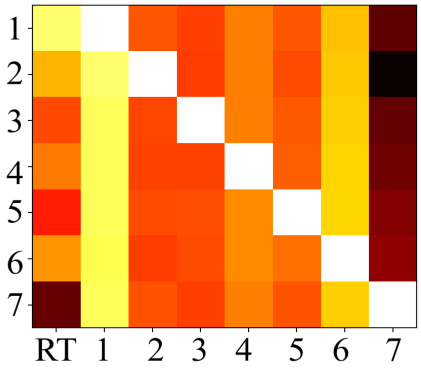
For Writing Quality, the structures are considerably different. Multiple root sentences are usually chosen with often a deeper tree. We do not constrain the tree to have a single head, since multi-headed RST discourse dependency trees are possible (although not as common). However, the multiple roots found in these structures do not correspond to multi-nuclear relations. Interestingly, the probabilities for the root are concentrated on sentences at the beginning of the document, although the model has no notion of order. It is less clear whether this structure could be construed as discourse, although it is decidedly not similar to RST.



Dependency tree: RT: [1], 1: [2, 3, 4, 5, 6, 7]

1. I have to say that I am very happy with my recent visit to Just Brakes
2. We needed our rear brakes replaced and I forgot that we used Just Brakes for our last repair and part of our service was still under warranty
3. The mechanic explained what we needed done and showed us the worn parts
4. My wife asked about an additional repair another shop said we needed , which they quoted at over 1400
5. Our Just Brakes mechanic said we did not need the repair
6. So I assume the other shop is dishonest or incompetent
7. Either way , thank you Just Brakes , you have earned a loyal customer

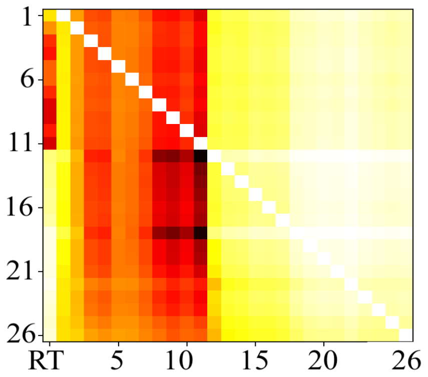
Table 4.3: Example of attention scores, derived dependency tree and text for a test sentence in Yelp 2017 (gold label=5). Root sentences are underlined. In this example the very light and very dark shadings suggest the model is very certain about the structure.



Dependency tree: RT: [7], 7: [1, 2, 3, 4, 5, 6]

1. madam speaker , i thank the gentleman from the great city of UNK , massachusetts , for yielding
2. madam speaker , i rise in opposition to the conference report on h r UNK , the so called usa patriot act , because we have not taken meaningful steps to eliminate or correct the most egregious sections of this act
3. in particular , it is UNK that the conference agreement does not include a meaningful judicial review mechanism for UNK UNK , under the foreign intelligence UNK act , as applied against u s citizens
4. given that the power that today 's UNK technology gives to government and given the broad powers that we have given to intelligence agencies under this act , the absence of post execution judicial review in today 's conference report constitutes one of its most critical shortcomings
5. madam speaker , in order to ensure that the powers granted by the patriot act are not susceptible to abuse , our government must always operate with meaningful oversight , checks and balances
6. after all , it is the maximum transparency and active judicial review which is our ultimate weapon in UNK both governmental abuse and overreaching by governments to restrict the individual freedoms of our citizen
7. <u>for these reasons , i ask my colleagues to oppose the this version of the patriot act reauthorization</u>

Table 4.4: Example of attention scores, derived dependency tree and text for a test sentence in the Congressional Votes dataset (gold label=N). Root sentences are underlined. In this example, the mixed shading in the first column suggests the model is less certain about the root, but more certain that all sentences should be children of sentence 7.



Dependency tree: RT: [8, 9, 11], 9: [26], 11: [1, 2, 3, 4, 5, 6, 7, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]

1. One week after an outbreak of Legionnaires ' disease at a Manhattan hospital was blamed for the death of a patient , a second medical center , Harlem Hospital Center , a few miles away , announced yesterday that it might have identified traces of the bacteria in one of its buildings
2. The bacteria have now been detected in as many as three city hospital buildings in two months , highlighting what experts said was a barely visible but potentially deadly problem plaguing hospitals in the city

. . .

7. Two of those patients later died , and an autopsy on one showed last week that Legionnaires ' disease had been the cause
8. On the same day that NewYork Presbyterian announced the results , the hospital disclosed that the water supply at a second building that it operates , the Greenberg Pavilion , on the Upper East Side , had also tested positive for the bacteria , though no cases have been diagnosed
9. According to the Centers for Disease Control and Prevention , up to 18,000 people in the United States contract Legionnaires ' disease each year , though it is rarely deadly unless those infected have weakened immune systems
10. The bacteria that causes it , which first became widely known in the United States after a large outbreak at a Philadelphia hotel in 1976 , are almost always present in the city 's water supply
11. Hospitals are particularly vulnerable to outbreaks
12. But Dr Victor Yu , the chief of infectious diseases at the Pittsburgh Veterans Administration Health Care center , who is an expert on Legionnaires ' disease , said that few test their water supplies routinely

. . .

18. One expert on Legionnaires ' disease , Dr Joseph S UNK , a professor of clinical medicine and pediatrics at Albert Einstein College of Medicine , said he would not be surprised if other hospitals came forward as well

. . .

26. " If I 'm working here and something is in the water , I have to be careful , " he said

Table 4.5: Example of attention scores, derived dependency tree and text for a test sentence in the Writing Quality dataset (gold label=very good). Root sentences are underlined. In this example, the beginning of the document is likely to be the root.

		Yelp 2017	Congressional Votes	Writing Quality	GUM
Tree height		2.253	2.162	3.162	101.5
Nodes	1	12.58%	14.59%	10.54%	1.57%
	2	74.15%	79.36%	57.4%	<1%
	3	12.51%	5.80%	18.4%	1.62%
	4	<1%	<1%	7.41%	3.21%
	5	<1%	0	3.16%	2.62%
	6	<1%	0	1.5%	1.42%
	7	0	0	1.09%	2.97%
	8	0	0	<1%	2.06%
	9	0	0	<1%	2.64%
	10	0	0	<1%	1.82%
	11	0	0	<1%	1.93%
	12	0	0	<1%	2.25%
	13	0	0	0	1.19%
	…	…	…	…	…
	109	0	0	0	<1%

Table 4.6: Average tree height and percentage of nodes at each depth of the tree for the derived dependency trees. For the GUM corpus, the dependency trees are derived from the gold-labeled RST trees.

Dependency Trees We provide a quantitative analysis in Table 4.6 to support the observations made in the previous section. For reference, we include statistics for dependency trees that are instead derived from RST trees on the GUM corpus (Zeldes, 2017).[5] While RST dependency trees are expected to be deeper in part because they are at the EDU instead of sentence-level, we also note they are not as top-heavy with a more even distribution of nodes at different depths. As surmised earlier, the trees on Yelp are very shallow, consistent with results in Liu and Lapata (2018). However, our shallow trees on the Congressional Votes are not consistent with the deeper trees found by Liu and Lapata (2018). We speculate we may have trained longer (explaining the better accuracy) and the model may have learned bet-

[5] We choose to analyze this corpus (instead of RST-DT) because it encompasses several genres beyond just news (interviews, travel guides, how-to guides, academic writing, biographies, fiction) and serves as a better point of comparison for the three datasets.

ter trees for the task. The only task achieving deeper trees is that of identifying very good writing.

Root Sentences Because the trees are so shallow, we analyze the root sentences in more depth to confirm our suspicions of lexical cues. We calculate the positive pointwise mutual information scores for a root word occurring in the root sentence as compared to all other sentences. Our results are listed in Table 4.7.

Yelp	uuu, sterne, star, rating, deduct, 0, edit, underwhelmed, update, allgemein
Congressional Votes	oppose, republican, majority, thank, gentleman, leadership, california, measure, president, vote
Writing Quality	valley, mp3, firm, capital, universal, venture, silicon, analyst, capitalist, street

Table 4.7: Top 10 words most associated with the root sentence (measured with PPMI).

In the Yelp dataset, words in the root sentence can express sentiment ('underwhelmed'), but also reflect cases where a Yelper explicitly mentions how many stars ('sterne' in German) they are going to give ('0'). The model interestingly learns the idiosyncratic rating system of a particularly prolific Yelper who uses 'uuu' to symbolize a star. While we do find sentiment-bearing words[6] to occur on average more frequently in the root sentence as compared to other sentences, this difference is marginally significant (t-test, p=0.057). This result suggests that more than just sentiment is being learned, as confirmed by the many non-sentiment words found in the word association list.

For the Congressional Votes, the list contains words that explicitly express their opinion ('oppose') but also politeness ('thank') and other terms used for stance-taking.

The list for Writing Quality revolves around tech, suggesting the model may be learning topics instead of actual good or bad writing. The topics were not controlled for and warrants further research into the validity of the dataset.

[6] Sentiment scores were calculated using the model in Hutto and Gilbert (2014), as packaged in NLTK.

Model	Yelp 2017
baseline	71.24
root-only	56.19
edu	70.73
deeper	71.04

Table 4.8: Accuracy on sentiment analysis using the baseline Liu and Lapata (2018) and its variants.

As a final analysis of root sentences, we implement the **root-only** model for the Yelp dataset, where we prune the original dataset to include only the root sentences. We hypothesize the **baseline** model only attends to the root sentences and thus expect this model to perform comparably to the original. However, as illustrated in Table 4.8, the accuracy drops considerably. Although this is a negative result, we understand this is likely because the biLSTM is making use of all the sentences, such that the structured attention does not have to. This observation guides our design decision to remove the biLSTM for the **deeper** model.

In summary, we found the model learns to identify lexical cues that benefit the task, and in the case of Writing Quality, possibly topics. This finding, together with the mostly shallow trees, leads us to conclude the model is not learning discourse.

4.6 Can we inject explicit discourse?

In this section, we attempt to encourage the model to learn structures more similar to RST discourse trees. As a first and straightforward step, we feed in EDUs instead of sentences in the **edu** model. While performance degrades slightly (see Table 4.8), we do not find the resulting dependency trees to be much different, as illustrated in 4.9. The average tree height is 2.365 (compared to 2.253) and nodes of depth 3 are only slightly more numerous (14.33% vs. 12.51%). We again attribute this to the presence of the biLSTM and other factors we attempt to address in the **deeper** model. In future work, we intend to re-evaluate this model *after* making the

		baseline	edu	deeper
Tree height		2.253	2.365	2.803
Nodes	1	12.58%	9.79%	15.67%
	2	74.15%	74.16%	51.9%
	3	12.51%	14.33%	26.25%
	4	<1%	<1%	5.43%
	5	<1%	<1%	<1%
	6	<1%	<1%	<1%
	7	0	<1%	<1%

Table 4.9: Average tree height and percentage of nodes at each depth of the tree for the baseline model and its variants.

changes to the network mentioned in the **deeper** model, to more fairly assess the impact of EDU-level inputs, and additionally evaluate on the other two tasks.

4.7 Can we learn deeper structures?

We have seen that the **baseline** and **edu** models do not learn deep structures and the structured attention seems to attend mainly to the root sentence. Here we attempt to address these shortcomings by implementing the **deeper** model (due to resource constraints, we evaluate only on Yelp 2017). This model removes the document-level biLSTM to place the burden on the structured attention for composing sentences. We additionally take into account information from the subtrees for the structured attention and use a sum weighted by the root scores in the pooling layer. The performance degrades slightly (see Table4.8), but critically the model learns deeper trees, as can be seen in Table 4.9. The average height is 2.803 (compared to 2.253) and there are considerably more nodes of depth 3 and 4.

A possibility we must consider is that these tasks don't require deeper structures. That is, there are sufficient patterns in the words of a sentence that don't necessitate composition over sentences. This is quite possible for sentiment analysis and Congressional debates prediction, as evidenced by the clear patterns we find in the texts. The challenge then is to find a task that is not susceptible to these pitfalls, but is also tractable for an NLP setting.

4.8 Chapter Summary

In this chapter, we performed a detailed analysis of the model in Liu and Lapata (2018), and several proposed variants of it. We were motivated by two questions: what is the model learning, and is it discourse? To answer these questions we analyzed the attention scores, the derived dependency trees and words in the root sentences. This analysis revealed that the model mostly attended to the root sentences, and specifically lexical cues that were helpful to each task. There was no discernible discourse in these structures. This finding prompted our next question of whether we could inject explicit structure. This model also failed to yield discourse-like structures, which prompted our next question. How can we induce deeper structures? We made a series of changes to the model to remove artifacts that were trying to take on duties of the structured attention, such as the biLSTM. We succeeded in showing that we could achieve deeper dependency trees, although we would not yet consider these to represent discourse. We speculate the tasks themselves do not require knowledge of discourse. Our challenge for future work is then to find a new task that makes more use of discourse.

This study opens many avenues of research to further explore latent representations of discourse, combined with explicit structure. In particular, the attention scores could be hard-wired to reflect the structure of the corresponding RST tree. These explicit structures could also be used just for initialization, then allowing the model to learn a derivative structure.

Chapter 5
Conclusion

In this report, we focused on the one hand on discourse as dictated by a specific linguistic theory (RST), and on the other hand on discourse purportedly learned by a model.

Our first study investigated segmentation of RST discourse units in the medical domain. We present a first, small-scale corpus of medical documents annotated with EDUs. Errors in the medical domain provide insights into the differences at lower linguistic levels for this domain, and suggest possible problem areas with the operationalization of RST.

Our second study explored the model of Liu and Lapata (2018) that claims to learn discourse structures by using structured attention that parses into a dependency tree. A careful analysis reveals the structures are not discourse and instead shows the model mostly learns lexical cues. We propose a variant of the model that is able to induce deeper structures on the task of sentiment analysis.

5.1 Future Work

In future work, we would like to understand whether errors in discourse segmentation, which undoubtedly affect the discourse trees, have a detectable effect on a downstream task.

For our analysis of latent structures, we would like to understand whether deeper structures can be induced on the other tasks, as well. More broadly, we would like to explore other tasks such as essay scoring or argumentation mining that purportedly make more use of discourse.

References

Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. An investigation on the influence of genres and textual organisation on the use of discourse relations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 454–468. Springer, 2014.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *SIGDIAL Workshop*, 2001.

Jihun Choi, Kang Min Yoo, and Sang goo Lee. Learning to compose task-specific tree structures. AAAI, 2017.

Iria Da Cunha, Silvia Fernández, Patricia Velázquez Morales, Jorge Vivaldi, Eric SanJuan, and Juan Manuel Torres-Moreno. A new hybrid summarizer based on vector space model, statistical physics and linguistics. In *Mexican International Conference on Artificial Intelligence*, pages 872–882. Springer, 2007.

Anita de Waard and Henk Pander Maat. Verb form indicates discourse segment type in biological research papers: Experimental evidence. *Journal of English for Academic Purposes*, 11(4):357–366, December 2012.

Vanessa Wei Feng and Graeme Hirst. Two-pass Discourse Segmentation with Pairing and Global Features. 2014.

Wei Vanessa Feng. *RST-style discourse parsing and its applications in discourse analysis*. PhD thesis, University of Toronto (Canada), 2015.

Seeger Fisher and Brian Roark. The utility of parse-derived features for automatic discourse segmentation. *ACL*, 2007.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, 2013.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, November 2017.

Clayton J Hutto and Eric Gilbert. VADER - A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *ICWSM*, 2014.

Yangfeng Ji and Jacob Eisenstein. Representation Learning for Text-level Discourse Parsing. *ACL*, 2014.

Yangfeng Ji and Noah A. Smith. 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. Association for Computational Linguistics, 2017.

Shafiq R Joty, Giuseppe Carenini, and Raymond T Ng. CODRA - A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 2015.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured Attention Networks. *ICLR*, 2017.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles, September 2016. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. Learning Structured Text Representations. *TACL*, 2018.

Annie Louis and Ani Nenkova. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *TACL*, 2013.

Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. *EMNLP*, 2015.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

Nikita Nangia and Samuel R Bowman. ListOps: A Diagnostic Dataset for Latent Tree Learning. *arXiv.org*, April 2018.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*, 2017.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

Anne Rimrott. The discourse structure of research articles abstracts–a rhetorical structure theory (RST) analysis. *Proceedings of the 22nd North West Linguistics Conference (NWLC) at Simon Fraser University*, pages 207–220, 2007.

Evan Sandhaus. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 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, 2013.

Radu Soricut and Daniel Marcu. Sentence Level Discourse Parsing using Syntactic and Lexical Information. *HLT-NAACL*, 2003.

Duyu Tang, Bing Qin, and Ting Liu. 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, 2015.

Nynke van der Vliet and Gisela Redeker. Complex sentences as leaky units in discourse parsing. *Proceedings of Constraints in Discourse*, 2011.

Adina Williams, Andrew Drozdov, and Samuel R Bowman. Do latent tree learning models identify meaningful structure in sentences? *arXiv preprint arXiv:1709.01121*, 2017.

Dani Yogatama and Noah A Smith. Linguistic structured sparsity in text categorization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 786–796, 2014.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. Learning to compose words into sentences with reinforcement learning. *ICLR*, 2017.

Amir Zeldes. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.