A Joint Model of Rhetorical Discourse Structure and Summarization

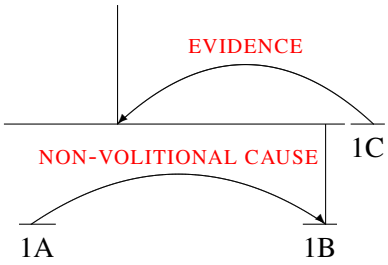
Naman Goyal and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{naman.goyal21 + jacobe}@gmail.com

Abstract

In Rhetorical Structure Theory, discourse units participate in asymmetric relationships, with one element acting as the *nucleus* and the other as the *satellite*. In the resulting tree-like nuclearity structure, the importance of each discourse unit can be measured by the number of relations in which it acts as the nucleus or as the satellite. Existing approaches to automatically parsing such structures suffer from two problems: they employ local inference techniques that do not capture document-level structural regularities, and they rely on annotated training data, which is expensive to obtain at the discourse level. We investigate the SampleRank structure learning algorithm as a potential solution to both problems. SampleRank allows us to incorporate arbitrary document-level features in a global stochastic inference algorithm. Furthermore, it enables the training of a joint model of discourse structure and summarization, which can be learned from document-level summaries alone, without discourse-level supervision. We obtain mixed results in the fully supervised case, and negative results for the joint model of discourse structure and summarization.

1 Introduction

Rhetorical structure theory (RST) is a hierarchical model of document-level organization, in which segments of text are linked in binary or multi-way discourse relations (Mann and Thompson, 1988). Many RST relations are asymmetric, containing a *nucleus* and a *satellite*. An example is shown in Figure 1, with unit 1B as the nucleus of its relationship with 1A, and then the combined unit 1A:B at the nucleus of its relationship with 1C. In any given discourse relation, the nucleus is more likely to relevant to a summary of the document (Marcu, 1999c), and its sentiment is more likely to be relevant to the overall document-level polarity (Heerschop et al., 2011; Bhatia et al., 2015). Thus, recovering this *nuclearity structure* is a key task for discourse parsing, with important practical applications.



[The more people you love,]^{1A} [the weaker you are.]^{1B} [You'll do things for them that you know you shouldn't do.]^{1C}

Figure 1: An example Rhetorical Structure Theory parse of a small segment of text.

All known RST discourse parsers take one of two approximations, which are well known in structure prediction. In dynamic programming approaches to discourse parsing, the feature space is locally restricted, allowing only features of discourse units that are sequentially adjacent (Joty et al., 2015), or adjacent in the discourse parse (Yoshida et al., 2014; Li et al., 2014). This makes exact inference possible, but at the cost of ignoring aspects of document structure that may be relevant for identify-

Proceedings of the Workshop on Structured Prediction for NLP, pages 25–34,
Austin, TX, November 5, 2016.

ing the correct parse. For example, we may prefer balanced nuclearity structures, or we may prefer to avoid left-branching structures, but these properties cannot be captured with local features. Alternatively, transition-based methods construct the discourse parse through a series of local decisions, typically driven by a classifier (Marcu, 1999b; Sagae, 2009; Ji and Eisenstein, 2014). While the classifier is free to examine any aspect of the document or the existing partial parse, the accuracy of such methods may be limited by search errors.

A second limitation of existing discourse parsers relates to the amount of available training data. Because discourse is a high-level linguistic phenomenon, relatively large amounts of text must be annotated to produce each training instance. In RST, the smallest possible components of each discourse relation are *elementary discourse units* (EDUs), which correspond roughly to clauses. A relatively long news article might feature only a few dozen discourse relations, yet it still requires considerable time for the annotator to read and understand. This suggests that it will be inherently difficult to train accurate discourse parsers using standard supervised learning techniques.

This paper proposes to solve both problems using SampleRank, a structure learning algorithm (Wick et al., 2011). SampleRank uses stochastic search to explore the space of possible outputs, updating its model after each sample. It imposes no limitations on the feature set; given an appropriate sampling distribution, it is capable of exploring the entire space of output configurations (in the limit).

Furthermore, SampleRank can be trained using indirect supervision, which provides a potential solution to the problem of limited training data for discourse parsing. Because discourse nuclearity structures are closely linked to other document-labeling tasks — such as summarization and sentiment analysis — it is in principle possible to use labels from those tasks as a supervision signal for discourse parsing itself. To do this, we link discourse structure and summarization using a *constraint* proposed by Hirao et al. (2013). SampleRank then explores the joint space of extractive summaries and discourse parses, scoring the summaries against automatically-obtained reference summaries, while simultaneously learning to produce discourse parses that are compatible with high-scoring summaries.

At this stage, we have obtained only mixed empirical results with the application of SampleRank to RST discourse parsing: SampleRank offers improvements on one metric for RST parsing in the supervised learning scenario, but it does not improve over a summarization baseline in the indirect supervision scenario. Nonetheless, we hope the ideas presented here will inspire further research in stochastic structure prediction for automated discourse structure analysis.

2 Discourse Parsing as Structure Prediction

We first describe a supervised discourse parser that uses SampleRank to escape the limitations of local features and local structure prediction. Our parser is designed to recover the nuclearity structure of a document, e.g., the unlabeled edges in Figure 1. The full discourse parsing task also requires predicting the nature of the relation between discourse units, e.g., ELABORATION or CONDITION, but we do not consider the relation prediction problem in this work. We also do not consider the problem of *discourse segmentation*, which involves splitting the text into *elementary discourse units*. Prior work shows that relatively simple classification-based approaches can achieve high accuracy on the discourse segmentation task (Hernault et al., 2010; Xuan Bach et al., 2012).

Let $d_i \in \mathcal{D}(x_i)$ represent the nuclearity structure for document i, where x_i represents both the text of the document and its segmentation into elementary discourse units. The set of possible nuclearity structures $\mathcal{D}(x_i)$ includes trees in which adjacent discourse units are related by either mononuclear (subordinating) or multinuclear (coordinating) discourse relations.[1] Each relation instantiates a larger discourse unit, which may then be related to its neighbors, until the entire document is covered by a connected nuclearity structure. Danlos (2008) offers a formal comparison of the representational capacity of RST and related discourse models.

[1] All relations shown in Figure 1 are mononuclear. An example of a multinuclear discourse relation is LIST.

We propose a log-linear probability model over discourse structures,

$$P(d \mid x) \propto \exp\left(\boldsymbol{\theta}^\top \boldsymbol{f}(d, x)\right), \quad (1)$$

where $\boldsymbol{f}(d, x)$ represents a vector of features and $\boldsymbol{\theta}$ represents a vector of weights. As noted above, prior work has largely focused on two restrictions to this model: either constraining the feature function $\boldsymbol{f}(\cdot)$ to consider only local phenomena, or using a local, transition-based approach to incrementally construct the discourse nuclearity structure.

Instead, we use stochastic search to identify the top-scoring discourse structure for any document. This enables the use of arbitrary features, while avoiding making premature commitments to local discourse structures. The SampleRank algorithm (Wick et al., 2011; Zhang et al., 2014) enables us to learn the weight vector $\boldsymbol{\theta}$ in the context of this stochastic inference algorithm. To use SampleRank, we must define three things:

- a feature function $\boldsymbol{f}(\cdot)$;
- a sampling distribution $q(\cdot)$;
- a scoring function $\omega(\cdot)$.

At each step in the algorithm, we sample a discourse structure $d' \sim q(d)$, where d is the previous discourse structure. This sample is then stochastically accepted or rejected, according to the Metropolis-Hastings algorithm: if the sample d' achieves higher likelihood $\ell(d')$ than the previous sample $\ell(d)$, then it is accepted; if not, the sample may still be accepted with probability $\frac{\ell(d')}{\ell(d)}$. When the probability $P(d \mid x)$ and scoring function $\omega(d)$ disagree, an update is made to $\boldsymbol{\theta}$ to try to align the probability with the scoring metric. For more on the details of the algorithm, see the original paper (Wick et al., 2011).

2.1 Features

We employ the following features for every internal node (discourse unit) of an RST tree:

Lexical Features These features capture the first word and last word of both the left and right EDU of internal node. We also add lexical features combined with nuclearity of the EDU.

Cluster Features These features include the Brown et al. (1992) cluster prefix for last and first word of both left and right EDU of internal node.

Syntactic Features These set of features employ POS tags for last and first word of both left and right EDU of internal node.

Sentence-Paragraph Features We also add two features if left and right EDU are in same sentence and if they are in same paragraph.

Text Organizational Features Each sample contains a complete nuclearity structure for the document, and we can compute global features of this structure. Specifically, we compute: whether the full RST tree is left sided, right sided or fully balanced; the sequential position of the overall root nucleus EDU in the document.

2.2 Sampling

The SampleRank algorithm proceeds by making a series of local changes to a complete discourse structure. These changes must preserve the validity of the structure (so that it is impossible to transition from a valid RST nuclearity structure to an invalid structure); they must also be *ergodic*, meaning that they enable a complete exploration of the space of valid RST trees for a given document.

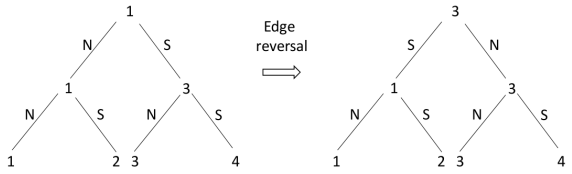
To facilitate stochastic exploration of the space of discourse parses, we convert the RST nuclearity structure to a representation proposed by Hirao et al. (2013), called *dependency discourse trees* (DEP-DT). This representation is a spanning tree over the elementary discourse units (EDUs) of a text. The relationship between RST nuclearity structures and dependency discourse trees is analogous to the relationship in syntax between context-free constituency structures and dependency grammar: just as syntactic constituents have a head element, each composite discourse unit has a most central elementary discourse unit. However, due to the more constrained nature of RST, it is possible to uniquely identify the original RST nuclearity structure from a DEP-DT.

The discourse proposal distribution q_{disc} governs the moves that chooses the next sample discourse tree from the current discourse tree. In RST

parse tree, a set of internal nodes represent relations between adjacent discourse units. Our sampler chooses any internal node with equal probability, and performs one of three possible alterations to the subtree defined by the internal node: edge polarity change, left rotate, and right rotate.

2.2.1 Edge polarity change

This moves changes the "polarity" of the chosen internal node. There are three possible polarities: $N-N$ (indicating a multinuclear relation), $N-S$ (indicating that the leftmost element is the nucleus), and $S-N$ (indicating that the rightmost element is the nucleus). Non-binary multinuclear relations are binarized. As an example, consider switching the polarity of the root node from $N-S$ to $S-N$:



2.2.2 Tree Rotations

A rotation is an operation that changes one binary tree into another. In a tree of n leaf nodes, there are $n-1$ possible rotations: one for each non-root internal node. The rotation corresponding to a node changes the structure of the tree near the node, but leaves the structure intact elsewhere. A rotation operation will keep the order of the leaf nodes intact, but it will change the depth for some nodes.

As shown in Figure 2, a right rotation operation on any internal node (Q) will consist of following operations

- Take the left child (P) of the chosen internal node (Q) and cut off its right subtree (B).
- Move it (P) to the place of the chosen internal node (Q) and attach that as its right child.
- Attach the removed subtree (B) from step 1 as the left child of the original chosen node (Q).

The left rotate is exactly the opposite of the above operation and can be described as following on internal node P:

- Take the right child (Q) of the chosen internal node (P) and cut off its right subtree (B).
- Move it (Q) to the place of the chosen internal node (P) and attach that as its left child.
- Attach the removed subtree (B) from step 1 as the right child of the original chosen node (P).

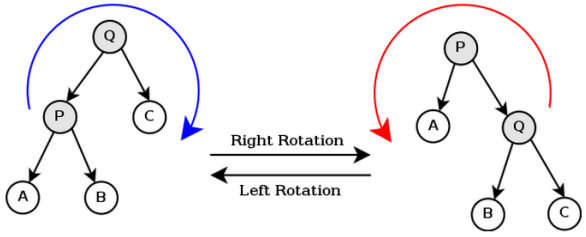


Figure 2: Left and right tree rotation. (Image `Tree_rotation.png` from `Ramasamy` at the English Language Wikipedia.)

To show that this sampler is ergodic, consider that any arbitrary n-node binary search tree can be transformed into any other arbitrary n-node binary search tree using O(n) rotations. We can convert any binary search tree with n nodes into a right-branching chain of length n using at most $O(n)$ right rotation operations. If a node in the tree has a left subtree then we perform a right rotation on that tree node. There can be $O(n)$ such nodes, so we need at most n right rotations. By using similar argument we can prove that a right-branching chain of length n can be converted into any binary search tree with n nodes using as most n left rotations. Combining these two transformations, it is possible to convert any arbitrary n-node binary search tree into any other arbitrary n-node binary search tree, using $O(n)$ rotations.

2.3 Objective function

RST trees are scored in terms of F1 on three properties (Abney et al., 1991; Marcu, 2000): **span** (do the subtrees in the response match the subtrees in the reference?), **nuclearity** (does each subtree have the same nucleus as in identical subtree in the reference?), and **relation** (is each discourse relation governing each span correctly identified?). These metrics form a cascade: every error on the span metric propagates to the nuclearity metric, and every error on the nuclearity metric propagates to the relation

metric. The relation metric is not relevant for this research, as we do not attempt to predict discourse relations. Therefore, we define the objective function as,

$$\omega(d) = \text{F1}_{span}(d, d^{gold}) + \text{F1}_{nuclearity}(d, d^{gold}). \quad (2)$$

This definition carries the usual advantage of SampleRank training, which is to optimize the desired objective, rather than a proxy such as log-likelihood.

3 Summaries as Supervision

The previous section describes how SampleRank can enable the training of an RST discourse parser with arbitrary features and (approximate) global inference. A further advantage of SampleRank is that training can directly target the F1 objective, rather than a log-likelihood or max-margin objective that may relate only tangentially to the true scoring function.

However, a second challenge for discourse parsing is the expense of obtaining labeled training data. In syntactic parsing, each sentence contains many syntactic dependencies; in contrast, in discourse parsing, each elementary discourse unit corresponds to only a single discourse dependency. This means that annotators produce an order-of-magnitude fewer discourse annotations for a given amount of text, making the creation of large discourse-annotated corpora difficult. The RST Treebank is the largest known dataset for discourse parsing, but it contains only a few hundred documents.

Prior work has frequently noted the connection between discourse nuclearity structure and summarization: for example, Marcu (1999c) shows that the nuclearity of a segment predicts its overall importance in the discourse, and Hirao et al. (2013) show that RST nuclearity trees can be exploited for single-document summarization in a constraint-based optimization framework. Summarization annotations are considerably easier to obtain than discourse parses, since they are often available "for free", in the form of bullet-point summaries of news articles (Marcu, 1999a; Svore et al., 2007).

We propose to exploit these annotations to train a discourse parser. We scrape a corpus of newspaper articles and summaries from the CNN website. We then introduce the summary s as an additional variable, while using the discourse parse d to constrain the space of possible summaries: specifically, the elements of the text that align with the summary must be close to the root of the RST tree. By training a model to produce a good summary, we simultaneously train a discourse parser to produce nuclearity structures that are compatible with the ground truth summaries. In this way, a discourse parser can be trained by indirect supervision.

Again, this model can be defined in a log-linear framework:

$$P(s \mid x) = \sum_{d \in \mathcal{D}(x)} P(s, d \mid x) \quad (3)$$

$$P(s, d \mid x) = \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(d, x) + \boldsymbol{\mu}^\top \boldsymbol{g}(s, d, x)) \times \delta(s \in \mathcal{C}(d, x)), \quad (4)$$

where s indicates a summary, $\mathcal{C}(d, x)$ indicates the set of summaries that are compatible with discourse parse d on text x, and $\delta(s \in \mathcal{C}(d, x))$ is an indicator function,

$$\delta(s \in \mathcal{C}(d, x)) = \begin{cases} 1, & s \in \mathcal{C}(d, x) \\ 0, & s \notin \mathcal{C}(d, x). \end{cases} \quad (5)$$

The vector $\boldsymbol{g}(\cdot)$ indicates a vector of features describing the summary, discourse parse, and text; $\boldsymbol{\mu}$ indicates a vector of weights on these features; and $\boldsymbol{\theta}$ and $\boldsymbol{f}(\cdot)$ are weights and features as in Equation 1.

We use a slightly modified version of SampleRank to learn in this setting. To do so, we first define the constraints, features, and scoring function. We then present our adaptation of SampleRank to this form of indirect supervision.

3.1 Summary Constraints

Hirao et al. (2013) propose to relate nuclearity to summarization, by constraining the set of summaries that are compatible with any discourse parse. Their method is based on converting nuclearity structures to a *dependency discourse tree* (DEP-DT), as described in § 2.2. Given this dependency discourse structure, we can express the following constraint on

permissible summaries:

$$\sum_{i=1}^{N} \ell_i s_i \leq L \quad (6)$$

$$\forall i : s_{\text{parent}(i)} \geq s_i \quad (7)$$

where N is the number of EDUs in the document; $\boldsymbol{s}$ is an N-dimensional binary vector that represents the summary, i.e. $s_i = 1$ indicates that the ith EDU is included in the summary; ℓ_i is the number of words of the ith EDU; and L is the maximum length of the summary in words. Constraint (6) ensures that the entire summary contains fewer than L words, and constraint (7) captures the connection to the discourse structure, ensuring that the summary is a rooted subtree of the dependency discourse tree. Thus, the elementary discourse unit i can be present in the summary only if all of its ancestors in the DEP-DT are also present.

Hirao et al. (2013) the performance of constraint-based summarization on the RST treebank, which includes paired summaries and discourse structures for 30 documents. They find that constraint-based summarization yields better ROUGE scores than two extractive baselines: a maximum-coverage summarizer, and a "LEAD" baseline of simply selecting the first few sentences. However, most of these gains are obtained using gold summaries. The improvements offered by automatically produced summaries are much more modest; for ROUGE-2, they do not rise to the level of statistical significance. Our approach is motivated by the idea that using the summarization task to train discourse parser may yield discourse parses that are better, particularly for the downstream task of summarization. To train our system, we gather a much larger dataset by scraping the CNN news website, where each news article is accompanied by a bullet point summary. This data is described in more detail in § 4.2.

3.2 Features

The feature vector $\boldsymbol{g}(s, d, x)$ includes features of the summary. We add the following simple summary features:

Depth-weighted term Frequency Many extractive summarization algorithms are based in part on term frequency, preferring sentences that cover some of the most important elements in the text (Mani and Maybury, 1999). We reward EDUs for containing high-frequency words, in proportion to their depth in the dependency discourse tree:

$$\psi_i = \sum_{i}^{N} s_i \frac{\sum_j^V x_{i,j} \sum_{i'}^N x_{i',j}}{\text{Depth}(i)}, \quad (8)$$

where s_i is an indicator of whether EDU i appears in the summary, V is the vocabulary size, $x_{i,j}$ is the count of word j in EDU i, and $\sum_{i'}^N x_{i',j}$ counts the term frequency over the entire document.

Summary EDU position Previous summarization research shows that the position of each sentence is an important factor in extractive summarization. We employ three positioned-based features: the minimum, maximum and average position of all EDUs appearing in the summary.

Many more summarization features are considered by Berg-Kirkpatrick et al. (2011), and these may be incorporated in the model in future work.

3.3 Summary proposal distribution

To use SampleRank to train from indirect supervision, we must augment the sample state to the tuple (s, d), where s is the summary and d is the discourse structure. The proposal distribution must therefore modify the summary as well as the discourse structure. Our proposal takes a stage-wise approach, first sampling a discourse structure $d \sim q_d(d^{\text{old}})$, and then sampling a summary conditioned on the discourse structure, $s \sim q_{s|d}(d)$, such that s is guaranteed to obey the constraints described above. The discourse structure proposal is unchanged from § 2.2; the summary proposal is as follows:

- Initialize the *summary frontier* to a list containing one element, the root of the dependency discourse tree.
- Repeat until the summary contains L tokens:
  - Sample an EDU from the current summary frontier, with uniform probability across the frontier.

– Add the sampled EDU node text to the summary, remove it from the frontier, and add its DEP-DT children to the frontier list.

The discourse structure sampler is unchanged and is not conditioned on the summaries, so the sampler is ergodic over the space of possible discourse structures for a given document. The summary sampler can generate any summary that meets the constraints for a given discourse structure, and is not conditioned on its prior state. Thus, the overall sampler is ergodic over the paired space of discourse structures and summaries that satisfy their constraints.

To compute the Hastings correction for the Metropolis-Hastings acceptance probability, it is necessary to compute the sampling probabilities. The probability of sampling any summary is equal to the product of probabilities of selecting each EDU at each stage of the sampling procedure, which is in turn based on the frontier size.

Algorithm 1 Sample Rank algorithm for learning discourse parsing and extract summarization from indirect supervision

for $e = 1$ to #epochs **do**
  for $i = 1$ to N **do**
    $d^{'} \sim q_d(\cdot \mid x_i, d_i)$
    $s^{'} \sim q_{s|d}(\cdot \mid x_i, d^{'})$
    $y^{'} \leftarrow \{d^{'}, s^{'}\}$
    $y^{+} \leftarrow arg\max_{y \in \{y_i, y^{'}\}} \omega(y)$
    $y^{-} \leftarrow arg\min_{y \in \{y_i, y^{'}\}} \omega(y)$
    $y_i \leftarrow$ *acceptOrReject*$(y^{'}, y_i; \boldsymbol{\theta}_t, \omega, q)$
    $\triangledown f \leftarrow f(x_i, y^{+}) - f(x_i, y^{-})$
    $\Delta\omega = \omega(y^{+}) - \omega(y^{-})$
    if $\Delta\omega \neq 0$ *and* $\boldsymbol{\theta}_t^{\top} \triangledown f < \Delta\omega$ **then**
      $\boldsymbol{\theta}_{t+1} \leftarrow update(\triangledown f, \Delta\omega, \boldsymbol{\theta}_t)$
      $t \leftarrow t + 1$
    end if
  end for
end for

3.4 Scoring function

In this setting, we receive no supervision on the discourse structure, only on the summary s. Our scoring function therefore can only quantify the summary quality, which we do using the ROUGE metric (Lin, 2004).

For completeness, Algorithm 1 presents our specialization of the SampleRank algorithm to learning joint discourse parsing and summarization from indirect summary-based supervision.

4 Evaluation

We evaluate the supervised model from § 2 on the RST parsing task, and the indirectly-supervised model § 3 on summarization.

4.1 Supervised evaluation

The supervised model is evaluated on supervised task of discourse parsing on RST-DT dataset (Carlson et al., 2002). The RST Discourse Treebank (RST-DT) consists of 385 documents, with 347 for train and 38 for testing in the standard split. We only focus on nuclearity and span prediction tasks. We use the same F1 score on span and nuclearity as our evaluation metrics defined in the section 2.3.

We compare our SampleRank approach with several competitive parsers from the literature: HILDA (Hernault et al., 2010), a bottom-up classification-driven parser; DPLP (Ji and Eisenstein, 2014), a shift-reduce parser that uses representation learning; and a condition random field (CRF) based parser with post-editing operations and a rich array of features (Feng and Hirst, 2014). SampleRank is competitive on the span metric, outperforming all systems except for the CRF approach, which employs rich linguistic features including syntax and entity transitions. On the nuclearity metric, the SampleRank-based parser does somewhat worse than these prior efforts.

4.2 Indirect supervision

We evaluate our indirectly supervised model on the task of summarization for CNN news document and summaries, using the data. The data is obtained by crawling the CNN news website for news articles and the summaries are obtained by the bullet sections. We collected 2000 such news documents and summaries. The CNN summaries are not necessarily extractive, so for supervised training, we link each summary bullet to a sentence in the original text with the highest ROUGE score. (This link from summary bullets to sentences is necessary to compute

	Span F1	Nuclearity F1
HILDA (Hernault et al., 2010)	83.0	68.4
DPLP basic features (Ji and Eisenstein, 2014)	79.4	68.0
DPLP representation learning (Ji and Eisenstein, 2014)	82.1	**71.1**
CRF + post-editing (Feng and Hirst, 2014)	**85.7**	71.0
SampleRank (this work)	84.2	65.3

Table 1: Evaluation of RST discourse parsing

the TKP constraints.) The average summary length in the CNN dataset is roughly 10% of the full document length.

We use ROUGE-1 and ROUGE-2 scores, as defined by Lin (2004), for scoring the summaries. § 4.2 presents the results, in comparison with a simple "LEAD" baseline, which selects the first n sentences of the document. The learning-based method was not able to outperform LEAD, a negative result.

We also apply the Tree Knapsack Problem (TKP) summarization algorithm (Hirao et al., 2013), which incorporates Rhetorical Structure Theory by producing summaries that obey the constraints elaborated in § 3.1, using the RST parses produced by supervised SampleRank training on the RST treebank. Even this method is not able to produce better scoring summaries than LEAD. Hirao et al. (2013) obtained slight improvements on ROUGE-1 over LEAD, using HILDA discourse parses on a dataset of 30 single-document summaries in the RST treebank. The CNN dataset may be less amenable to discourse-driven summarization than the RST data, or the difference may be explained HILDA's superior performance on nuclearity metric.

5 Related Work

Early work on RST discourse parsing focused on local classifiers (Marcu, 1999b; Hernault et al., 2010), with more recent work exploring structure prediction techniques such as sequence labeling (Joty et al., 2015), chart parsing (Li et al., 2014), and minimum spanning tree (Feng and Hirst, 2014). A parallel line of research has considered incremental discourse parsing techniques such as shift-reduce (Sagae, 2009; Ji and Eisenstein, 2014). Muller et al. (2012) apply more advanced search-based algorithms for transition-based discourse parsing in the framework of Segmented Discourse Representation Theory (SDRT). Our proposed approach has the advantage of allowing arbitrary features, and avoiding local search errors; however, stochastic search is not guaranteed to fully explore the search space in any finite amount of time.

We are unaware of prior work on indirect supervision for discourse parsing from downstream tasks. A somewhat related line of work has used explicitly labeled discourse relations as a source of supervision for the classification of implicit discourse relations. Marcu and Echihabi (2002) were the first to explore this approach, working in the context of RST. Sporleder and Lascarides (2008) suggest that informational differences between explicit and implicit discourse relations limit the possible efficacy of this approach. More recent work has treated these two relation types as separate domains, obtaining good results by applying domain adaptation techniques (Braud and Denis, 2014; Ji et al., 2015).

Recent work has applied a number of machine learning techniques to summarization, with particularly relevant work focusing on syntactically-motivated sentence compression (Berg-Kirkpatrick et al., 2011). The combination of the proposed approach with abstractive summarization via sentence compression might yield better results on summarization metrics. Discourse structure has also been linked to sentence compression (Sporleder and Lapata, 2005), suggesting another intriguing direction for future work. Other recent machine learning approaches have employed neural attentional mechanisms for sentence summarization (Rush et al., 2015), but to our knowledge such structure-free discriminatively trained approaches have not been applied on the document level.

	ROUGE-1		ROUGE-2	
	F score	Recall	F score	Recall
LEAD	0.2818	0.2569	0.1154	0.1042
SampleRank, trained on CNN summaries (this work)	0.2317	0.2304	0.0858	0.0851
TKP+SampleRank trained on RST treebank	0.2731	0.2730	0.0967	0.0963

Table 2: Evaluation of joint summarization and discourse parsing algorithm

6 Discussion

This paper proposes a new structure learning approach for discourse parsing, based on the SampleRank algorithm. This approach has the potential to address two major problems with existing discourse parsing algorithms: (1) use of local features or incremental decoding algorithms, and (2) lack of sufficient labeled data. We find some advantages in the supervised setting, with good results on span identification, but relatively poor results on nuclearity. It is possible that fine-tuning the training objective could better balance between these two metrics. We then showed how SampleRank can learn a model that jointly parses the discourse nuclearity structure and produces an extractive summary, using only summary-document pairs as training data. Unfortunately the resulting summarizer fails to outperform a simple baseline. A natural next step would be to design more expressive features for capturing summarization quality, and to learn a joint model from both labeled discourse parses and summaries.

Acknowledgments

This work is supported by a Google Faculty Research award. Thanks to the reviewers for their helpful feedback, to Yangfeng Ji, for help with the features, and to Gongbo Zhang, for helping to build the summary dataset.

References

Steven Abney, S Flickenger, Claudia Gdaniec, C Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Frederick Jelinek, Judith Klavans, Mark Liberman, et al. 1991. Procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the workshop on Speech and Natural Language*, pages 306–311. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 481–490, Portland, OR.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon, September.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

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

Laurence Danlos. 2008. Strong generative capacity of RST, SDRT and discourse dependency DAGSs. In Anton Benz and Peter Kühnlein, editors, *Constraints in Discourse*, pages 69–95. Benjamins.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 511–521, Baltimore, MD.

Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1061–1070. ACM.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem.

In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 1515–1520, Seattle, WA.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon, September.

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

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Inderjeet Mani and Mark T Maybury. 1999. *Advances in automatic text summarization*, volume 293. MIT Press.

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

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 368–375.

Daniel Marcu. 1999a. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–144. ACM.

Daniel Marcu. 1999b. A decision-based approach to rhetorical parsing. In *SIGIR*, pages 365–372.

Daniel Marcu. 1999c. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pages 123–136.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December. The COLING 2012 Organizing Committee.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 379–389, Lisbon, September.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 81–84, Paris, France, October. Association for Computational Linguistics.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 257–264. Association for Computational Linguistics.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

Krysta Marie Svore, Lucy Vanderwende, and Christopher JC Burges. 2007. Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 448–457.

Michael Wick, Khashayar Rohanimanesh, Kedar Bellare, Aron Culotta, and Andrew McCallum. 2011. Samplerank: Training factor graphs with atomic gradients. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 777–784, Seattle, WA.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 160–168, Seoul, South Korea, July. Association for Computational Linguistics.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Yuan Zhang, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Amir Globerson. 2014. Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 197–207, Baltimore, MD.