Evaluating the Impact of a Hierarchical Discourse Representation on Entity Coreference Resolution Performance

Sopan Khosla James Fiacco Carolyn Rosé
Language Technologies Institute
Carnegie Mellon University, USA
{sopank,jfiacco,cprose}@cs.cmu.edu

Abstract

Recent work on entity coreference resolution (CR) follows current trends in Deep Learning applied to embeddings and relatively simple task-related features. SOTA models do not make use of hierarchical representations of discourse structure. In this work, we leverage automatically constructed discourse parse trees within a neural approach and demonstrate a significant improvement on two benchmark entity coreference-resolution datasets. We explore how the impact varies depending upon the type of mention.

1 Introduction

Historically, theories of discourse coherence (Chafe, 1976; Hobbs, 1979; Grosz and Sidner, 1986; Clark and Brennan, 1991) have offered elaborate expositions on how the patterns of anaphoric references in discourse are constrained by limitations in human capacity to manage attention and resolve ambiguity. Hobbs (1979) acknowledges that these human limitations have meant that coreference resolution in natural text can be achieved with relatively high accuracy using a combination of recency and simple semantic constraints. State-of-the-art neural approaches for coreference resolution (Lee et al., 2017; Joshi et al., 2019, 2020) have therefore not surprisingly shown strong performance relying on surface-level features and local-context (i.e., extracted from a small text window around the mention). Traditional approaches, on the other hand, make an attempt to formally model the process of managing attention, for example, the stack in Grosz and Sidner (1986)'s model. Their stack-based model suggests specific places where recency might fail while a more explicit model of discourse structure might make a correct prediction, for example, where an anaphor and a nearby potential (but incorrect) antecedent are in adjacent but separate discourse segments. Because of the potential existence of such cases, we hypothesize that formally incorporating a representation of discourse structure would have a small but non-random positive impact on the ability to correctly resolve anaphoric references. This effect might vary depending upon the semantic informativeness of alternative types of anaphoric expressions, since they impose different constraints on where their antecedent can be located within a hierarchical discourse structure. There is also a danger that the level of accuracy with which the hierarchical structure of discourse can be obtained in practice might reduce the positive impact still further.

The contribution of this paper is an empirical investigation of the impact of including a representation of the hierarchical structure of discourse within a neural entity coreference approach. To this end, we leverage a state-of-the-art RST discourse-parser to convert a flat document into a tree-like structure from which we can derive features that model the structural constraints. We embed this representation within an architecture that is enabled to learn to use this information deferentially depending upon the type of mention. The results demonstrate that this level of nuance enables a small but significant improvement in coreference accuracy, even with automatically constructed RST trees.

2 Related Work

Though recency is the strongest predictor for coreference resolution (CR), prior work in CR has benefited from the inclusion of semantic features such as type-information on top of the surface and syntax-level features. Soon et al. (2001); Bengtson and Roth (2008) used dictionaries like WordNet to extract the semantic class for a noun. More recently, Khosla and Rose (2020) showed that adding NER style type-information to Lee et al. (2017) substantially improves performance across multiple datasets.

Discourse-level features have been successfully

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1645–1651
June 6–11, 2021. ©2021 Association for Computational Linguistics

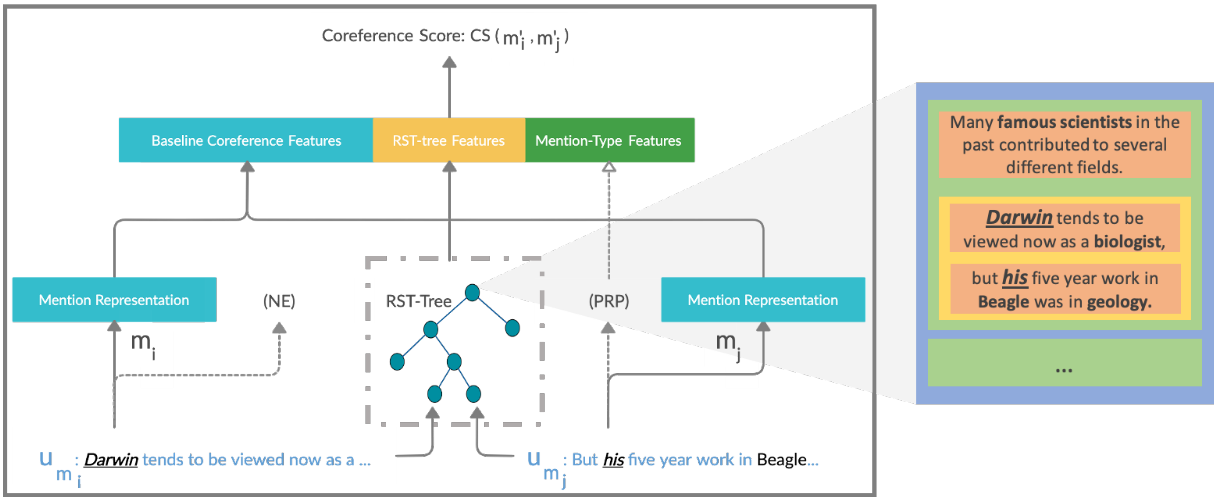
Coreference Score: CS (m'ᵢ , m'ⱼ)
Baseline Coreference Features
RST-tree Features
Mention-Type Features
Mention Representation
(NE)
RST-Tree
(PRP)
Mention Representation
mᵢ
mⱼ
u mᵢ : Darwin tends to be viewed now as a ...
u mⱼ : But his five year work in Beagle...
Many famous scientists in the past contributed to several different fields.
Darwin tends to be viewed now as a biologist,
but his five year work in Beagle was in geology.
...

Figure 1: Schematic diagram of our discourse-informed neural architecture. Discourse (yellow) and mention-type features (green) are concatenated with baseline features (blue) to obtain the mention-pair representation for scoring.

employed in multiple downstream NLP tasks like summarization (Louis et al., 2010), sentiment analysis (Somasundaran et al., 2009), and student writing evaluation (Burstein et al., 2013). For coreference resolution, Cristea et al. (1999) showed that the potential of natural language systems to correctly determine co-referential links, can be increased by exploiting the hierarchical structure of texts. Their discourse model was informed by Vein Theory (Fox, 1987), which identifies chains of elementary discourse units, over discourse structure trees that are built according to the RST (Mann and Thompson, 1987) requirements. Haghighi and Klein (2010) proposed an entity-centered model that leveraged discourse features like dependency-parse tree distance, sentence distance, and the syntactic positions (subject, object, and oblique) of the mention and antecedent to perform coreference.

In this work, we use Yu et al. (2018)'s RST parser to convert documents into RST discourse-structure trees (Mann and Thompson, 1987; Taboada and Mann, 2006). From these trees, we derive distance and coverage-based features to model the discourse-level structural constraints, which are passed as input to a neural-network based coreference resolver. To our knowledge, ours is the first work that tries to explicitly incorporate discourse-level constraints for coreference resolution in a neural setting.

3 Model

In this section, we explain how we introduce discourse-level features into a neural CR system.

3.1 Baseline

We leverage Lee et al. (2017) as our baseline. We replace the word-embeddings with a BERT encoder. A preprocessing step for CR is to identify the mentions within the text that need to be resolved. Following Bamman et al. (2020) and Khosla and Rose (2020), we remove this possible source of error from our evaluation of entity coreference accuracy by using gold-standard mentions.

The baseline model's prediction of coreference for a pair of mentions, $\mathbf{S}\left(m_i, m_j\right)$, is computed as follows. The representations of the two mentions m_i and m_j along with their element-wise product $(m_i \odot m_j)$ and other features like distance between the mentions (d_m), and distance between the sentences that contain the mentions (d_s), are joined together and passed through a fully-connected layer $\mathbf{F}$ (blue boxes in Figure 1).

$$mm_{ij} = [m_i; m_j; m_i \odot m_j; d_m; d_s; ...]$$

$$\mathbf{S}\left(m_i, m_j\right) = \mathbf{F}\left(mm_{ij}\right)$$

3.2 Incorporating Discourse-level Features

By incorporating a representation of the hierarchical discourse structure into the representation that is input to the neural model, we seek to add the capability for reasoning that is not possible in the baseline for each mention-pair (mm_{ij}). None of the features included in the baseline distinguish between pairs that occur within the same or different discourse segments, for example. The closest feature in the baseline that approximates document-level relationships is d_s, since it can be assumed that mentions are less likely to occur within the same segment the further apart they are in the discourse. But this is not universally true.

RST (Mann and Thompson, 1987) offers a theoretical framework in which documents can be parsed into trees that capture the hierarchical discourse structure of the text. In this work, we incorporate structural features from such discourse trees, obtained automatically from Yu et al. (2018). We concatenate three structural features, extracted from the discourse-tree of the document, with mm_{ij} to model these constraints (as shown in Figure 1). We use binarized RST-trees to represent the discourse hierarchy and relationships within each document. Discourse-units identified by the parser occur at the leaves (l) of the output tree.

Consider the document under consideration doc and its RST-tree t_{doc}. For the current mention m_j and candidate mention m_i, and the position of the smallest discourse-unit they belong to in the tree ($l_{u_{m_j}}$ and $l_{u_{m_i}}$ respectively):

DistLCA ($\mathbf{d^j_{lca}}$) encodes the distance between $l_{u_{m_j}}$ and $LCA(l_{u_{m_i}}, l_{u_{m_j}})$. This feature provides information about the amount of generality required to have the two mentions in the same discourse subtree. The smaller the *DistLCA*, the closer the two mentions are assumed to be in the discourse.

LeafCoverageLCA ($\mathbf{lc_{lca}}$) encodes the number of sentences that are covered by the discourse subtree with $LCA(l_{u_{m_i}}, l_{u_{m_j}})$ as its root. This feature captures the coverage of the level of discourse that encloses both mentions. The larger the *LeafCoverageLCA*, the more the document area that needs to be covered to include both mentions.

WordCoverageLCA ($\mathbf{wc_{lca}}$) encodes the number of words that are covered by the discourse subtree with $LCA(l_{u_{m_i}}, l_{u_{m_j}})$ as its root. This feature is analogous to *LeafCoverageLCA* but operates on word-level rather than the discourse-unit-level.

3.3 Mention Types and Cognitive Load

Across different types of anaphoric mentions, depending upon how much information about the antecedent is made apparent, there are differences with respect to the cognitive load imposed on the reader. Because this places differential constraints on the interpretation process, we hypothesize that enabling the model to learn different strategies depending upon the mention-type will be advantageous. We divide mentions into three types ($type$) motivated by the above-mentioned intuition: (i) pronouns (low lexical information, high cognitive load on the reader), (ii) named-entities (already grounded mentions), and (iii) all other noun phrases. A mention is put in the second category if it contains at least one named-entity as predicted by an off-the-shelf NER system.[1] To identify pronouns, we compare the mention against a manually curated list of English pronouns. Ultimately, the discourse and mention-type features are concatenated with mm_{ij} and passed through a fully-connected layer for scoring (Figure 1).

$$\mathbf{S}(m_i, m_j) = \mathbf{F}([mm_{ij}; d^j_{lca}; lc_{lca}; wc_{lca}; type_j])$$

4 Experimental Setup

In this section, we describe the datasets and evaluation metrics we use in our experiments.

4.1 Datasets

We gauge the benefits of using RST-tree features on two state-of-the-art entity CR datasets discussed below. Since, our off-the-shelf RST parser (Yu et al., 2018) is trained on news articles, the choice of datasets is motivated by the attempt at reducing the distribution shift between training and inference while ensuring that the parser was trained on different data than we are using for testing. We use the English subset of **OntoNotes** (Pradhan et al., 2012). The corpus contains multiple sub-genres ranging from news articles to telephone conversations. We also evaluate our approach on a subset of the RST sub-genre of the ARRAU corpus (Poesio et al., 2018) (**A-RST(gt)**), which contains RST ground-truth parse-tree annotations in the RST Discourse-Treebank (Carlson et al., 2003). Following Yu et al. (2018), we keep 347 A-RST(gt) articles for training (out of which we set aside 22 articles for development), and 38 articles for testing. Although ARRAU also annotates bridging (Clark, 1975) and abstract anaphora (Webber, 1991), in this work, we only focus on entity anaphora.

4.2 Evaluation Metrics

Both OntoNotes and A-RST(gt) are input to the system in the CoNLL 2012 format. We evaluate the systems on the F1-score for MUC, B3, and CEAF metrics using the CoNLL-2012 official scripts. However, we only show the average F1-score of the above-mentioned metrics in this

[1] https://demo.allennlp.org/named-entity-recognition

Model	OntoNotes	A-RST(gt)
Lee et al. (2017)	83.36	85.80
+ type	83.70	85.95
+ disc	83.63	86.19
+ disc + type	**83.89**	**86.51**
+ disc(gt)	-	86.41
+ disc(gt) + type	-	**86.70**
+ disc(gt) + type $-\mathbf{d_s}$	-	86.66

Table 1: Performance (Avg. F1) of discourse-informed model variants (gold-mentions) on OntoNotes and A-RST(gt). Underlined numbers represent scores that are significantly different from the baseline ($p < 0.01$).[2]

paper for brevity. We report the mean score of 5 independent runs with different seeds.[3]

5 Results

Ground-truth RST-Trees: To establish an upper-bound for the improvement through introduction of the discourse-tree features, we use features extracted from ground-truth trees. We evaluate the upper-bound performance on A-RST(gt) as it contains documents with annotations for coreference as well as RST-structures. Our results show that incorporating ground-truth tree features along with the mention's type (**+ disc(gt) + type**) gives a boost of 0.90 Avg. F1 ($p < 0.01$) over the baseline (Table 1), suggesting that discourse-level features are beneficial on A-RST(gt). Furthermore, we also find that removing d_s from this discourse-informed model does not cause a statistically-significant drop in performance. We believe that this happens because when discourse-structure features are included in the model, the signal from d_s becomes redundant and sub-optimal.

Predicted RST-Trees: In our second set of experiments we use discourse-trees extracted using Yu et al. (2018)'s RST-parser. As shown in Table 1, adding predicted discourse-tree features improves over the baseline on both datasets, with A-RST(gt) corpus witnessing the highest absolute gain of 0.71 Avg. F1 points. Please note that the results are statistically significant with $p < 0.01$.[4] The relative improvement on OntoNotes is smaller than A-RST(gt) (0.53 absolute Avg. F1 points). This could partially be explained by the fact that the RST-parser is trained on news articles, and therefore, might not generalize well on conversational sub-genres of OntoNotes like tc or bc.

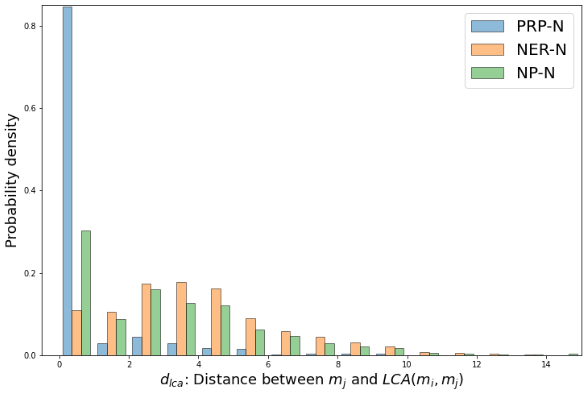

Figure 2: Distribution of d_{lca} for the three different categories of anaphoric mention-pairs.

Ablation Study: To evaluate the contribution of each feature separately, we also perform an ablation study (Table 1). On A-RST(gt), we find that the *type* feature by itself does not provide a considerable boost over the baseline. Use of RST-tree based structural features, on the other hand, shows statistically significant improvements ($p < 0.01$), however, the jump is small (from 85.80 to 86.19). Our final model which includes both RT-tree features and *type* gives the best results. **+ disc + type** performs much better than **+ disc** on both datasets (improvement of 0.32 Avg. F1 points on A-RST(gt) and 0.26 points on Onto) suggesting that the use of *type* as a feature enhances the discriminative power of discourse-tree features.

Mention Type Analysis: To study the influence of different mention-types on the discriminative power of discourse features, we analyze the distribution of d_{lca} across different mention-pair categories in the A-RST(gt) training set.

Setup. To this end, we firstly extract relevant coreferent mention-pairs from the ground-truth clusters. To create a pair for each mention m_j, we choose the mention m_i that belongs to the same cluster (C) as m_j, occurs before it in the document ($i < j$), and is the closest instance of C to m_j. Pairs created using this algorithm do not have other supporting mentions from the same cluster in between them. We then extract three types of mention-pairs from these relevant pairs for our analysis: (i) m_j is a pronoun and m_i is not a pronoun (**PRP-N**); (ii) m_j contains a named entity and m_i is not a pronoun (**NE-N**); and (iii) m_j is neither a pronoun

[2] We leave the evaluation of the impact of including RST structural features in the end-to-end CR setting as future work.

[3] Refer to Appendix A for hyperparameter values and Appendix B for detailed results.

[4] We performed a one-tailed t-test to evaluate significance.

nor contains a named entity, m_i is not a pronoun, and m_i, m_j have no lexical overlap (**NP-N**).

Results. Figure 2 shows that there is indeed a dependence between d_{lca} and mention-pair type. Most of the PRP-N pairs have a $d_{lca} < 5$ even though the the full RST-tree of a document can be as deep as 24 levels. This corroborates our intuition that anaphors with higher ambiguity occur closer to their antecedents in the discourse. For NP-N, we find that 90% of the pairs have $d_{lca} < 8$, whereas, d_{lca} can go as large as 10 for NE-N. This trend explains, at least partially, the difference between the performance of discourse-informed models with and without the mention-type feature.

6 Conclusion

In this paper, we show that a representation of hierarchical discourse structure is beneficial for entity coreference resolution. Our proposed discourse-informed model observes small but statistically significant improvements over a state-of-the-art neural baseline on two coreference resolution datasets. Our analysis shows that the impact of the representation on performance is related to the cognitive load imposed by the type of anaphoric mention.

While the model proposed in this work could serve as a useful baseline for the benefits of including discourse structure-based features in neural coreference resolution models, we realize that there is potential for achieving additional improvements by including more complex constraints (e.g. Right Frontier Constraint (Asher et al., 2003)). We plan to study the affect of such features in future work.

Acknowledgements

We thank the anonymous NAACL reviewers for their insightful comments. We are also grateful to the members of the TELEDIA group at LTI, CMU for the invaluable feedback. This work was funded in part by NSF grants 1949110, 1822831, 1546393, and funding from Dow Chemical.

References

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 44–54.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013. Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52.

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

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic.*

H. H. Clark. 1975. Bridging. In *Proceedings of TINLAP*.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*.

Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. 1999. Discourse structure and co-reference: An empirical study. In *The relation of discourse/dialogue structure and reference*.

Barbara A. Fox. 1987. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge Studies in Linguistics. Cambridge University Press.

Barbara Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S Weld. 2019. Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5807–5812.

Sopan Khosla and Carolyn Rose. 2020. Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Sadat Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, et al. 2018. Anaphora resolution with the arrau corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, page 1–40, USA. Association for Computational Linguistics.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Maite Taboada and William C Mann. 2006. Applications of rhetorical structure theory. *Discourse studies*, 8(4):567–588.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive processes*, 6(2):107–135.

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

Appendix

A Hyperparameters

Hyperparameter	**Value**
BERT	base-cased
BERT weights	freeze
BiLSTM hidden dim	200
d_{lca} embedding-size	20
lc_{lca} embedding-size	20
wc_{lca} embedding-size	20
$type$ embedding-size	20
FC-layer 1 size	150
FC-layer 2 size	150
Dropout	0.2

Table 2: Hyperparameter values for our model. Reader is referred to `https://github.com/dbamman/lrec2020-coref` for the implementation of the baseline model.

B Detailed Results

Model	**OntoNotes**			**A-RST(gt)**		
	MUC	**B^3**	**CEAF**	**MUC**	**B^3**	**CEAF**
Lee et al. (2017)	90.8	82.3	77.1	79.0	89.0	89.3
+ type	91.1	82.7	77.3	79.3	89.0	89.4
+ disc	91.0	82.5	77.4	79.5	89.2	89.7
+ disc + type	91.2	82.7	77.8	79.7	89.7	90.1
- d_s	-	-	-	79.4	88.5	89.3
- d_s + disc(gt) + type	-	-	-	80.2	89.6	90.2
+ disc(gt)	-	-	-	79.8	89.5	89.9
+ disc(gt) + type	-	-	-	80.2	89.8	90.1

Table 3: Detailed Performance (F1 score) of discourse-informed model variants (gold-mentions) on OntoNotes and A-RST(gt).

C Results on Validation Set

Model	**OntoNotes**	**A-RST(gt)**
Lee et al. (2017)	83.42	85.98
+ type	84.01	86.15
+ disc	83.91	86.40
+ disc + type	**84.38**	**86.74**
+ disc(gt)	-	86.68
+ disc(gt) + type	-	**87.02**

Table 4: Performance (Avg. F1) of discourse-informed model variants (gold-mentions) on OntoNotes and A-RST(gt) validation set.

D Computational Infrastructure

All our experiments are performed on a single Nvidia GeForce GTX 1080 Ti GPU. Training took

20-22 hours on OntoNotes, and 5-7 hours on A-RST(gt). We trained the models for 100 epochs with an early-stopping criteria on the Avg. F1 performance on the validation set.