Incorporating Centering Theory into Neural Coreference Resolution

Haixia Chai and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
{haixia.chai, michael.strube}@h-its.org

Abstract

In recent years, transformer-based coreference resolution systems have achieved remarkable improvements on the CoNLL dataset. However, how coreference resolvers can benefit from discourse coherence is still an open question. In this paper, we propose to incorporate centering transitions derived from centering theory in the form of a graph into a neural coreference model. Our method improves the performance over the SOTA baselines, especially on pronoun resolution in long documents, formal well-structured text, and clusters with scattered mentions.[1]

1 Introduction

Coreference resolution is the task to find all expressions which refer to the same entity. The coreferential mentions could occur anywhere in the discourse. In recent years, many transformer-based models (Joshi et al., 2019, 2020; Kirstain et al., 2021) achieved improvements on the CoNLL benchmark (Pradhan et al., 2012). In contrast to using transformers such as BERT (Devlin et al., 2019) which learn the text input sequentially in limited chunks, how knowledge about the structure of discourse can benefit coreference resolution is less explored in the neural NLP era.

Coreference plays an essential role in discourse coherence. A referring expression using a reduced linguistic form (e.g., pronoun) indicates a referential relation to its antecedent in previous utterances. The referring expression connects utterances and contributes to discourse coherence implicitly. On the other hand, coreference resolution can benefit from a coherent discourse. It has long been acknowledged that coherence structure can impose constraints on referential accessibility from a linguistic perspective (Asher and Lascarides, 2003). Centering theory (Joshi and Weinstein, 1981; Grosz et al., 1983, 1995; Walker et al., 1998) is a method to formally describe discourse coherence by using attentional state (i.e., the focus of attention of the participants at each utterance of the discourse). Figure 1 shows how the coherence structure of an example text is built by means of tracking the changes in the local attentional state.[2] By applying centering theory, Gordon and Scearce (1995) investigate how local coherence influences the interpretation of ambiguous pronouns. From reading-time experiments, they observe that utterances with pronouns were read faster in the centering `continue` than in the `shift` status, while utterances with noun phrases containing rich lexical information were read more quickly in the centering `shift` than in the `continue` status. We conjecture that this pattern could contribute to coreference resolution.

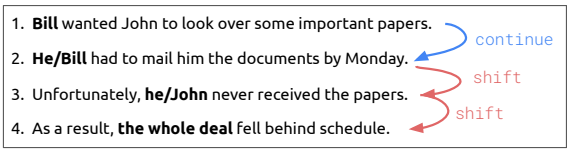

Figure 1: An example text shows how foci change sentence by sentence. The words in bold are the focus of each sentence. The arrows indicate centering transitions with two different transition types, `continue` and `shift`.

In this work, we explore the effect of changes in attentional state in the discourse on entity coreference resolution in a neural approach. Inspired by Jeon and Strube (2020), we capture the most salient mentions of each sentence as centers to compute the local centering transition relations in accordance with centering theory. We then extend the coherence structure globally in the form of a graph. It makes the centering transitions available between any two sentences. Lastly, we fuse the novel discourse structure into a neural coreference model (Kirstain et al., 2021). From the results, our

[1] Our code and model are publicly available at: https://github.com/HaixiaChai/CT-Coref

[2] The example is based on Gordon and Scearce (1995).

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2996 - 3002

proposed method improves the SOTA models up to 80.9 F1 score. Our extensive analysis shows that our approach performs better on pronoun resolution in long documents, formal well-structured text such as magazine and newswire genres, and documents with scattered mentions of clusters. Overall, we observe that incorporating discourse structure derived from centering theory can benefit coreference resolution.

2 Related Work

Discourse for Coreference. In early work, some structural features of the discourse were used in machine learning approaches, such as sentential position and distance between two mentions in sentences, phrases or mentions (Soon et al., 2001; Sapena et al., 2013). In the deep learning period, only a few researchers incorporated discourse information into a coreference model to our knowledge. Recently, Khosla et al. (2021) use rhetorical structure theory (RST) (Mann and Thompson, 1988) to capture the hierarchical discourse structure of documents, from which they encode three distance features for the candidate and query mentions on different levels (i.e., word-level, discourse-unit-level and discourse subtree). Held et al. (2021) apply discourse coherence (Grosz, 1977, 1978; Grosz and Sidner, 1986) to cross-document coreference resolution. They retrieve candidate mentions by modeling the attentional state within a latent embedding space as a set of nearest neighbors for a query mention. As a pruning method, these neighborhoods constrain the search space for their following pairwise classifier. Different from the approaches above, we use centering theory to extract centering transitions between sentences through attentional state. We then use these relations to construct a discourse structure that shows how centers change as discourse proceeds dynamically.

Centering Theory. Since centering theory is a linguistic theory, a great number of early works (Walker et al., 1994; Di Eugenio, 1998; Turan, 1998; Strube and Hahn, 1999) were dedicated to transform it to a computational one for various languages. Lately, Jeon and Strube (2020) is the first work that applies centering theory in a neural model for discourse coherence explicitly. They capture the relationships sentence by sentence for assessing text coherence. In the coreference resolution task, coreferent mentions could occur anywhere in the discourse rather than the adjacent sentence only. Thus, we propose a fully connected centering transitions graph in our model. In addition, Jeon and Strube compute the centers of each sentence on a token-level, while we do it on the span-level.

	$C_b(s_i) \approx C_b(s_{i-1})$	$C_b(s_i) \neq C_b(s_{i-1})$	No $C_b(s_i)$
$C_b(s_i) \approx C_p(s_i)$	Continue	Shift	None
$C_b(s_i) \neq C_p(s_i)$	Retain		

Table 1: Centering transition relations. For instance, `continue` indicates that the center of utterance s_i is similar to the one in its previous utterance s_{i-1}.

3 Model

3.1 Baseline

We study the model proposed by Kirstain et al. (2021) as our baseline. It is a *start-to-end* (**s2e**) coreference resolution model that only considers boundary points of a span to compute the mention and antecedent scores without additional handcrafted features. Similar to the method of Lee et al. (2018), they measure how likely a candidate mention c is to be an antecedent of a query mention q by a scoring function $f(c, q)$. The function is the addition of two mention scores $f_m(c)$, $f_m(q)$ and an antecedent score $f_a(c, q)$. Our model is based on this scoring function.[3]

3.2 Incorporating Centering Transitions

According to centering theory, we formulate centering transitions among utterances — sentences specifically — in our approach. Figure 2 shows our model architecture.

Centering Theory. Centering theory describes the local coherence and its relationship to attentional state within a discourse segment. From each utterance, one can extract (1) a set of *forward-looking centers* (C_f) ranked according to their prominence, (2) a single *backward-looking center* (C_b) connected with one of the C_f of the immediately preceding utterance, and (3) a *preferred center* (C_p) which is the most salient center in C_f. Following Jeon and Strube (2020), Table 1 presents all relations of centering transition at the local level. The relationships between discourse segments and

[3]For more details, we refer to the original Kirstain et al. (2021) paper.

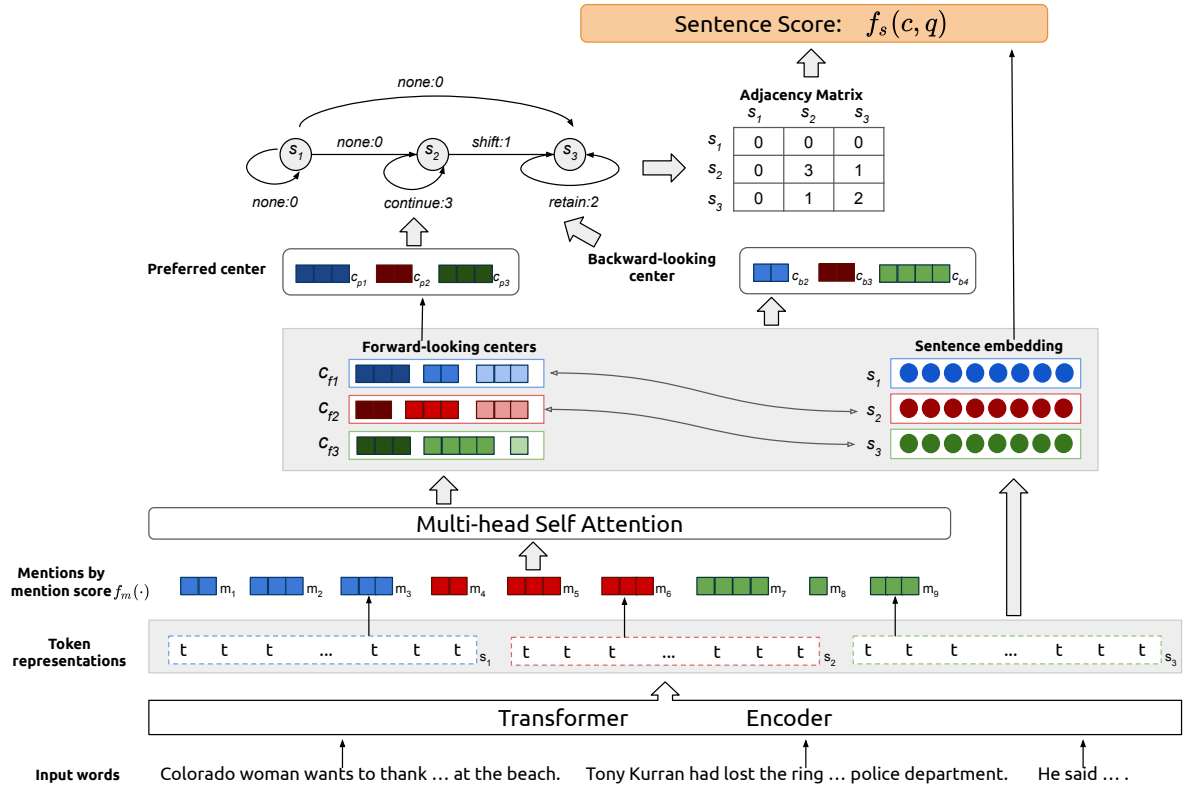

Figure 2: The figure shows our model architecture incorporating centering transitions for the sentence score. There are three example sentences in blue, red and green colors. A string of squares refers to a mention comprising a different number of tokens. The mention with a darker color indicates that it is a more salient center in a sentence.

utterances then provide the fundamental structural centering relations for discourse.

Scoring Function. To combine the structure of the centering transitions with the coreference model, we add sentence score $f_s(c, q)$ to the scoring function as shown below. The last item measures the relationship of the encoded sentences where two examined mentions are located. Especially, if the query mention is a singleton, we set the scoring function to 0.

$$f(c, q) = f_m(c) + f_m(q) + f_a(c, q) + f_s(c, q)$$

Centering Transitions. Having the mention scores $f_m(\cdot)$, we use top λn mentions for further processing of centering transitions (where n is the number of input tokens). Inspired by Jeon and Strube (2020), the remaining mentions with their positions in each sentence are encoded and fed into a multi-head self-attention matrix — $softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$ — to compute the attention score (Vaswani et al., 2017). Q and K stand for the observed mentions of the sentence. From the ranked diagonal elements of the self-attention matrix, we take the top m mentions as C_f, and the first most salient mention as C_p. As for C_b of the sentence s_i, we select the mention from C_f of its previous adjacent sentence s_{i-1}, which has the highest semantic similarity with the current sentence s_i. Here, we simply use the averaged token representations as the sentence embedding $\mathbf{e}_{s_i}$. Finally, we generate centering transition relations (i.e., `continue`, `retain` and `shift`) between each two adjacent sentences by computing cosine similarity according to the rules in Table 1.

Sentence Score. Sometimes, a candidate mention is more than one sentence away from the query mention. Thus, we apply centering transitions not only at a local level but also to all other sentences in discourse globally. Treating sentences as nodes and transition relations as edges, each sentence can be encoded with its neighbouring nodes weighted by the edges which are connected to it, including self-connections. Then, we calculate the sentence score for each pair of mentions by using the embeddings of the sentences $\mathbf{e}_s$ where the candidate and query mentions belong to.

$$f_s(c, q) = \mathbf{A} \cdot \mathbf{e}_{s_c} \cdot \mathbf{B} \cdot \mathbf{e}_{s_q}$$

Model	MUC			B^3			$CEAF_e$			LEA			CoNLL
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
c2f	85.7	85.3	85.5	79.5	**78.7**	79.1	76.8	75.0	75.9	76.2	75.7	75.9	80.2
s2e	86.5	85.1	85.8	80.3	77.9	79.1	76.8	**75.4**	76.1	78.3	75.8	77.0	80.3
s2e + se	87.0	85.1	86.0	80.5	78.2	79.3	77.6	74.9	76.2	78.4	76.1	77.2	80.5
s2e + se_ct	**87.2**	**85.3**	**86.3**	**80.7**	78.6	**79.6**	**78.2**	75.2	**76.7**	**78.7**	**76.5**	**77.6**	**80.9**

Table 2: Performance on the test set of the English OntoNotes 5.0 dataset. **c2f** refers to Joshi et al. (2020).

In the above formula, $\mathbf{A}$ refers to an adjacency matrix, which represents the centering transitions between sentences by numerical values (i.e., `continue`: 3, `retain`: 2, `shift`: 1, and `none`: 0). They are induced from the above parts. Then, we use a bilinear product over the resulting representations with trainable parameter $\mathbf{B}$ to compute the sentence score.[4] We examine this setting in our experiment **s2e+se_ct**. When $\mathbf{A}$ is an identity matrix (i.e., the matrix with ones on the main diagonal and zeros elsewhere), the aggregation over $\mathbf{e_s}$ does not occur. We use this setting **s2e+se** as our simple baseline system for comparative evaluation.

4 Experiments

Settings. We train and evaluate our models on the English OntoNotes 5.0 dataset (Pradhan et al., 2012). The results are reported using the CoNLL F1 score — the average of MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998), CEAFe (Luo, 2005) — and LEA (Moosavi and Strube, 2016).

In our experiments, we examine two models, **s2e+se** and **s2e+se_ct**, as explained in §3. We have 8 self-attention heads for the attention mechanisms, select the top 5 mentions for C_f of each sentence, and set the threshold of cosine similarity as 0.8 for computing the centering transition relations. Following the baseline, We also use Longformer (Beltagy et al., 2020) as our pretrained model, which can process a sequence up to 4096 tokens. We set the other parameters the same as the baseline (Kirstain et al., 2021).[5] All our experiments are performed on a single NVIDIA Tesla V100 32G GPU.

Results. Table 2 shows our results. The model **s2e+se_ct** achieves the best result with 80.9 F1 on OntoNotes. Though both examined models outperform the baselines only by a small margin, it suggests that incorporating centering transitions is helpful to some extent for coreference resolution. To thoroughly utilize the discourse structure, a graph or tree-based coreference model would be a promising research direction. One option would be latent trees which have been explored by Björkelund and Kuhn (2014) and Martschat and Strube (2015) for providing a more reliable basis for coreference resolution before the neural NLP era.

Analyses. First, we check the performance of our model for pronoun resolution on: (1) GAP dataset (Webster et al., 2018); and (2) OntoNotes test dataset in which we only keep the resolved clusters containing pronouns in both gold and system outputs. Table 3 shows the marginal differences between the baseline and our model on the GAP benchmark. GAP is a gender-balanced corpus of ambiguous pronouns sampled from Wikipedia, in which most of examples are short texts. We compute the distributions of lengths of examples by sentence on both GAP and OntoNotes. As shown in the Table 4, the large majority of examples in GAP are 2-4 sentences texts, while the test set of OntoNotes has many documents longer than 5 sentences. The experiment on OntoNotes in Table 5 shows that our model outperforms the baseline across all evaluation metrics. Overall, the two observed results suggest that our model involving centering transition relations between sentences can improve pronoun resolution especially on long documents.

	Masc	Fem	Bias	Overall
s2e	**91.9**	**88.2**	0.96	**90.1**
s2e + se_ct (ours)	91.8	87.8	0.96	89.9

Table 3: F1 scores of the examined coreference resolvers running on the test set of the GAP dataset.

[4] We only adopt adjacency matrix to aggregate e_{s_c} rather than to aggregate both e_{s_c} and e_{s_q}, as the former performs better based on our experiments.

[5] https://github.com/yuvalkirstain/s2e-coref

	Number of Sentences								
	1	2	3	4	5-9	10-20	21-40	41-60	61+
GAP	70	515	**878**	433	104	-	-	-	-
OntoNotes	-	8	10	13	70	**94**	90	46	35

Table 4: Distributions of document length of the two datasets. The bold-faced numbers present the peak text length of each dataset.

	MUC	B^3	$CEAF_e$	LEA	CoNLL F1
s2e	87.8	80.8	79.5	79.3	82.7
s2e + se_ct (ours)	**88.4**	**81.6**	**80.1**	**80.1**	**83.4**

Table 5: F1 scores of the examined models for pronoun resolution on the test set of OntoNotes.

Second, we investigate how models perform on different genres. The CoNLL-2012 data contains broadcast conversation (bc), broadcast news (bn), magazine genre (mz), newswire genre (nw), pivot text (pt), telephone conversation (tc), and web data (wb) genres. In Table 6, we find that our method gets the most improvements on mz and nw genres, in which text is always formal well-structured. In contrast, tc and wb are the most challenging genres for our approach, where disfluent and ungrammatical segments and sentences may occur. Therefore, we summarize that discourse structure information is more beneficial for narrative text than less-formal text like conversation and web data. Resolving coreference in noisy user-generated text such as text on social media platforms is even harder (Chai et al., 2020).

	bc	bn	mz	nw	pt	tc	wb
s2e	78.2	83.2	84.1	74.0	88.2	81.8	77.9
s2e + se_ct (ours)	78.7	83.1	85.0	74.9	89.0	80.7	77.2
Improvement	0.5	-0.1	**0.9**	**0.9**	0.8	-1.1	-0.7

Table 6: Performance of the examined models on the test set for genres. The bottom line shows the improvement over the baseline by our method.

Finally, we observe the effect of maximum sentence distance d_{c_i} between any two mentions of each cluster on models. We take the average of all d_c in the same document as the sentence distance of it. Figure 3 depicts that our method performs better when mentions have a distance of more than six sentences. So, utilizing centering transitions globally is helpful for resolving clusters where mentions are more scattered. Our method captures how centers change between not only adjacent sentences but also non-adjacent sentences in the discourse. This is specially designed for coreference resolution based on centering theory. Meanwhile, we observe that it is difficult for both systems to resolve coreference on documents with long sentence distances (i.e., 12+ sentences).

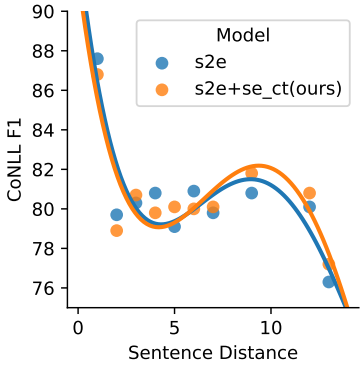

Figure 3: Performance on sentence distance with regression lines.

5 Conclusion

We present a neural coreference model incorporating discourse structure information based on centering theory. The model captures the centering transition relationships between sentences. Each sentence is encoded with all neighbour sentences in a weighted graph. Our approach outperforms the baseline with 80.9 F1 score. Especially, it helps resolving pronoun in long documents, text in formal genres and clusters with scattered mentions.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments and suggestions. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies PhD. scholarship.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation*, pages 563–566, Granada, Spain.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution

with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.

Haixia Chai, Wei Zhao, Steffen Eger, and Michael Strube. 2020. Evaluation of coreference resolution systems under adversarial attacks. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 154–159, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Di Eugenio. 1998. Centering in Italian. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, chapter 7, pages 115–138. Oxford University Press, Oxford.

Peter C Gordon and Kimberly A Scearce. 1995. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, 23(3):313–323.

Barbara J. Grosz. 1978. Focusing in dialog. In *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '78, page 96–103, USA. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics, 12, 175-204*.

Barbara Jean Grosz. 1977. *The Representation and Use of Focus in Dialogue Understanding.* Ph.D. thesis. AAI7731381.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2020. Centering-based neural coherence modeling with hierarchical discourse segments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.

Aravind K. Joshi and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the IJCAI, Vancouver, CA*, pages 385–387.

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 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Sopan Khosla, James Fiacco, and Carolyn Rosé. 2021. Evaluating the impact of a hierarchical discourse representation on entity coreference resolution performance. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1645–1651, Online. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

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

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

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*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2013. A constraint-based hypergraph partitioning approach to coreference resolution. *Computational Linguistics*, 39(4):847–884.

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.

Michael Strube and Udo Hahn. 1999. Functional centering–grounding referential coherence on information structure. *Computational Linguistics*, 25(3):309–344.

Umit Turan. 1998. Ranking forward-looking centers in Turkish: Universal and language-specific properties. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, chapter 8, pages 139–160. Oxford University Press, Oxford.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. *Centering Theory in Discourse*. Oxford University Press.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.