

# Incorporating Singletons and Mention-based Features in Coreference Resolution via Multi-task Learning for Better Generalization

Yilun Zhu<sup>♣</sup>, Siyao Peng<sup>♡</sup>, Sameer Pradhan<sup>♣</sup><sup>◇</sup>, and Amir Zeldes<sup>♣</sup>

<sup>♣</sup>Department of Linguistics, Georgetown University

<sup>♡</sup>Center for Information and Language Processing (CIL), LMU Munich

<sup>♣</sup>Linguistic Data Consortium, University of Pennsylvania

<sup>◇</sup>cemantix.org

{yz565, Amir.Zeldes}@georgetown.edu, siyaopeng@cis.lmu.de, pradhan@cemantix.org

## Abstract

Previous attempts to incorporate a mention detection step into end-to-end neural coreference resolution for English have been hampered by the lack of singleton mention span data as well as other entity information. This paper presents a coreference model that learns singletons as well as features such as entity type and information status via a multi-task learning-based approach. This approach achieves new state-of-the-art scores on the OntoGUM benchmark (+2.7 points) and increases robustness on multiple out-of-domain datasets (+2.3 points on average), likely due to greater generalizability for mention detection and utilization of more data from singletons when compared to only coreferent mention pair matching.<sup>1</sup>

## 1 Introduction

Coreference is a linguistic phenomenon that occurs when two or more expressions in a text refer to the same entity (e.g. *the Vice President... She*). Conceptually, resolving coreference takes two steps: identifying all mention candidates from a text as opposed to non-referring expressions, and linking identified mentions into clusters. However, in a given document, some mentions are never referred back to: these are called singletons, i.e. mentions that, unlike non-referring expressions, could be referred back to in principle, but are not involved in any coreference relations in context. Singletons are important to coreference resolution since they represent true negatives in cluster linking (Kübler and Zhekova, 2011), but also to how humans understand discourse from a theoretical perspective (Grosz et al., 1995), since they also constitute mentioned entities (i.e. clusters of size 1).

However, due to the lack of singleton annotation in the most frequently used coreference dataset for

English, i.e. OntoNotes V5.0 (Weischedel et al., 2011; Pradhan et al., 2013), previous attempts have either ignored singletons (Lee et al., 2017, 2018; Wu et al., 2020; Dobrovolskii, 2021) or incorporated pseudo-singletons into the model (Wu and Gardner, 2021; Toshniwal et al., 2021). The first approach is commonly used in contemporary end-to-end (e2e) systems which train directly on detecting coreferring mentions, but causes problems in that models cannot differentiate singleton spans from non-referring or random/meaningless spans, i.e. penalizing these two types equally. Though e2e has achieved significant progress on OntoNotes, it does not align with linguistic theories on how humans resolve the task. The second approach attempts to amend the model with pseudo-singletons by predicting non-coreferring mentions, but the accuracy gap between gold and generated singletons is unknown and ultimately leads to degradation.

Previous work has also shown that recent coreference models struggle with domain generalization (Moosavi and Strube, 2017; Zhu et al., 2021). To alleviate the problem, Moosavi and Strube (2018) proposed a novel algorithm to incorporate linguistic features and showed improvement in out-of-domain (OOD) data. Subramanian and Roth (2019) applied adversarial training to improve generalization. However, the first approach requires carefully designed linguistic features, and both papers evaluated generalization only on one single-genre dataset, limiting the validity of the results.

To tackle these challenges, we introduce a novel coreference model. Our contributions can be summarized as follows: First, we propose a multi-task learning (MTL) based neural coreference model with constrained mention detection, which jointly learns several mention-based tasks, including singleton detection, entity type recognition, and information status classification. Second, experiments demonstrate that the proposed model achieves new

<sup>1</sup>The code is publicly available at <https://github.com/yilunzhu/coref-mtl>.

state-of-the-art performance on the OntoGUM test set. Third, we show that our model outperforms strong baselines on two OOD datasets, showing it generalizes more reliably to unseen data than plain e2e. We release all code and provide a system that detects and links all mentions, including singletons, and outputs predicted entity types.

## 2 Related Work

**MTL for coreference** Multitask learning (Caruana, 1997; Collobert and Weston, 2008) uses a single model with shared parameters trained to perform multiple tasks, with potential benefits arising from synergies between related objectives. Previous work has investigated the use of MTL for coreference by harnessing related pre-training tasks. Yu and Poesio (2020); Kobayashi et al. (2022) applied an MTL framework to a more specific bridging resolution problem, with standard coreference resolution as the additional task. Luan et al. (2018) used MTL with coreference resolution, entity recognition, and relation extraction for scientific knowledge graph construction. Lu and Ng (2021) used five MTL tasks for event coreference resolution.

**Neural coreference resolution** The e2e approach jointly learns mention detection and coreferent pair scoring (Lee et al., 2017), and achieved SOTA scores on the OntoNotes test set before several extensions were proposed. Lee et al. (2018); Kantor and Globerson (2019) improved span representations to improve pair matching. Joshi et al. (2020) added better pre-trained language models to gain additional score boosting. Wu et al. (2020) adapted a question-answering framework into the task and improved both span detection and coreference matching scores. Dobrovolskii (2021) also improved performance by initially matching coreference links via words instead of spans.

## 3 Methods

### 3.1 Model

Let  $N$  be the number of possible spans in a document  $D$ . The coreference task can be formulated as assigning an antecedent span  $y_i$  for each span  $i$ , where the set of possible antecedents for each span  $i$  contains a dummy antecedent  $\epsilon$  and all preceding spans:  $\mathcal{Y}(i) = \{\epsilon, 1, \dots, i - 1\}$ .

$$s(i, j) = \begin{cases} 0, & j = \epsilon \\ s_m(i) + s_m(j) + s_c(i, j), & j \neq \epsilon \end{cases}$$

where  $s_m(i)$  and  $s_m(j)$  are the mention scores that determine how likely the selected text span is a mention candidate. Previous work utilizes a scoring function to measure how likely the span is a coreference markable. However, singletons in the training data are ignored and thus weaken the model’s generalization capability. Therefore, our proposed model uses two scoring functions to represent the distributions of markables and mentions better. The mention scoring function uses two feed-forward networks fed by the representation of each span: one part is a markable score that calculates the score of the span being a coreferent markable in the document; the other is the mention candidate score that determines how likely a span is a mention candidate. The formula is represented as follows:

$$\begin{aligned} s_m(i) &= \beta_1 \cdot s_{\text{markable}(i)} + \beta_2 \cdot s_{\text{mention}(i)} \\ s_{\text{markable}(i)} &= w_{\text{markable}} \cdot \text{FFNN}(g_i) \\ s_{\text{mention}(i)} &= w_{\text{mention}} \cdot \text{FFNN}(g_i) \end{aligned}$$

where  $\cdot$  denotes a dot product, FFNN denotes a feed-forward neural network,  $\beta_1$  and  $\beta_2$  denote model parameters that adjust the weights of markable scores and mention candidate scores, and  $g_i$  denotes the represented embeddings of the span (we use the same span representing method as in Lee et al. (2017)). The two scoring functions are computed via two standard feed-forward neural networks. The purpose of this design is to prevent random text spans being fed to the pair-matching step. Following the e2e approach (Lee et al., 2017, 2018; Joshi et al., 2020), we concatenate the boundary representations, the soft head vector and an additional feature vector  $\phi$  containing speaker information, and feed the resulting vector into separate feed-forward neural networks to calculate markable scores and mention candidate scores.

In addition to the main pair-matching task, our model adds three mention-based tasks: a (possibly singleton) mention span detection task, entity type recognition, and information status classification (see below). For each task, the span vector is fed into a separate feed-forward network for classification. Each task is assigned a weight to calculate the total loss score:

$$\mathcal{L}_{\text{total}} = \sum_{c=1}^C \mathcal{W}_c \cdot \mathcal{L}_c$$

where  $\mathcal{W}_c$  is the weight for task  $c$ . See Appendix A for an overview of the model architecture.

	Markble Detection			MUC			B <sup>3</sup>			CEAF <sub>φ4</sub>			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
<b>In-domain - ONTOGUM</b>													
Joshi et al. (2019)	<b>91.0</b>	71.9	80.3	<b>83.3</b>	69.7	75.9	70.8	59.2	64.5	70.5	45.8	55.5	65.5
MTL (sg)	90.2	75.0	<b>81.9</b>	82.7	72.8	77.4	70.4	63.1	66.5	71.5	49.2	58.3	67.6
MTL (sg+ent)	90.0	<b>75.1</b>	<b>81.9</b>	82.8	<b>72.9</b>	<b>77.6</b>	<b>71.2</b>	<b>63.6</b>	<b>67.2</b>	<b>71.9</b>	<b>50.2</b>	<b>59.1</b>	<b>68.2</b>
MTL (sg+ent+infs.)	90.0	75.0	81.8	82.1	72.3	76.9	70.0	62.3	65.9	70.0	48.6	57.3	66.9
<b>Out-of-domain - ONTONOTES</b>													
Joshi et al. (2019)	<b>83.9</b>	76.9	80.3	<b>77.6</b>	72.7	75.1	66.9	60.6	63.6	<b>64.3</b>	54.5	59.0	65.9
MTL (sg+ent)	82.2	<b>80.2</b>	<b>81.2</b>	77.0	<b>76.1</b>	<b>76.5</b>	<b>67.1</b>	<b>64.0</b>	<b>65.5</b>	63.6	<b>59.5</b>	<b>61.5</b>	<b>67.8</b>
<b>Out-of-domain - WIKICOREF</b>													
Joshi et al. (2019)	79.9	58.8	67.7	73.7	60.1	66.2	66.4	43.4	52.4	56.6	31.6	40.5	53.0
MTL (sg+ent)	<b>80.4</b>	<b>60.0</b>	<b>68.7</b>	<b>74.5</b>	<b>61.8</b>	<b>67.5</b>	<b>67.8</b>	<b>45.3</b>	<b>54.4</b>	<b>59.0</b>	<b>33.0</b>	<b>42.4</b>	<b>55.6</b>

Table 1: Comparison between Joshi et al. (2019) and our model on test sets of both in-domain (OntoGUM 8.0) and out-of-domain datasets (OntoNotes and WikiCoref). The overall F1 score is the average of F1s from three evaluation metrics MUC, B<sup>3</sup>, and CEAF<sub>φ4</sub>. All models are trained on OntoGUM.

### 3.2 Task Selection

Since OntoNotes does not contain singletons, we choose a corpus for which singleton information is available but follows the same annotation scheme as OntoNotes. The OntoGUM corpus (Zhu et al., 2021) is an adapted version of the GUM corpus (Zeldes, 2017), a multi-layer corpus with a range of annotations at the word level (part-of-speech, morphology), phrase level (phrase trees, entity recognition, and linking), dependency level (Universal Dependencies syntax) and document-level (discourse parses and coreference). Although OntoGUM uses the same singleton-free coreference scheme as OntoNotes, information about singletons can be recovered from the original GUM corpus. We therefore select three annotations from GUM and investigate whether they are helpful for coreference resolution on OntoGUM: nested mention span detection, entity type, and information status.

**Mention detection** As outlined in Section 1, we integrate gold nested mentions, including singletons (sg), into our model to improve mention detection and coreference. The task aims to recognize meaningful referential text spans and makes more information available to the model than the plain e2e approach that only trains on coreferring mentions (~39% of mentions in GUM are singletons).

**Entity type** GUM assigns one of ten entity types (ent) to each mention – person, organization, etc. (see Figure 2 in Appendix D). Since a cluster usually has one entity type, this feature instructs the model regarding which mentions belong to the same semantic class.

**Information status** Information status (infs.) indicates how an entity was introduced into discourse, e.g. new, previously mentioned or inferrable from other mentions (Prince, 1981). Each mention is assigned one of six labels (see Appendix C). This task is expected to inform the model about the likelihood and how an entity was previously introduced.

## 4 Experiments

### 4.1 Datasets

OntoGUM (Zhu et al., 2021) is a coreference dataset following the same annotation scheme as OntoNotes. This paper adds other layers to the coreference annotation, such as mention spans (including singletons), aligned entity types, and information status, automatically extracted from the GUM corpus. We train the model with GUM v8.0, which includes 193 documents across 12 written and spoken genres with ~180K tokens.

We also evaluate our model on two OOD datasets of the same annotation scheme: OntoNotes and WikiCoref. OntoNotes includes richly annotated documents with layers including syntax, propositions, named entities, word senses, and coreference, but no singleton mentions or aligned (non-named) entity types (Pradhan et al., 2013). Its test set includes 348 documents with 170K tokens. WikiCoref (Ghaddar and Langlais, 2016) is a manually annotated corpus from English Wikipedia, containing 30 documents with ~60K tokens.

### 4.2 Baseline

Combining the e2e approach with a contextualized language model (LM) and span masking is one of the best models on OntoNotes. Following Joshi

et al. (2020), we use large SpanBERT embeddings as the LM and the improved coarse-to-fine (Lee et al., 2018) SOTA model as our baseline model (see Appendix B for implementation details).

### 4.3 Task Weights

The task weights are a list of parameters that controls the relative importance of various tasks in our model, which are optimized via hyperparameter search on the OntoGUM dev set to achieve the best performance. In the optimal setting with 2 auxiliary tasks, the loss weight for the major task coreference relation identification is set to 0.4 and the weights for singleton detection and entity type recognition are set to 0.2 each. The weights are 0.15 for each auxiliary task when information status is added to training.

### 4.4 Results

**In-domain Evaluation** We train the model on OntoGUM and evaluate it in-domain. As shown in the first part of Table 1, our model with the best setting improves average F1 by 2.7 points and achieves new SOTA performance on the OntoGUM benchmark, indicating the benefit of the MTL tasks. We also note that recall scores of both mention detection and coreference matching show a significant increase by 3.2 and 4.0 points, respectively, which suggests that the MTL approach helps the model capture more non-trivial markable spans and coreference relations than the baseline model, with little or no precision cost. In addition, though information status contributes to the result as a sole auxiliary task (see Table 2), it is harmful when training with other tasks.

**Out-of-domain Evaluation** To test the robustness of our model, we evaluate on two OOD datasets sharing the same annotation scheme with OntoGUM. The second part of Table 1 shows that our best in-domain model with mention detection and entity type as auxiliary tasks outperforms the baseline model on both datasets by 2.3 points on average. For OntoNotes, though our model has slightly lower precision, the recall results in substantially better performance; for WikiCoref, our model performs better on both precision and recall. These results indicate that the knowledge gained from the multiple mention-based tasks can be transferred to unseen text types, and is likely a combination of more training data (since singletons include instances not considered by the baseline

training) and the learning of features distinguishing non-mentions from mentions and ones corresponding to semantic types.

### 4.5 Ablation Study

To show the importance of each task in our model, we ablate each task in the architecture and report the average F1 on the OntoGUM development set. In Table 2, singleton scores and the mention detection task contribute 1.3 points to the final result, indicating that this feature is the most important one.

	Avg. F1	$\Delta$
Base model	67.0	
w/ singleton detection (=sg)	68.3	+1.3
w/ sg + entity type (=et)	68.7	+0.4
w/ sg + et + information status	67.8	-0.9

Table 2: Comparison of various tasks included in the coreference model on the OntoGUM development data.

With the addition of the nested entity type recognition task, the model brings a smaller increase (0.4 points) to the final result. There could be several reasons for this: one is that the LM has already learned entity types latently, so giving this as an explicit feature is redundant; the other reason is that the baseline model rarely groups mentions with different entity types into clusters so that entity type features can only correct few errors.

When only integrating information status into the model, the result (avg. F1 67.6) outperforms the baseline model, showing the effectiveness of this type of information. However, when all three tasks are incorporated, the overall score (67.8) is lower than excluding information status classification (68.7), which shows that information status is redundant when other mention-based features are specified.

## 5 Error analysis

We conduct quantitative and qualitative error analyses to illustrate how our model differs from the baseline. Firstly we conduct a quantitative analysis following Lu and Ng (2020), who classify resolution errors into 13 classes. Following their approach, we merge coreference errors into 6 groups. Table 3 displays the distribution of errors observed in the OntoGUM development set. These errors are present in the baseline e2e model but correctly resolved by our proposed MTL model (e2e errors) or vice versa (mtl errors).



Error type	mtl errors		e2e errors	
Pronouns				
- 1st & 2nd person pronouns	6	3.6%	12	5.0%
- 3rd person pronouns	20	12.1%	68	28.3%
Definiteness				
- Definite nouns	63	38.2%	98	40.8%
- Indefinite nouns	13	7.9%	13	5.4%
Proper nouns	23	14.0%	19	7.9%
Others	40	24.2%	30	12.5%
Total	165	100.0%	240	100.0%

Table 3: Number and percentage of errors by class that are produced by e2e but avoided by the MTL model (e2e errors) and produced by the MTL model but resolved by the e2e model (mtl errors).

The majority of mtl errors involve definite nominals, revealing the challenge of resolving cherry-picked cases that must be memorized within a multi-genre context. However, our proposed model demonstrates its ability to correctly identify relations when multiple clusters are involved. Furthermore, nearly 16% of resolved errors are associated with pronouns, indicating that our model is more capable of accurately identifying coreference relationships within the context of third-person pronouns and demonstrates a slight improvement in handling pronouns in dialogue, particularly first and second-person pronouns.

We also observe that our proposed model reduces errors across nearly all types compared to the baseline model, particularly in the case of third-person pronouns. This result suggests that integrating entity type recognition and mention detection in the MTL framework enables accurate recognition of noun-pronoun relations, particularly for pronouns that do not provide explicit entity type information, e.g., *it*. Additionally, the MTL model demonstrates improved error avoidance with definite nouns. These findings highlight the enhanced performance of our proposed model in identifying coreference relations within the local context.

We also identify several errors that illustrate the impact of singleton detection and entity type recognition. Examples in Table 4 demonstrate how including singletons and mention-based features improves the retrieval of accurate mention spans and enhances coreference relationships. The first three examples highlight how entity-type recognition contributes to resolution by avoiding type mismatches. In example (1), the pressure from entity type recognition likely aids in identifying *Harrow* as a school (an ORGANIZATION). In example (2), the MTL model recognizes *it* as an EVENT, thereby correctly creating two distinct groups and

#### Entity type errors

- 1 he did represent [**the school**]<sub>1</sub> during the very first Eton v [**Harrow**]<sub>1</sub> cricket match
- 2 Who cut [**the grass**]<sub>1</sub>? Marlena did [**it**]<sub>2</sub>. Marlena did [**it**]<sub>2</sub> a long time ago, but [**it**]<sub>1</sub> hasn't been watered. [**It**]<sub>1</sub>'s dying.
- 3 I made [**noises**]<sub>1</sub> with **my heels** but [**they**]<sub>1</sub> were too loud so I stopped.

#### Singleton errors

- 4 The main reason attributed for the pollution of Athens is because the city is enclosed by mountains in [**a basin which does not let the smog leave**]<sub>1</sub> ... have greatly contributed to better atmospheric conditions in [**the basin**]<sub>1</sub>.
- 5 This means that if [**the govt**]<sub>1</sub> decided to print 1 quadrillion dollars in the span of a week ... we're loaning [**the US govt**]<sub>1</sub> **the very money it prints**

Table 4: A qualitative analysis of OntoGUM dev errors that appear in the e2e model but are avoided by our MTL model. MTL predictions (gold) are represented by [brackets]<sub>x</sub>. E2e predictions (errors) are highlighted in colored text and each color in an example denotes a coreference cluster.

avoiding coreference with *the grass* (a PLANT entity). Similarly, example (3) presents pressure to recognize that *they* is not an inanimate OBJECT, so it correctly prefers *noises* as the antecedent. Examples (4) and (5) illustrate how mention detection identifies missing mentions in the baseline model or improves boundary recognition. These representative examples provide valuable insights into the significance of incorporating singletons and auxiliary mention-based tasks into a coreference model.

## 6 Conclusion

This paper presents a neural coreference model that connects singletons and other mention-based features to coreference relation matching via an MTL architecture, which (1) outperforms a strong baseline and achieves new SOTA results on OntoGUM and (2) beats the baseline model on two unseen datasets. The results show the effect of singletons and mention features and indicate improvements in model robustness when transferring to unseen data rather than overfitting distributions in the training data. In addition, our resulting system can output all mentions (incl. singletons) with entity types out-of-the-box, which benefits a series of downstream applications such as Entity Linking, Dialogue Systems, Machine Translation, Summarization, and more, since our single model already outputs typed spans for all entities mentioned in a text (see Figure 2b in Appendix D for an illustration).

## Limitations

In this work, we have experimented with training our model on OntoGUM. Due to the lack of singletons and other mention-based annotations, we do not train the model on the most frequently used and one of the largest coreference datasets. Thus the proposed model has not been tested on a large-scale dataset and compared with other coreference models on OntoNotes.

We evaluate the model on two English OOD datasets to investigate the model generalization. Several coreference datasets in other languages share the same annotation scheme as OntoGUM, such as Arabic (Pradhan et al., 2013), and Chinese (Pradhan et al., 2013). The proposed model needs to be evaluated on datasets in other languages and demonstrate the model generalization across languages. However, this would require singleton annotated data in those languages as well. With recent releases such as CorefUD (Nedoluzhko et al., 2022) promoting standardization of multilingual coreference annotations and singleton annotations, we are hopeful that such experiments will be possible in the near future.

## References

- Rich Caruana. 1997. [Multitask learning](#). *Mach. Learn.*, 28(1):41–75.
- Ronan Collobert and Jason Weston. 2008. [A unified architecture for natural language processing: Deep neural networks with multitask learning](#). In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, pages 160—167, New York, NY, USA. Association for Computing Machinery.
- Vladimir Dobrovolskii. 2021. [Word-level coreference resolution](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Abbas Ghaddar and Phillippe Langlais. 2016. [Wiki-Coref: An English coreference-annotated corpus of Wikipedia articles](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).
- 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.
- Yufang Hou. 2020. [Bridging anaphora resolution as question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. [SpanBERT: Improving pre-training by representing and predicting spans](#). *CoRR*, abs/1907.10529.
- 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.
- Ben Kantor and Amir Globerson. 2019. [Coreference resolution with entity equalization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.
- Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022. [Constrained multi-task learning for bridging resolution](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 759–770, Dublin, Ireland. Association for Computational Linguistics.
- Sandra Kübler and Desislava Zhekova. 2011. [Singletons and coreference resolution evaluation](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria. Association for Computational Linguistics.
- 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, Copenhagen, Denmark. 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.
- Jing Lu and Vincent Ng. 2020. [Conundrums in entity coreference resolution: Making sense of the state of the art](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.
- Jing Lu and Vincent Ng. 2021. [Constrained multi-task learning for event coreference resolution](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

- 4504–4514, Online. Association for Computational Linguistics.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.
- Nafise Sadat Moosavi and Michael Strube. 2017. [Lexical features in coreference resolution: To be used with caution](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.
- Nafise Sadat Moosavi and Michael Strube. 2018. [Using linguistic features to improve the generalization capability of neural coreference resolvers](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Brussels, Belgium. Association for Computational Linguistics.
- Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. [CorefUD 1.0: Coreference meets Universal Dependencies](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. [Towards robust linguistic analysis using OntoNotes](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.
- Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Syntax and semantics: Vol. 14. Radical Pragmatics*, pages 223–255. Academic Press, New York.
- Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. [Bridging resolution: Task definition, corpus resources and rule-based experiments](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Sanjay Subramanian and Dan Roth. 2019. [Improving generalization in coreference resolution via adversarial training](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. [On generalization in coreference resolution](#). In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.
- Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. [CorefQA: Coreference resolution as query-based span prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.
- Zhaofeng Wu and Matt Gardner. 2021. [Understanding mention detector-linker interaction in neural coreference resolution](#). In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 150–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Juntao Yu and Massimo Poesio. 2020. [Multitask learning-based neural bridging reference resolution](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Amir Zeldes. 2017. [The GUM corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.
- Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. [OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres](#). 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 461–467, Online. Association for Computational Linguistics.

## A Model architecture overview

Figure 1 shows the architecture of the model proposed in this paper.

## B Implementation details

We use Pytorch and the pre-trained SpanBERT-large (Joshi et al., 2020) model from HuggingFace<sup>2</sup> for token representations. Experiments run on Nvidia RTX A6000 GPUs with 64GB RAM. Following previous work (Lee et al., 2018; Joshi

<sup>2</sup><https://huggingface.co/>

et al., 2020), we use a batch size of 1 document for training and evaluation. The coreference task uses the same loss strategy as the baseline model (Joshi et al., 2020) and each auxiliary task uses Cross Entropy loss. We use AdamW to optimize coreference loss and Adam to optimize auxiliary loss. We train 14,500 steps with `task_learning_rate` of 0.0003 for baselines and our models.

## C Information Status

There are six types of information status in the data:


- *new* (first, unmediated mention of an entity)
- *given:active* (subsequent mention after a recent previous mention)
- *given:inactive* (subsequent mention of a non-recently mentioned entity)
- *accessible:inferrable* (new entity whose existence could be inferred from other mentions, e.g. via bridging anaphora (Roesiger et al., 2018; Hou, 2020), as in *a house ... [the door]*)
- *accessible:commonground* (entities accessible to speakers in the situation, e.g. *pass [the salt]!*)
- *accessible:aggregate* (new entities referring back to multiple entities, i.e. split antecedents as in *Kim ... Yun ... [they]*).

When information status is included in the auxiliary tasks, our model is trained to predict the label for each mention.

## D Sample data

Figure 2b shows the extent of annotations available to the MTL model for training, compared to the data restricted to coreferring pairs in Figure 2a, as used by the baseline e2e approach. Since OntoNotes-style data, such as OntoGUM, does not contain singletons, mention types or information status, only coreferring mentions and their spans can be used for learning by the baseline model. The information in the bottom panel, by contrast, is much richer and covers all referring expressions with information status and one of ten entity types: ABSTRACT, ANIMAL, EVENT, OBJECT, ORGANIZATION, PERSON, PLACE, PLANT, SUBSTANCE and TIME.

While each information type (coreference, mention boundaries, mention types, and information status) is not totally predictable from others, they

overlap to some extent and exhibit different information densities: mention boundaries are available for many spans and are densely attested. Mention types are available for each mention, but some types are rare, e.g. abstract mentions marked by  in Figure 2 are the most common. Information status is mostly predictable from coreference, e.g. singletons and chain-initial mentions are *new*, and chain-medial or final mentions are given (*given:active* if recently mentioned, otherwise *given:inactive*). Accessible mentions are less trivial and comparatively rare (about 7.1% of mentions in GUM), indicating whether they are accessible in the common ground, their identity is inferrable from some other mention (*accessible:inferrable*), or by aggregating information from multiple mentions (*accessible:aggregate*). This information could help systems to learn whether a span is likely to have an antecedent.



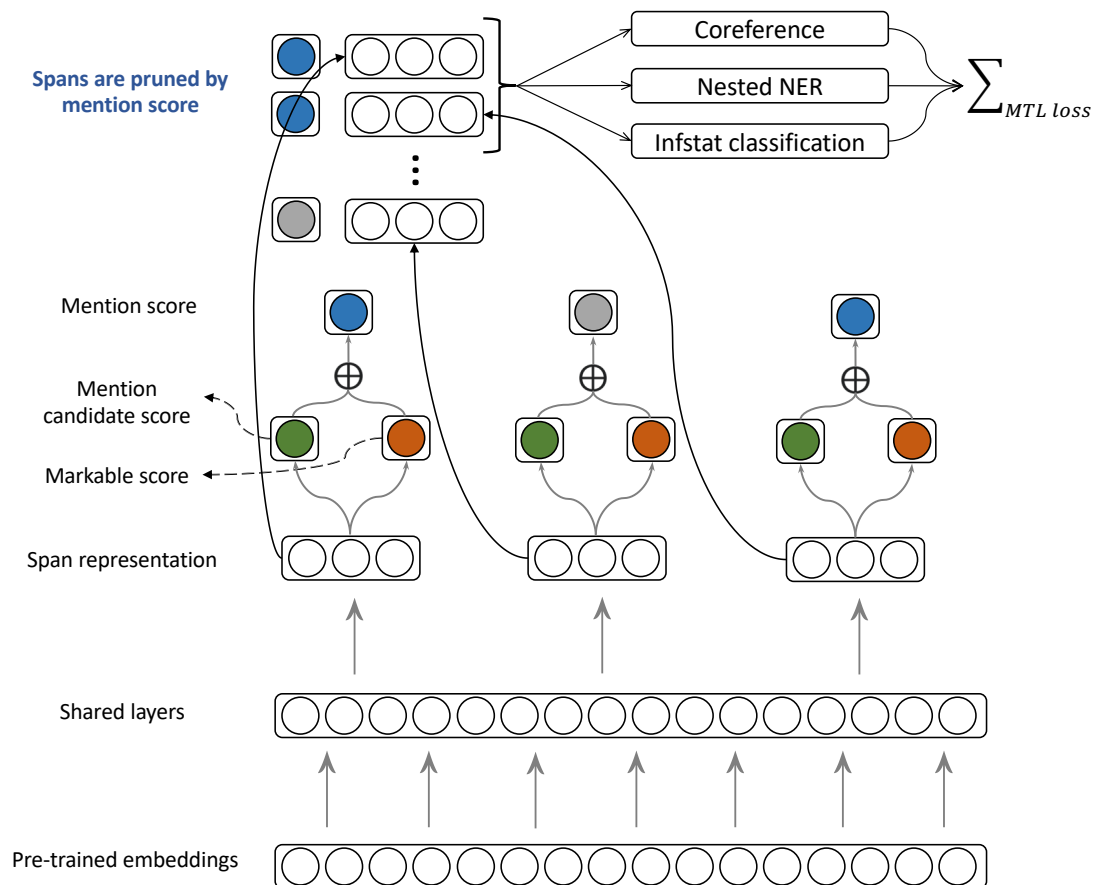


Figure 1: An overview of the proposed MTL model architecture. Only selected spans with high mention scores (in blue) are considered in the three auxiliary tasks.

New Zealand begins process to consider changing national flag design Thursday May 7, 2015

On Tuesday, the New Zealand government announced the start of a public process to suggest designs for a new national flag, and determine whether their citizens would prefer a different national flag over the current one.

The current flag of New Zealand. The current New Zealand flag is partially based on the United Kingdom's flag; the new one would be unique to New Zealand. The government's Flag Consideration Project has planned a number of conferences and roadshows as part of this process...

(a) Information available to the baseline e2e model.

New Zealand begins process to consider changing national flag design

Thursday May 7, 2015 On Tuesday, the New Zealand government announced the start of a public process to suggest designs for a new national flag, and determine whether their citizens would prefer a different national flag over the current one.

The current flag of New Zealand. The current New Zealand flag is partially based on the United Kingdom's flag; the new one would be unique to New Zealand. The government's Flag Consideration Project has planned a number of conferences and roadshows as part of this process, with the first meeting set to take place in Christchurch

★ Information status: accessible-inferrable	👤 ABSTRACT	👤 PERSON
★ Information status: accessible-commonground	🐾 ANIMAL	📍 PLACE
📦 Co-referring mention group (by color)	🔔 EVENT	🌱 PLANT
🔗 Opposition pair (by color)	📦 OBJECT	🧪 SUBSTANCE
📦 Singletons	🏢 ORGANIZATION	🕒 TIME

(b) Information available to the MTL model for the same document.

Figure 2: Training data from an OntoGUM article in the news genre.