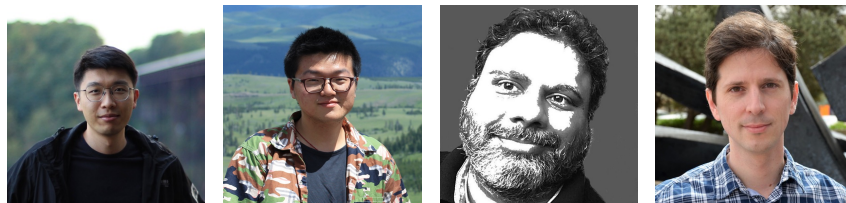


Incorporating Singletons and Mention-based Features in Coreference Resolution via Multi-task Learning for Better Generalization

Yilun Zhu¹, Siyao Peng², Sameer Pradhan^{3,4}, Amir Zeldes¹



¹ *GEORGETOWN UNIVERSITY*

²



³   **Penn**
UNIVERSITY of PENNSYLVANIA
⁴ **cemantix.org**

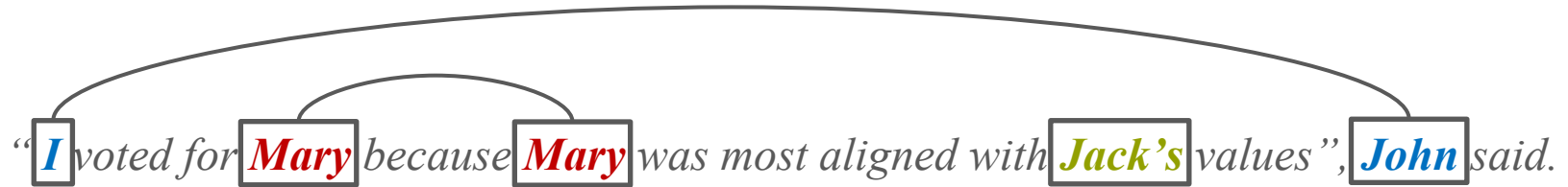
Coreference Resolution and Mentions: An Example

*“**I** voted for **Mary** because **Mary** was most aligned with **Jack’s** values”, **John** said.*

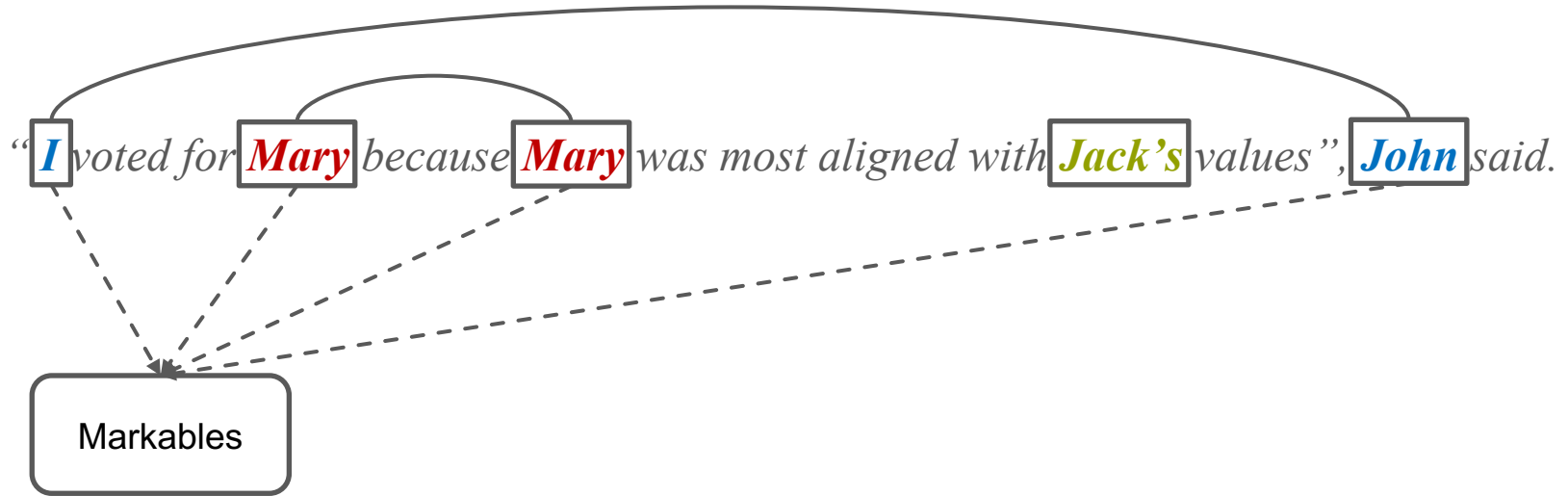
Coreference Resolution and Mentions: An Example

“**I** voted for **Mary** because **Mary** was most aligned with **Jack’s** values”, **John** said.

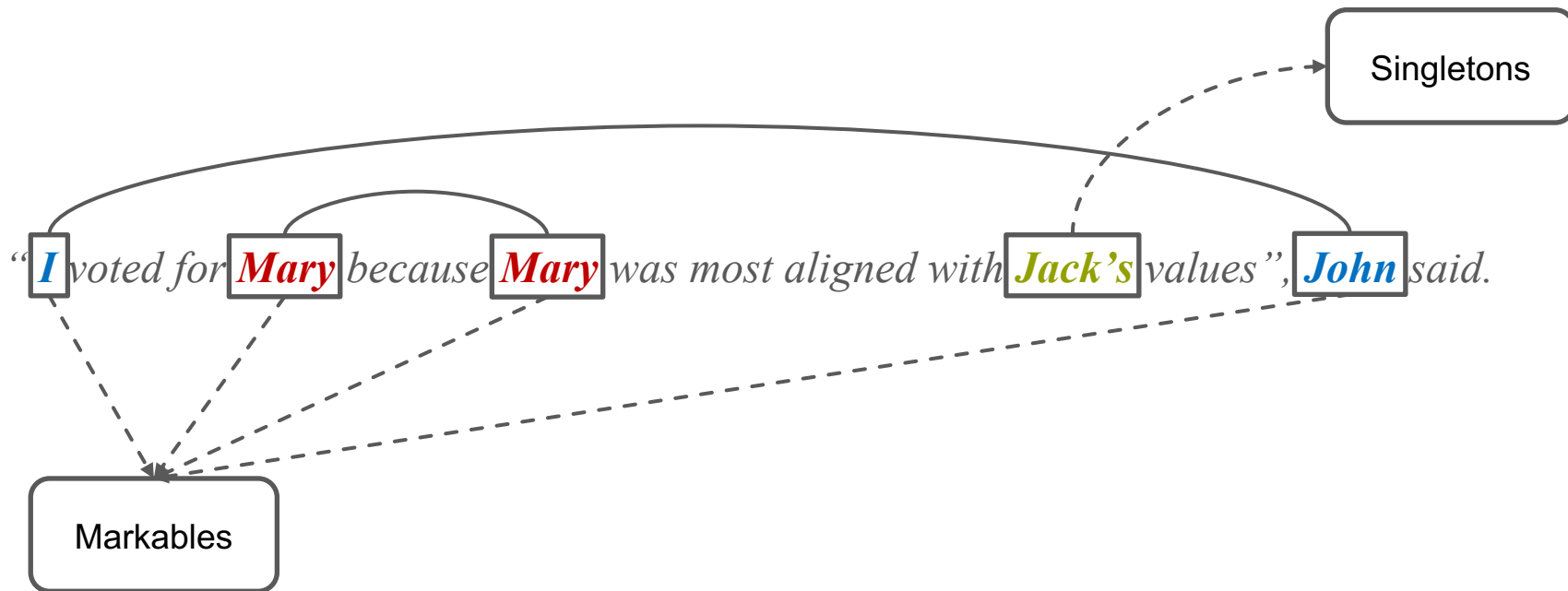
Coreference Resolution and Mentions: An Example



Coreference Resolution and Mentions: An Example



Coreference Resolution and Mentions: An Example



Why Are Singletons Important?

- How humans understand discourse from a theoretical perspective (Grosz et al., 1995)
 - Singletons constitute mentioned entities (i.e. clusters of size 1)
- Represent true negatives in cluster linking (Kübler and Zhekova, 2011)

However...

- Dataset - OntoNotes
 - Lack singleton annotation

Greek court rules worship of ancient Greek deities is legal Monday, March 27, 2006 Greek court worshippers of the ancient Greek religion may now formally associate and worship at archeological sites. Pri the religion was banned from conducting public worship at archeological sites by the Greek Ministry of the religion was relatively secretive. The Greek Orthodox Church, a Christian denomination, is extremely crit worshippers of the ancient deities. Today, about 100,000 Greeks worship the ancient gods, Poseidon, Aphrodite, and Athena. The Greek Orthodox Church estimates that number is clo Many neo-pagan religions, such as Wicca, use aspects of ancient Greek religions in their pra instead focuses exclusively on the ancient religions, as far as the fragmentary nature of the surviving source ma

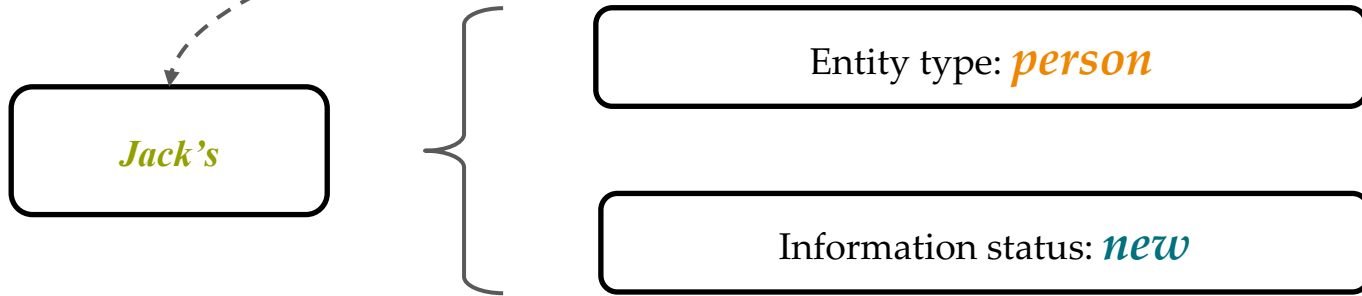
- Coreference Models

End-to-end (Lee et al., 2017, Lee et al., 2018, Joshi et al., 2020, Dobrovolskii, 2021, etc.) & *Seq2seq* (Bohnet et al., 2023)

- Models cannot differentiate singleton spans from non-referring or random/meaningless spans, thus penalizing these two types equally
- Do not align with linguistic theories on how humans resolve the task

Features Other Than Singletons?

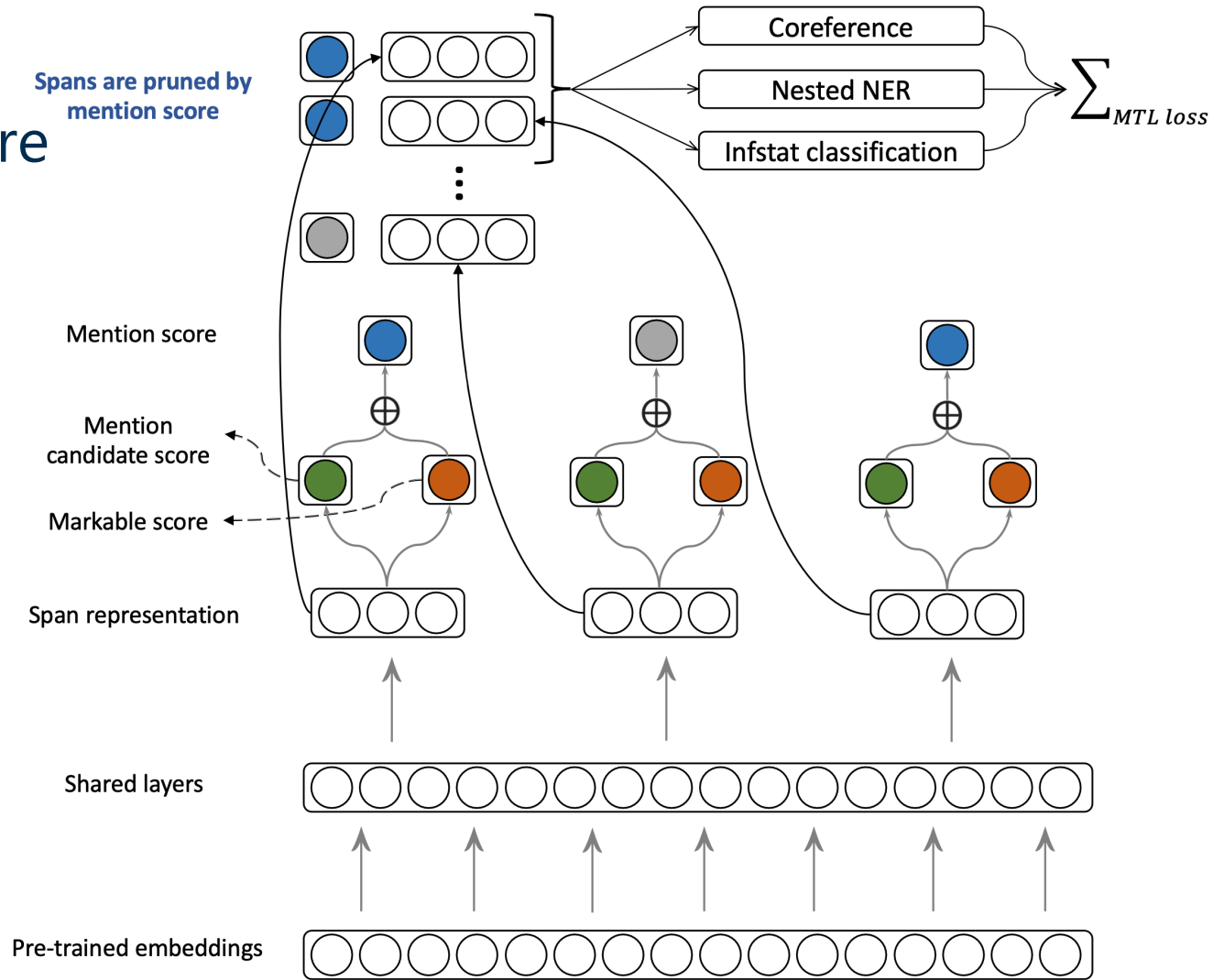
“*I* voted for *Mary* because *Mary* was most aligned with **Jack's** values”, *John* said.



Datasets

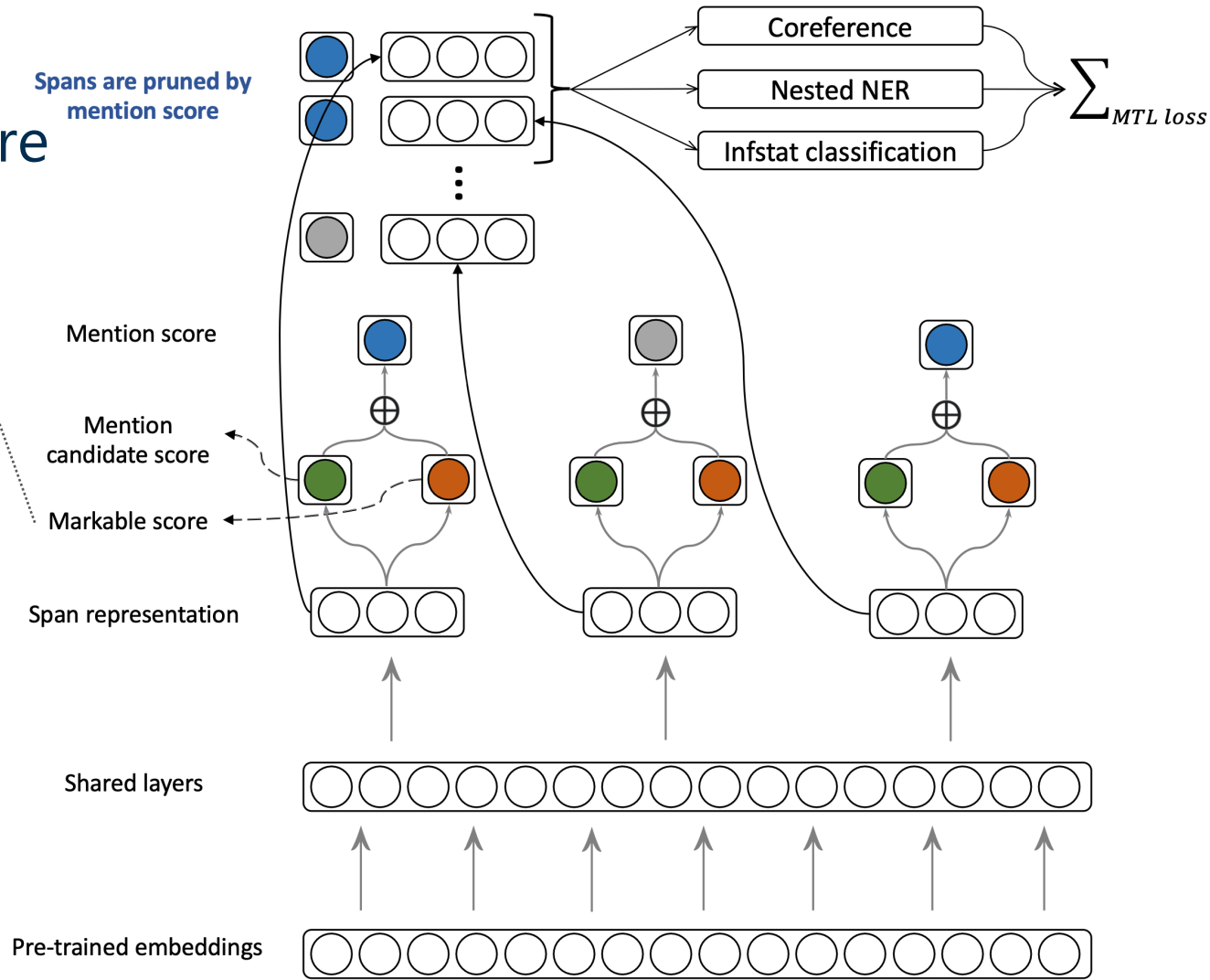
- In-domain
 - OntoGUM (Zhu et al., 2021): dataset has multi-layer annotations of
 - Singletons & Markables
 - Mention-based annotations
 - Entity types (abstract, animal, event, object, organization, person, place, plant, substance, time)
 - Information status (new, given:active, given:inactive, accessible:inferrable, accessible:commonground, accessible:aggregate)
 - Coreference relations following OntoNotes guidelines
- Out-of-domain
 - OntoNotes V5.0 (Weischedel et al., 2011; Pradhan et al., 2013)
 - WikiCoref (Ghaddar and Langlais, 2016)

Model Architecture



Model Architecture

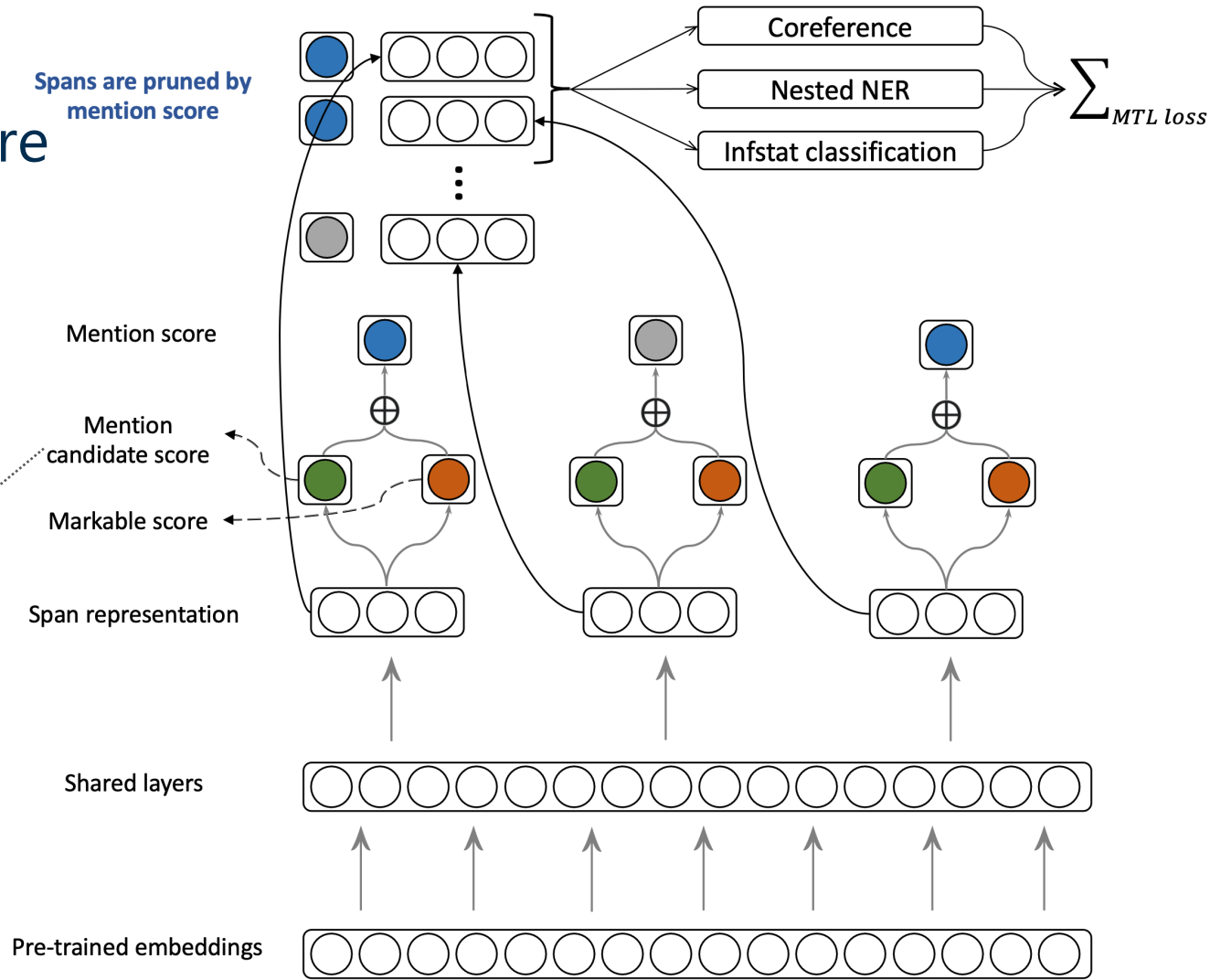
Markable score
How likely a span is coreferred in the document?



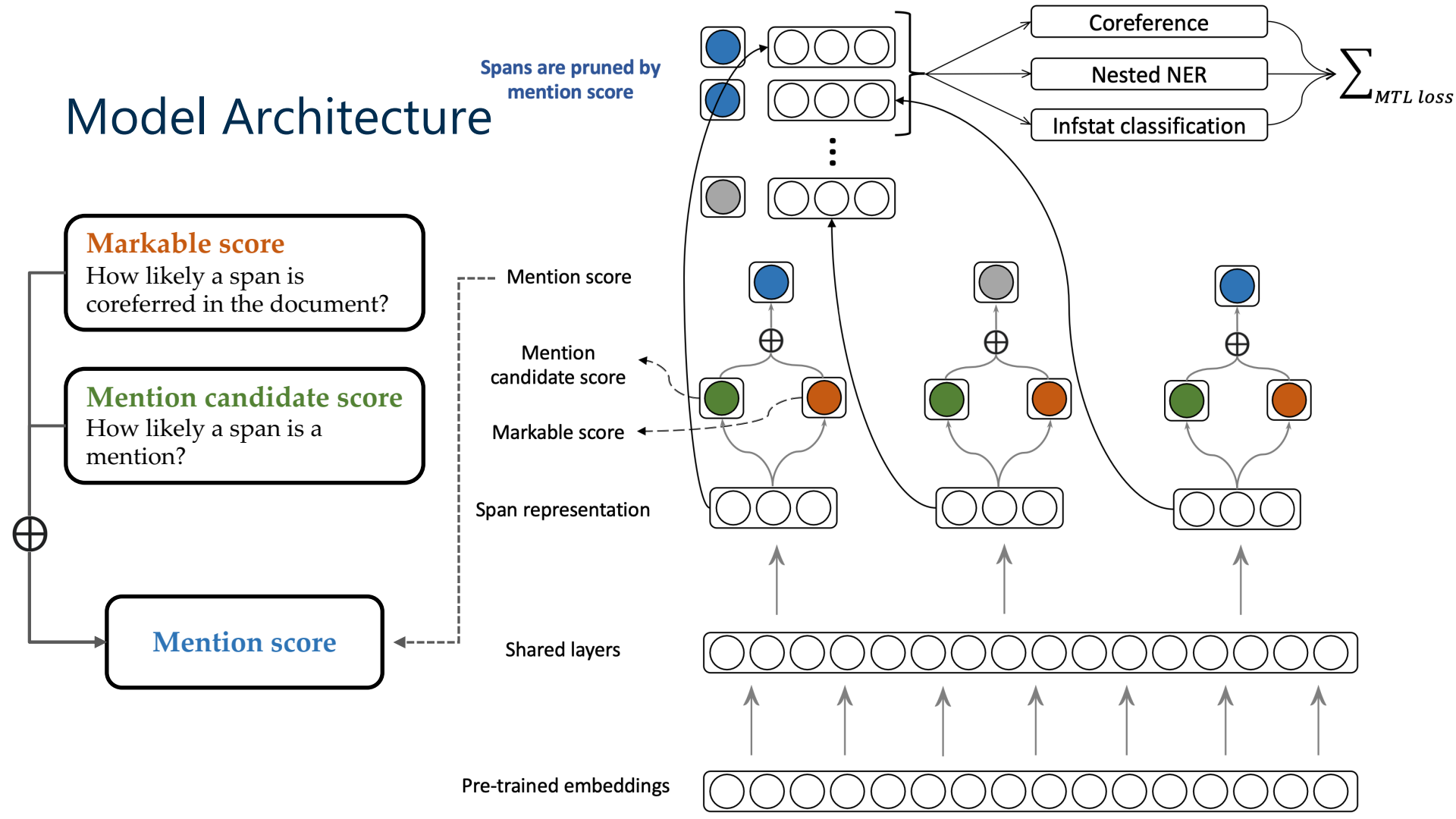
Model Architecture

Markable score
How likely a span is coreferred in the document?

Mention candidate score
How likely a span is a mention?



Model Architecture



Model Architecture

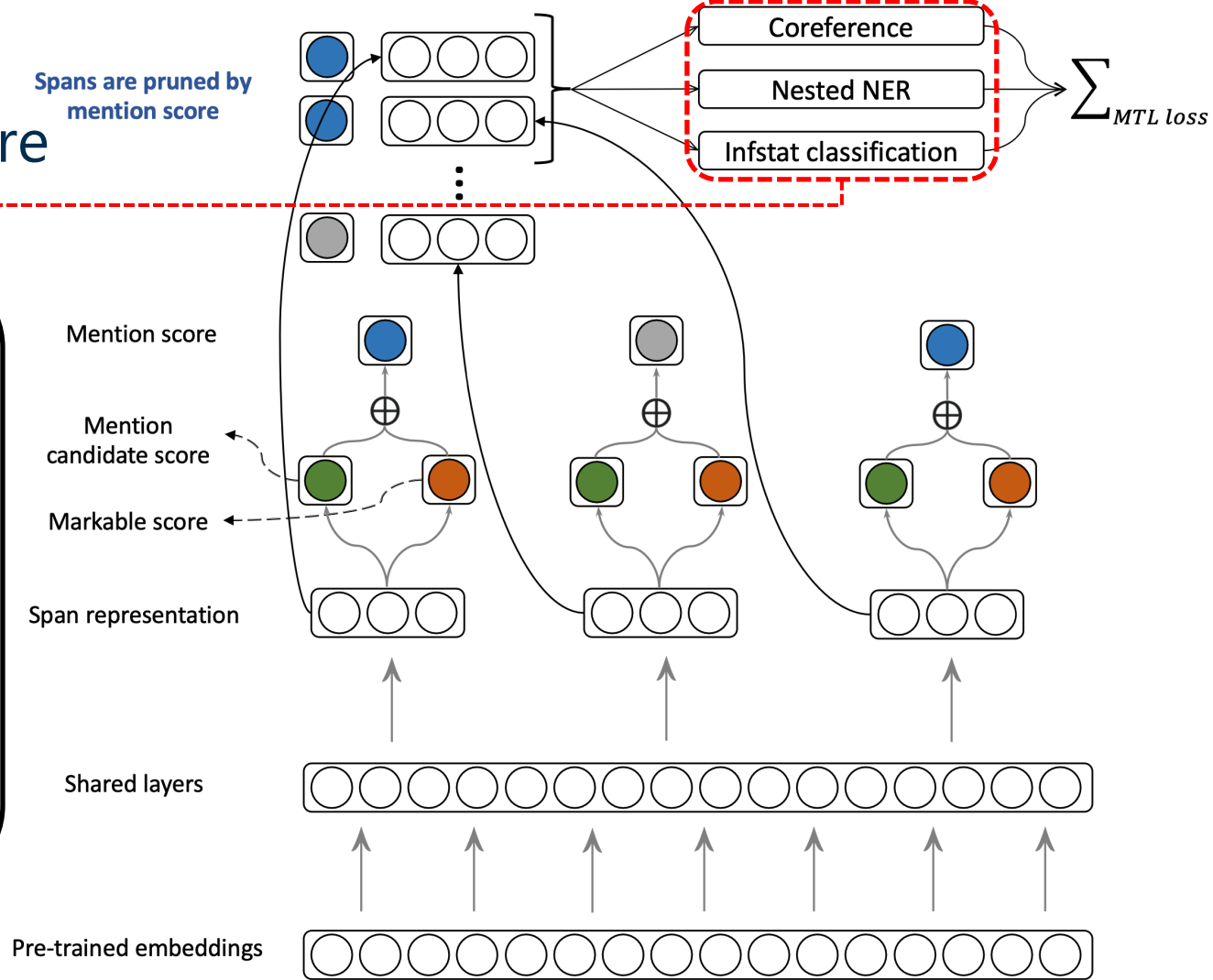
- Coreference linking

- Entity type recognition:

Assign a **non-pruned span** one of the ten types

- Infstat classification:

Assign a **non-pruned span** one of the six classes

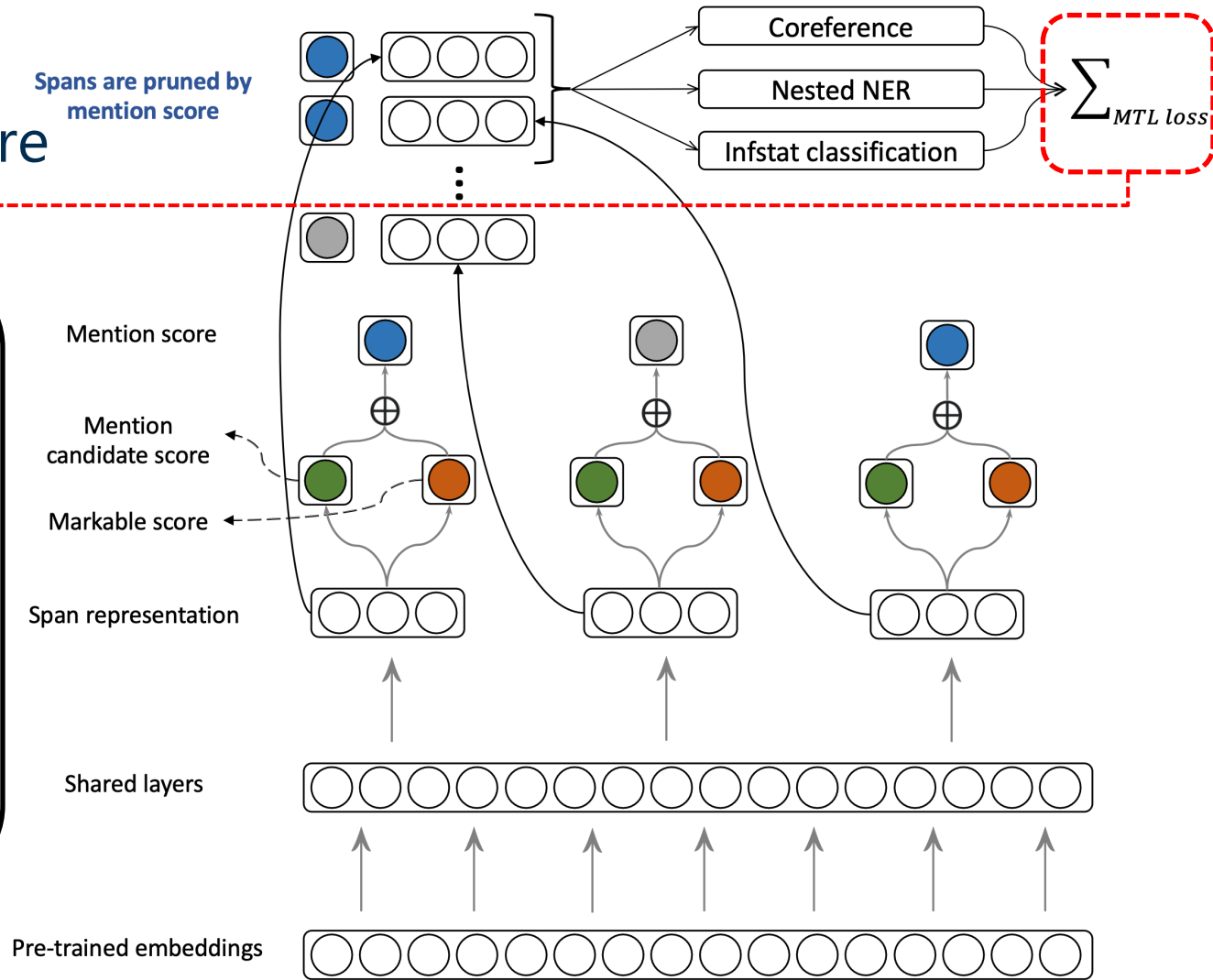


Model Architecture

- Task weights for MTL loss

$$\mathcal{L}_{total} = \sum_{t=1}^T \omega_t * \mathcal{L}_t + \mathcal{L}_{markable} + \mathcal{L}_{mention}$$

Hyperparameter tuning to find the best weights



Experiments & Results

	Markble Detection			MUC			B ³			CEAF _{ϕ_4}			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
In-domain - ONTOGUM													
Joshi et al. (2019)	91.0	71.9	80.3	83.3	69.7	75.9	70.8	59.2	64.5	70.5	45.8	55.5	65.5
MTL (sg)	90.2	75.0	81.9	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	75.1	81.9	82.8	72.9	77.6	71.2	63.6	67.2	71.9	50.2	59.1	68.2
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
Out-of-domain - ONTONOTES													
Joshi et al. (2019)	83.9	76.9	80.3	77.6	72.7	75.1	66.9	60.6	63.6	64.3	54.5	59.0	65.9
MTL (sg+ent)	82.2	80.2	81.2	77.0	76.1	76.5	67.1	64.0	65.5	63.6	59.5	61.5	67.8
Out-of-domain - WIKICOREF													
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)	80.4	60.0	68.7	74.5	61.8	67.5	67.8	45.3	54.4	59.0	33.0	42.4	55.6

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, B³, and CEAF _{ϕ_4} . All models are trained on OntoGUM.

Experiments & Results

	Markble Detection			MUC			B ³			CEAF _{ϕ_4}			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
In-domain - ONTOGUM													
Joshi et al. (2019)	91.0	71.9	80.3	83.3	69.7	75.9	70.8	59.2	64.5	70.5	45.8	55.5	65.5
MTL (sg)	90.2	75.0	81.9	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	75.1	81.9	82.8	72.9	77.6	71.2	63.6	67.2	71.9	50.2	59.1	68.2
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
Out-of-domain - ONTONOTES													
Joshi et al. (2019)	83.9	76.9	80.3	77.6	72.7	75.1	66.9	60.6	63.6	64.3	54.5	59.0	65.9
MTL (sg+ent)	82.2	80.2	81.2	77.0	76.1	76.5	67.1	64.0	65.5	63.6	59.5	61.5	67.8
Out-of-domain - WIKICOREF													
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)	80.4	60.0	68.7	74.5	61.8	67.5	67.8	45.3	54.4	59.0	33.0	42.4	55.6

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, B³, and CEAF _{ϕ_4} . All models are trained on OntoGUM.

Best setting:
MTL-sg+ent

Improve the
baseline
model by
**2.7 points in
domain**

Experiments & Results

	Markble Detection			MUC			B ³			CEAF _{ϕ_4}			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
In-domain - ONTOGUM													
Joshi et al. (2019)	91.0	71.9	80.3	83.3	69.7	75.9	70.8	59.2	64.5	70.5	45.8	55.5	65.5
MTL (sg)	90.2	75.0	81.9	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	75.1	81.9	82.8	72.9	77.6	71.2	63.6	67.2	71.9	50.2	59.1	68.2
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
Out-of-domain - ONTONOTES													
Joshi et al. (2019)	83.9	76.9	80.3	77.6	72.7	75.1	66.9	60.6	63.6	64.3	54.5	59.0	65.9
MTL (sg+ent)	82.2	80.2	81.2	77.0	76.1	76.5	67.1	64.0	65.5	63.6	59.5	61.5	67.8
Out-of-domain - WIKICOREF													
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)	80.4	60.0	68.7	74.5	61.8	67.5	67.8	45.3	54.4	59.0	33.0	42.4	55.6

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, B³, and CEAF _{ϕ_4} . All models are trained on OntoGUM.

Best setting:
MTL-sg+ent

Improve the
baseline
model by
2.3 points
out domain
by average

Expected Outputs

Joint predictions of

- Mention spans
- Coreference relations
- Entity types
- Information status

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-inferable
- ★ Information status: accessible-commonground
- Co-referring mention group (by color)
- Apposition pair (by color)
- Singletons

ABSTRACT	PERSON
ANIMAL	PLACE
EVENT	PLANT
OBJECT	SUBSTANCE
ORGANIZATION	TIME

Error Analysis

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).

Analysis following Lu and Ng (2020)

Error types	MTL vs baseline
- Pronouns	26 vs 80
- Definiteness	76 vs 111
- Proper nouns	23 vs 19
- Others	40 vs 30

Conclusion

- We propose a MTL based neural coreference model with constrained mention detection, which jointly learns several mention-based tasks
- Achieve new SOTA performance on the OntoGUM test set
- Demonstrate better generalization on two OOD datasets
- Release our code at: <https://github.com/yilunzhu/coref-mtl>

Appendix A: Ablation Study

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

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

Singletons (sg) and entity types (et)	+1.7
Sg, ent, and information status	+0.8

Appendix B: Example Errors

GOLD: by [brackets]
ERROR: in colored text

- Example 1-3
Entity-type recognition contributes to resolution
by avoiding type mismatches

- Example 4-5
Mention detection identifies missing mentions in
the baseline model or improves boundary
recognition

Entity type errors

- 1 he did represent [**the school**]₁ during the very first Eton v [**Harrow**]₁ cricket match
- 2 Who cut [**the grass**]₁? Marlina did [**it**]₂. Marlina did [**it**]₂ a long time ago, but [**it**]₁ hasn't been watered. [**It**]₁'s dying.
- 3 I made [**noises**]₁ with **my heels** but [**they**]₁ 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**]₁ ... have greatly contributed to better atmospheric conditions in [**the basin**]₁.
- 5 This means that if [**the govt**]₁ decided to print 1 quadrillion dollars in the span of a week ... we're loaning [**the US govt**]₁ **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]_x. E2e predictions (errors) are highlighted in colored text and each color in an example denotes a coreference cluster.