Recent Computational Approaches to Coreference Resolution

Milan Straka
Institute of Formal and Applied Linguistics
Charles University

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Education and early loves

Byron received his early formal education at Aberdeen Grammar School, and in August 1799 entered the school of Dr. William Glennie, in Dulwich. [17]

Placed under the care of a Dr. Bailey, he was encouraged to exercise in moderation but not restrain himself from "violent" bouts in an attempt to overcompensate for his deformed foot.

His mother interfered with his studies, often withdrawing him from school, with the result that he lacked discipline and his classical studies were neglected.

In 1801, he was sent to Harrow, where he remained until July 1805. [6]

An undistinguished student and an unskilled cricketer, he did represent the school during the very first Eton v Harrow cricket match at Lord 's in 1805. [19]

His lack of moderation was not restricted to physical exercise.

Byron fell in love with Mary Chaworth, whom he met while at school, [6] and she was the reason he refused to return to Harrow in September 1803.

His mother wrote, " He has no indisposition that I know of but love, desperate love, the worst of all maladies in my opinion. In short, the boy is distractedly in love with Miss Chaworth." [6]

In Byron 's later memoirs, " Mary Chaworth is portrayed as the first object of his adult sexual feelings." [20]

Byron finally returned in January 1804, [6] to a more settled period which saw the formation of a circle of emotional involvements with other Harrow boys, which he recalled with great vividness: " My school friendships were with me passions (for I was always violent)." [21]

Model Zoo

e2e: End-to-end Neural Coreference Resolution

Lee et al. (2017)

e2e: End-to-end Neural Coreference Resolution

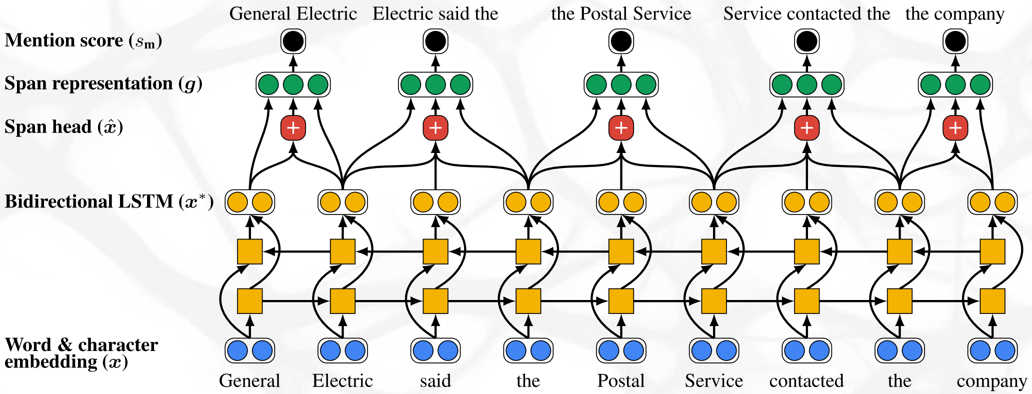


Figure 1 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

- Every possible span considers all preceding spans and ε as antecedents.

e2e: End-to-end Neural Coreference Resolution

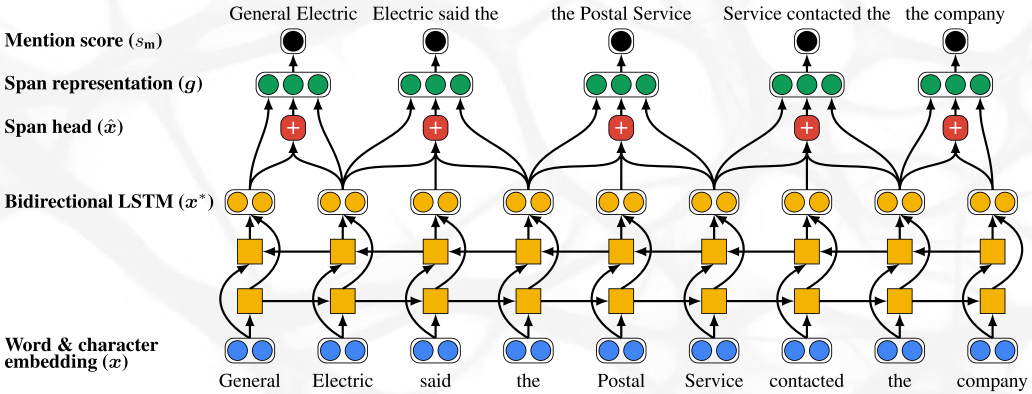


Figure 1 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

- Every possible span considers all preceding spans and ε as antecedents.
- For a span $i = (\text{start}(i), \text{end}(i))$, the score of span j being an antecedent of span i is computed as

$$s(i, j) = \begin{cases} 0 \text{ if } j = \varepsilon, \\ s_m(i) + s_m(j) + s_a(i, j) \text{ otherwise.} \end{cases}$$

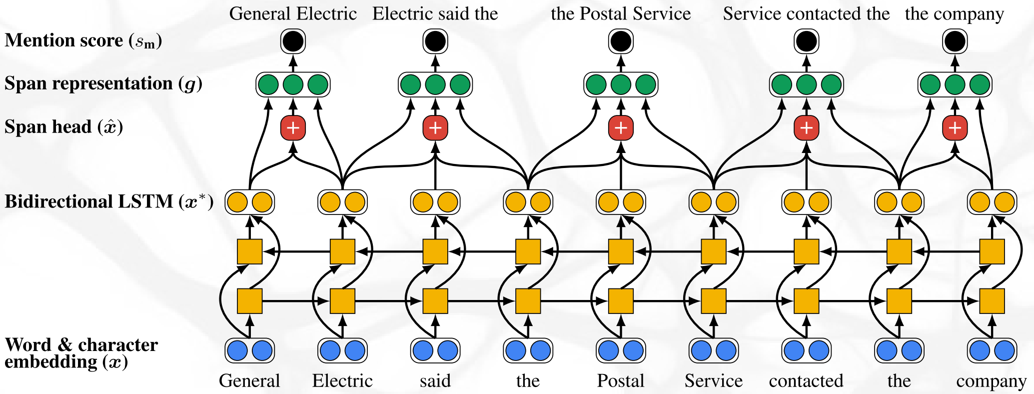


Figure 1 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

- Span is represented as

$$\boldsymbol{g}_i = \left[\boldsymbol{x}_{\text{start}(i)}, \boldsymbol{x}_{\text{end}(i)}, \textit{soft head} \sum\nolimits_{t=\text{start}(i)}^{\text{end}(i)} \alpha_t \boldsymbol{x}_t, \textit{span features } \varphi(i)\right].$$

e2e: End-to-end Neural Coreference Resolution

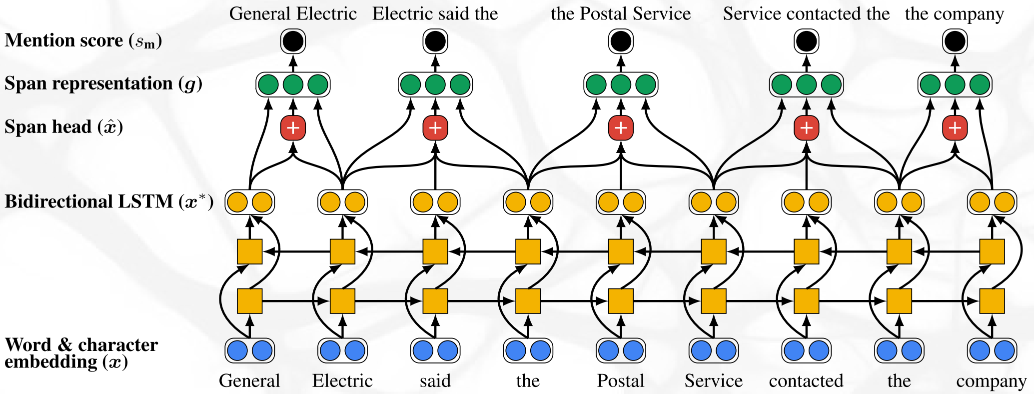


Figure 1 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

- Span is represented as

$$\boldsymbol{g}_i = \left[\boldsymbol{x}_{\text{start}(i)}, \boldsymbol{x}_{\text{end}(i)}, \textit{soft head } \sum\nolimits_{t=\text{start}(i)}^{\text{end}(i)} \alpha_t \boldsymbol{x}_t, \textit{span features } \varphi(i)\right].$$

- Mention score $s_m(i) = f_m(\boldsymbol{g}(i))$,

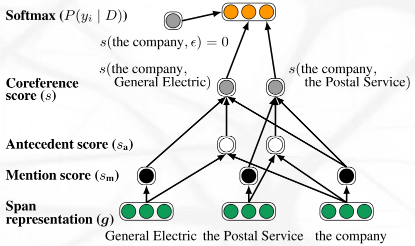


Figure 2 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

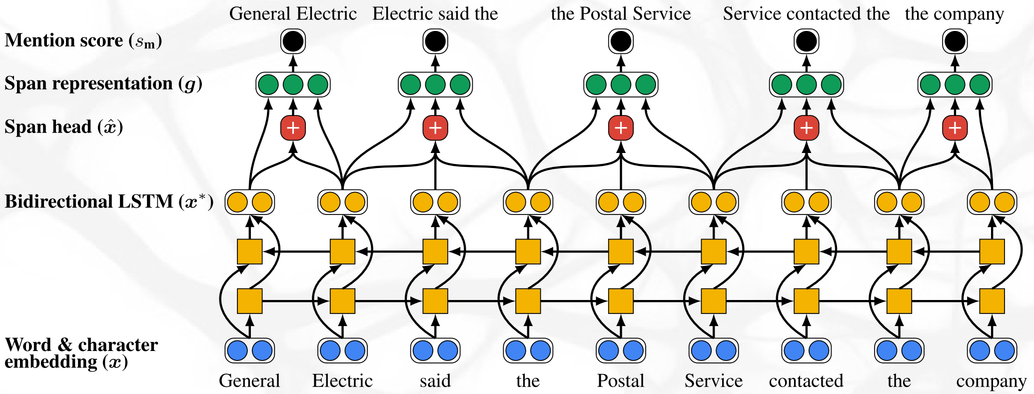


Figure 1 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

- Span is represented as

$$\boldsymbol{g}_i = \left[\boldsymbol{x}_{\text{start}(i)}, \boldsymbol{x}_{\text{end}(i)}, \textit{soft head} \sum\nolimits_{t=\text{start}(i)}^{\text{end}(i)} \alpha_t \boldsymbol{x}_t, \textit{span features } \varphi(i)\right].$$

- Mention score $s_m(i) = f_m(\boldsymbol{g}(i))$,
- antecedent score $s_a(i, j) = f_a([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \odot \boldsymbol{g}_j, \varphi(i, j)])$.

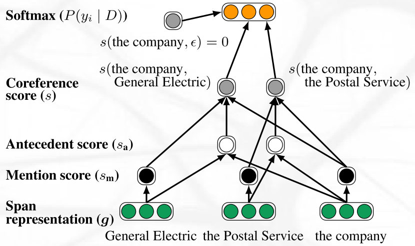


Figure 2 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

e2e: End-to-end Neural Coreference Resolution

- However, there are up to $\mathcal{O}(n^4)$ span-span combinations.

e2e: End-to-end Neural Coreference Resolution

- However, there are up to $\mathcal{O}(n^4)$ span-span combinations.
  - consider spans to a maximum length $L = 10$;

e2e: End-to-end Neural Coreference Resolution

- However, there are up to $\mathcal{O}(n^4)$ span-span combinations.
  - consider spans to a maximum length $L = 10$;
  - keep only λn spans for $\lambda = 0.4$ with maximum $s_m(i)$;

e2e: End-to-end Neural Coreference Resolution

- However, there are up to $\mathcal{O}(n^4)$ span-span combinations.
  - consider spans to a maximum length $L = 10$;
  - keep only λn spans for $\lambda = 0.4$ with maximum $s_m(i)$;
  - for each span, consider up to $K = 250$ nearest mentions.

- However, there are up to $\mathcal{O}(n^4)$ span-span combinations.
  - consider spans to a maximum length $L = 10$;
  - keep only λn spans for $\lambda = 0.4$ with maximum $s_m(i)$;
  - for each span, consider up to $K = 250$ nearest mentions.

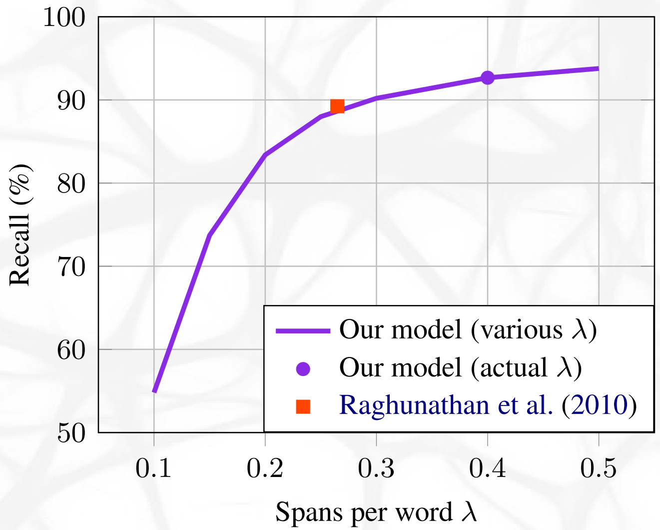


Figure 3 of "End-to-end Neural Coreference Resolution", Lee et al. (2017)

Model Results

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM
Lee et al. (2017)	e2e	$67.2_{\varnothing}$

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM
Lee et al. (2017)	e2e	$67.2_{\varnothing}$
Lee et al. (2018)	e2e	70.4_{ELMo}

c2f: Higher-order Coreference Resolution with Coarse-to-fine Inference

Lee et al. (2018)

c2f: Higher-order CR with Coarse-to-fine Inference

- Scoring function is extended by assing $s_c(i,j)$:

$$s(i,j) = \begin{cases} 0 \text{ if } j = \varepsilon, \\ s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j) \text{ otherwise,} \end{cases}$$

c2f: Higher-order CR with Coarse-to-fine Inference

- Scoring function is extended by assing $s_c(i,j)$:

$$s(i,j) = \begin{cases} 0 \ \text{if} \ j = \varepsilon, \\ s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j) \ \text{otherwise}, \end{cases}$$

where

$$s_c(i,j) = \boldsymbol{g}_i^T \boldsymbol{W}_c \boldsymbol{g}_j \approx (\boldsymbol{W}_q \boldsymbol{g}_i)^T (\boldsymbol{W}_k \boldsymbol{g}_j).$$

c2f: Higher-order CR with Coarse-to-fine Inference

- Scoring function is extended by assing $s_c(i,j)$:

$$s(i,j) = \begin{cases} 0 \ \text{ if } \ j = \varepsilon, \\ s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j) \ \text{ otherwise,} \end{cases}$$

where

$$s_c(i,j) = \boldsymbol{g}_i^T \boldsymbol{W}_c \boldsymbol{g}_j \approx (\boldsymbol{W}_q \boldsymbol{g}_i)^T (\boldsymbol{W}_k \boldsymbol{g}_j).$$

- Two-step pruning:
  1. keep λn spans with highest $s_m(i)$ and maximum length $L = 30$,

c2f: Higher-order CR with Coarse-to-fine Inference

- Scoring function is extended by assing $s_c(i,j)$:

$$s(i,j) = \begin{cases} 0 & \text{if } j = \varepsilon, \\ s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j) & \text{otherwise,} \end{cases}$$

where

$$s_c(i,j) = \boldsymbol{g}_i^T \boldsymbol{W}_c \boldsymbol{g}_j \approx (\boldsymbol{W}_q \boldsymbol{g}_i)^T (\boldsymbol{W}_k \boldsymbol{g}_j).$$

- Two-step pruning:
  1. keep λn spans with highest $s_m(i)$ and maximum length $L = 30$,
  2. keep $K = 50$ top antecedents according to $s_m(i), s_m(j), s_c(i,j)$.

c2f: Higher-order CR with Coarse-to-fine Inference

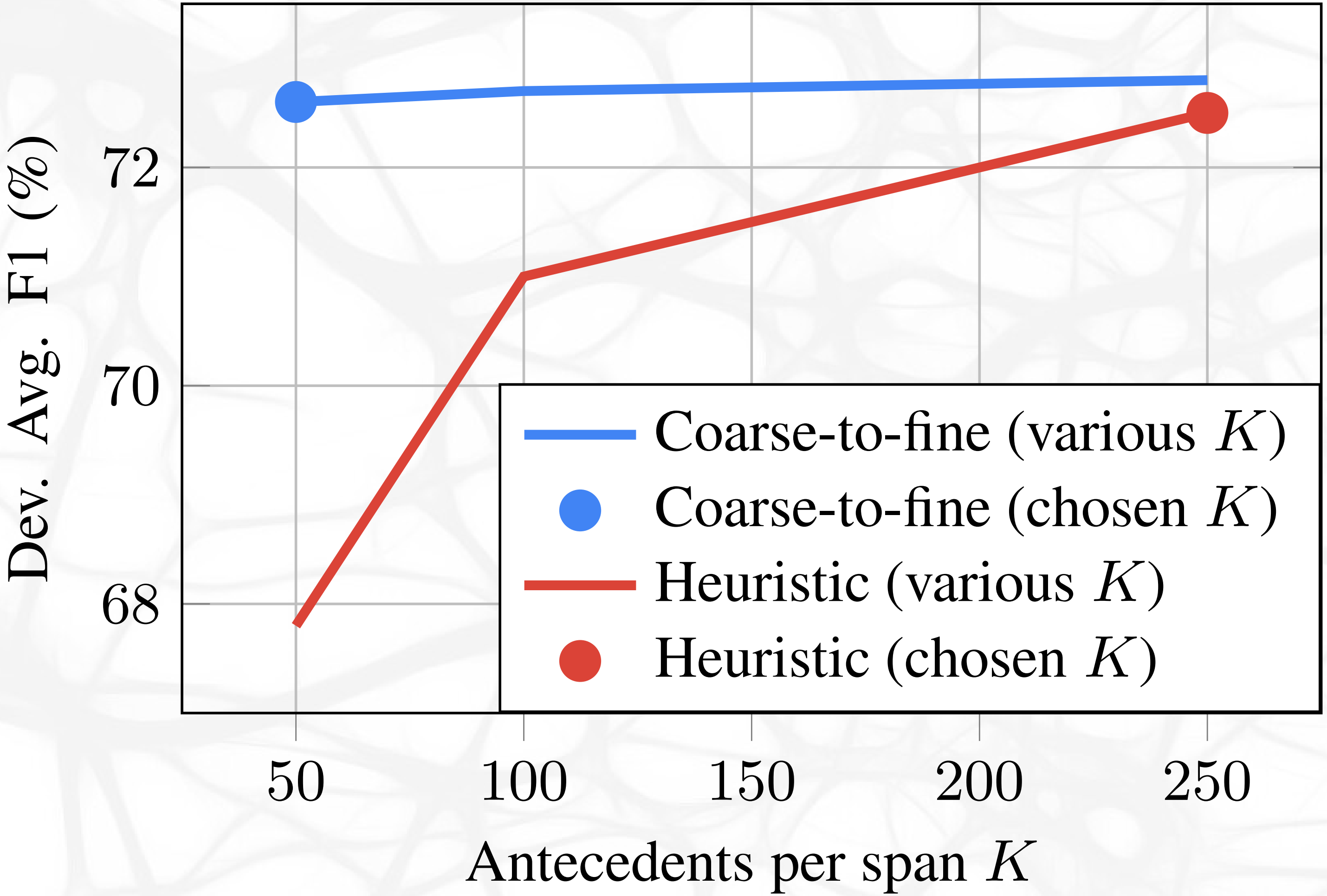


Figure 2 of "Higher-order CR with Coarse-to-fine Inference", Lee et al. (2018)

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM
Lee et al. (2017)	e2e	$67.2_{\varnothing}$
Lee et al. (2018)	e2e	70.4_{ELMo}
Lee et al. (2018)	c2f	73.0_{ELMo}

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM
Lee et al. (2017)	e2e	$67.2_{\emptyset}$
Lee et al. (2018)	e2e	70.4_{ELMo}
Lee et al. (2018)	c2f	73.0_{ELMo}
Joshi et al. (2019)	c2f	73.9_{BERT}

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}

SpanBERT: Improving Pre-training by Representing and Predicting Spans

Joshi et al. (2020)

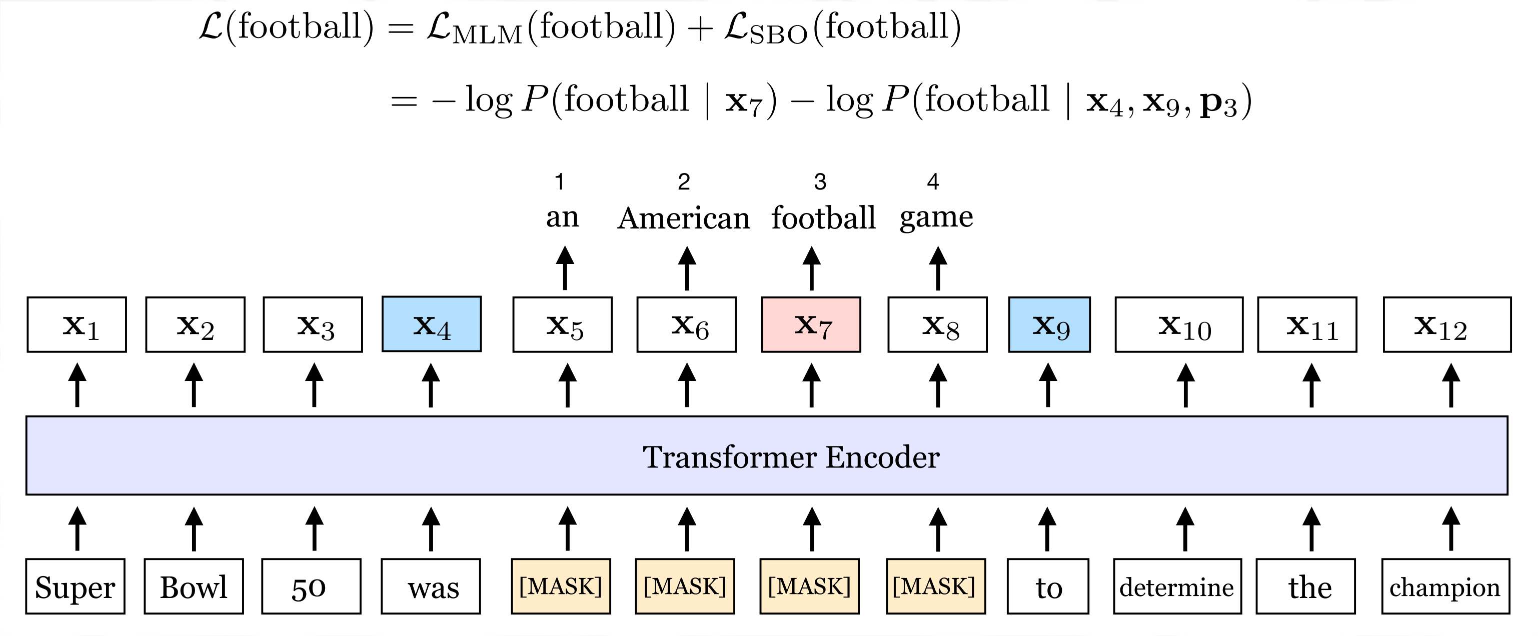


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

Figure 1 of "SpanBERT: Improving Pre-training by Representing and Predicting Spans", Joshi et al. (2020)

SpanBERT: Improving Pre-training by Repr. and Pred. Spans

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$

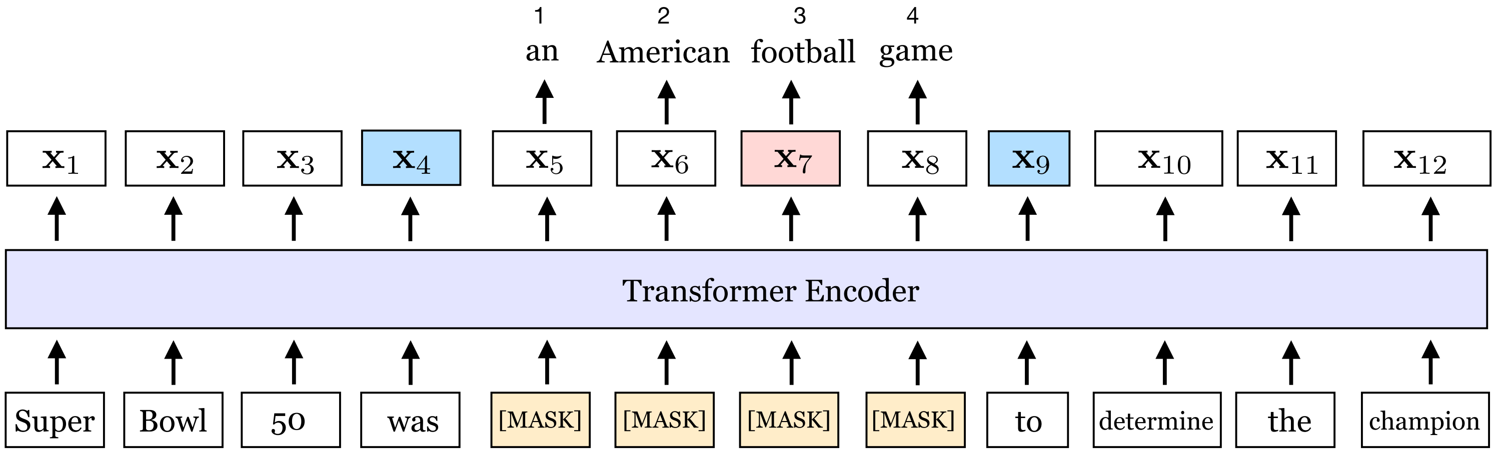


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

Figure 1 of "SpanBERT: Improving Pre-training by Representing and Predicting Spans", Joshi et al. (2020)

- MLM, Span Boundary Objective, no NSP (single segment like RoBERTa)

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}

s2e: Coreference Resolution without Span Representations

Kirstain et al. (2021)

s2e: Coreference Resolution without Span Representations

- A span is represented purely using its starting and ending token

$$\boldsymbol{m}^s = f_m^s(\boldsymbol{x}), \qquad \boldsymbol{m}^e = f_m^e(\boldsymbol{x}).$$

s2e: Coreference Resolution without Span Representations

- A span is represented purely using its starting and ending token

$$\boldsymbol{m}^s = f_m^s(\boldsymbol{x}), \qquad \boldsymbol{m}^e = f_m^e(\boldsymbol{x}).$$

- Mention score for a mention from token i to token j is then

$$s_m(i,j) = \boldsymbol{v}_s^T \boldsymbol{m}_i^s + \boldsymbol{v}_e^T \boldsymbol{m}_j^e + (\boldsymbol{m}_i^s)^T \boldsymbol{W}_m \boldsymbol{m}_j^e.$$

s2e: Coreference Resolution without Span Representations

- A span is represented purely using its starting and ending token

$$\boldsymbol{m}^s = f_m^s(\boldsymbol{x}), \qquad \boldsymbol{m}^e = f_m^e(\boldsymbol{x}).$$

- Mention score for a mention from token i to token j is then

$$\boldsymbol{s}_m(i,j) = \boldsymbol{v}_s^T \boldsymbol{m}_i^s + \boldsymbol{v}_e^T \boldsymbol{m}_j^e + (\boldsymbol{m}_i^s)^T \boldsymbol{W}_m \boldsymbol{m}_j^e.$$

- Mention score is computed for all spans, and only λn are kept.
  - Maximum span length L is used for its inductive bias.

s2e: Coreference Resolution without Span Representations

- A span is represented purely using its starting and ending token

$$\boldsymbol{m}^s = f_m^s(\boldsymbol{x}), \qquad \boldsymbol{m}^e = f_m^e(\boldsymbol{x}).$$

- Mention score for a mention from token i to token j is then

$$s_m(i,j) = \boldsymbol{v}_s^T \boldsymbol{m}_i^s + \boldsymbol{v}_e^T \boldsymbol{m}_j^e + (\boldsymbol{m}_i^s)^T \boldsymbol{W}_m \boldsymbol{m}_j^e.$$

- Mention score is computed for all spans, and only λn are kept.
  - Maximum span length L is used for its inductive bias.
- Antecedent score is $s_a(i_1, j_1, i_2, j_2) = [\boldsymbol{a}_{i_1}^s, \boldsymbol{a}_{j_1}^e]^T \boldsymbol{W}_a [\boldsymbol{a}_{i_2}^s, \boldsymbol{a}_{j_2}^e]$ for

$$\boldsymbol{a}^s = f_a^s(\boldsymbol{x}), \qquad \boldsymbol{a}^e = f_a^e(\boldsymbol{x}).$$

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}
Kirstain et al. (2021)	s2e		80.3_{Longf}

LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution

Otmazgin et al. (2023)

Manual classification of links into 6 classes:

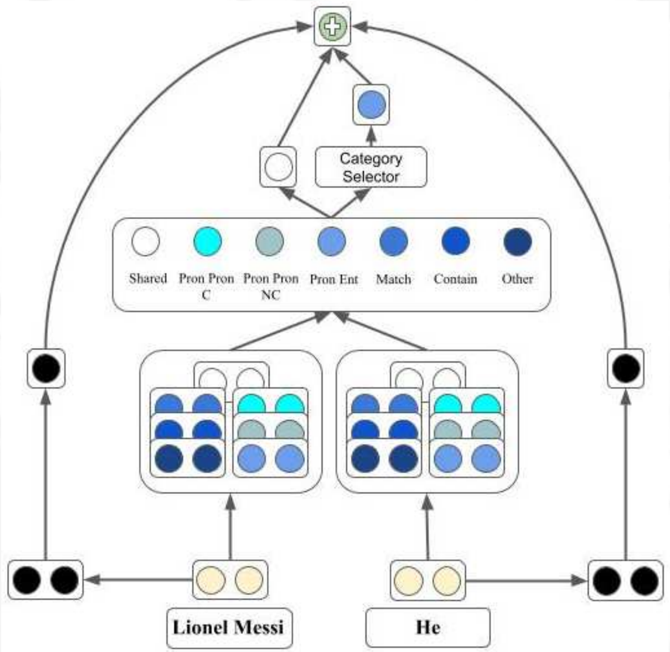


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,

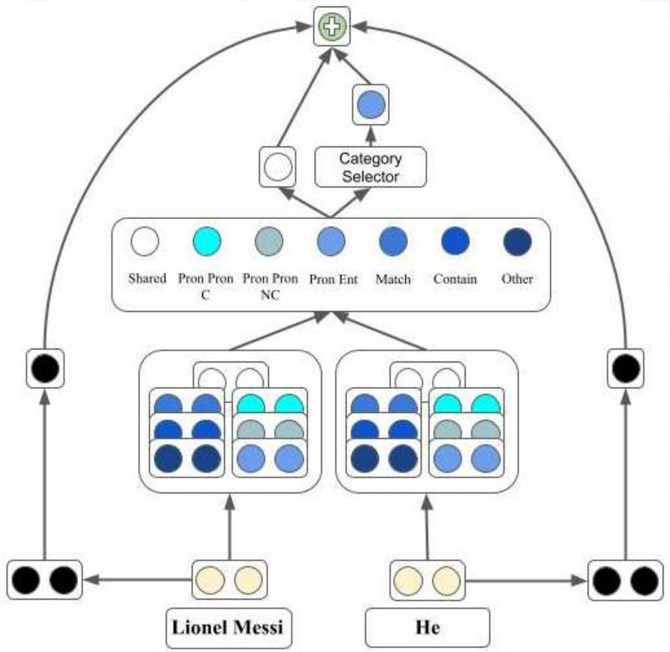


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,

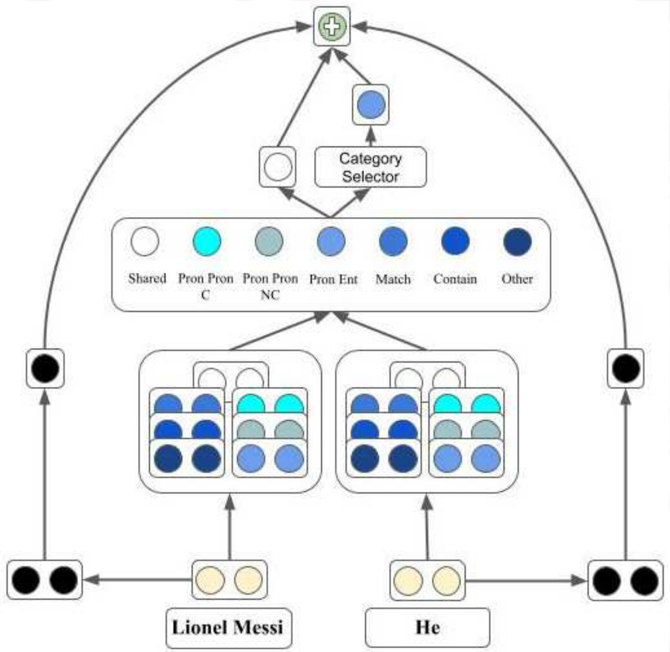


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,

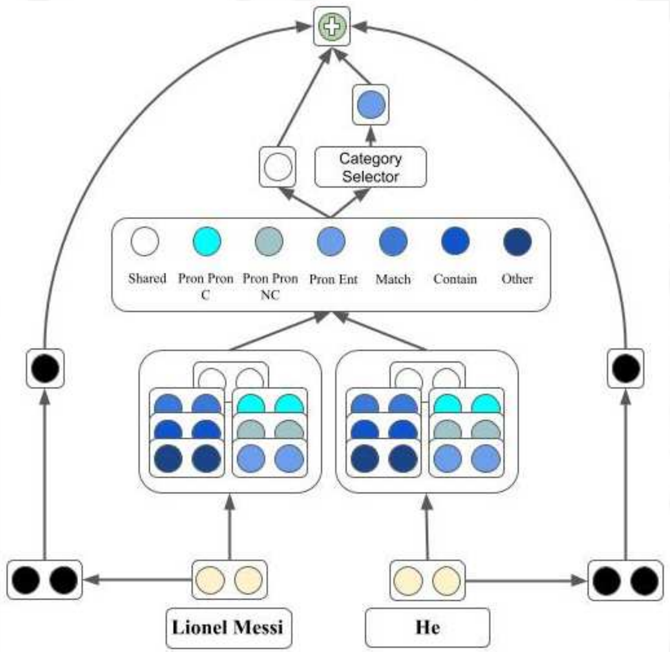


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,
- MATCH: exact forms,

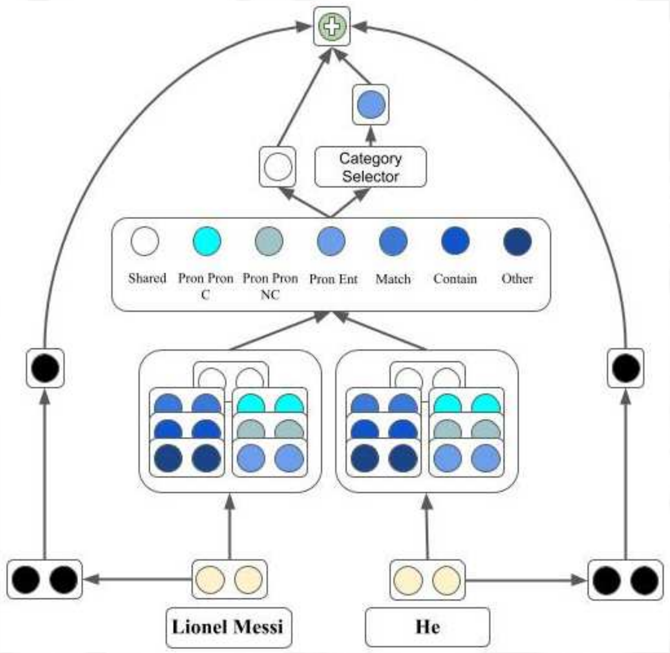


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,
- MATCH: exact forms,
- CONTAINS: one form containing other,

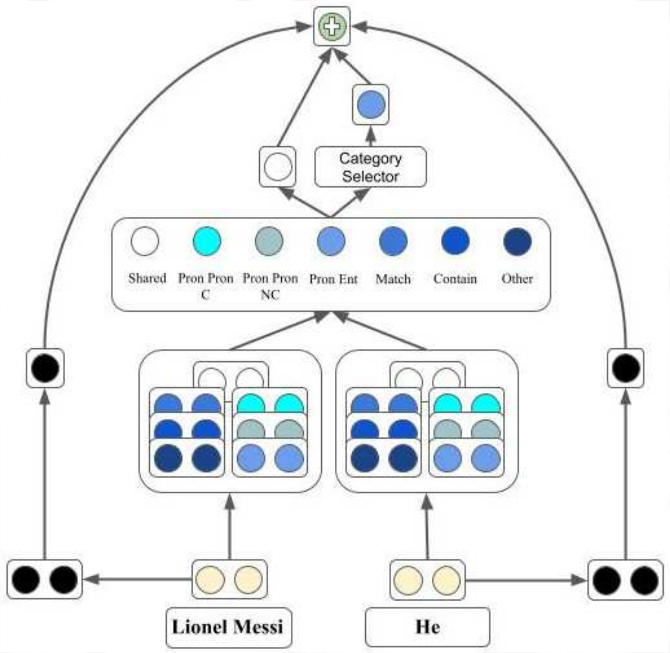


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,
- MATCH: exact forms,
- CONTAINS: one form containing other,
- OTHER.

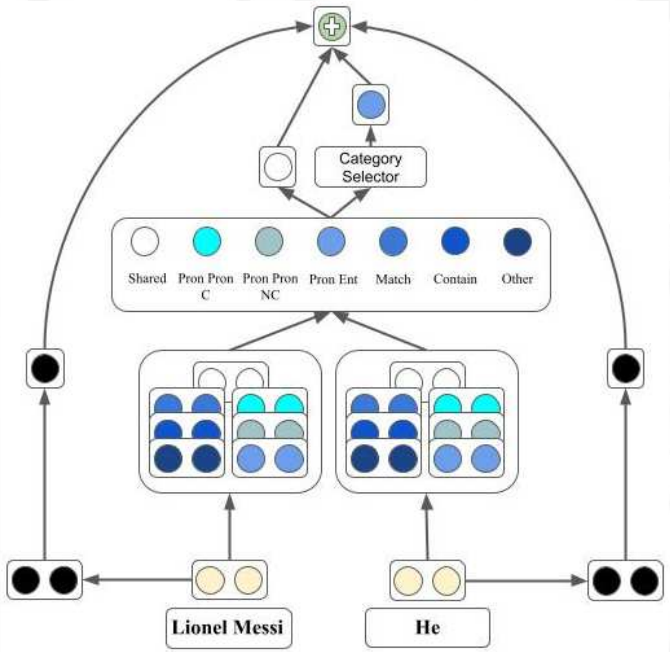


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,
- MATCH: exact forms,
- CONTAINS: one form containing other,
- OTHER.

Create seven antecedent scores – a generic one, and one for every link class.

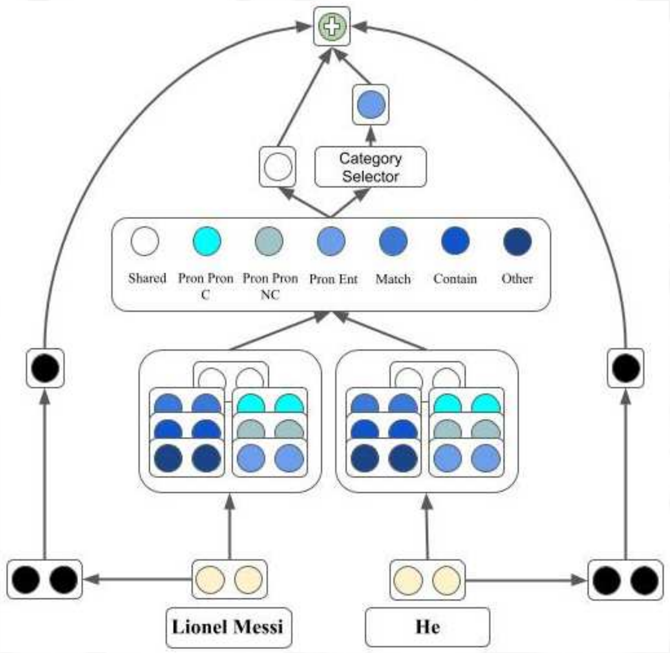


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

LingMess: Linguistically Informed Multi Expert Scorers for CR

Manual classification of links into 6 classes:

- PRON-PRON-C: compatible pronouns,
- PRON-PRON-NC: non-compatible pronouns,
- ENT-PRON: pronoun and non-pronoun,
- MATCH: exact forms,
- CONTAINS: one form containing other,
- OTHER.

Create seven antecedent scores – a generic one, and one for every link class.

Final antecedent score is a sum of the generic antecedent score and the score of the corresponding class-specific score.

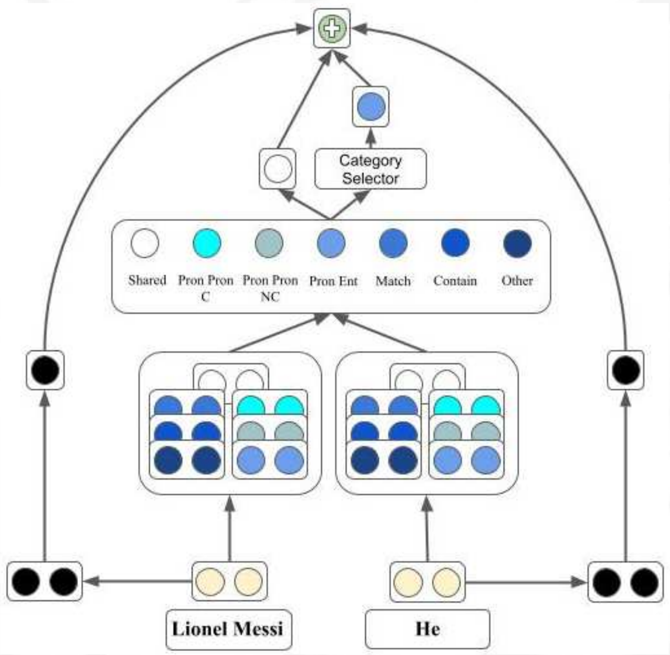


Figure 1: Architecture of our multi expert model. Given two spans *"Lionel Messi"* and *"He"*, we sum four scores: individual mention scores (black), f_m(*"Lionel Messi"*), f_m(*"He"*), and pairwise scores, shared antecedent score (white) f_a(*"Lionel Messi"*, *"He"*) and the relevant "expert" score (blue) $f_a^{\text{PRON-ENT}}$(*"Lionel Messi"*, *"He"*).

Figure 1 of "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution", Otmazgin et al. (2023)

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}
Kirstain et al. (2021)	s2e		80.3_{Longf}
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}

WL: Word-Level Coreference Resolution

Dobrovolskii (2021)

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.
  - Syntactic head is used by the author.

- Represent each span by its **head**.
  - Syntactic head is used by the author.
- We start by computing token representation

$$\boldsymbol{t} = \boldsymbol{W}_A \boldsymbol{x}.$$

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.
  - Syntactic head is used by the author.
- We start by computing token representation

$$\boldsymbol{t} = \boldsymbol{W}_A \boldsymbol{x}.$$

- We then compute bilinear (coarse) antecedent score

$$s_c(i, j) = \boldsymbol{t}_i^T \boldsymbol{W}_C \boldsymbol{t}_j,$$

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.
  - Syntactic head is used by the author.
- We start by computing token representation

$$\boldsymbol{t} = \boldsymbol{W}_A \boldsymbol{x}.$$

- We then compute bilinear (coarse) antecedent score

$$s_c(i, j) = \boldsymbol{t}_i^T \boldsymbol{W}_C \boldsymbol{t}_j,$$

and keep the k most likely antecedent for every mention.

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.
  - Syntactic head is used by the author.
- We start by computing token representation

$$\boldsymbol{t} = \boldsymbol{W}_A \boldsymbol{x}.$$

- We then compute bilinear (coarse) antecedent score

$$s_c(i, j) = \boldsymbol{t}_i^T \boldsymbol{W}_C \boldsymbol{t}_j,$$

  and keep the k most likely antecedent for every mention.
- Finally, we compute $s(i, j) = s_c(i, j) + s_a(i, j)$ for $s_a(i, j) = f_a([\boldsymbol{t}_i, \boldsymbol{t}_j, \boldsymbol{t}_i \odot \boldsymbol{t}_j, \varphi(i, j)])$;

WL: Word-Level Coreference Resolution

- Represent each span by its **head**.
  - Syntactic head is used by the author.
- We start by computing token representation

$$\boldsymbol{t} = \boldsymbol{W}_A \boldsymbol{x}.$$

- We then compute bilinear (coarse) antecedent score

$$s_c(i, j) = \boldsymbol{t}_i^T \boldsymbol{W}_C \boldsymbol{t}_j,$$

  and keep the k most likely antecedent for every mention.
- Finally, we compute $s(i, j) = s_c(i, j) + s_a(i, j)$ for $s_a(i, j) = f_a([\boldsymbol{t}_i, \boldsymbol{t}_j, \boldsymbol{t}_i \odot \boldsymbol{t}_j, \varphi(i, j)])$; $s_a(i, j) < 0$ implies no link.

- Heads are extended into spans by a span extraction module:

	WL F1	SA	SL F1
wl + RoBERTa	83.11	97.16	80.72
-BCE	83.05	97.11	80.60
wl + SpanBERT	82.52	97.13	80.14
-BCE	82.32	97.10	79.99
wl + BERT	77.55	96.20	74.80
wl + Longformer	82.98	97.14	80.56
JOSHI-REPLICA	n/a	n/a	79.74
+RoBERTa	n/a	n/a	78.65

Table 2: Model comparisons on the OntoNotes 5.0 development dataset (best out of 20 epochs). **WL F1** means word-level CoNLL-2012 F1 score, i.e. the coreference metric on the word-level dataset; **SA** is the span extraction accuracy or the percentage of correctly predicted spans; **SL F1** is the span-level CoNLL-2012 F1 score, the basic coreference metric.

Table 2 of "Word-Level Coreference Resolution", Vladimir Dobrovolski (2021)

WL: Word-Level Coreference Resolution

- Heads are extended into spans by a span extraction module:
  - the head token representation is concatenated to all token representations,

	WL F1	**SA**	**SL F1**
wl + RoBERTa	83.11	97.16	80.72
-BCE	83.05	97.11	80.60
wl + SpanBERT	82.52	97.13	80.14
-BCE	82.32	97.10	79.99
wl + BERT	77.55	96.20	74.80
wl + Longformer	82.98	97.14	80.56
JOSHI-REPLICA	n/a	n/a	79.74
+RoBERTa	n/a	n/a	78.65

Table 2: Model comparisons on the OntoNotes 5.0 development dataset (best out of 20 epochs). **WL F1** means word-level CoNLL-2012 F1 score, i.e. the coreference metric on the word-level dataset; **SA** is the span extraction accuracy or the percentage of correctly predicted spans; **SL F1** is the span-level CoNLL-2012 F1 score, the basic coreference metric.

Table 2 of "Word-Level Coreference Resolution", Vladimir Dobrovolski (2021)

- Heads are extended into spans by a span extraction module:
  - the head token representation is concatenated to all token representations,
  - passed through a feed forward network,

	WL F1	SA	SL F1
wl + RoBERTa	83.11	97.16	80.72
-BCE	83.05	97.11	80.60
wl + SpanBERT	82.52	97.13	80.14
-BCE	82.32	97.10	79.99
wl + BERT	77.55	96.20	74.80
wl + Longformer	82.98	97.14	80.56
JOSHI-REPLICA	n/a	n/a	79.74
+RoBERTa	n/a	n/a	78.65

Table 2: Model comparisons on the OntoNotes 5.0 development dataset (best out of 20 epochs). **WL F1** means word-level CoNLL-2012 F1 score, i.e. the coreference metric on the word-level dataset; **SA** is the span extraction accuracy or the percentage of correctly predicted spans; **SL F1** is the span-level CoNLL-2012 F1 score, the basic coreference metric.

Table 2 of "Word-Level Coreference Resolution", Vladimir Dobrovolski (2021)

- Heads are extended into spans by a span extraction module:
  - the head token representation is concatenated to all token representations,
  - passed through a feed forward network,
  - passed through a 1D convolution with kernel size 3,

	WL F1	SA	SL F1
wl + RoBERTa	83.11	97.16	80.72
-BCE	83.05	97.11	80.60
wl + SpanBERT	82.52	97.13	80.14
-BCE	82.32	97.10	79.99
wl + BERT	77.55	96.20	74.80
wl + Longformer	82.98	97.14	80.56
JOSHI-REPLICA	n/a	n/a	79.74
+RoBERTa	n/a	n/a	78.65

Table 2: Model comparisons on the OntoNotes 5.0 development dataset (best out of 20 epochs). **WL F1** means word-level CoNLL-2012 F1 score, i.e. the coreference metric on the word-level dataset; **SA** is the span extraction accuracy or the percentage of correctly predicted spans; **SL F1** is the span-level CoNLL-2012 F1 score, the basic coreference metric.

Table 2 of "Word-Level Coreference Resolution", Vladimir Dobrovolski (2021)

- Heads are extended into spans by a span extraction module:
  - the head token representation is concatenated to all token representations,
  - passed through a feed forward network,
  - passed through a 1D convolution with kernel size 3,
  - the resulting 2 outputs for every token are logits of that token being the starting or ending token of the span.

	WL F1	**SA**	**SL F1**
wl + RoBERTa	83.11	97.16	80.72
-BCE	83.05	97.11	80.60
wl + SpanBERT	82.52	97.13	80.14
-BCE	82.32	97.10	79.99
wl + BERT	77.55	96.20	74.80
wl + Longformer	82.98	97.14	80.56
JOSHI-REPLICA	n/a	n/a	79.74
+RoBERTa	n/a	n/a	78.65

Table 2: Model comparisons on the OntoNotes 5.0 development dataset (best out of 20 epochs). **WL F1** means word-level CoNLL-2012 F1 score, i.e. the coreference metric on the word-level dataset; **SA** is the span extraction accuracy or the percentage of correctly predicted spans; **SL F1** is the span-level CoNLL-2012 F1 score, the basic coreference metric.

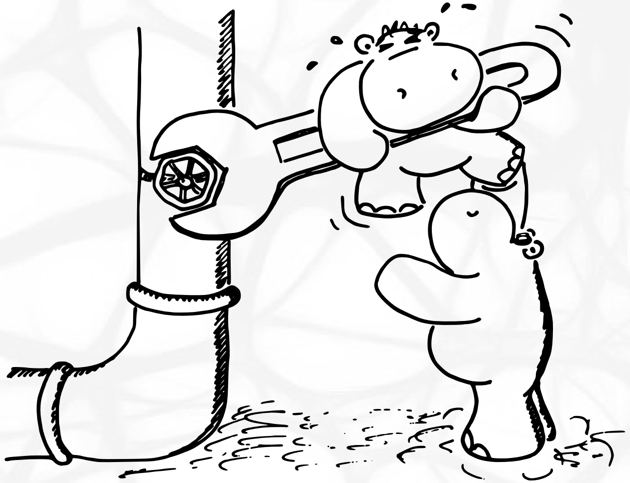
Table 2 of "Word-Level Coreference Resolution", Vladimir Dobrovolski (2021)

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}
Kirstain et al. (2021)	s2e		80.3_{Longf}
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}
Dobrovolskii (2021)	WL		81.0_{RoBE}

CAW: Conjunction-Aware Word-level Coreference Resolution

D'Oosterlinck et al. (2023)



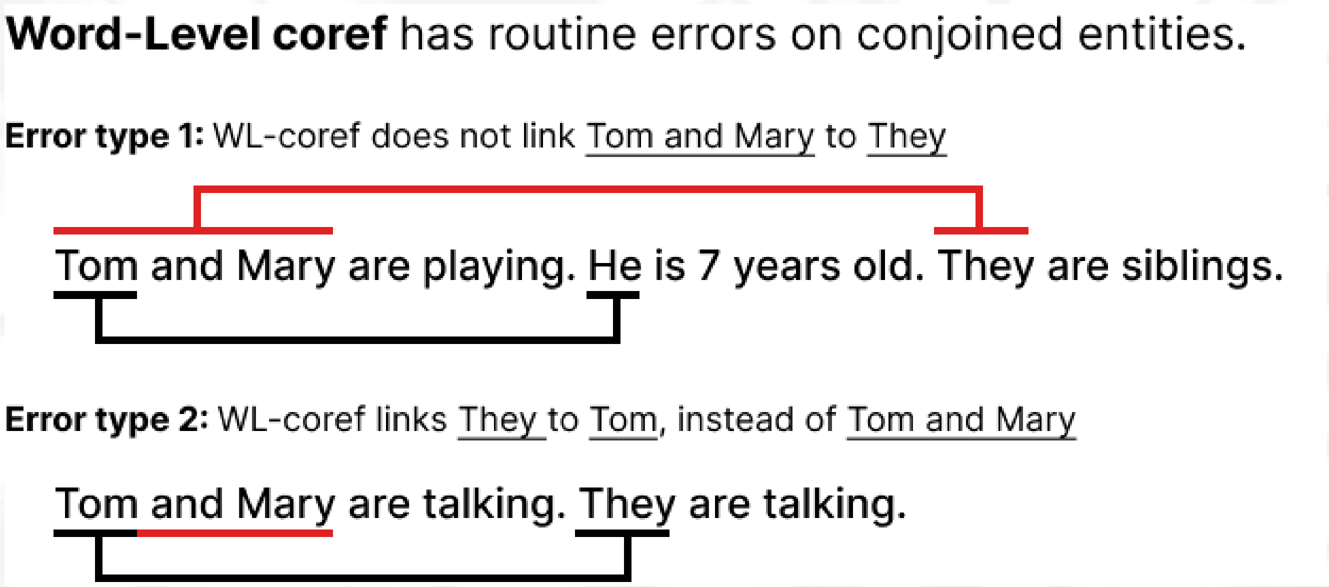


Figure 1: We identify two types of failure cases for `WL-coref` when processing conjoined mentions. Our simple solution, `CAW-coref`, addresses these errors.

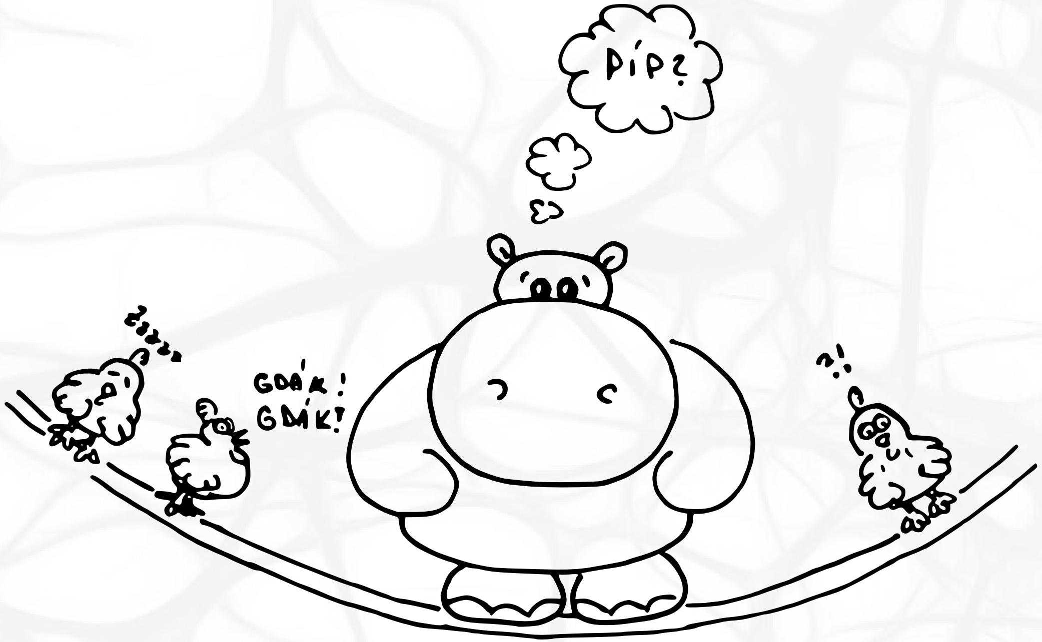
Figure 1 of "CAW-coref: Conjunction-Aware Word-level Coreference Resolution", D'Oosterlinck et al. (2023)

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}
Kirstain et al. (2021)	s2e		80.3_{Longf}
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}
Dobrovolskii (2021)	WL		81.0_{RoBE}
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}

ASP: Autoregressive Structured Prediction with Language Models

Liu et al. (2022)



ASP: Autoregressive Structured Prediction with LMs

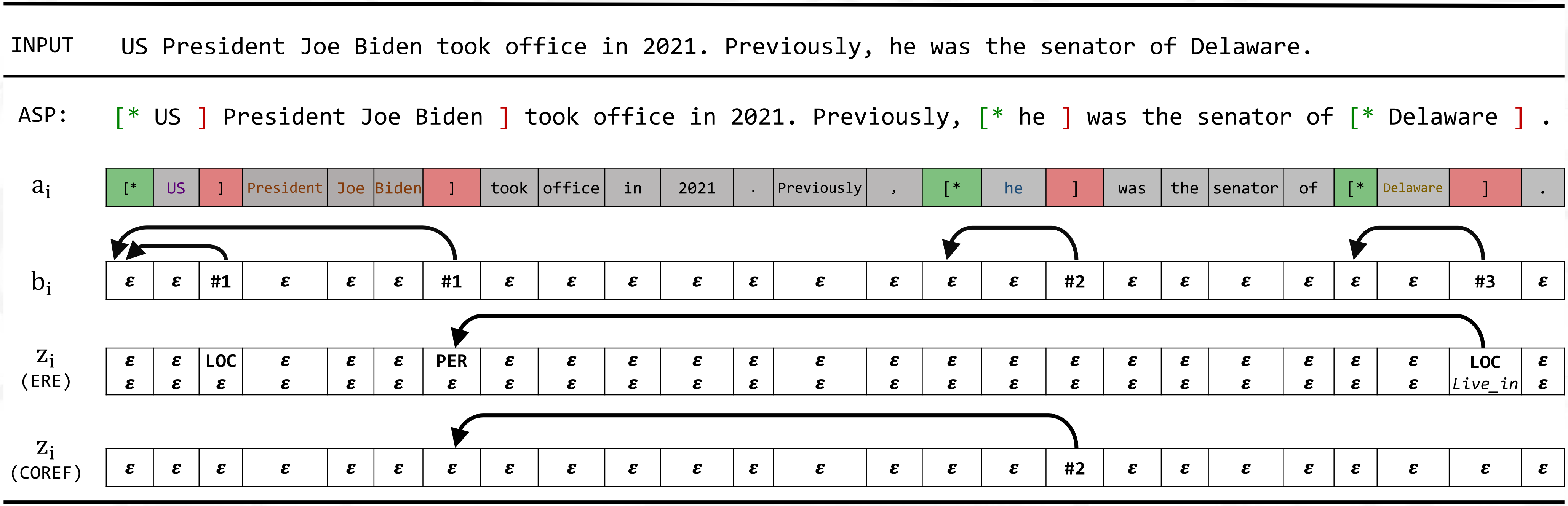


Figure 1: Illustration of the target outputs of our framework on coreference resolution (**COREF**) and end-to-end relation extraction (**ERE**). The lower part illustrates the decoding process of our model. The actions $\boldsymbol{y}_i$ are color-coded as], [* and copy. The structure random variables z_i are presented along with coreference links or relation links. We present words in the copy cells merely as an illustration.

Figure 1 of "Autoregressive Structured Prediction with Language Models", Liu et al. (2022)

ASP: Autoregressive Structured Prediction with LMs

At each step, the output consists of a triple:

At each step, the output consists of a triple:

- an action `[*, copy, ]`;

At each step, the output consists of a triple:

- an action `[*`, `copy`, `]`;
- if the action is `]`, a pointer to some previous `[*`;

ASP: Autoregressive Structured Prediction with LMs

At each step, the output consists of a triple:

- an action [∗, copy,];
- if the action is], a pointer to some previous [∗;
- if the action is], a pointer to an antecedent represented by its], or to ε.

At each step, the output consists of a triple:

- an action `[*`, `copy`, `]`;
- if the action is `]`, a pointer to some previous `[*`;
- if the action is `]`, a pointer to an antecedent represented by its `]`, or to ε.

The local probabilities are computed using a softmax over a dynamic set with a parametrized scoring function.

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M
Lee et al. (2017)	e2e	$67.2_{\varnothing}$	
Lee et al. (2018)	e2e	70.4_{ELMo}	
Lee et al. (2018)	c2f	73.0_{ELMo}	
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}
Joshi et al. (2020)	c2f		79.6_{SpanB}
Kirstain et al. (2021)	s2e		80.3_{Longf}
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}
Dobrovolskii (2021)	WL		81.0_{RoBE}
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B
Lee et al. (2017)	e2e	$67.2_{\emptyset}$		
Lee et al. (2018)	e2e	70.4_{ELMo}		
Lee et al. (2018)	c2f	73.0_{ELMo}		
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}	
Joshi et al. (2020)	c2f		79.6_{SpanB}	
Kirstain et al. (2021)	s2e		80.3_{Longf}	
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}	
Dobrovolskii (2021)	WL		81.0_{RoBE}	
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}	
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ∼350M	xl PLM ∼3B	xxl PLM ∼11B
Lee et al. (2017)	e2e	$67.2_{\varnothing}$			
Lee et al. (2018)	e2e	70.4_{ELMo}			
Lee et al. (2018)	c2f	73.0_{ELMo}			
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}		
Joshi et al. (2020)	c2f		79.6_{SpanB}		
Kirstain et al. (2021)	s2e		80.3_{Longf}		
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}		
Dobrovolskii (2021)	WL		81.0_{RoBE}		
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}		
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}

Model OntoNotes English Results

Paper	Model	∅/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	$67.2_{\emptyset}$				1
Lee et al. (2018)	e2e	70.4_{ELMo}				1
Lee et al. (2018)	c2f	73.0_{ELMo}				1
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}			1
Joshi et al. (2020)	c2f		79.6_{SpanB}			1
Kirstain et al. (2021)	s2e		80.3_{Longf}			1
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}			1
Dobrovolskii (2021)	WL		81.0_{RoBE}			1
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}			1
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}	$\mathcal{O}(n)$

seq2seq: Coreference Resolution through a seq2seq Transition-Based System

Bohnet et al. (2023)

Input: *Speaker-A* I still have n't gone to that fresh French restaurant by your house
Prediction: SHIFT: next sentence

Input: *Speaker-A* I_2 still have n't gone to that fresh French restaurant by your house *Speaker-A* I_{17} 'm like dying to go there
Prediction:
A $I_{17} \rightarrow I_2$
B SHIFT: next sentence

Input: *Speaker-A* [1 I] still have n't gone to that fresh French restaurant by your house *Speaker-A* [1 I] 'm like dying to go there *Speaker-B* You mean the one right next to the apartment
Prediction:
A You → [1
B the apartment → your house
C the one right next to the apartment → that fresh French restaurant by your house
D SHIFT: next sentence

Input: *Speaker-A* [1 I] still have n't gone to [3 that fresh French restaurant by [2 your house]] *Speaker-A* [1 I] 'm like dying to go there *Speaker-B* [1 You] mean [3 the one right next to [2 the apartment]] *Speaker-B* yeah yeah yeah
Prediction: SHIFT: next sentence

Figure 1: Example of one of our transition-based coreference systems, the *Link-Append* system. The system processes a single sentence at a time, using an input encoding of the prior sentences annotated with coreference clusters, followed by the new sentence. As output, the system makes predictions that link mentions in the new sentence to either previously created coreference clusters (e.g., "You → [1") or when a new cluster is created, to previous mentions (e.g., "the apartment → your house"). The system predicts "SHIFT" when processing of the sentence is complete. Note in the figure we use the word indices 2 and 17 to distinguish the two incidences of "I" in the text.

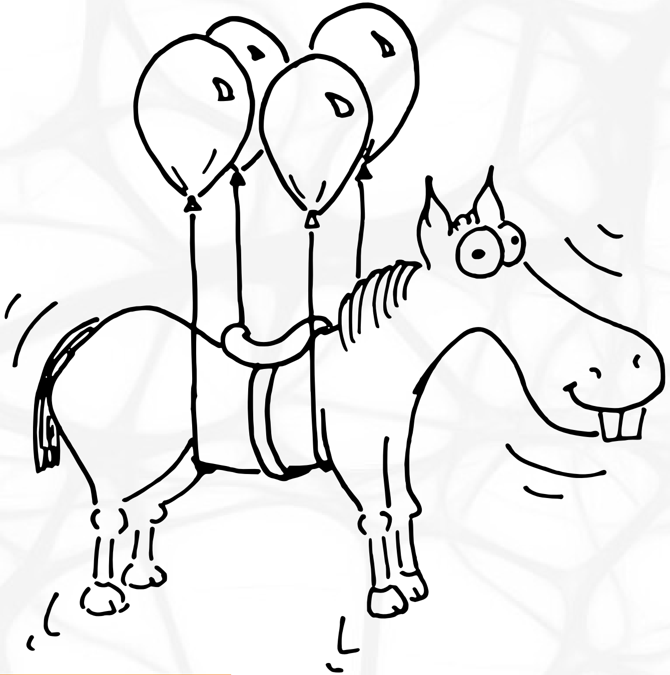
Figure 1 of "Coreference Resolution through a seq2seq Transition-Based System", Bohnet et al. (2023)

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	$67.2_{\emptyset}$				1
Lee et al. (2018)	e2e	70.4_{ELMo}				1
Lee et al. (2018)	c2f	73.0_{ELMo}				1
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}			1
Joshi et al. (2020)	c2f		79.6_{SpanB}			1
Kirstain et al. (2021)	s2e		80.3_{Longf}			1
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}			1
Dobrovolskii (2021)	WL		81.0_{RoBE}			1
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}			1
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}	$\mathcal{O}(n)$
Bohnet et al. (2023)	seq2seq			$78.0^{\text{dev}}_{\text{mT5}}$	83.3_{mT5}	$\mathcal{O}(n)$

CorefQA: Coreference Resolution as Query-based Span Prediction

Wu et al. (2020)



CorefQA: CR as Query-based Span Prediction

Original Passage
In addition , *many people* were poisoned when **toxic gas** was released. *They* were poisoned and did not know how to protect *themselves* against **the poison**.
Our formulation
Q1: Who were poisoned when toxic gas was released?
A1: [*They*, *themselves*]
Q2: What was released when many people were poisoned?
A2: [**the poison**]
Q3: Who were poisoned and did not know how to protect themselves against the poison?
A3: [*many people*, *themselves*]
Q4: Whom did they not know how to protect against the poison?
A4: [*many people*, *They*]
Q5: They were poisoned and did not know how to protect themselves against what?
A5: [**toxic gas**]

Figure 1: An illustration of the paradigm shift from coreference resolution to query-based span prediction. Spans with the same format represent coreferent mentions.

Figure 1 of "CorefQA: Coreference Resolution as Query-based Span Prediction", Wu et al. (2020)

CorefQA: CR as Query-based Span Prediction

- Using SpanBERT and representing each span by its starting and ending token, compute mention scores and keep the top-scoring λn mentions for $\lambda = 0.2$ and maximum length $L = 10$.

CorefQA: CR as Query-based Span Prediction

- Using SpanBERT and representing each span by its starting and ending token, compute mention scores and keep the top-scoring λn mentions for $\lambda = 0.2$ and maximum length $L = 10$.
- For a mention, we compute the antecedent score $s_a(i|j)$ by
  - constructing a *context-query* input for SpanBERT,

CorefQA: CR as Query-based Span Prediction

- Using SpanBERT and representing each span by its starting and ending token, compute mention scores and keep the top-scoring λn mentions for $\lambda = 0.2$ and maximum length $L = 10$.
- For a mention, we compute the antecedent score $s_a(i|j)$ by
  - constructing a *context-query* input for SpanBERT,
  - using BIO encoding to represent the antecedent (and possibly several of them); an antecedent ε is represented using all O-s.

CorefQA: CR as Query-based Span Prediction

- Using SpanBERT and representing each span by its starting and ending token, compute mention scores and keep the top-scoring λn mentions for $\lambda = 0.2$ and maximum length $L = 10$.
- For a mention, we compute the antecedent score $s_a(i|j)$ by
  - constructing a *context-query* input for SpanBERT,
  - using BIO encoding to represent the antecedent (and possibly several of them); an antecedent ε is represented using all O-s.
- To handle bidirectionality, the final antecedent score is computed as

$$s(i, j) = s_a(i|j) + s_a(j|i).$$

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	$67.2_{\varnothing}$				1
Lee et al. (2018)	e2e	70.4_{ELMo}				1
Lee et al. (2018)	c2f	73.0_{ELMo}				1
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}			1
Joshi et al. (2020)	c2f		79.6_{SpanB}			1
Kirstain et al. (2021)	s2e		80.3_{Longf}			1
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}			1
Dobrovolskii (2021)	WL		81.0_{RoBE}			1
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}			1
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}	$\mathcal{O}(n)$
Bohnet et al. (2023)	seq2seq			$78.0^{\text{dev}}_{\text{mT5}}$	83.3_{mT5}	$\mathcal{O}(n)$
Wu et al. (2020)	CorefQA	$79.9^{+\text{QA}}_{\text{SpanB}}$	$83.1^{+\text{QA}}_{\text{SpanB}}$			$\mathcal{O}(n)$

CorPipe: Winning System of CRAC 22 and 23

Straka and Straková (2022), Straka (2023)

CorPipe: Winning System of CRAC 22 and 23

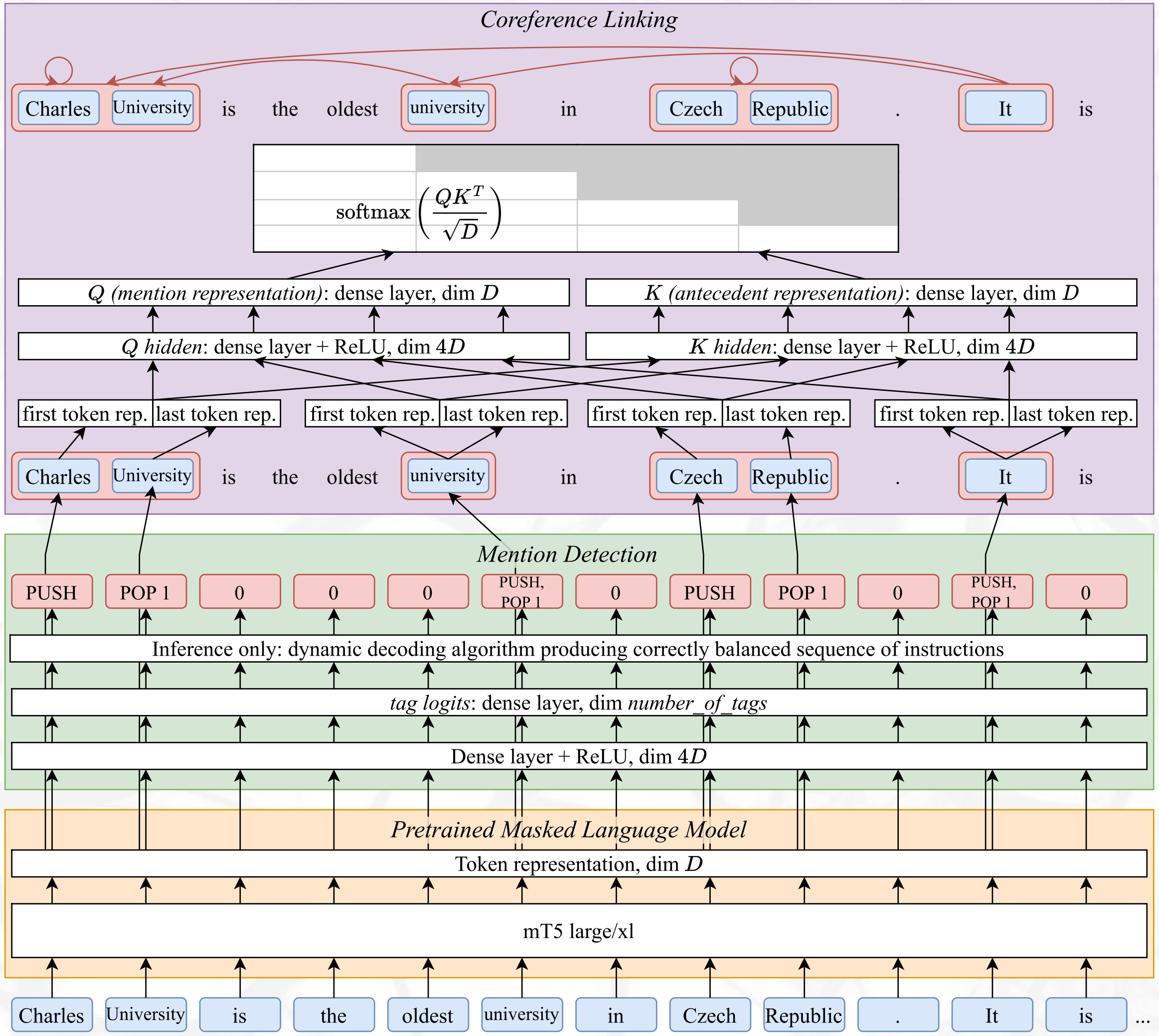


Figure 1 of "ÚFAL CorPipe at CRAC 2023: Larger Context Improves Multilingual Coreference Resolution", Straka (2023)

Model OntoNotes English Results

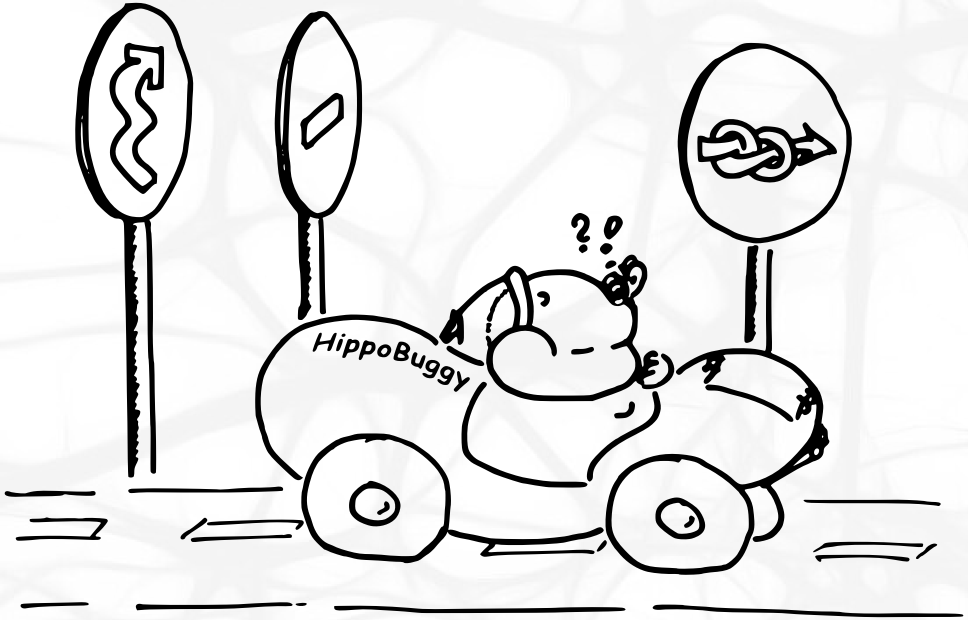
Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	$67.2_{\varnothing}$				1
Lee et al. (2018)	e2e	70.4_{ELMo}				1
Lee et al. (2018)	c2f	73.0_{ELMo}				1
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}			1
Joshi et al. (2020)	c2f		79.6_{SpanB}			1
Kirstain et al. (2021)	s2e		80.3_{Longf}			1
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}			1
Dobrovolskii (2021)	WL		81.0_{RoBE}			1
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}			1
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}	$\mathcal{O}(n)$
Bohnet et al. (2023)	seq2seq			$78.0^{\text{dev}}_{\text{mT5}}$	83.3_{mT5}	$\mathcal{O}(n)$
Wu et al. (2020)	CorefQA	$79.9^{+\text{QA}}_{\text{SpanB}}$	$83.1^{+\text{QA}}_{\text{SpanB}}$			$\mathcal{O}(n)$
	CorPipe		80.7_{T5}	82.0_{FT5}		1

Model OntoNotes English Results

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	$67.2_{\varnothing}$				1
Lee et al. (2018)	e2e	70.4_{ELMo}				1
Lee et al. (2018)	c2f	73.0_{ELMo}				1
Joshi et al. (2019)	c2f	73.9_{BERT}	76.9_{BERT}			1
Joshi et al. (2020)	c2f		79.6_{SpanB}			1
Kirstain et al. (2021)	s2e		80.3_{Longf}			1
Otmazgin et al. (2023)	LingMess/s2e		81.4_{Longf}			1
Dobrovolskii (2021)	WL		81.0_{RoBE}			1
D'Oosterlinck et al. (2023)	CAW/WL		81.6_{RoBE}			1
Liu et al. (2022)	ASP	76.6_{T5}	79.3_{T5}	82.2_{FT5}	82.5_{FT5}	$\mathcal{O}(n)$
Bohnet et al. (2023)	seq2seq			$78.0^{\text{dev}}_{\text{mT5}}$	83.3_{mT5}	$\mathcal{O}(n)$
Wu et al. (2020)	CorefQA	$79.9^{+\text{QA}}_{\text{SpanB}}$	$83.1^{+\text{QA}}_{\text{SpanB}}$			$\mathcal{O}(n)$
	CorPipe		80.7_{T5}	82.0_{FT5}		1
	CorPipe		77.2_{mT5}	78.9_{mT5}		1

Multiple Languages – 17 CorefUD Treebanks

Uniqueness of Mention Heads Across CorefUD



Uniqueness of Mention Heads Across CorefUD

Treebank	Unique mention heads
ca_ancora	99.19%
cs_pcedt	98.72%
cs_pdt	98.64%
de_parcorfull	99.73%
de_potsdamcc	97.43%
en_gum	98.74%
en_parcorfull	99.58%
es_ancora	99.22%
fr_democrat	97.99%
hu_korkor	99.22%
hu_szegedkoref	99.52%
lt_lcc	99.60%
no_bokmaalnarc	95.47%
no_nynorsknarc	95.39%
pl_pcc	95.16%
ru_rucor	99.97%
tr_itcc	99.42%

Uniqueness of Mention Heads Across CorefUD

Treebank	Unique mention heads
pl_pcc	95.16%
no_nynorsknarc	95.39%
no_bokmaalnarc	95.47%
de_potsdamcc	97.43%
fr_democrat	97.99%
cs_pdt	98.64%
cs_pcedt	98.72%
en_gum	98.74%
ca_ancora	99.19%
es_ancora	99.22%
hu_korkor	99.22%
tr_itcc	99.42%
hu_szegedkoref	99.52%
en_parcorfull	99.58%
lt_lcc	99.60%
de_parcorfull	99.73%
ru_rucor	99.97%

Uniqueness of Mention Heads Across CorefUD

Treebank	Unique mention heads	Unique or double head
pl_pcc	95.16%	99.59%
no_nynorsknarc	95.39%	99.95%
no_bokmaalnarc	95.47%	99.95%
de_potsdamcc	97.43%	99.84%
fr_democrat	97.99%	99.96%
cs_pdt	98.64%	99.93%
cs_pcedt	98.72%	99.95%
en_gum	98.74%	99.98%
ca_ancora	99.19%	99.99%
es_ancora	99.22%	100.00%
hu_korkor	99.22%	100.00%
tr_itcc	99.42%	100.00%
hu_szegedkoref	99.52%	100.00%
en_parcorfull	99.58%	100.00%
lt_lcc	99.60%	99.97%
de_parcorfull	99.73%	100.00%
ru_rucor	99.97%	100.00%

Uniqueness of Mention Heads Across CorefUD

Treebank	Unique mention heads	Unique or double head	Unique, double, triple
pl_pcc	95.16%	99.59%	99.96%
no_nynorsknarc	95.39%	99.95%	100.00%
no_bokmaalnarc	95.47%	99.95%	100.00%
de_potsdamcc	97.43%	99.84%	99.95%
fr_democrat	97.99%	99.96%	100.00%
cs_pdt	98.64%	99.93%	99.97%
cs_pcedt	98.72%	99.95%	100.00%
en_gum	98.74%	99.98%	100.00%
ca_ancora	99.19%	99.99%	100.00%
es_ancora	99.22%	100.00%	100.00%
hu_korkor	99.22%	100.00%	100.00%
tr_itcc	99.42%	100.00%	100.00%
hu_szegedkoref	99.52%	100.00%	100.00%
en_parcorfull	99.58%	100.00%	100.00%
lt_lcc	99.60%	99.97%	100.00%
de_parcorfull	99.73%	100.00%	100.00%
ru_rucor	99.97%	100.00%	100.00%

Training on Multiple Treebanks

Training on Multiple Treebanks

- Training a single multilingual model improves performance of all treebanks
  - CorPipe 23, mT5-large

Configuration	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu korko	hu szege	lt	no bookm	no nynor	pl	ru	tr
Single Multilingual Model	**74.8**	**81.6**	**80.3**	**79.0**	**69.7**	**75.4**	**76.8**	**66.0**	**82.8**	**70.3**	**69.5**	**69.8**	**77.9**	**81.5**	**81.7**	**77.1**	**75.2**	**57.2**
Per-Corpus Models	–3.7	–1.4	–0.5	–0.4	–7.7	–3.3	–1.6	–7.6	–1.5	–2.0	–9.1	–1.0	–3.0	–2.3	–2.9	–1.0	–2.0	–15.8
Joint Czech Model			–0.1	–0.3														
Joint German Model					–4.8	–3.9												
Joint English Model							–1.9	–4.5										
Joint Parcorfull Model					–4.4			–2.5										
Joint Hungarian Model											–5.9	–1.1						
Joint Norwegian Model														–1.3	–1.8			
Zero-Shot Multilingual Models	–13.2	–4.8	–24.2	–16.0	–13.7	–10.6	–14.4	–13.8	–1.9	–5.4	–15.1	–15.0	–23.4	–14.3	–18.0	–17.5	–15.5	–0.8

Table 6: Ablation experiments evaluated on the development sets (CoNLL score in %) using the mT5-large model with context size 2560. We report the average of best 5 out of 7 runs, using for every corpus the single epoch achieving the highest average 5-run score.

- Training a single multilingual model improves performance of all treebanks
  - CorPipe 22, RemBERT

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
G) EFFECT OF SEVERAL LANGUAGE-SPECIFIC BASE PRETRAINED MODELS														
XLM-R base individual	68.7	71.4	75.7	73.9	65.7	62.0	71.2	63.2	75.6	63.1	61.5	73.4	69.8	65.6
mBERT (Devlin et al., 2019)	-2.8	-1.5	-3.0	-3.4	-3.3	+0.4	-2.8	-1.1	-1.8	-1.1	-2.7	-7.5	-4.4	-3.6
BERTa (Armengol-Estapé et al., 2021)		+1.3												
RobeCzech (Straka et al., 2021)			+2.0	+2.8										
gBERT (Chan et al., 2020)					-9.9	+5.3								
SpanBERT (Joshi et al., 2020)							-0.4	-2.4						
BETO (Cañete et al., 2020)									+0.4					
CamemBERT (Martin et al., 2020)										-0.2				
HuBERT (Nemeskey, 2020)											+3.6			
LitLatBERT (Ulčar and Robnik-Šikonja, 2021)												+2.7		
HerBERT (Mroczkowski et al., 2021)													+1.6	
RuBERT (Kuratov and Arkhipov, 2019)														+0.2
XLM-R large individual	+4.0	+4.6	+3.1	+4.1	+0.0	+6.9	+1.0	**+7.8**	+3.8	+3.3	+7.4	-0.8	+5.8	+4.8
RemBERT individual	-0.0	+4.9	+3.1	+3.1	-15.2	+0.0	+2.6	-18.3	+3.9	+3.8	+3.3	-4.3	+5.0	+4.3
XLM-R large multilingual	+6.1	**+6.1**	+2.1	+3.2	+8.0	**+16.2**	+4.1	+7.7	+5.0	+4.8	+6.9	**+4.6**	+5.1	+6.9
RemBERT multilingual	**+6.6**	+6.0	**+3.6**	**+4.4**	**+10.6**	+14.5	**+4.3**	+6.1	**+5.5**	**+5.1**	**+7.7**	+3.5	**+6.0**	**+9.0**

- Training a single multilingual model improves performance of all treebanks
  - CorPipe 22, XLM-R-large: slight reduction for the largest treebanks

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
G) Effect of Several Language-specific base Pretrained Models														
XLM-R base individual	68.7	71.4	75.7	73.9	65.7	62.0	71.2	63.2	75.6	63.1	61.5	73.4	69.8	65.6
mBERT (Devlin et al., 2019)	-2.8	-1.5	-3.0	-3.4	-3.3	+0.4	-2.8	-1.1	-1.8	-1.1	-2.7	-7.5	-4.4	-3.6
BERTa (Armengol-Estapé et al., 2021)		+1.3												
RobeCzech (Straka et al., 2021)			+2.0	+2.8										
gBERT (Chan et al., 2020)					-9.9	+5.3								
SpanBERT (Joshi et al., 2020)							-0.4	-2.4						
BETO (Cañete et al., 2020)									+0.4					
CamemBERT (Martin et al., 2020)										-0.2				
HuBERT (Nemeskey, 2020)											+3.6			
LitLatBERT (Ulčar and Robnik-Šikonja, 2021)												+2.7		
HerBERT (Mroczkowski et al., 2021)													+1.6	
RuBERT (Kuratov and Arkhipov, 2019)														+0.2
XLM-R large individual	+4.0	+4.6	+3.1	+4.1	+0.0	+6.9	+1.0	**+7.8**	+3.8	+3.3	+7.4	-0.8	+5.8	+4.8
RemBERT individual	-0.0	+4.9	+3.1	+3.1	-15.2	+0.0	+2.6	-18.3	+3.9	+3.8	+3.3	-4.3	+5.0	+4.3
XLM-R large multilingual	+6.1	**+6.1**	+2.1	+3.2	+8.0	**+16.2**	+4.1	+7.7	+5.0	+4.8	+6.9	**+4.6**	+5.1	+6.9
RemBERT multilingual	**+6.6**	+6.0	**+3.6**	**+4.4**	**+10.6**	+14.5	**+4.3**	+6.1	**+5.5**	**+5.1**	**+7.7**	+3.5	**+6.0**	**+9.0**

Training on Multiple Treebanks

- Training a single base-sized multilingual model makes performance of larger treebanks worse

Training on Multiple Treebanks

- CorPipe 23 & 22: Surprisingly, the mixing ratios do not matter much

Configuration	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu korko	hu szege	lt	no bookm	no nynor	pl	ru	tr
MIX RATIO WEIGHTS OF INDIVIDUAL CORPORA IN PERCENTS																		
Logarithmic		*8.1*	*10.0*	*9.4*	*1.0*	*3.2*	*6.6*	*1.0*	*8.3*	*7.4*	*2.6*	*5.8*	*3.4*	*7.2*	*6.9*	*8.6*	*6.2*	*4.2*
Uniform		*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*	*5.9*
Square Root		*8.4*	*14.0*	*11.7*	*1.4*	*2.4*	*5.6*	*1.4*	*8.8*	*6.9*	*2.0*	*4.6*	*2.5*	*6.5*	*6.0*	*9.5*	*5.1*	*3.1*
Linear		*8.7*	*24.4*	*17.0*	*0.2*	*0.7*	*3.9*	*0.2*	*9.6*	*5.9*	*0.5*	*2.6*	*0.8*	*5.3*	*4.5*	*11.3*	*3.2*	*1.2*
B) AVERAGE OF 5 RUNS USING FOR EVERY RUN THE SINGLE EPOCH ACHIEVING THE HIGHEST SCORE ACROSS ALL CORPORA																		
Logarithmic	74.8	81.7	79.9	78.6	71.5	**76.2**	**76.6**	67.9	82.8	70.4	68.3	69.4	78.0	81.4	81.5	76.9	74.6	55.5
w/o corpus id	–0.2	+0.0	+0.1	+0.2	–1.9	–0.3	–0.3	–0.9	–0.2	–0.4	+0.0	–0.2	–0.2	+0.1	–0.2	+0.3	+1.0	–0.3
Uniform	–0.6	–0.4	–1.1	–0.9	+0.1	–1.0	–0.8	–6.7	–0.4	–0.2	**+1.0**	**+0.1**	–0.2	–0.1	+0.2	–0.1	+0.5	+0.0
w/o corpus id	–0.6	–0.7	–0.6	–0.5	+1.0	–1.6	–0.5	–0.6	–0.1	–0.6	+0.3	–0.5	–0.9	–0.1	–1.3	–0.5	+0.8	–3.0
Square Root	–0.2	–0.1	+0.8	+0.7	–2.5	–0.2	–0.1	–4.2	–0.1	+0.0	+0.9	–0.4	+0.2	+0.3	+0.0	**+0.4**	+1.5	+0.4
w/o corpus id	+0.1	–0.2	+0.6	+0.6	**+1.3**	–2.1	–0.2	–0.7	+0.2	**+0.1**	+0.0	–0.4	–0.1	+0.2	+0.1	+0.1	+1.2	**+1.1**
Linear	**+0.3**	**+0.2**	**+1.1**	**+1.1**	–0.7	–1.9	–0.2	**+3.8**	**+0.5**	–0.1	–0.7	–0.1	+0.3	–0.4	**+0.3**	+0.1	**+1.6**	+0.0
w/o corpus id	+0.1	+0.0	+1.0	+1.0	–2.1	–2.5	–0.2	+1.3	+0.2	–0.1	+0.4	–0.5	**+0.5**	**+0.4**	+0.3	+0.4	+1.0	+0.8

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results on Arabic OntoNotes
  - only 359 training documents, compared to 1,940 English ones

Paper	Method	Arabic only	Arabic & English	Arabic & Chinese
Min (2021)	e2e, mBERT-base	46.8	56.4	

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results on Arabic OntoNotes
  - only 359 training documents, compared to 1,940 English ones

Paper	Method	Arabic only	Arabic & English	Arabic & Chinese
Min (2021)	e2e, mBERT-base	46.8	56.4	
Min (2021)	e2e, GigaBERT-base	62.1	64.6	

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results on Arabic OntoNotes
  - only 359 training documents, compared to 1,940 English ones

Paper	**Method**	**Arabic only**	**Arabic & English**	**Arabic & Chinese**
Min (2021)	e2e, mBERT-base	46.8	56.4	
Min (2021)	e2e, GigaBERT-base	62.1	64.6	
	CorPipe, mT5-large	64.1	66.1	65.9

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results on Arabic OntoNotes
  - only 359 training documents, compared to 1,940 English ones

Paper	**Method**	**Arabic only**	**Arabic & English**	**Arabic & Chinese**
Min (2021)	e2e, mBERT-base	46.8	56.4	
Min (2021)	e2e, GigaBERT-base	62.1	64.6	
	CorPipe, mT5-large	64.1	66.1	65.9
Bohnet et al. (2022)	seq2seq, mT5-xxl		68.7	

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results also on Chinese OntoNotes

Paper	Method	Chinese only	Chinese & English	Chinese & Arabic
Xia and Durme (2021)	ICoref, XLM-R-large	63.2	69.0	

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results also on Chinese OntoNotes

Paper	Method	Chinese only	Chinese & English	Chinese & Arabic
Xia and Durme (2021)	ICoref, XLM-R-large	63.2	69.0	
	CorPipe, mT5-large	70.3	71.6	70.2

Training on Multiple Treebanks, OntoNotes Arabic

- Similar results also on Chinese OntoNotes

Paper	Method	Chinese only	Chinese & English	Chinese & Arabic
Xia and Durme (2021)	ICoref, XLM-R-large	63.2	69.0	
	CorPipe, mT5-large	70.3	71.6	70.2
Bohnet et al. (2022)	seq2seq, mT5-xxl		74.3	

Language-specific vs Multilingual PLMs

Language-specific vs Multilingual PLMs

- For same-sized PLMs & individual treebanks, the results are mixed.

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
G) EFFECT OF SEVERAL LANGUAGE-SPECIFIC BASE PRETRAINED MODELS														
XLM-R base individual	68.7	71.4	75.7	73.9	65.7	62.0	71.2	63.2	75.6	63.1	61.5	73.4	69.8	65.6
mBERT (Devlin et al., 2019)	-2.8	-1.5	-3.0	-3.4	-3.3	+0.4	-2.8	-1.1	-1.8	-1.1	-2.7	-7.5	-4.4	-3.6
BERTa (Armengol-Estapé et al., 2021)		+1.3												
RobeCzech (Straka et al., 2021)			+2.0	+2.8										
gBERT (Chan et al., 2020)					-9.9	+5.3								
SpanBERT (Joshi et al., 2020)							-0.4	-2.4						
BETO (Cañete et al., 2020)									+0.4					
CamemBERT (Martin et al., 2020)										-0.2				
HuBERT (Nemeskey, 2020)											+3.6			
LitLatBERT (Ulčar and Robnik-Šikonja, 2021)												+2.7		
HerBERT (Mroczkowski et al., 2021)													+1.6	
RuBERT (Kuratov and Arkhipov, 2019)														+0.2
XLM-R large individual	+4.0	+4.6	+3.1	+4.1	+0.0	+6.9	+1.0	**+7.8**	+3.8	+3.3	+7.4	-0.8	+5.8	+4.8
RemBERT individual	-0.0	+4.9	+3.1	+3.1	-15.2	+0.0	+2.6	-18.3	+3.9	+3.8	+3.3	-4.3	+5.0	+4.3
XLM-R large multilingual	+6.1	**+6.1**	+2.1	+3.2	+8.0	**+16.2**	+4.1	+7.7	+5.0	+4.8	+6.9	**+4.6**	+5.1	+6.9
RemBERT multilingual	**+6.6**	+6.0	**+3.6**	**+4.4**	**+10.6**	+14.5	**+4.3**	+6.1	**+5.5**	**+5.1**	**+7.7**	+3.5	**+6.0**	**+9.0**

- For same-sized PLMs & multilingual training, the results are mostly worse.

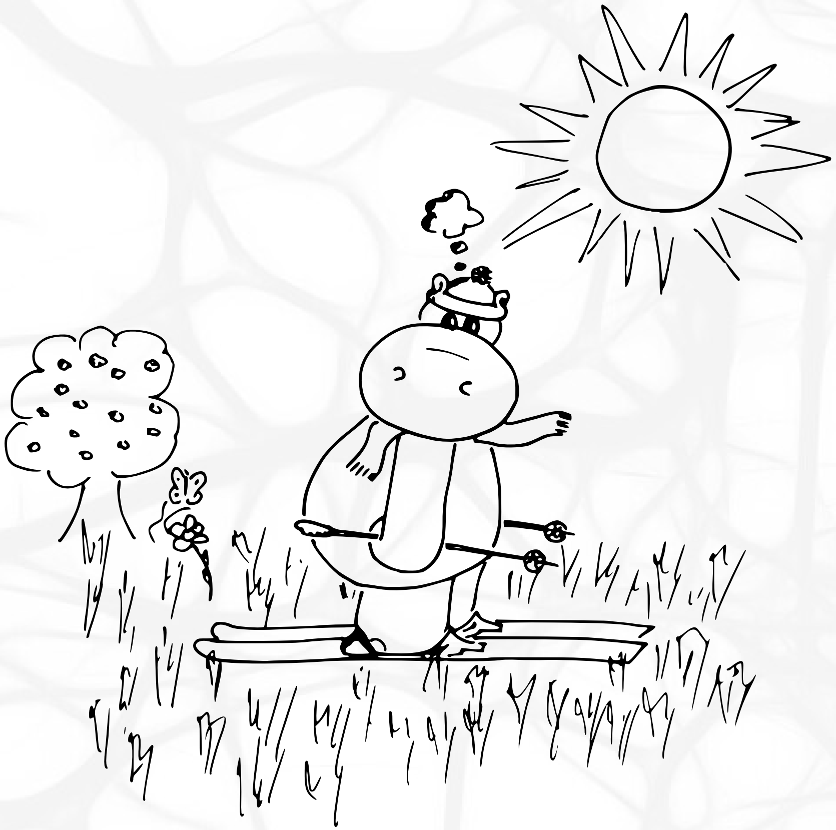
Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
G) Effect of Several Language-specific base Pretrained Models														
XLM-R base individual	68.7	71.4	75.7	73.9	65.7	62.0	71.2	63.2	75.6	63.1	61.5	73.4	69.8	65.6
mBERT (Devlin et al., 2019)	-2.8	-1.5	-3.0	-3.4	-3.3	+0.4	-2.8	-1.1	-1.8	-1.1	-2.7	-7.5	-4.4	-3.6
BERTa (Armengol-Estapé et al., 2021)		+1.3												
RobeCzech (Straka et al., 2021)			+2.0	+2.8										
gBERT (Chan et al., 2020)					-9.9	+5.3								
SpanBERT (Joshi et al., 2020)							-0.4	-2.4						
BETO (Cañete et al., 2020)									+0.4					
CamemBERT (Martin et al., 2020)										-0.2				
HuBERT (Nemeskey, 2020)											+3.6			
LitLatBERT (Ulčar and Robnik-Šikonja, 2021)												+2.7		
HerBERT (Mroczkowski et al., 2021)													+1.6	
RuBERT (Kuratov and Arkhipov, 2019)														+0.2
XLM-R large individual	+4.0	+4.6	+3.1	+4.1	+0.0	+6.9	+1.0	**+7.8**	+3.8	+3.3	+7.4	-0.8	+5.8	+4.8
RemBERT individual	-0.0	+4.9	+3.1	+3.1	-15.2	+0.0	+2.6	-18.3	+3.9	+3.8	+3.3	-4.3	+5.0	+4.3
XLM-R large multilingual	+6.1	**+6.1**	+2.1	+3.2	+8.0	**+16.2**	+4.1	+7.7	+5.0	+4.8	+6.9	**+4.6**	+5.1	+6.9
RemBERT multilingual	**+6.6**	+6.0	**+3.6**	**+4.4**	**+10.6**	+14.5	**+4.3**	+6.1	**+5.5**	**+5.1**	**+7.7**	+3.5	**+6.0**	**+9.0**
C) Effect of Multilingual Data and the Pretrained Model														
XLM-R base multilingual	73.3	75.8	76.0	75.0	73.4	74.1	73.1	**75.4**	78.4	66.1	65.2	**78.0**	72.1	71.7
XLM-R base individual	-4.6	-4.4	-0.3	-1.1	-7.8	-12.1	-1.9	-12.2	-2.8	-3.0	-3.8	-4.6	-2.3	-6.1

Language-specific vs Multilingual PLMs

- Base-sized language-specific PLMs worse than large-sizes multilingual.

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
G) EFFECT OF SEVERAL LANGUAGE-SPECIFIC BASE PRETRAINED MODELS														
XLM-R base individual	68.7	71.4	75.7	73.9	65.7	62.0	71.2	63.2	75.6	63.1	61.5	73.4	69.8	65.6
mBERT (Devlin et al., 2019)	-2.8	-1.5	-3.0	-3.4	-3.3	+0.4	-2.8	-1.1	-1.8	-1.1	-2.7	-7.5	-4.4	-3.6
BERTa (Armengol-Estapé et al., 2021)		+1.3												
RobeCzech (Straka et al., 2021)			+2.0	+2.8										
gBERT (Chan et al., 2020)					-9.9	+5.3								
SpanBERT (Joshi et al., 2020)							-0.4	-2.4						
BETO (Cañete et al., 2020)									+0.4					
CamemBERT (Martin et al., 2020)										-0.2				
HuBERT (Nemeskey, 2020)											+3.6			
LitLatBERT (Ulčar and Robnik-Šikonja, 2021)												+2.7		
HerBERT (Mroczkowski et al., 2021)													+1.6	
RuBERT (Kuratov and Arkhipov, 2019)														+0.2
XLM-R large individual	+4.0	+4.6	+3.1	+4.1	+0.0	+6.9	+1.0	**+7.8**	+3.8	+3.3	+7.4	-0.8	+5.8	+4.8
RemBERT individual	-0.0	+4.9	+3.1	+3.1	-15.2	+0.0	+2.6	-18.3	+3.9	+3.8	+3.3	-4.3	+5.0	+4.3
XLM-R large multilingual	+6.1	**+6.1**	+2.1	+3.2	+8.0	**+16.2**	+4.1	+7.7	+5.0	+4.8	+6.9	**+4.6**	+5.1	+6.9
RemBERT multilingual	**+6.6**	+6.0	**+3.6**	**+4.4**	**+10.6**	+14.5	**+4.3**	+6.1	**+5.5**	**+5.1**	**+7.7**	+3.5	**+6.0**	**+9.0**

Zero-shot Evaluation of Unseen Language



Zero-shot Evaluation of Unseen Language

- CorPipe 22 unseen language performance comparable to the shared task baseline (c2f + mBERT)

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
F) Zero-shot Evaluation of a Multilingual Model														
Multilingual XLM-R base	73.3	75.8	76.0	75.0	73.4	74.1	73.1	**75.4**	78.4	66.1	65.2	**78.0**	72.1	71.7
Zero-shot XLM-R base	-17.1	-11.1	-28.6	-23.8	-13.3	-13.8	-19.8	-18.5	-6.8	-7.6	-16.1	-23.8	-24.6	-15.1
Multilingual RemBERT	**+1.9**	**+1.6**	**+3.3**	**+3.3**	**+2.9**	**+2.4**	**+2.4**	-6.1	**+2.7**	**+2.0**	**+4.0**	-1.2	**+3.7**	**+2.9**
Zero-shot RemBERT	-12.5	-6.7	-23.7	-20.6	-11.1	-7.5	-15.6	-9.8	-2.8	-8.3	-10.5	-20.0	-18.3	-7.2
Multilingual RemBERT	**75.3**	**77.4**	**79.3**	**78.3**	**76.3**	**76.5**	**75.5**	**69.3**	**81.1**	**68.1**	**69.2**	**76.8**	**75.8**	**74.6**
Zero-shot RemBERT	-14.4	-8.3	-27.0	-23.8	-14.0	-9.9	-18.0	-3.7	-5.6	-10.4	-14.5	-18.8	-22.0	-10.2

Model	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
Baseline to RemBERT	-11,0	-13,3	-9,1	-10,7	-20,9	-19,1	-9,1	-12,0	-15,5	-13,6	-10,5	-8,1	-12,2	-11,9

Zero-shot Evaluation of Unseen Language

- CorPipe 23 unseen language performance slightly better than the shared task baseline (c2f + mBERT)

Configuration	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu korko	hu szege	lt	no bookm	no nynor	pl	ru	tr
Single Multilingual Model	**74.8**	**81.6**	**80.3**	**79.0**	**69.7**	**75.4**	**76.8**	**66.0**	**82.8**	**70.3**	**69.5**	**69.8**	**77.9**	**81.5**	**81.7**	**77.1**	**75.2**	**57.2**
Per-Corpus Models	–3.7	–1.4	–0.5	–0.4	–7.7	–3.3	–1.6	–7.6	–1.5	–2.0	–9.1	–1.0	–3.0	–2.3	–2.9	–1.0	–2.0	–15.8
Joint Czech Model			–0.1	–0.3														
Joint German Model					–4.8	–3.9												
Joint English Model							–1.9	–4.5										
Joint Parcorfull Model					–4.4			–2.5										
Joint Hungarian Model											–5.9	–1.1						
Joint Norwegian Model														–1.3	–1.8			
Zero-Shot Multilingual Models	–13.2	–4.8	–24.2	–16.0	–13.7	–10.6	–14.4	–13.8	–1.9	–5.4	–15.1	–15.0	–23.4	–14.3	–18.0	–17.5	–15.5	–0.8

Table 6: Ablation experiments evaluated on the development sets (CoNLL score in %) using the mT5-large model with context size 2560. We report the average of best 5 out of 7 runs, using for every corpus the single epoch achieving the highest average 5-run score.

Model	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu korko	hu szege	lt	no bookm	no nynor	pl	ru	tr
Baseline to Multilingual	-17,8	-16,3	-12,6	-13,8	-25,6	-18,3	-13,7	-30,8	-15,9	-15,0	-14,2	-6,2	-11,8	-12,5	-41,0	-12,0	-9,4	-34,5

Zero-shot Evaluation of Unseen Language

ÚFAL

- OntoNotes demonstrates similar behavior, with largest decrease on unseen Chinese

Model	English	Arabic	Chinese
CorPipe, mT5-large, individual treebanks	77.2	64.1	70.3

Zero-shot Evaluation of Unseen Language

- OntoNotes demonstrates similar behavior, with largest decrease on unseen Chinese

Model	English	Arabic	Chinese
CorPipe, mT5-large, individual treebanks	77.2	64.1	70.3
CorPipe, mT5-large, unseen language	61.7	54.1	48.3

Unseen Language Example on en_gum, Byron Biography

Education and early **loves**

Byron received **his** early formal **education** at Aberdeen Grammar **School**, and in **August** 1799 entered the **school** of **Dr.** William Glennie, in **Dulwich** . [17]

Placed under the **care** of a **Dr.** Bailey , **he** was encouraged to exercise in **moderation** but not restrain **himself** from "violent" **bouts** in an **attempt** to overcompensate for **his** deformed **foot** .

His mother interfered with **his studies**, often withdrawing **him** from **school**, with the **result** that **he** lacked **discipline** and **his** classical **studies** were neglected .

In **1801** , **he** was sent to **Harrow**, where **he** remained until **July** 1805. [6]

An undistinguished **student** and an unskilled **cricketer** , **he** did represent the **school** during the very first **Eton** v **Harrow** cricket **match** at **Lord** 's in 1805. [19]

His lack of **moderation** was not restricted to physical **exercise**.

Byron fell in **love** with **Mary** Chaworth, **whom he** met while at **school** , [6] and **she** was the **reason he** refused to return to **Harrow** in **September** 1803 .

His mother wrote, " **He** has no **indisposition** that **I** know of but **love**, desperate **love**, the worst of all **maladies** in **my opinion** . In short, the **boy** is distractedly in **love** with **Miss** Chaworth ." [6]

In **Byron** 's later **memoirs**, " **Mary** Chaworth is portrayed as the first **object** of **his** adult sexual **feelings** ." [20]

Byron finally returned in **January** 1804, [6] to a more settled **period** which saw the **formation** of a **circle** of emotional **involvements** with other Harrow **boys** , which **he** recalled with great **vividness**: " **My** school **friendships** were with **me passions** (for **I** was always violent)." [21]

Questions?