ÚFAL CorPipe at CRAC 2023: Larger Context Improves Multilingual 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

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks
    - usable also on unseen languages

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks
    - usable also on unseen languages
    - source code released, pre-trained model being released

CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks
    - usable also on unseen languages
    - source code released, pre-trained model being released
  - supports documents of unbounded size (still uses quadratic attention)

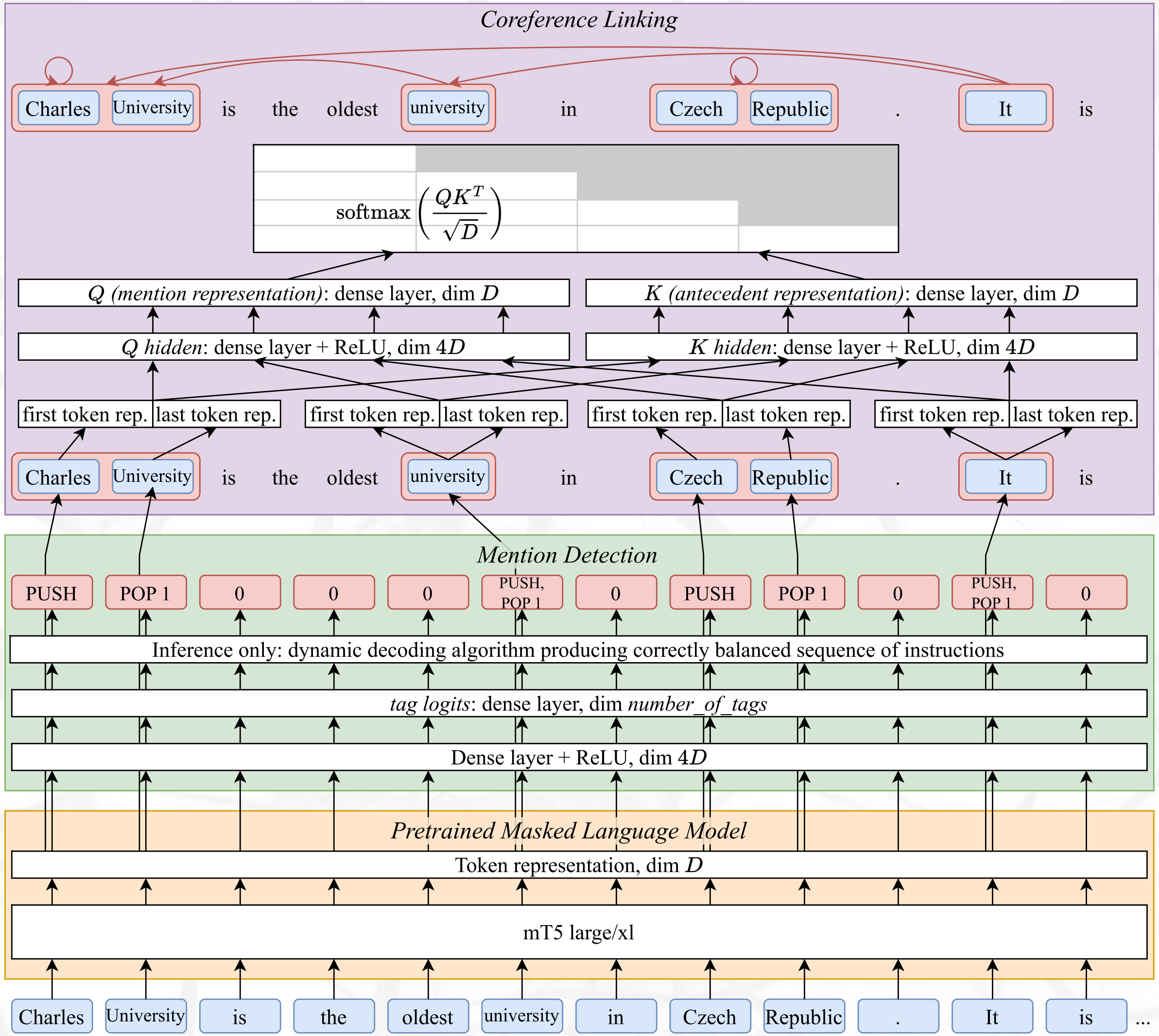
CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks
    - usable also on unseen languages
    - source code released, pre-trained model being released
  - supports documents of unbounded size (still uses quadratic attention)
  - supports ensembling using multiple GPUs in parallel

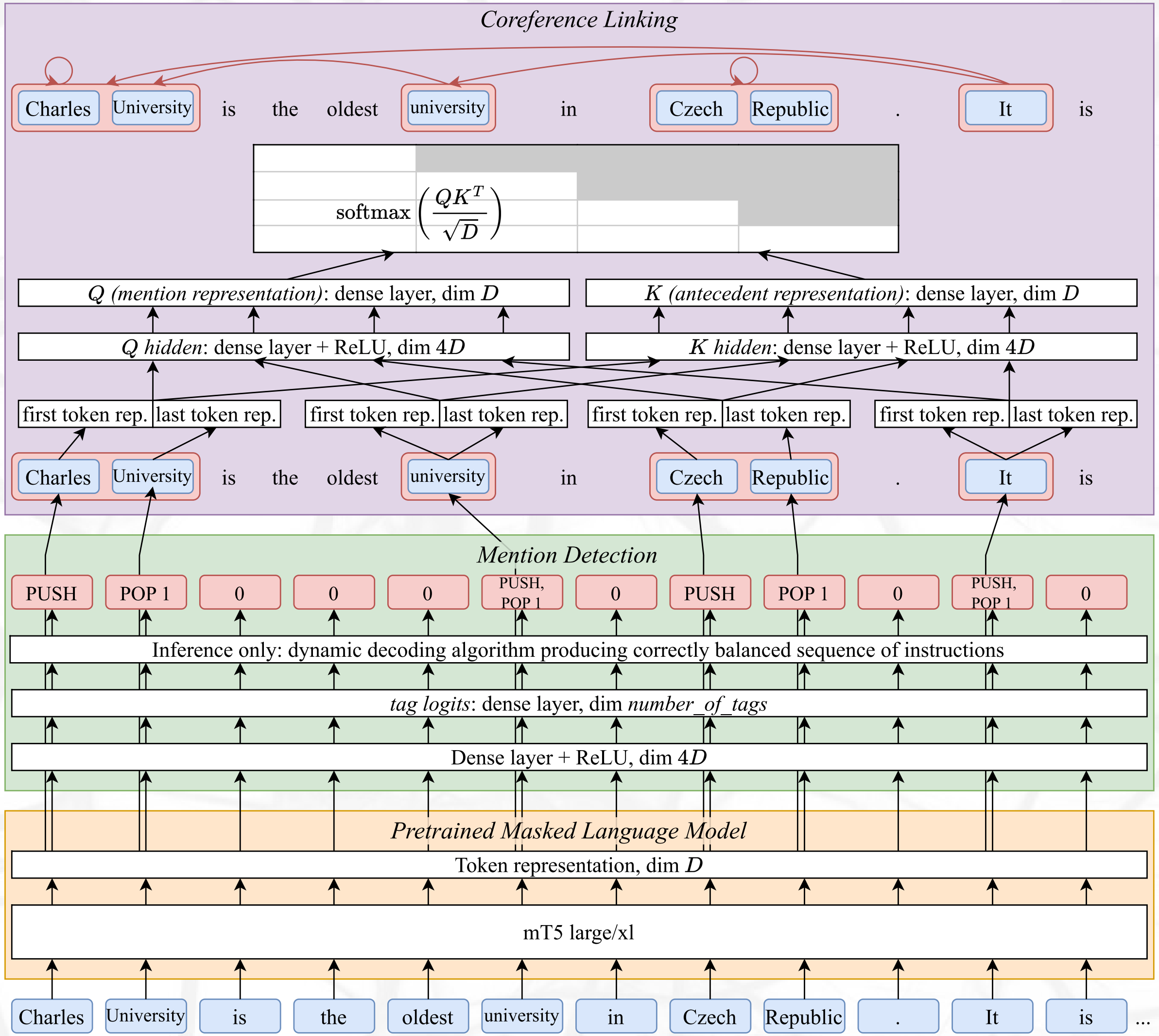
CorPipe 23

- winning entry of the CRAC 2023 Shared Task on Multilingual Coreference Resolution
  - a slight evolution of the CorPipe system from CRAC 2022
- distinguishing features:
  - a single multilingual model for all 17 treebanks
    - usable also on unseen languages
    - source code released, pre-trained model being released
  - supports documents of unbounded size (still uses quadratic attention)
  - supports ensembling using multiple GPUs in parallel
  - can generate singleton mentions

CorPipe 23 Architecture

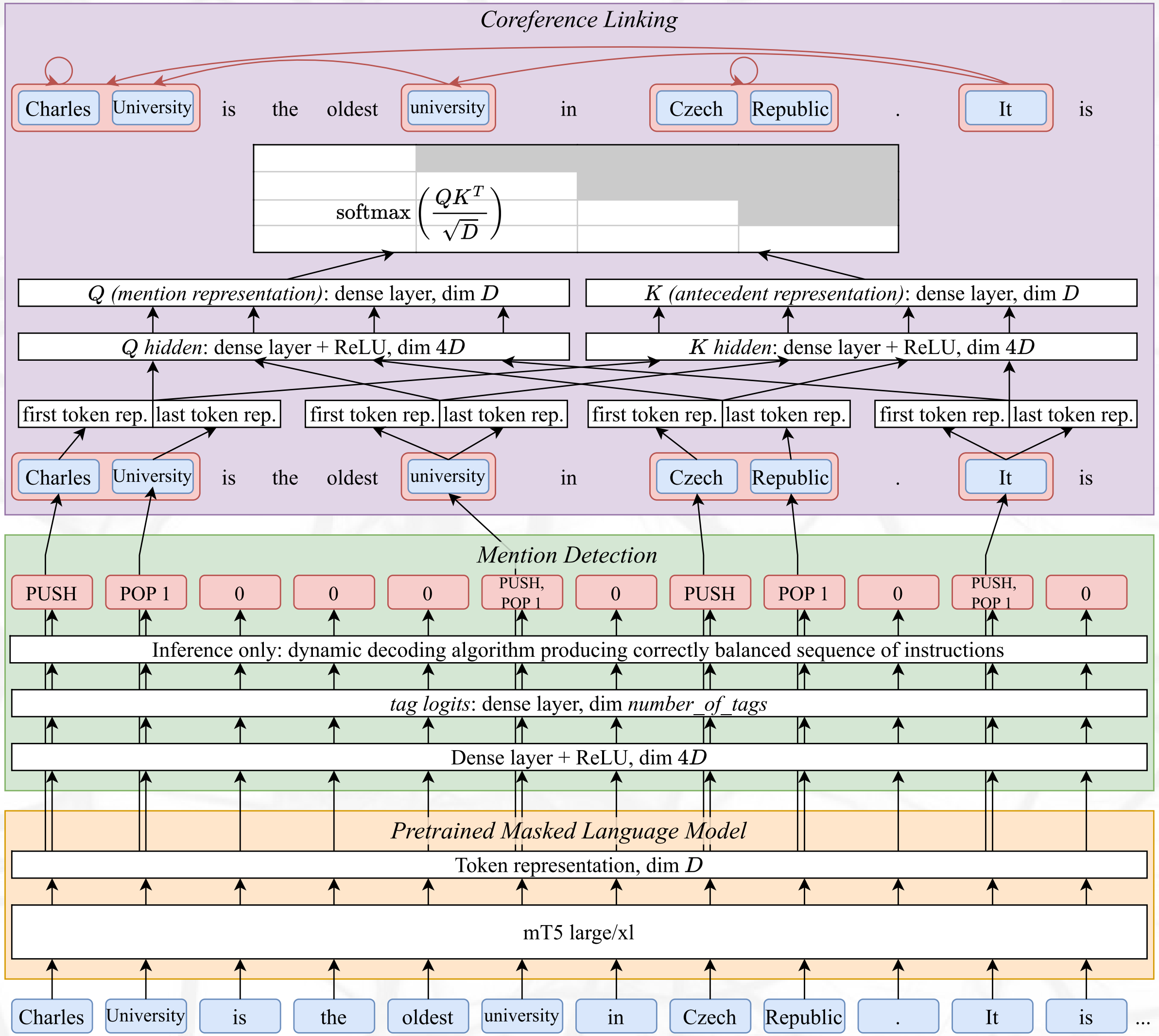


We first detect mentions with an extension of a BIO encoding.



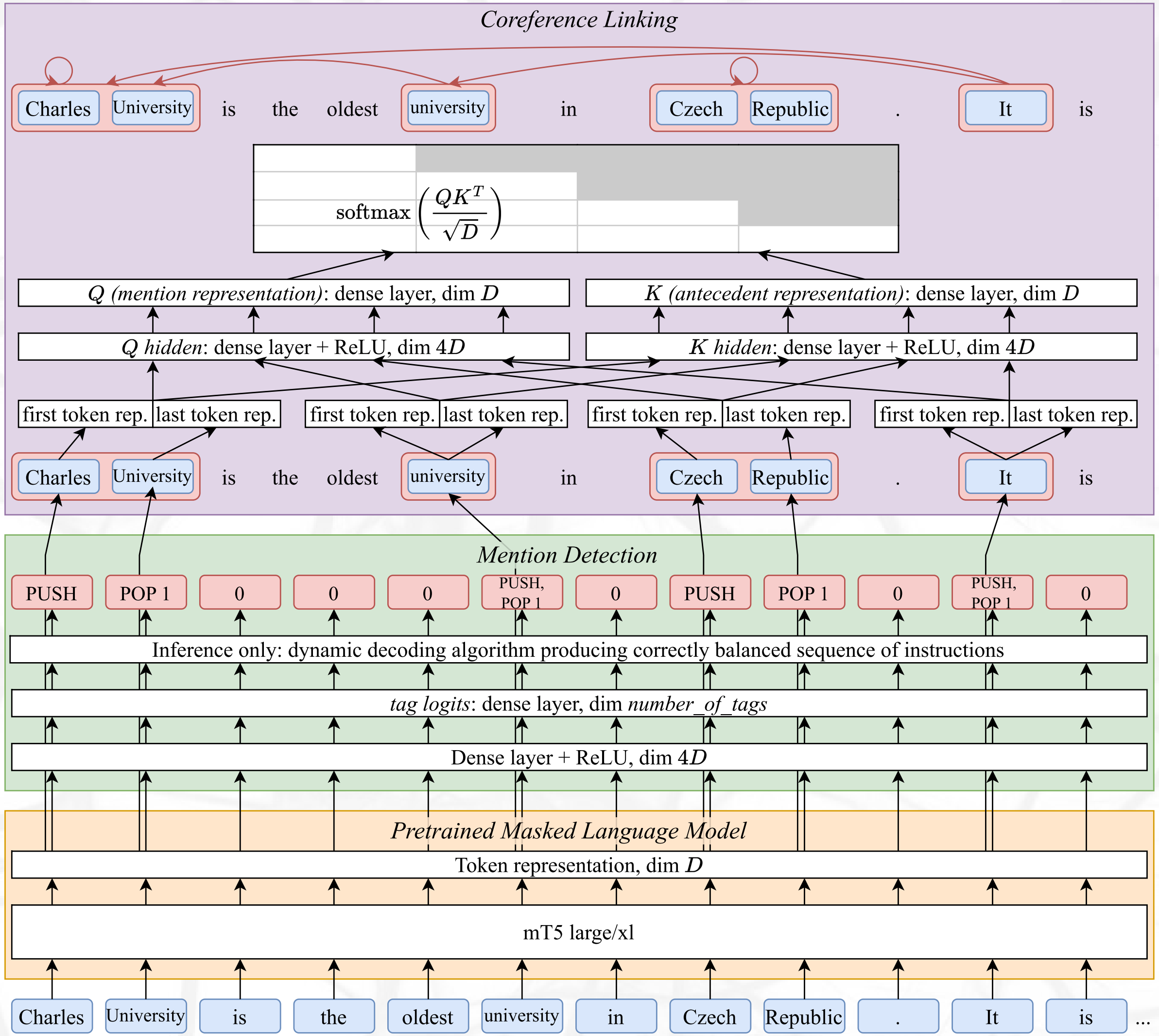
We first detect mentions with an extension of a BIO encoding.

- In every token, we
  - `PUSH` starting mentions to the stack,



We first detect mentions with an extension of a BIO encoding.

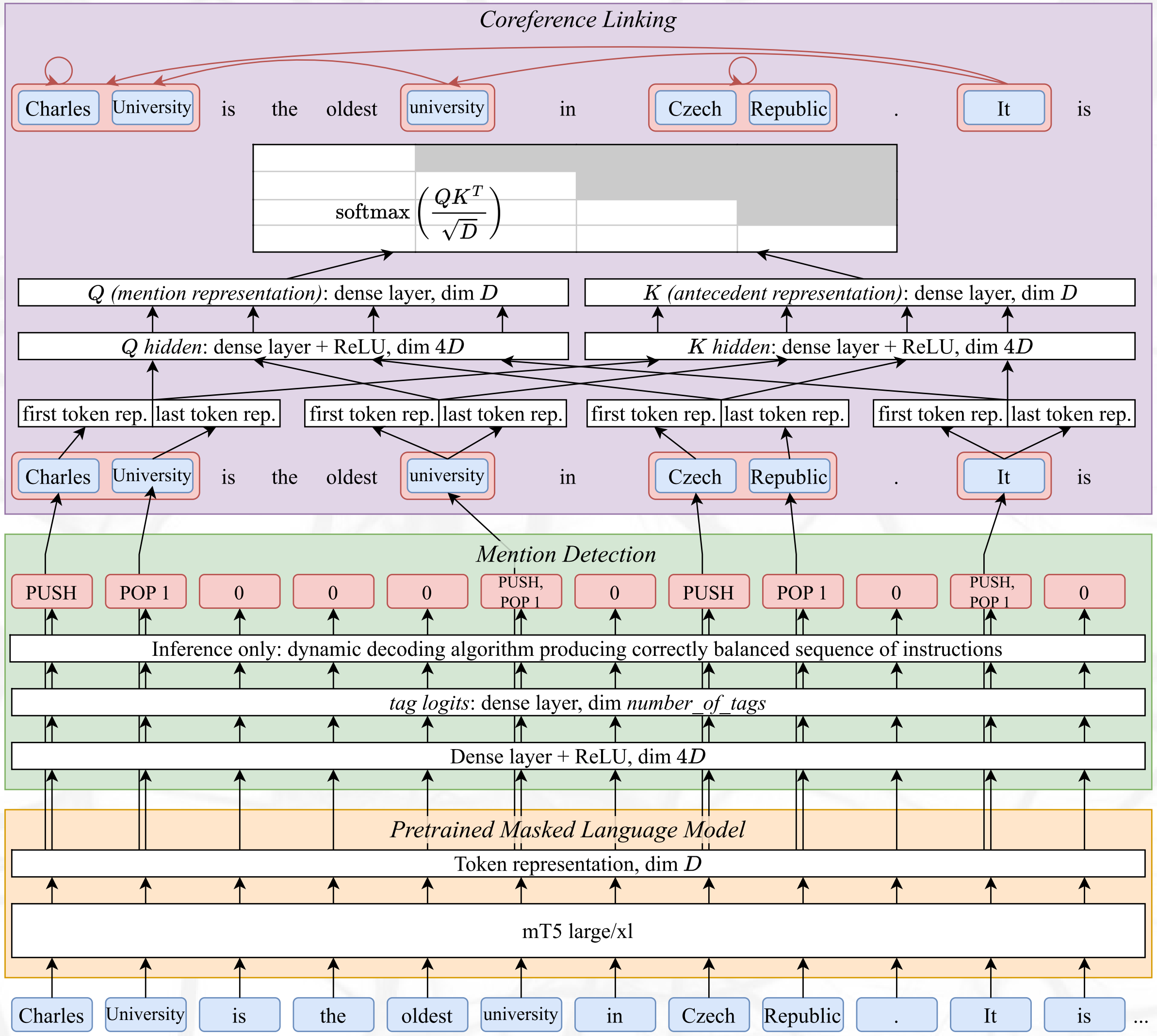
- In every token, we
  - PUSH starting mentions to the stack,
  - POP(i) every mention ending in stack.



We first detect mentions with an extension of a BIO encoding.

- In every token, we
  - `PUSH` starting mentions to the stack,
  - `POP(i)` every mention ending in stack.

Because the mentions can be crossing, the `POP` instructions is parametrized by the stack index of the ending mention.

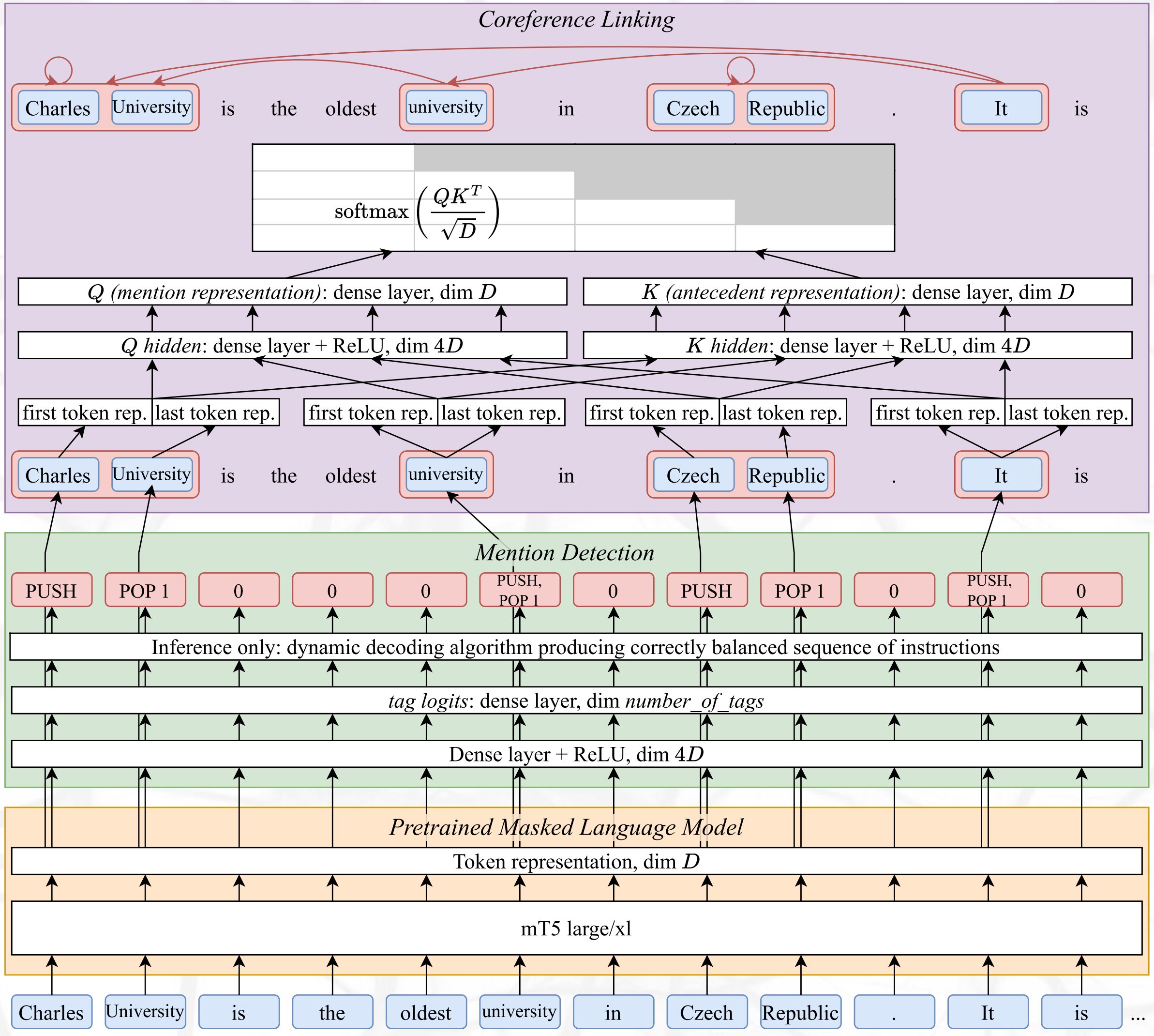


CorPipe 23 Architecture

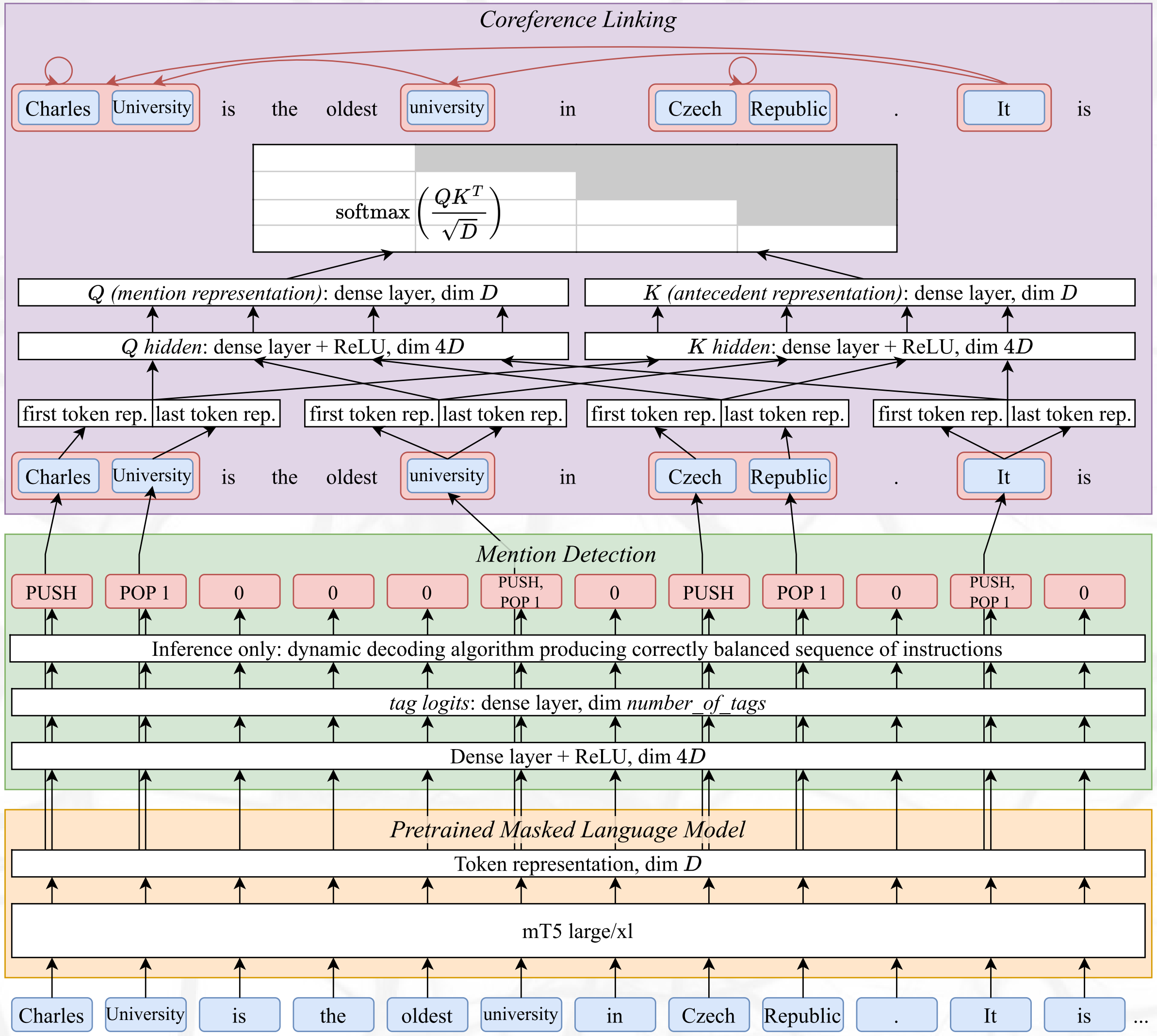
We first detect mentions with an extension of a BIO encoding.

- In every token, we
  - `PUSH` starting mentions to the stack,
  - `POP(i)` every mention ending in stack.

  Because the mentions can be crossing, the `POP` instructions is parametrized by the stack index of the ending mention.
- We considered CRF, but no gain & difficult ensembling.

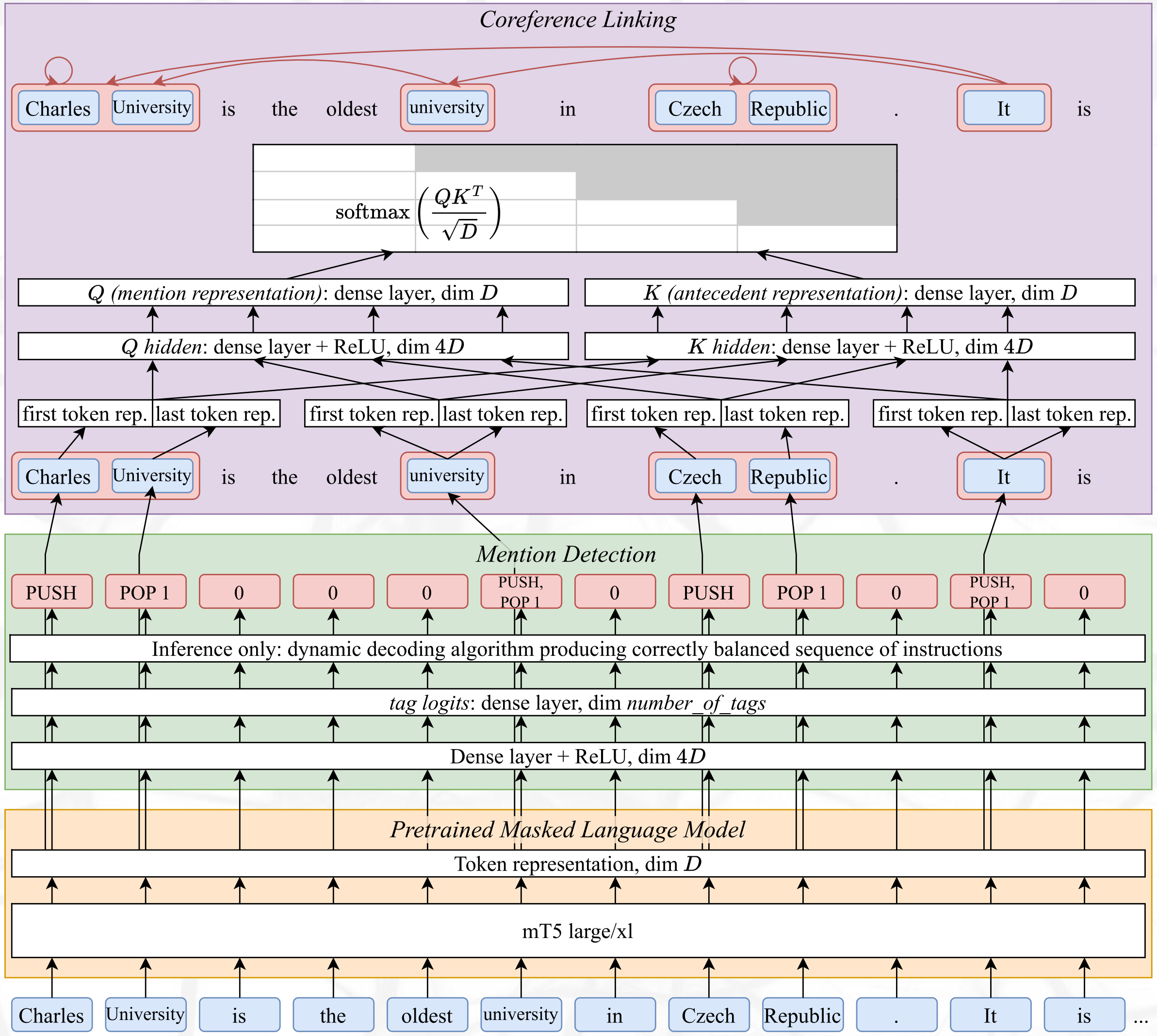


We train an antecedent predictor by minimizing multilabel classification loss with all preceding antecedents as targets, with the distribution computed using self-attention.



We train an antecedent predictor by minimizing multilabel classification loss with all preceding antecedents as targets, with the distribution computed using self-attention.

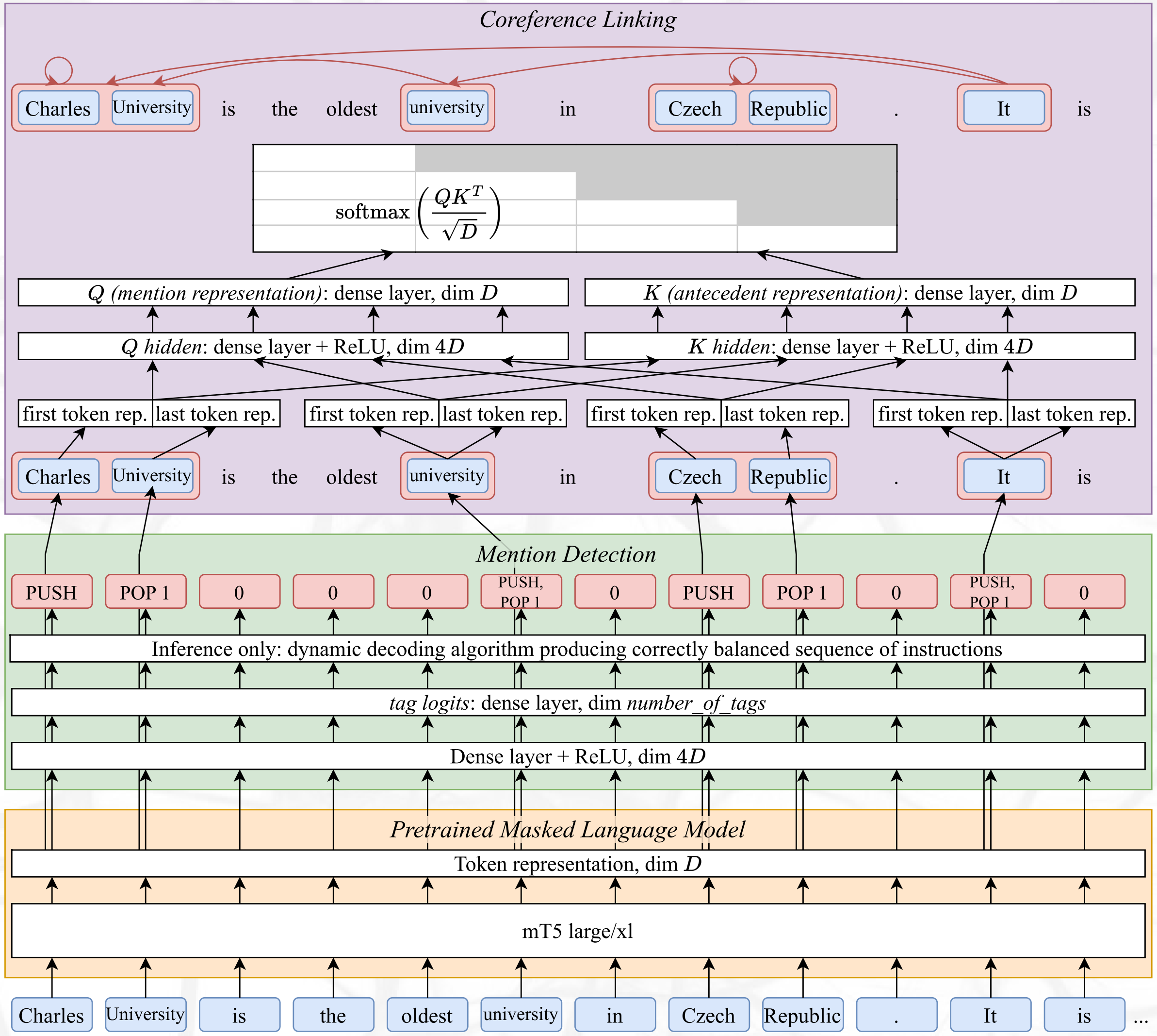
- If a mention has no antecedent, we link it to **itself.**



We train an antecedent predictor by minimizing multilabel classification loss with all preceding antecedents as targets, with the distribution computed using self-attention.

- If a mention has no antecedent, we link it to **itself.**

During prediction, we predict only the most probable link.

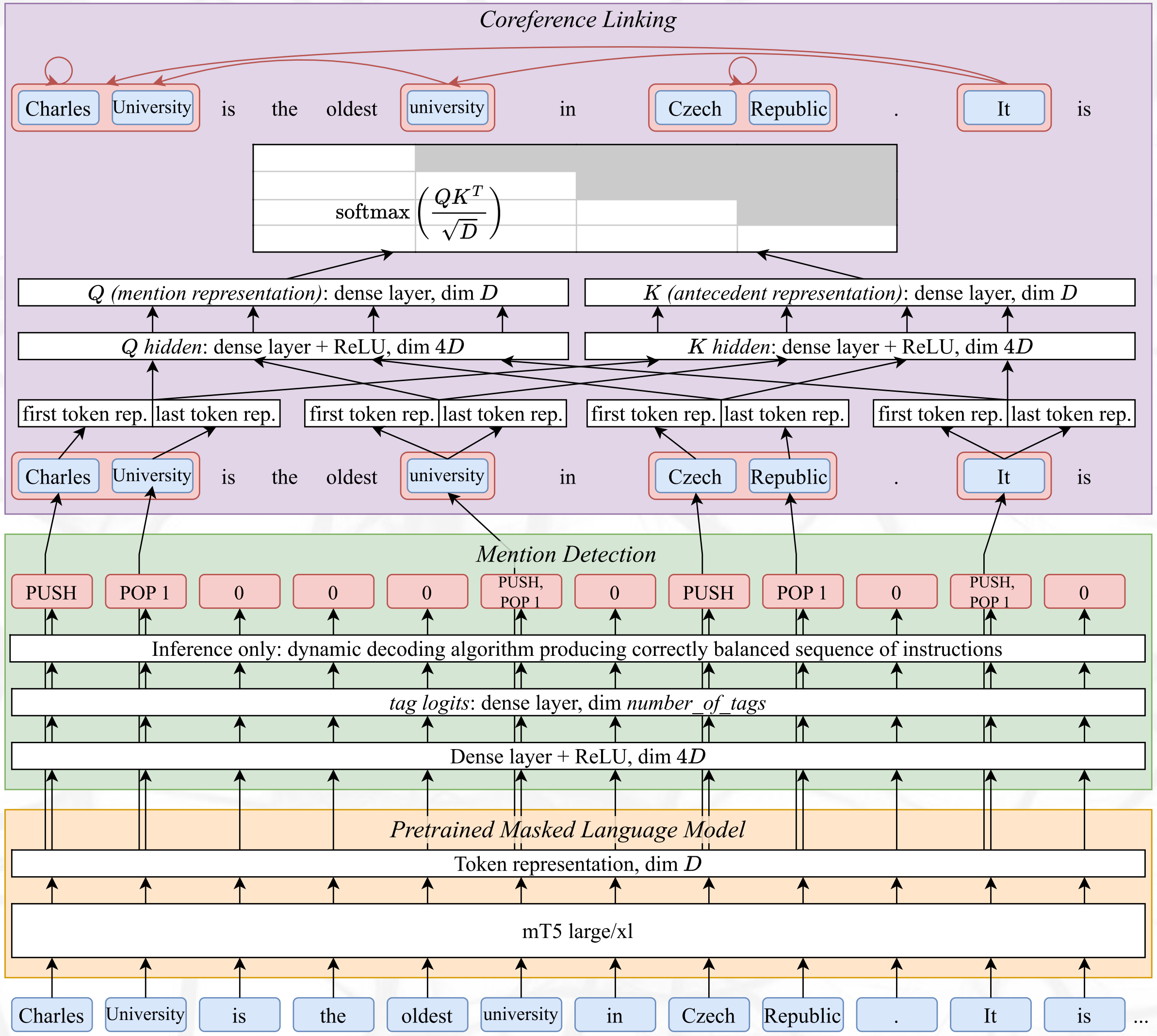


We train an antecedent predictor by minimizing multilabel classification loss with all preceding antecedents as targets, with the distribution computed using self-attention.

- If a mention has no antecedent, we link it to **itself.**

During prediction, we predict only the most probable link.

Both tasks are trained jointly using a shared encoder.



CorPipe 23 Training

We train multilingual models on all 17 treebanks of CorefUD 1.1.

CorPipe 23 Training

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:
  - *uniform*: each corpus has the same probability;

CorPipe 23 Training

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:
  - *uniform*: each corpus has the same probability;
  - *linear*: proportionally to corpus size;

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:
  - *uniform*: each corpus has the same probability;
  - *linear*: proportionally to corpus size;
  - *square root*: proportionally to square root of the corpus size;

CorPipe 23 Training

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:
  - *uniform*: each corpus has the same probability;
  - *linear*: proportionally to corpus size;
  - *square root*: proportionally to square root of the corpus size;
  - *logarithmic*: proportionally to the corpus size logarithm.

We train multilingual models on all 17 treebanks of CorefUD 1.1.

- Because the sizes range from tiny (457 sentences) to large (almost 40k sentences), we consider sub-/over-sampling the individual datasets, sampling each batch proportionally to *mix ratios*:
  - *uniform*: each corpus has the same probability;
  - *linear*: proportionally to corpus size;
  - *square root*: proportionally to square root of the corpus size;
  - *logarithmic*: proportionally to the corpus size logarithm.
- The data might or might not get a corpus id subword indicating the origin of the document.

CorPipe 23 Training

We train models with either

- mT5-large ($\sim$560M parameters),or
- mT5-xl ($\sim$1.8G parameters) encoders.

CorPipe 23 Training

We train models with either

- mT5-large (∼560M parameters),or
- mT5-xl (∼1.8G parameters) encoders.

We use the Adafactor optimizer using 5e-4 learning rate with slanted triangual schedule.

CorPipe 23 Training

We train models with either

- mT5-large (~560M parameters),or
- mT5-xl (~1.8G parameters) encoders.

We use the Adafactor optimizer using 5e-4 learning rate with slanted triangual schedule.

Our default configuration is to train for 15 epochs, each comprising 8k steps.

We train models with either

- mT5-large (~560M parameters),or
- mT5-xl (~1.8G parameters) encoders.

We use the Adafactor optimizer using 5e-4 learning rate with slanted triangual schedule.

Our default configuration is to train for 15 epochs, each comprising 8k steps.

- The large-sized models are trained on a single 40GB A100 GPU for 10 hours, with a maximum possible batch size 8.

CorPipe 23 Training

We train models with either

- mT5-large (~560M parameters),or
- mT5-xl (~1.8G parameters) encoders.

We use the Adafactor optimizer using 5e-4 learning rate with slanted triangual schedule.

Our default configuration is to train for 15 epochs, each comprising 8k steps.

- The large-sized models are trained on a single 40GB A100 GPU for 10 hours, with a maximum possible batch size 8.
- The xl-sized models are trained on four 40GB A100 GPUs for 20 hours, with a maximum possible batch size 12.

CorPipe 23 Training

We train models with either

- mT5-large (~560M parameters),or
- mT5-xl (~1.8G parameters) encoders.

We use the Adafactor optimizer using 5e-4 learning rate with slanted triangual schedule.

Our default configuration is to train for 15 epochs, each comprising 8k steps.

- The large-sized models are trained on a single 40GB A100 GPU for 10 hours, with a maximum possible batch size 8.
- The xl-sized models are trained on four 40GB A100 GPUs for 20 hours, with a maximum possible batch size 12.

During training, we consider segments of up to 512 subwords; during inference, we scale up to 2560 subwords (mT5 uses relative positional encodings).

CorPipe 23 Training

We trained 30 models differing in size (large or xl) and several hyperparameters (learning rate, batch size, updates per epoch).

CorPipe 23 Training

We trained 30 models differing in size (large or xl) and several hyperparameters (learning rate, batch size, updates per epoch).

For each treebank we consider the best-performing checkpoint of every model after every epoch.

CorPipe 23 Training

We trained 30 models differing in size (large or xl) and several hyperparameters (learning rate, batch size, updates per epoch).

For each treebank we consider the best-performing checkpoint of every model after every epoch.

Optionally, we perform ensembling of the trained models, by averaging the predicted distributions.

CorPipe 23 Training

We trained 30 models differing in size (large or xl) and several hyperparameters (learning rate, batch size, updates per epoch).

For each treebank we consider the best-performing checkpoint of every model after every epoch.

Optionally, we perform ensembling of the trained models, by averaging the predicted distributions.

- During inference, models are loaded on individual GPUs and executed in parallel.

CorPipe 23 Training

We trained 30 models differing in size (large or xl) and several hyperparameters (learning rate, batch size, updates per epoch).

For each treebank we consider the best-performing checkpoint of every model after every epoch.

Optionally, we perform ensembling of the trained models, by averaging the predicted distributions.

- During inference, models are loaded on individual GPUs and executed in parallel.

Our main submission for every corpus is an ensemble of 3 best checkpoints.

Official CRAC 2023 Results per Treebank

System	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
ÚFAL CorPipe	**74.90** **1**	**82.59** **1**	**79.33** **1**	**79.20** **1**	**72.12** **1**	**71.09** **1**	**76.57** **1**	**69.86** **1**	**83.39** **1**	**69.82** **1**	**68.92** **1**	**69.47** **1**	**75.87** **1**	**78.74** **1**	**78.77** **1**	**79.54** **1**	**82.46** **1**	**55.63** **1**
Anonymous	70.41 2	79.51 2	75.88 2	76.39 2	64.37 3	68.24 5	72.29 2	59.02 3	80.52 2	66.13 2	64.65 3	66.25 2	70.09 2	75.32 2	73.33 2	77.58 2	80.19 2	47.22 2
Ondfa	69.19 3	76.02 3	74.82 3	74.67 3	71.86 2	69.37 3	71.56 3	61.62 2	77.18 3	60.32 4	66.38 2	65.75 4	68.52 3	72.39 4	70.91 4	76.90 3	76.50 4	41.52 4
McGill	65.43 4	71.75 4	67.67 7	70.88 4	41.58 7	70.20 2	66.72 4	47.27 4	73.78 4	65.17 3	60.74 4	65.93 3	65.77 6	73.73 3	72.43 3	76.14 4	77.28 3	45.28 3
DeepBlueAI	62.29 5	67.55 7	70.38 4	69.93 5	48.81 5	63.90 7	63.58 6	43.33 5	69.52 5	55.69 6	54.38 5	63.14 5	66.75 4	69.86 6	68.53 5	73.11 5	74.41 5	36.14 8
DFKI-Adapt	61.86 6	68.21 6	68.72 5	67.34 6	52.52 4	69.28 4	65.11 5	36.87 7	69.19 6	58.96 5	51.53 7	58.56 6	66.01 5	70.05 5	68.21 6	67.98 6	72.48 6	40.67 5
Morfbase	59.53 7	68.23 5	64.89 8	64.74 8	39.96 9	64.87 6	62.80 8	40.81 6	69.01 7	53.18 8	52.91 6	56.41 7	64.08 7	68.17 7	66.35 7	67.88 7	68.53 8	39.22 6
BASELINE†	56.96 8	65.26 8	67.72 6	65.22 7	44.11 6	57.13 9	63.08 7	35.19 8	66.93 8	55.31 7	40.71 9	55.32 8	63.57 8	65.10 9	65.78 8	66.08 8	69.03 7	22.75 9
DFKI-MPrompt	53.76 9	55.45 9	60.39 9	56.13 9	40.34 8	59.75 8	57.83 9	34.32 9	58.31 9	52.96 9	44.53 8	48.79 9	56.52 9	65.12 8	62.99 9	61.15 9	61.96 9	37.44 7

Table 1: Official results of CRAC 2023 Shared Task on the test set (CoNLL score in %). The system † is described in Pražák et al. (2021); the rest in Žabokrtský et al. (2023).

Official CRAC 2023 Results: Metrics, CorPipe 23 Variants

System	Head-match	Partial-match	Exact-match	+Singletons
ÚFAL CorPipe	**74.90 (1)**	**73.33 (1)**	**71.46 (1)**	**76.82 (1)**
Anonymous	70.41 (2)	69.23 (2)	67.09 (2)	73.20 (2)
Ondfa	69.19 (3)	68.93 (3)	53.01 (8)	68.37 (3)
McGill	65.43 (4)	64.56 (4)	63.13 (3)	68.23 (4)
DeepBlueAI	62.29 (5)	61.32 (5)	59.95 (4)	54.51 (5)
DFKI-Adapt	61.86 (6)	60.83 (6)	59.18 (5)	53.94 (6)
Morfbase	59.53 (7)	58.49 (7)	56.89 (6)	52.07 (7)
BASELINE	56.96 (8)	56.28 (8)	54.75 (7)	49.32 (8)
DFKI-MPrompt	53.76 (9)	51.62 (9)	50.42 (9)	46.83 (9)

Table 2: Official results of CRAC 2023 Shared Task on the test set with various metrics in %.

Official CRAC 2023 Results: Metrics, CorPipe 23 Variants

System	Head-match	Partial-match	Exact-match	+Singletons
ÚFAL CorPipe	**74.90 (1)**	**73.33 (1)**	**71.46 (1)**	**76.82 (1)**
Anonymous	70.41 (2)	69.23 (2)	67.09 (2)	73.20 (2)
Ondfa	69.19 (3)	68.93 (3)	53.01 (8)	68.37 (3)
McGill	65.43 (4)	64.56 (4)	63.13 (3)	68.23 (4)
DeepBlueAI	62.29 (5)	61.32 (5)	59.95 (4)	54.51 (5)
DFKI-Adapt	61.86 (6)	60.83 (6)	59.18 (5)	53.94 (6)
Morfbase	59.53 (7)	58.49 (7)	56.89 (6)	52.07 (7)
BASELINE	56.96 (8)	56.28 (8)	54.75 (7)	49.32 (8)
DFKI-MPrompt	53.76 (9)	51.62 (9)	50.42 (9)	46.83 (9)

Table 2: Official results of CRAC 2023 Shared Task on the test set with various metrics in %.

Submission	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
Original CorPipe 2022	70.3	79.9	76.0	76.8	63.3	**72.6**	72.3	57.6	81.2	65.4	66.2	65.4	68.6	75.4	73.6	79.0	78.4	42.5
Single mT5 large model	+2.6	+2.2	+2.1	+0.8	+6.7	–1.2	+1.6	+4.0	+0.9	+0.1	+1.6	+3.3	**+7.4**	**+3.5**	+2.2	–0.5	+2.4	+7.6
Single mT5 xl model	+2.7	+2.0	+2.0	+1.5	+2.7	–3.0	+2.9	+6.8	+1.6	+2.6	–0.7	**+4.1**	+4.7	+3.3	+3.7	–0.3	+2.6	+10.3
Per-treebank best mT5 model	+3.4	+2.6	+1.7	+1.6	**+13.1**	–4.1	+3.2	+10.3	+1.2	+3.3	–0.2	+2.0	+6.6	+3.0	+4.2	–0.8	+3.8	+7.6
Per-treebank 3-model ensemble	**+4.6**	**+2.7**	**+3.3**	**+2.4**	+8.8	–1.5	**+4.3**	**+12.3**	**+2.2**	**+4.4**	**+2.7**	+4.1	+7.3	+3.3	**+5.2**	**+0.5**	**+4.1**	**+13.1**
Per-treebank 8-model ensemble	***+4.9***	***+3.3***	***+3.3***	***+2.7***	*+7.7*	*–0.8*	*+4.2*	***+13.4***	***+2.3***	*+3.2*	***+3.3***	***+5.4***	***+7.8***	***+4.2***	***+5.4***	***+0.8***	***+4.2***	***+14.0***

Table 3: Official results of ablation experiments on the test set (CoNLL score in %). The 8-model ensemble (in italics) was evaluated during the post-competition phase.

Comparing Context Sizes on Dev

ÚFAL

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
A) CONTEXT SIZES FOR THE MT5-LARGE MODEL																		
mT5-large 512	72.8	78.1	78.1	76.9	**70.7**	**75.4**	75.6	**67.4**	80.3	68.6	**70.6**	67.3	77.4	77.8	78.7	75.8	71.1	48.6
mT5-large 256	–5.9	–8.8	–4.0	–5.3	–7.1	–3.2	–5.3	–11.7	–6.0	–4.1	–2.9	–4.5	–8.6	–6.4	–6.4	–4.8	–6.7	–4.6
mT5-large 384	–1.6	–2.9	–1.3	–1.8	–0.6	–0.3	–2.0	–1.6	–2.2	–1.3	–1.4	–1.1	–2.7	–2.4	–2.6	–1.2	–2.0	–1.5
mT5-large 768	+1.2	+2.5	+1.2	+1.5	–0.7	**+0.0**	+0.9	–1.4	+1.5	+1.3	–0.6	+2.1	+0.4	+2.7	+2.2	+0.4	+2.7	+3.3
mT5-large 1024	+1.6	+3.2	+1.8	+1.9	–1.0	**+0.0**	+1.1	–1.4	+2.1	+1.7	–1.1	+2.3	**+0.5**	+3.5	+2.6	+0.7	+3.6	+4.7
mT5-large 1536	+1.9	+3.3	+2.2	+2.1	–1.0	**+0.0**	+1.2	–1.4	+2.4	+1.5	–1.1	+2.4	**+0.5**	**+3.8**	**+3.1**	+1.0	+4.1	+6.8
mT5-large 2048	+2.0	+3.5	+2.2	**+2.1**	–1.0	**+0.0**	**+1.2**	–1.4	+2.5	**+2.0**	–1.1	+2.4	**+0.5**	+3.8	+3.0	+1.2	+4.1	+7.4
mT5-large 2560	**+2.0**	**+3.5**	**+2.2**	+2.1	–1.0	**+0.0**	**+1.2**	–1.4	**+2.5**	+1.7	–1.1	**+2.5**	**+0.5**	+3.7	+3.0	**+1.3**	+4.1	**+8.6**
mT5-large 4096	+1.7	+3.4	+2.1	+2.0	–1.0	**+0.0**	**+1.2**	–1.4	+2.5	+1.5	–1.1	**+2.5**	**+0.5**	+3.7	+2.8	+1.2	**+4.4**	+3.1
B) CONTEXT SIZES FOR THE MT5-XL MODEL																		
mT5-xl 512	73.3	77.5	78.4	77.2	**73.9**	76.1	75.4	72.9	80.1	68.4	**70.3**	67.2	77.2	77.7	78.3	76.1	71.3	47.6
mT5-xl 256	–6.1	–8.6	–3.9	–5.4	–9.2	–3.7	–5.8	–9.6	–5.7	–4.9	–2.8	–4.6	–10.1	–6.1	–6.5	–4.7	–6.7	–4.7
mT5-xl 384	–1.7	–2.6	–1.3	–1.9	–2.4	+0.1	–1.6	–0.4	–2.2	–1.5	–1.6	–1.2	–2.5	–2.2	–2.3	–1.3	–2.5	–0.6
mT5-xl 768	+1.1	+2.2	+1.3	+1.7	–4.4	+0.1	+1.3	+0.9	+1.7	+1.5	–1.3	+1.9	**+1.5**	+2.6	+2.2	+0.5	+2.6	+2.4
mT5-xl 1024	+1.5	+3.2	+1.9	+2.3	–4.4	**+0.1**	+1.5	**+1.0**	+2.3	+2.1	–1.5	+2.1	+1.2	+3.3	+2.9	+0.8	+3.9	+3.2
mT5-xl 1536	+1.8	+3.4	+2.4	+2.6	–4.4	**+0.1**	+1.7	**+1.0**	+2.7	**+2.1**	–1.5	+2.2	+1.2	**+3.8**	+3.5	+1.1	+5.2	+3.5
mT5-xl 2048	+1.8	**+3.5**	+2.6	**+2.6**	–4.4	+0.1	+1.7	**+1.0**	+2.8	+2.1	–1.5	**+2.2**	+1.2	+3.7	**+3.9**	+1.3	**+5.5**	+3.6
mT5-xl 2560	**+1.9**	+3.4	+2.6	+2.6	–4.4	+0.1	+1.7	**+1.0**	+2.8	+2.0	–1.5	+2.2	+1.2	+3.7	+3.6	**+1.4**	+5.3	**+5.7**
mT5-xl 4096	+1.7	+3.5	**+2.6**	+2.5	–4.4	+0.1	**+1.7**	**+1.0**	**+2.8**	+1.8	–1.5	+2.2	+1.2	+3.6	+3.6	+1.4	+5.3	+2.6

Table 4: Ablation experiments evaluated on the development sets (CoNLL score in %). 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. The runs in italics use largest context length not exceeding 512 subwords when tokenized with the mT5 tokenizer.

Comparing Context Sizes on Dev

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
D) COMPARISON OF PRETRAINED LANGUAGE MODELS WITH DIFFERENT CONTEXT SIZES																		
mT5-large 512	72.8	78.1	78.1	76.9	70.7	75.4	75.6	67.4	80.3	68.6	**70.6**	67.3	77.4	77.8	78.7	75.8	71.1	48.6
mT5-base 512	–3.9	–4.2	–4.1	–4.5	–3.8	–5.2	–3.8	+1.2	–3.6	–3.3	–8.3	–3.8	–1.6	–3.3	–3.0	–4.3	–4.6	–7.1
XLM-R-base 256	–7.3	–10.0	–6.6	–8.0	–15.1	–5.5	–7.1	–9.8	–7.6	–4.6	–4.4	–4.7	–8.0	–6.3	–8.5	–6.5	–6.9	–5.3
XLM-R-base 384	–4.0	–5.2	–5.0	–5.6	–3.2	–4.1	–5.0	–2.2	–4.9	–2.9	–5.3	–2.8	–2.6	–3.8	–5.2	–3.8	–3.9	–2.5
XLM-R-base 512	–1.9	–2.8	–3.4	–4.0	–0.5	–3.9	–3.5	+2.4	–2.6	–1.5	–2.8	–1.7	+0.9	–1.8	–2.3	–3.3	–0.8	–2.3
XLM-R-base mT5-512	*–3.4*	*–4.9*	*–5.0*	*–5.6*	*–3.4*	*–4.1*	*–4.4*	*–0.6*	*–4.6*	*–2.3*	*–5.0*	*–3.5*	*+0.1*	*–2.9*	*–3.9*	*–3.6*	*–2.3*	*–2.2*
XLM-R-large 256	–3.9	–6.0	–2.8	–3.5	–7.6	–2.1	–3.9	–2.3	–4.1	–2.6	–2.3	–0.7	–7.6	–3.8	–5.0	–2.4	–4.6	–5.3
XLM-R-large 384	–0.7	–1.0	–0.6	–0.5	–1.6	+0.2	+0.0	+1.6	–1.3	+0.1	–2.1	+1.5	–2.5	–1.2	–1.8	+0.0	–0.9	–3.4
XLM-R-large 512	+1.1	+1.2	+0.7	+0.9	+1.5	+0.8	+0.8	+2.7	+0.9	+1.7	–0.9	**+2.7**	**+1.0**	+1.2	+1.0	+0.6	+2.1	–0.8
XLM-R-large mT5-512	*–0.1*	*–0.9*	*–0.6*	*–0.6*	*+0.5*	*+0.4*	*+0.0*	*+2.3*	*–0.9*	*+0.8*	*–2.1*	*+0.8*	*–0.7*	*+0.2*	*–0.4*	*+0.3*	*+0.5*	*–3.0*
RemBERT 256	–4.9	–7.3	–2.4	–3.9	–4.2	+1.0	–4.5	–4.7	–5.4	–3.0	–5.9	–3.5	–9.9	–5.8	–6.3	–3.1	–4.1	–11.3
RemBERT 384	–1.5	–1.9	–0.1	–0.8	+1.1	**+2.8**	–1.5	+0.8	–1.9	–0.3	–5.3	–1.1	–3.6	–2.6	–2.0	–0.1	–0.4	–9.5
RemBERT 512	+0.2	+0.7	+1.2	+0.7	+3.4	+2.5	+0.1	+4.2	+0.5	+1.0	–3.3	+0.0	–1.1	+0.0	+0.0	+0.9	+2.2	–10.0
RemBERT mT5-512	*–0.6*	*–1.0*	*+0.1*	*–0.6*	***+5.4***	*+2.6*	*–0.5*	*+2.3*	*–1.3*	*+0.4*	*–5.4*	*–0.3*	*–1.2*	*–1.0*	*–0.5*	*+0.7*	*+0.5*	*–10.5*
mT5-large 768	+1.2	+2.5	+1.2	+1.5	–0.7	+0.0	+0.9	–1.4	+1.5	+1.3	–0.6	+2.1	+0.4	+2.7	+2.2	+0.4	+2.7	+3.3
mT5-large 2560	+2.0	**+3.5**	+2.2	+2.1	–1.0	+0.0	+1.2	–1.4	+2.5	+1.7	–1.1	+2.5	+0.5	**+3.7**	+3.0	+1.3	+4.1	**+8.6**
mT5-xl 512	+0.5	–0.6	+0.3	+0.3	+3.2	+0.7	–0.2	+5.5	–0.2	–0.2	–0.3	–0.1	–0.2	–0.1	–0.4	+0.3	+0.2	–1.0
mT5-xl 2560	**+2.4**	+2.8	**+2.9**	**+2.9**	–1.2	+0.8	**+1.5**	**+6.5**	**+2.6**	**+1.8**	–1.8	+2.1	+1.0	+3.6	**+3.2**	**+1.7**	**+5.5**	+4.7

Table 4: Ablation experiments evaluated on the development sets (CoNLL score in %). 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. The runs in italics use largest context length not exceeding 512 subwords when tokenized with the mT5 tokenizer.

Comparing Mixing Strategies on Dev

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*
A) AVERAGE OF 5 RUNS USING FOR EVERY CORPUS THE SINGLE EPOCH ACHIEVING THE HIGHEST AVERAGE 5-RUN SCORE																		
Logarithmic	74.8	81.6	80.3	79.0	69.7	75.4	**76.8**	66.0	82.8	70.3	69.5	**69.7**	77.9	81.5	81.7	77.1	75.2	**57.2**
w/o corpus id	–0.2	**+0.2**	–0.1	+0.1	–0.4	+0.1	–0.3	–0.2	+0.0	+0.0	–0.2	–0.3	+0.5	**+0.2**	–0.4	**+0.2**	+0.2	–2.4
Uniform	–0.3	–0.1	–1.2	–0.9	+1.7	+0.0	–0.8	–4.2	–0.3	+0.1	**+0.2**	–0.4	**+1.0**	+0.0	–0.1	+0.0	–0.2	–0.1
w/o corpus id	–0.4	–0.4	–0.7	–0.6	+2.3	+0.3	–0.8	+1.5	–0.1	–0.4	–1.3	–0.5	–0.7	–0.4	–1.3	–0.5	–0.2	–3.0
Square Root	+0.0	+0.2	+0.5	+0.4	–0.2	**+0.9**	–0.6	–2.1	–0.1	+0.1	–0.7	–0.1	+0.8	+0.1	–0.2	+0.2	+0.9	–0.7
w/o corpus id	+0.2	+0.1	+0.4	+0.3	**+2.7**	–0.9	–0.3	+1.1	+0.1	+0.0	–0.4	–0.2	+0.1	+0.1	–0.1	+0.1	+0.5	–0.7
Linear	**+0.4**	+0.1	**+0.8**	**+0.7**	+0.6	–0.1	–0.2	**+4.8**	**+0.3**	**+0.4**	–0.9	–0.4	+0.6	–0.3	**+0.1**	+0.2	**+1.1**	–0.3
w/o corpus id	+0.0	+0.0	+0.7	+0.6	–2.0	–1.4	–0.8	+4.0	+0.3	–0.1	–0.4	–0.9	+0.4	+0.1	–0.1	+0.2	+0.7	–0.8
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

Table 7: 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.

Evaluating Ensembling on Dev

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
A) Ensembles for the mT5-large Model for Various Context Sizes																		
Average of 5 runs, 512	72.8	78.1	78.1	76.9	70.7	75.4	75.6	**67.4**	80.3	68.6	70.6	67.3	77.4	77.8	78.7	75.8	71.1	48.6
Ensemble of 5 runs, 512	+1.0	+0.8	+0.8	+0.7	**+3.1**	**+1.3**	+0.5	–0.4	+0.8	+0.6	**+1.2**	+0.7	+1.6	+0.9	+0.9	+1.0	+1.5	+0.8
Average of 5 runs, 768	+1.2	+2.5	+1.2	+1.5	–0.7	+0.0	+0.9	–1.4	+1.5	+1.3	–0.6	+2.1	+0.4	+2.7	+2.2	+0.4	+2.7	+3.3
Average of 5 runs, 2560	+2.0	+3.5	+2.2	+2.1	–1.0	+0.0	+1.2	–1.4	+2.5	+1.7	–1.1	+2.5	+0.5	+3.7	+3.0	+1.3	+4.1	+8.6
Ensemble of 5 runs, 2560	**+3.3**	**+4.3**	**+3.0**	**+3.0**	+2.3	**+1.3**	**+1.3**	–0.8	**+3.6**	**+2.5**	+1.1	**+3.5**	**+1.8**	**+4.6**	**+3.5**	**+2.3**	**+6.3**	**+11.5**
B) Ensembles for the mT5-xl Model for Various Context Sizes																		
Average of 5 runs, 512	73.3	77.5	78.4	77.2	73.9	76.1	75.4	72.9	80.1	68.4	70.3	67.2	77.2	77.7	78.3	76.1	71.3	47.6
Ensemble of 5 runs, 512	+0.8	+1.1	+0.9	+0.8	–2.3	**+0.2**	+0.8	**+1.9**	+1.1	+1.1	+0.9	+1.8	+1.6	+1.1	+0.8	+1.0	+1.3	+0.3
Average of 5 runs, 768	+1.1	+2.2	+1.3	+1.7	–4.4	+0.1	+1.3	+0.9	+1.7	+1.5	–1.3	+1.9	+1.5	+2.6	+2.2	+0.5	+2.6	+2.4
Average of 5 runs, 2560	+1.9	+3.4	+2.6	+2.6	–4.4	+0.1	+1.7	+1.0	+2.8	+2.0	–1.5	+2.2	+1.2	+3.7	+3.6	+1.4	+5.3	+5.7
Ensemble of 5 runs, 2560	**+3.5**	**+4.9**	**+3.6**	**+3.7**	**+2.4**	**+0.2**	**+2.3**	+1.1	**+3.6**	**+3.3**	**+1.3**	**+4.0**	**+3.0**	**+4.1**	**+5.0**	**+2.5**	**+7.1**	**+7.6**

Table 8: Ablation experiments evaluated on the development sets (CoNLL score in %). We report the average/ensemble of best 5 out of 7 runs, using for every corpus the single epoch achieving the highest average score.

The Effect of Multilingual Data and Zero-shot Evaluation

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.

Questions?