CorPipe at CRAC 2025: Evaluating Multilingual Encoders for Multilingual Coreference Resolution

Milan Straka
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague, Czech Republic
straka@ufal.mff.cuni.cz

Abstract

We present CorPipe 25, the winning entry to the CRAC 2025 Shared Task on Multilingual Coreference Resolution. This fourth iteration of the shared task introduces a new LLM track alongside the original unconstrained track, features reduced development and test sets to lower computational requirements, and includes additional datasets. CorPipe 25 represents a complete reimplementation of our previous systems, migrating from TensorFlow to PyTorch. Our system significantly outperforms all other submissions in both the LLM and unconstrained tracks by a substantial margin of 8 percentage points. The source code and trained models are publicly available at https://github.com/ufal/crac2025-corpipe.

1 Introduction

Coreference resolution seeks to identify and cluster multiple references to the same entity within text. The CRAC 2025 Shared Task on Multilingual Coreference Resolution (Novák et al., 2025a) represents the fourth iteration of this shared task, designed to advance research in multilingual coreference resolution across diverse languages and domains. Building upon the CorefUD 1.3 collection, this year's task introduces several notable changes: a new LLM track that relies on large language models (LLMs) for coreference resolution, reduced development and test sets (minidev and minitest) to lower computational demands, and the inclusion of additional datasets expanding language coverage.

As in the previous year, the submitted systems must also predict the *empty nodes*, which represent elided elements that are not explicitly present in the surface text but are necessary for coreference analysis. Empty nodes are especially important in pro-drop languages (like Slavic and Romance languages), where pronouns can be dropped from a sentence when they can be inferred, for example according to verb morphology, as in the Czech example *"Řekl, že nepřijde"*, translated as *"(He) said that (he) won't come"*.

CorPipe 25, our submission to the CRAC 2025 Shared Task, represents a complete reimplementation of our previous winning systems (Straka, 2024, 2023; Straka and Straková, 2022), transitioning from TensorFlow to PyTorch while preserving the architecture that has proven successful. Our system employs a three-stage pipeline approach: first predicting empty nodes,[1] then detecting mentions, and finally performing coreference linking through antecedent maximization on the identified spans. As in previous CorPipe versions, mention detection and coreference linking are trained jointly using a shared pretrained encoder model, and all models are fully multilingual, trained across all available corpora.

Our contributions are as follows:

- We present the winning entry to the CRAC 2025 Shared Task, surpassing other participants in both tracks by a substantial margin of 8 percentage points.
- We provide a complete reimplementation of CorPipe in PyTorch. The reimplementation enables us to leverage more pretrained multilingual models, allowing us to perform an evaluation of various models and providing insights into their relative performance for coreference resolution across diverse languages.
- We present performance comparisons between TensorFlow and PyTorch implementations, demonstrating the practical benefits of the migration.
- The CorPipe 25 source code is released at https://github.com/ufal/crac2025-corpipe under an open-source license. Three pretrained multilingual models of different sizes are also released, under the CC BY-NC-SA licence.

[1] Our empty node prediction system was provided to all participants as a baseline implementation.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 130–139
Suzhou, China, November 9, 2025.

2 Related Work

Neural Coreference Resolution Neural coreference resolution has been dominated by span-based approaches since the seminal work of Lee et al. (2017), who introduced an end-to-end neural model that jointly performs mention detection and coreference resolution. This approach was further refined by Lee et al. (2018) with coarse-to-fine inference, significantly improving both efficiency and accuracy. Joshi et al. (2020) demonstrated substantial improvements by incorporating SpanBERT (Joshi et al., 2019), a pretrained model specifically designed for span prediction tasks.

Alternative paradigms have emerged to address the limitations of span-based methods. Wu et al. (2020) formulated coreference as a question-answering task, while Liu et al. (2022) introduced a specialized autoregressive system and Bohnet et al. (2023) employed a text-to-text paradigm. However, all these architectures must evaluate the trained model repeatedly during processing of a single sentence.

Word-Level Coreference Resolution A significant departure from span-based approaches came with Dobrovolskii (2021), who proposed word-level coreference resolution, which represents mentions by their head-words only. The approach has been extended by D'Oosterlinck et al. (2023) with CAW-coref, which introduces conjunction-aware handling to better manage complex mention structures. More recently, Liu et al. (2024) proposed MSCAW-coref that aims to work in a multilingual setting and accounts for singleton mentions. This approach has been adopted by Stanza (Qi et al., 2020), a widely-used Python natural language processing toolkit.

Multilingual Coreference Resolution The CRAC shared tasks on multilingual coreference resolution (Žabokrtský et al., 2022, 2023; Novák et al., 2024, 2025a) have been instrumental in advancing the field, providing standardized evaluation framework, the CorefUD dataset (Novák et al., 2025b), and a multilingual baseline (Pražák et al., 2021).

Previous versions of CorPipe have participated in all CRAC shared tasks, evolving from basic multilingual models (Straka and Straková, 2022) to incorporating larger contexts (Straka, 2023) and performing zero mention prediction from raw text (Straka, 2024).

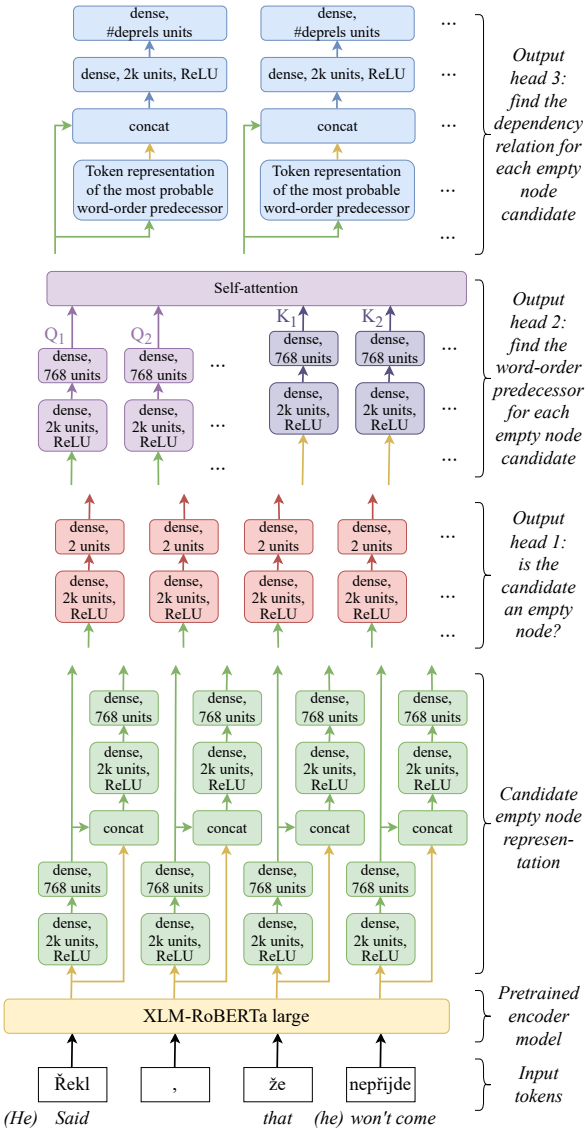

Figure 1: The system architecture of the empty node prediction baseline. Every ReLU activation is followed by a dropout layer with a dropout rate of 50%.

3 Architecture

Our system is essentially a PyTorch reimplementation of CorPipe 24 (Straka, 2024).

Empty Nodes Baseline First, empty nodes are predicted using a baseline system that was available to all shared task participants. The architecture of this system is illustrated in Figure 1.

Our approach for empty node prediction focuses on generating the essential information required for coreference evaluation: the word order position (determined by which input word the empty node follows), along with the dependency head and dependency relation. We do not predict forms or lemmas, even when available in training data. The model operates non-autoregressively, predicting up

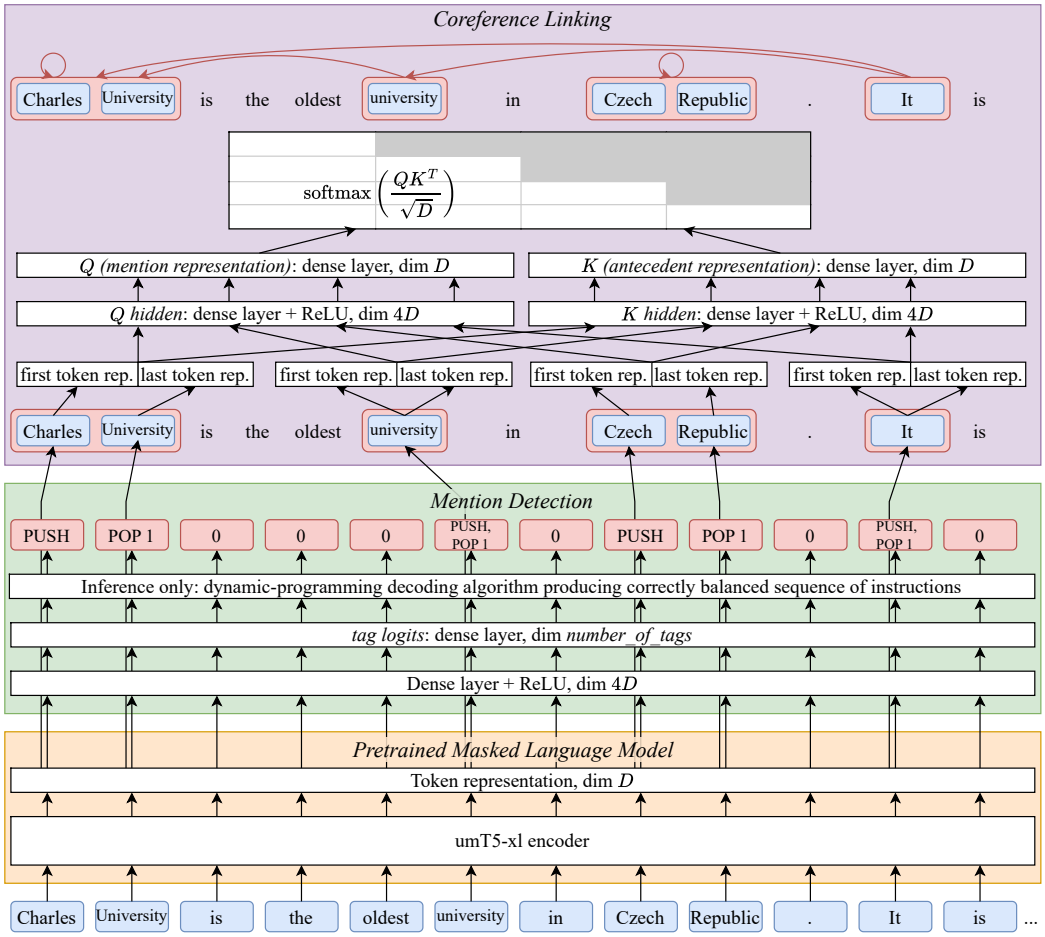

Figure 2: The CorPipe 25 model architecture.

to two empty nodes per input word, with each input word serving as the potential dependency head.

The architecture processes tokenized input through a XLM-RoBERTa-large (Conneau et al., 2020), representing each word by its first subword embedding. For each word, we generate two empty node candidates: the first through a dense-ReLU-dropout-dense module (768→2k→768 units), and the second by concatenating the first candidate with the input word representation and applying an analogous transformation. The candidates are processed by three heads, each following its own 2k-unit ReLU layer and dropout: (1) binary classification for empty node existence, (2) self-attention for word order position selection, and (3) dependency relation classification using the candidate representation concatenated with the embedding of the most likely word preceding it.

Training employs a single multilingual model with Adam optimizer (Kingma and Ba, 2015) for 20 epochs of 5 000 batches (64 sentences each). The learning rate linearly increases to 1e-5 in the first epoch and then decays to zero in the rest of the training following cosine decay (Loshchilov and Hutter, 2017). Sentences are sampled from all empty node corpora, proportionally to the square root of corpus size. Training required 19 hours on a single L40 GPU with 48GB RAM.

The source code is released under the MPL license at `https://github.com/ufal/crac2025_empty_nodes_baseline`, together with the full set of hyperparameters used. The trained model is available under the CC BY-SA-NC license at `https://www.kaggle.com/models/ufal-mff/crac2025_empty_nodes_baseline/`. Finally, the minidev and minitest sets of the CRAC 2025 Shared Task with predicted empty nodes are available to all participants.

Coreference Resolution Once the empty nodes have been predicted, we employ coreference resolution system based on CorPipe 23 from Straka (2023). The architectural overview is shown in Figure 2 and summarized below; detailed implementation specifics are available in the referenced work.

Our model processes documents sentence-by-sentence. To maximize available context for each sentence, we expand it with preceding tokens and

Model	Params	Batch Size	Learning Rate	Train Time
mT5 base	264M	8	6e-4	4h
umT5 base	269M	8	6e-4	4h
mT5 large	538M	8	6e-4	9.5h
mT5 xl	1593M	6	5e-4	22.5h
umT5 xl	1605M	6	5e-4	22.5h
mT5 xxl	5393M	6	5e-4	33h
umT5 xxl	5417M	6	5e-4	33h

Table 1: Properties of mT5 encoder models used. The training time is measured for 15 epochs 10k updates each using a single A100 GPU, with the exception of the xxl models, which are trained using a single H100 GPU.

at most 50 subsequent tokens, constrained by the maximum segment length (512 or 2 560 tokens). Input tokens first pass through a pretrained multilingual encoder. Subsequently, we predict coreference mentions using an enhanced BIO encoding scheme that handles potentially overlapping span sets. Each identified mention is then encoded as a concatenation of its boundary tokens (first and last), and coreference links are established through a self-attention mechanism that determines the most probable antecedent for each mention (including self-reference utilized by first entity mentions).

We employ different segment sizes during training versus inference: training always uses 512-token segments, while inference leverages extended 2 560-token segments (with the exception of two PROIEL corpora always using 512 tokens), exploiting relative positional encoding capabilities for improved long-range context modeling.

Training For the shared task submission, we train 13 multilingual models based on umT5-xl (Chung et al., 2023), differing only in random initialization and whether we express corpus size during sampling using sentences or words. The sentences are sampled proportionally to the square root of the corpus size; for ablations, we consider also values of this *sampling ratio* different from 0.5.

Every model is trained for 15 epochs with 10k batches each, with every batch consisting of 6 sentences. The model is trained using the AdaFactor optimizer (Shazeer and Stern, 2018). The learning rate follows a warmup schedule: linear increase to 5e-4 during the initial 10% of training, followed by a cosine decay (Loshchilov and Hutter, 2017) to 0. The model trains for 22.5 hours on a single A100 GPU with 40GB RAM. For ablation experiments, we also consider other umT5 and mT5 (Xue et al., 2021) models, whose properties and corresponding hyperparameters are summarized in Table 1.

For each model, we save checkpoints after every epoch, obtaining a pool of $13 \cdot 15$ checkpoints.

System	Head-match	Partial-match	Exact-match	With Singletons
UNCONSTRAINED				
CorPipeEnsemble	**75.84 1**	**74.90 1**	**72.76 1**	**78.33 1**
CorPipeBestDev	75.06 2	74.08 2	71.97 2	77.63 2
CorPipeSingle	74.75 3	73.74 3	71.53 3	77.43 3
Stanza	67.81 4	67.03 4	64.68 4	70.64 4
GLaRef-Propp	61.57 5	60.72 5	58.43 5	65.28 5
BASELINE-GZ	58.18 6	57.75 6	56.48 6	49.88 6
BASELINE	56.01 7	55.58 7	54.24 7	47.88 7
LLM				
GLaRef-CRAC25	62.96 1	61.66 1	58.98 1	65.61 1
NUST-FewShot	61.74 2	61.14 2	56.34 2	63.44 2
PUXCRAC2025	60.09 3	59.68 3	55.22 3	54.77 4
UWB	59.84 4	59.55 4	38.81 4	62.77 3

Table 2: Official results of CRAC 2025 Shared Task on the minitest set with various metrics in %.

4 Shared Task Results

In the shared task, teams were permitted to submit up to three systems. We selected the following configurations based on our checkpoint selection strategy:

- **CorPipeSingle**, a single best-performing checkpoint selected based on overall minidev performance across all corpora;
- **CorPipeBestDev**, employing corpus-specific optimal checkpoints selected individually based on minidev performance for each corpus from the pool of $13 \cdot 15$ checkpoints;
- **CorPipeEnsemble**, an ensemble of 5 best-performing checkpoints based on overall minidev performance across all corpora.

The first configuration CorPipeSingle corresponds to practical deployment, where a single model handles all corpora, while the others aim at maximizing performance.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
UNCONSTRAINED																							
CorPipeEnsemble	**75.81**	**82.91**	**77.11**	**80.71**	**65.51**	**73.01**	**76.11**	81.81	**84.51**	**76.31**	**71.81**	**74.51**	69.81	**77.71**	**68.61**	**71.01**	**69.91**	**77.21**	**78.21**	**76.31**	**80.21**	84.23	71.22
CorPipeBestDev	75.12	82.03	76.32	80.42	62.83	72.63	75.92	81.32	83.83	75.92	69.93	74.33	68.32	77.52	68.32	70.52	69.32	76.02	77.12	74.02	79.92	**84.81**	70.43
CorPipeSingle	74.83	82.52	76.23	80.13	63.02	72.82	75.23	80.83	84.12	75.83	70.32	74.42	66.13	76.53	67.33	69.73	68.93	75.83	76.23	73.63	79.43	84.22	**71.61**
Stanza	67.84	79.54	72.74	75.14	40.84	67.34	69.04	74.84	80.44	67.54	62.55	54.94	62.14	74.24	60.04	64.64	67.74	72.84	72.44	71.74	73.04	80.84	47.85
GLaRef-Propp	61.65	68.16	61.76	66.66	39.15	61.25	61.95	70.05	69.17	65.15	66.14	51.35	58.85	69.55	50.95	60.15	60.66	57.67	67.15	66.35	68.06	71.55	44.37
BASELINE-GZ†	58.26	68.85	69.55	67.95	29.56	55.76	61.67	66.06	71.05	63.86	55.06	29.46	31.06	66.86	47.16	54.37	64.35	65.35	62.56	63.06	68.15	67.66	51.74
BASELINE†	56.07	68.07	56.97	63.07	26.37	55.76	61.76	66.06	70.56	63.86	55.06	28.57	31.06	66.86	43.27	54.56	50.37	65.35	62.56	63.06	66.57	67.66	45.96
LLM																							
GLaRef-CRAC25	63.01	73.52	65.11	71.31	58.22	59.62	58.74	69.04	74.41	66.72	60.42	65.81	44.03	56.44	52.51	59.83	63.03	62.53	64.74	61.64	72.51	68.83	56.22
NUST-FewShot	61.72	60.94	51.44	54.34	58.51	48.74	69.82	70.42	61.84	71.91	57.63	57.92	**80.21**	71.32	43.53	52.34	66.02	59.24	72.82	68.92	70.82	71.42	39.03
PUXCRAC2025	60.13	68.03	56.93	63.03	43.73	57.43	61.73	69.13	70.53	63.83	61.51	47.93	45.32	66.83	50.62	61.62	50.34	65.31	65.23	63.03	66.53	67.64	56.11
UWB	59.84	79.21	61.02	68.22	25.34	67.61	73.61	**84.01**	73.62	58.64	49.14	47.64	0.04	75.81	38.94	67.31	68.31	63.42	73.81	72.01	64.54	80.11	24.34

Table 3: Official results of CRAC 2025 Shared Task on the minitest set (CoNLL score in %). The systems † are described in Pražák et al. (2021); the rest in Novák et al. (2025a).

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) CORPIPE SINGLE MODELS																							
Single mT5-large model	72.84	80.1	74.6	78.0	58.5	67.2	73.3	77.4	82.0	72.1	68.5	71.2	67.9	76.3	67.3	68.0	69.8	74.4	75.2	74.0	77.5	81.2	67.7
Single umT5-base model	–3.54	–2.7	–1.1	–2.9	–5.0	–5.2	–2.3	–4.6	–3.5	–0.9	–1.8	–6.3	–8.9	–3.6	–5.8	–4.3	–2.0	–1.5	–2.0	–3.5	–3.0	–3.5	–3.8
	69.27	77.4	73.5	75.1	53.5	62.0	71.0	72.8	78.6	71.2	66.7	64.9	59.0	72.7	61.5	63.7	67.8	72.9	73.2	70.4	74.5	77.8	63.9
Single umT5-xl model	+1.96	+2.4	+1.6	+2.1	+4.5	+5.6	+1.9	+3.4	+2.0	+3.7	+1.8	+3.2	–1.8	+0.2	–0.0	+1.7	–0.9	+1.4	+1.0	–0.4	+1.9	+3.0	+3.9
	74.75	82.5	76.2	80.1	63.0	72.8	75.2	80.8	84.1	75.8	70.3	74.4	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	71.6
Single mT5-xxl model	+3.16	**+3.0**	**+2.9**	**+3.3**	**+10.8**	**+6.1**	+3.6	+5.0	+2.4	+3.6	**+4.0**	+6.2	**+6.5**	+0.4	–5.1	**+3.9**	–0.5	**+3.2**	+1.7	+1.5	+2.7	+3.7	+1.9
	76.04	**83.1**	**77.5**	**81.3**	**69.3**	**73.3**	76.9	82.4	84.4	75.7	**72.4**	77.5	**74.4**	76.7	62.2	**71.9**	69.3	**77.6**	76.9	75.4	80.2	84.9	69.6
Single umT5-xxl model	**+3.46**	+3.0	+2.7	+3.2	+9.1	+4.5	**+3.8**	**+6.8**	+2.4	**+5.4**	+3.0	**+6.8**	+4.8	+1.4	–1.2	+3.2	–0.6	+0.6	**+3.1**	**+2.3**	**+2.8**	**+3.8**	**+4.7**
	76.26	83.1	77.3	81.2	67.6	71.7	**77.1**	**84.2**	84.4	**77.5**	71.5	**78.1**	72.7	77.7	66.1	71.2	69.2	75.0	**78.3**	**76.3**	**80.3**	**85.0**	**72.4**
B) CORPIPE ENSEMBLE MODELS																							
Single umT5-xl model	74.75	82.5	76.2	80.1	63.0	72.8	75.2	80.8	84.1	75.8	70.3	74.4	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	71.6
5 umT5-xl models	+1.05	+0.4	+0.9	+0.6	+2.5	+0.2	+0.8	+1.0	+0.4	+0.5	+1.5	+0.1	+3.7	+1.2	**+1.3**	+1.3	+1.0	+1.4	**+2.0**	**+2.7**	+0.8	–0.0	–0.4
	75.84	82.9	77.1	80.7	65.5	73.0	76.1	81.8	84.5	76.3	71.8	74.5	69.8	77.7	**68.6**	71.0	69.9	77.2	**78.2**	**76.3**	80.2	84.2	71.2
3 mT5-xxl models	+2.15	+1.4	+1.4	+1.0	+7.0	**+2.1**	+2.0	+2.9	+1.0	+1.0	+2.4	+6.1	**+8.9**	+1.3	–0.1	+2.8	+1.1	**+1.6**	+0.5	+2.3	+1.4	+0.6	–0.6
	76.93	83.9	77.6	81.1	70.0	**74.9**	77.3	83.7	85.1	76.7	72.7	80.5	**75.0**	77.8	67.2	72.5	70.0	**77.4**	76.7	75.9	80.8	84.8	71.0
3 umT5-xxl models	+2.05	+1.1	+1.5	**+1.7**	+5.8	+1.0	+2.0	+2.9	+0.9	**+2.2**	+2.5	+3.9	+7.9	+1.3	–0.5	**+3.5**	+1.2	–1.0	+1.1	+2.3	+1.9	**+1.8**	**+0.2**
	76.80	83.6	77.7	**81.8**	68.8	73.8	77.2	83.7	85.0	**78.0**	72.8	78.3	74.0	77.8	66.8	**73.2**	70.1	74.8	77.3	75.9	81.3	**86.0**	**71.8**
3 mT5-xxl models +	**+2.45**	**+1.7**	**+1.8**	+1.5	**+7.2**	+0.6	**+2.5**	**+3.3**	**+1.3**	+1.8	**+2.6**	**+6.2**	+8.9	**+1.6**	+0.5	+2.7	**+1.6**	+0.1	+1.3	+2.5	**+2.1**	+1.6	+0.2
+3 umT5-xxl models	**77.20**	**84.2**	**78.0**	81.6	**70.2**	73.4	**77.8**	**84.1**	**85.4**	77.6	**72.9**	**80.6**	75.0	**78.1**	67.8	72.4	**70.5**	75.9	77.5	76.1	**81.5**	85.8	71.8
C) CORPIPE PER-CORPUS BEST MODELS																							
Single umT5-xl model	74.75	**82.5**	76.2	80.1	**63.0**	**72.8**	75.2	80.8	**84.1**	75.8	**70.3**	**74.4**	66.1	76.5	67.3	69.7	68.9	75.8	76.2	73.6	79.4	84.2	**71.6**
Per-corpus best umT5-xl model	**+0.35**	–0.5	**+0.1**	**+0.3**	–0.2	–0.2	**+0.7**	**+0.5**	–0.3	**+0.2**	–0.4	–0.1	**+2.2**	**+1.0**	**+1.0**	**+0.8**	**+0.4**	**+0.2**	**+0.9**	**+0.4**	**+0.5**	**+0.6**	–1.2
	75.06	82.0	**76.3**	**80.4**	62.8	72.6	**75.9**	**81.3**	83.8	**75.9**	69.9	74.3	**68.3**	**77.5**	**68.3**	**70.5**	**69.3**	**76.0**	**77.1**	**74.0**	**79.9**	**84.8**	70.4

Table 4: Additional experiments on the CorefUD 1.3 minitest set (CoNLL score in %). The models in italics are post-competition submissions (i.e., submitted after the shared task deadline).

The official results of the CRAC 2025 Shared Task are summarized in Table 3 showing the CoNLL score and individual corpora performance, and in Table 2 showing four metrics across all corpora. All CorPipe 25 configurations substantially surpass all other participants, by 7 percent points for CorPipeSingle and 8 for CorPipeEnsemble. The CorPipeBestDev configuration only marginally outperforms CorPipeSingle, which we attribute to the exclusion of the two smallest corpora this year.

We evaluate additional mT5 and umT5 models on the minitest in Table 4. The xxl-sized models provide a boost of more than 1 percent point over the xl size; the ensemble of 3 mT5-xxl and umT5-xxl models provide an additional 1 percent point gain, achieving the best performance of 77.2%, a 1.4 percent point increase compared to the best competition submission.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) SUBMITTED CRAC25 SYSTEMS																							
CorPipeEnsemble	**76.51**	**84.1**	**76.9**	**81.1**	**64.2**	77.9	**77.5**	80.0	**85.1**	**79.6**	**72.5**	76.1	**66.8**	**82.0**	**69.7**	**73.1**	**69.4**	**81.6**	**80.1**	79.7	**80.3**	**80.0**	65.5
CorPipeSingle	75.69	83.2	75.9	80.2	62.7	76.9	76.5	**80.1**	84.2	79.0	71.9	**76.2**	66.0	80.6	68.1	71.9	67.6	80.2	79.2	**80.3**	79.2	78.3	**66.7**
Stanza	69.37	80.3	72.8	74.5	38.0	**78.0**	70.7	73.0	79.5	69.8	63.2	54.1	63.6	78.9	65.3	68.6	64.9	78.8	74.9	75.3	74.1	78.4	49.5
GLaRef-Propp	62.96	68.9	61.9	62.9	40.0	64.0	65.2	72.1	68.8	69.0	65.0	54.1	57.8	72.1	52.5	60.4	60.6	74.2	69.8	70.3	65.0	67.8	42.8
BASELINE-GZ	58.64	70.5	68.0	67.4	27.7	57.9	65.0	66.6	71.7	65.4	56.3	29.8	23.8	69.9	49.9	59.0	63.0	69.3	66.1	66.8	65.6	63.4	47.1
BASELINE	56.39	69.9	57.3	63.2	24.1	57.9	65.0	66.6	71.3	65.4	56.3	27.0	23.8	69.9	46.6	58.3	48.3	69.3	66.1	66.8	64.1	63.4	40.1
B) CORPIPE SINGLE MODELS																							
mT5-large	73.26	81.3	73.8	77.0	57.7	75.3	74.1	75.9	81.7	74.9	69.7	72.1	65.2	79.7	66.4	68.7	67.7	80.0	77.2	77.5	76.8	76.2	62.8
mT5-base	−4.43 68.83	−3.7 77.6	−3.4 70.4	−4.3 72.7	−5.0 52.7	−7.8 67.5	−2.6 71.5	−5.3 70.6	−3.1 78.6	−3.7 71.2	−4.2 65.5	−9.1 63.0	−8.1 57.1	−3.6 76.1	−5.7 60.7	−3.7 65.0	−2.2 65.5	−2.8 77.2	−1.9 75.3	−3.1 74.4	−3.8 73.0	−5.4 70.8	−4.8 58.0
umT5-base	−3.38 69.88	−2.6 78.7	−1.3 72.5	−2.2 74.8	−4.5 53.2	−6.4 68.9	−2.9 71.2	−3.6 72.3	−1.5 80.2	−1.8 73.1	−3.2 66.5	−7.8 64.3	−9.3 55.9	−2.9 76.8	−2.5 63.9	−3.4 65.3	−2.2 65.5	−5.3 74.7	−1.4 75.8	−1.6 75.9	−2.6 74.2	−3.0 73.2	−2.5 60.3
XLM-RoBERTa-base	−5.23 68.03	−6.2 75.1	−5.0 68.8	−5.3 71.7	−6.5 51.2	−4.4 70.9	−3.9 70.2	−7.4 68.5	−4.5 77.2	−4.4 70.5	−4.2 65.5	−8.3 63.8	−14.7 50.5	−2.5 77.2	−1.3 65.1	−3.1 65.6	−6.0 61.7	−2.7 77.3	−4.2 73.0	−4.0 73.5	−5.0 71.8	−5.6 70.6	−5.9 56.9
XLM-RoBERTa-large	−1.36 71.90	−1.6 79.7	−2.3 71.5	−0.7 76.3	+0.6 58.3	+0.1 75.4	+0.1 74.2	−2.4 73.5	−1.0 80.7	−1.4 73.5	−1.1 68.6	−0.5 71.6	−6.6 58.6	+0.2 79.9	+1.2 67.6	+1.2 69.9	−3.1 64.6	−2.0 78.0	−2.4 74.8	−1.5 76.0	−1.8 75.0	−2.2 74.0	−2.9 59.9
RemBERT	−1.84 71.42	−2.5 78.8	−2.2 71.6	−0.8 76.2	+0.3 58.0	+1.0 76.3	−0.4 73.7	−1.4 74.5	−1.1 80.6	−0.4 74.5	−1.4 68.3	−1.6 70.5	−4.7 60.5	+0.2 79.9	−1.0 65.4	−2.1 66.6	−4.4 63.3	−5.2 74.8	−2.3 74.9	−2.5 75.0	−1.5 75.3	−2.7 73.5	−3.5 59.3
InfoXLM-large	−1.44 71.82	−1.9 79.4	−1.9 71.9	−0.4 76.6	+1.3 59.0	+0.0 75.3	−0.3 73.8	−3.0 72.9	−1.0 80.7	−0.8 74.1	−1.1 68.6	−1.4 70.7	−7.3 57.9	+1.0 80.7	+0.5 66.9	+0.4 69.1	−3.3 64.4	−1.5 78.5	−1.6 75.6	−1.9 75.6	−1.7 75.1	−3.3 72.9	−2.4 60.4
T5Gemma-large-ul2	−3.13 70.13	−2.5 78.8	−1.2 72.7	−3.2 73.8	−6.1 51.6	−0.9 74.4	+0.7 74.8	+1.3 77.2	−0.6 81.1	−1.2 73.7	+0.5 70.2	−9.5 62.6	−6.0 59.2	−4.4 75.3	−5.5 60.9	−5.2 63.5	−1.7 66.0	−7.2 72.8	−1.8 75.4	−2.9 74.6	−2.9 73.9	−2.2 74.0	−6.6 56.2
T5Gemma-xl-ul2	−0.55 72.71	+0.0 81.3	+0.4 74.2	−0.9 76.1	−1.0 56.7	+1.1 76.4	+2.9 77.0	+5.5 81.4	+1.1 82.8	+1.5 76.4	+1.3 71.0	−2.2 69.9	−0.1 65.1	−2.2 77.5	−5.6 60.9	−3.0 65.7	−0.5 67.2	−3.3 76.7	+0.0 77.2	−1.3 76.2	−1.5 75.3	+0.5 76.7	−4.8 58.0
T5Gemma-xl-ul2-it	−0.07 73.19	+0.1 81.4	+0.8 74.7	−0.7 76.3	−0.2 57.5	+2.8 78.1	+2.9 77.0	+5.4 81.3	+0.9 82.6	+1.5 76.4	+2.2 71.9	−1.9 70.2	−0.7 64.5	−1.5 78.2	−3.4 63.0	−1.0 67.7	−0.4 67.3	−2.8 77.2	+0.2 77.4	−0.6 76.9	−0.9 75.9	+0.2 76.5	−4.6 58.2
T5Gemma-xl-prefixlm	−0.50 72.76	−1.1 80.2	+1.0 74.8	−0.2 76.8	−1.9 55.8	+1.4 76.7	+2.3 76.4	+5.4 81.3	+0.9 82.6	+0.5 75.4	+2.5 72.2	−1.7 70.4	+0.0 65.2	−1.3 78.4	−5.1 61.3	−2.8 65.9	+0.5 68.2	−5.3 74.7	+0.6 77.8	−1.3 76.2	−0.8 76.0	+0.9 77.1	−5.2 57.5
T5Gemma-xl-prefixlm-it	−1.89 71.37	−1.6 79.7	+0.3 74.1	−2.0 75.0	−5.4 52.3	+0.5 75.8	+1.9 76.0	+3.3 79.2	−1.2 80.5	+0.1 75.0	+0.8 70.5	−4.8 67.3	−3.4 61.9	−3.1 76.6	−5.4 61.0	−2.4 66.3	−0.7 67.0	−5.8 74.2	−1.7 75.5	−2.0 75.5	−1.8 75.0	−1.4 74.8	−5.7 57.1
T5Gemma-2B-ul2	+1.16 74.42	+1.3 82.5	+2.0 75.8	+1.4 78.4	+2.1 59.8	+1.7 77.0	+3.1 77.2	**+7.6** **83.5**	+1.5 83.2	+3.1 78.0	+1.8 71.5	+1.3 73.4	+2.9 68.1	+0.2 79.9	−4.1 62.3	−0.2 68.5	−0.7 67.0	−3.0 77.0	+1.8 79.0	+0.6 78.1	+1.0 77.8	+2.0 78.2	−1.9 60.9
mT5-xl	+0.16 73.42	−0.2 81.1	+0.2 74.0	+0.3 77.3	+1.4 59.1	−1.0 74.3	+0.1 74.2	+1.3 77.2	+0.4 82.1	+0.5 75.4	−0.6 69.1	+0.0 72.1	+0.9 66.1	−0.3 79.4	+1.6 68.0	+0.2 68.9	+0.1 67.8	−1.0 79.0	+0.0 77.2	−0.6 76.9	+0.4 77.2	+0.6 76.8	−0.8 62.0
umT5-xl	+2.40 75.66	+2.1 83.4	+2.5 76.3	+3.2 80.2	+5.0 62.7	+1.9 77.2	+2.8 76.9	+3.2 79.1	+2.3 84.0	+3.9 78.8	+2.1 71.8	+3.5 75.6	+0.3 65.5	+0.9 80.6	+1.7 68.1	+3.1 71.8	+0.6 68.3	**+0.3** **80.3**	+2.3 79.5	+1.9 79.4	+2.7 79.5	+2.4 78.6	**+4.0** **66.8**
mT5-xxl	+3.54 76.80	+2.4 83.7	+2.8 76.6	**+4.1** **81.1**	**+10.2** **67.9**	+2.0 77.3	+3.0 77.1	+5.8 81.7	+2.4 84.1	+3.8 78.7	+2.8 72.5	**+8.0** **80.1**	**+8.1** **73.3**	+1.6 81.3	**+2.4** **68.8**	+2.6 71.3	+0.8 68.5	+0.0 80.0	+2.0 79.2	**+2.9** **80.4**	+3.4 80.2	+3.3 79.5	+3.6 66.4
umT5-xxl	**+3.77** **77.03**	**+2.5** **83.8**	**+3.1** **76.9**	+3.9 80.9	+8.7 66.4	**+3.6** **78.9**	**+3.6** **77.7**	+6.5 82.4	**+2.8** **84.5**	**+4.9** **79.8**	**+3.7** **73.4**	+7.2 79.3	+6.0 71.2	**+1.7** **81.4**	+1.9 68.3	**+3.2** **71.9**	**+1.8** **69.5**	+0.3 80.3	**+3.2** **80.4**	+2.4 79.9	**+3.7** **80.5**	**+4.4** **80.6**	+3.8 66.6

Table 5: Ablations experiments on the CorefUD 1.3 minidev set (CoNLL score in %). The results are averages of 3 or more runs and for every run the epoch with best average score over the whole CorefUD is used.

5 Ablations Experiments

We perform a series of ablation experiments on the CorefUD 1.3 minidev set (to avoid overfitting on the minitest set). The presented results are averages of 3 or more runs, and for every run the epoch with the best average score across all corpora is used.

For reference, the minidev scores of the systems submitted to the CRAC 2025 Shared Task are summarized in Table 5.A.

The first set of experiments evaluates the impact of different models beyond the mT5 and umT5 families. Notably, we also evaluate the XLM-RoBERTa-base and XLM-RoBERTa-large models (Conneau et al., 2020), the RemBERT model (Chung et al., 2021), InfoXLM-large (Chi et al., 2021), and several variants of the recently introduced T5Gemma model (Zhang et al., 2025).

The results are summarized in Table 5.B. The umT5 models consistently outperform the mT5 ones, which is why we used them in the official submission.[2] The mT5 and umT5 models outperform the other evaluated models, particularly because they support longer contexts (Table 6.C and Straka, 2023, Table 4). When restricting the context to 512 tokens, XLM-RoBERTa-large model achieves the best performance, surpassing both InfoXLM-large and RemBERT. Finally, the recently introduced T5Gemma encoder-decoder model adapted from the Gemma decoder-only model seems to lag behind the umT5 models of corresponding sizes, despite supporting longer contexts too.

[2] In this context, it is unfortunate that the umT5-large model has not been released as it would likely outperform the mT5-large model, which is a size very suitable for deployment.

System	Avg	ca	cs pced	cs pdt	cu	de pots	en gum	en litb	es	fr anco	fr demo	grc	hbo	hi hdtb	hu kork	hu szeg	ko	lt	no bokm	no nyno	pl	ru	tr
A) Cross-Lingual Zero-Shot Evaluation of mT5-large Model																							
Single mT5-large Model	**73.26**	**81.3**	**73.8**	**77.0**	**57.7**	**75.3**	**74.1**	**75.9**	**81.7**	**74.9**	**69.7**	**72.1**	**65.2**	**79.7**	**66.4**	**68.7**	**67.7**	**80.0**	**77.2**	**77.5**	**76.8**	**76.2**	**62.8**
Zero-Shot Multilin. Models	−14.21	−4.2	−16.8	−15.1	−18.5	−13.4	−16.2	−13.0	−7.4	−30.5	−9.3	−8.0	−16.3	−13.5	−2.5	−12.2	−20.5	−24.2	−14.4	−18.4	−8.3	−14.8	−14.9
	59.05	77.1	57.0	61.9	39.2	61.9	57.9	62.9	74.3	44.4	60.4	64.1	48.9	66.2	63.9	56.5	47.2	55.8	62.8	59.1	68.5	61.4	47.9
B) Cross-Lingual Zero-Shot Evaluation of umT5-xl Model																							
Single umT5-xl Model	**75.66**	**83.4**	**76.3**	**80.2**	**62.7**	**77.2**	**76.9**	**79.1**	**84.0**	**78.8**	**71.8**	**75.6**	**65.5**	**80.6**	**68.1**	**71.8**	**68.3**	**80.3**	**79.5**	**79.4**	**79.5**	**78.6**	**66.8**
Zero-Shot Multilin. Models	−14.39	−4.6	−17.4	−15.5	−19.3	−13.5	−18.2	−14.9	−2.0	−34.3	−6.5	−6.3	−21.2	−12.6	−1.9	−14.9	−19.3	−24.0	−15.7	−18.4	−8.5	−16.9	−10.7
	61.27	78.8	58.9	64.7	43.4	63.7	58.7	64.2	82.0	44.5	65.3	69.3	44.3	68.0	66.2	56.9	49.0	56.3	63.8	61.0	71.0	61.7	56.1
C) Various Segment Sizes of mT5-large Model																							
Segment 2560	**73.26**	**81.3**	**73.8**	77.0	**57.7**	**75.3**	**74.1**	**75.9**	**81.7**	**74.9**	**69.7**	**72.1**	65.2	79.7	66.4	68.7	**67.7**	**80.0**	**77.2**	**77.5**	**76.8**	**76.2**	**62.8**
Segment 1024	−0.31	−0.1	−0.7	**+0.0**	**+0.0**	**+0.0**	−0.2	−0.1	−0.3	−0.7	−0.7	**+0.0**	**+1.7**	**+0.1**	**+0.2**	**+0.0**	−1.0	+0.0	−0.3	−0.7	−1.2	−0.9	−2.0
	72.95	81.2	73.1	**77.0**	**57.7**	**75.3**	73.9	75.8	81.4	74.2	69.0	**72.1**	**66.9**	**79.8**	**66.6**	**68.7**	66.7	80.0	76.9	76.8	75.6	75.3	60.8
Segment 512	−2.54	−4.2	−2.8	−2.2	**+0.0**	−0.2	−2.0	−4.8	−2.5	−2.2	−3.3	**+0.0**	−3.5	−1.0	−1.1	−2.1	−4.0	−1.5	−3.5	−3.2	−3.2	−4.6	−4.0
	70.72	77.1	71.0	74.8	**57.7**	75.1	72.1	71.1	79.2	72.7	66.4	**72.1**	61.7	78.7	65.3	66.6	63.7	78.5	73.7	74.3	73.6	71.5	58.8
D) Various Segment Sizes of umT5-xl Model																							
Segment 2560	**75.66**	**83.4**	**76.3**	80.2	**62.7**	**77.2**	**76.9**	**79.1**	**84.0**	**78.8**	**71.8**	**75.6**	65.5	**80.6**	68.1	71.8	**68.3**	**80.3**	79.5	**79.4**	**79.5**	**78.6**	**66.8**
Segment 1024	−0.45	−0.2	−0.9	**+0.0**	**+0.0**	−0.1	−0.4	−0.2	−0.5	−1.8	−0.4	**+0.0**	**+0.2**	+0.0	**+0.6**	**+0.1**	−1.1	−0.1	**+0.0**	+0.0	−1.1	−0.9	−3.1
	75.21	83.2	75.4	**80.2**	**62.7**	77.1	76.5	78.9	83.5	77.0	71.4	**75.6**	**65.7**	80.6	**68.7**	**71.9**	67.2	80.2	**79.5**	79.4	78.4	77.7	63.7
Segment 512	−2.30	−3.2	−2.7	−1.6	**+0.0**	−0.1	−1.9	−3.8	−2.1	−3.6	−1.7	**+0.0**	−5.1	−0.5	+0.2	−1.7	−4.4	−1.8	−3.1	−2.0	−2.3	−3.7	−5.3
	73.36	80.2	73.6	78.6	**62.7**	77.1	75.0	75.3	81.9	75.2	70.1	**75.6**	60.4	80.1	68.3	70.1	63.9	78.5	76.4	77.4	77.2	74.9	61.5
E) Various Sampling Ratios of mT5-large Model																							
Ratio 4/8	73.26	81.3	73.8	77.0	57.7	**75.3**	74.1	75.9	81.7	74.9	**69.7**	72.1	65.2	79.7	66.4	68.7	67.7	80.0	77.2	77.5	76.8	76.2	62.8
Ratio 0/8	−0.23	+0.1	−1.5	−0.5	+0.7	−0.4	−0.1	−0.4	+0.1	−0.7	−0.5	**+0.0**	−0.1	+0.5	−0.6	+0.2	−0.1	−1.0	+0.8	+0.5	+0.1	−0.7	−1.6
	73.03	81.4	72.3	76.5	58.4	74.9	74.0	75.5	81.8	74.2	69.2	**72.1**	65.1	80.2	65.8	68.9	67.6	79.0	78.0	78.0	76.9	75.5	61.2
Ratio 1/8	−0.18	−0.3	−1.0	−0.4	+0.9	−1.9	+0.0	−0.1	−0.1	−0.3	−0.5	−0.7	+0.8	+0.3	+1.0	+0.2	−0.2	−1.8	+0.2	+0.5	−0.2	+0.3	−0.6
	73.08	81.0	72.8	76.6	58.6	73.4	74.1	75.8	81.7	74.6	69.2	71.4	66.0	80.0	67.4	68.9	67.5	78.2	77.4	78.0	76.6	76.5	62.2
Ratio 2/8	−0.36	**+0.4**	−0.5	−0.2	+0.1	−0.9	−0.6	+0.1	+0.2	−0.6	−1.3	−1.8	+0.1	−0.2	−0.7	+0.1	−0.5	−1.7	+0.0	+0.0	+0.2	+0.4	−0.6
	72.90	**81.7**	73.3	76.8	57.8	74.4	73.5	76.0	81.9	74.3	68.4	70.3	65.3	79.5	65.7	68.8	67.2	78.3	77.2	77.5	77.0	76.6	62.2
Ratio 3/8	−0.24	+0.1	−0.2	−0.4	**+1.3**	−1.8	−0.3	−0.3	+0.6	−0.1	−0.5	−1.6	−0.2	−0.1	−0.4	−0.8	**+0.3**	−0.9	+0.0	−0.5	−0.4	+0.6	+0.1
	73.02	81.4	73.6	76.6	**59.0**	73.5	73.8	75.6	82.3	74.8	69.2	70.5	65.0	79.6	66.0	67.9	**68.0**	79.1	77.2	77.0	76.4	76.8	62.9
Ratio 5/8	**+0.09**	+0.3	−0.6	+0.4	−2.6	−0.7	**+0.5**	**+0.7**	**+0.7**	+0.2	−0.6	−3.1	+1.2	**+0.6**	**+2.1**	**+0.8**	−0.3	−0.2	**+1.0**	+0.4	**+0.5**	+0.4	**+0.2**
	73.35	81.6	73.2	77.4	55.1	74.6	**74.6**	**76.7**	**82.4**	75.1	69.1	69.0	66.4	**80.3**	**68.5**	**69.5**	67.4	79.8	**78.2**	77.9	**77.3**	76.6	**63.0**
Ratio 6/8	−0.31	+0.0	+0.3	+0.4	+0.2	−2.5	−0.6	−0.6	+0.2	**+0.8**	−1.0	−1.9	−0.4	−0.2	−0.4	−0.1	−0.1	**+0.5**	−0.1	−0.1	−0.3	−0.9	+0.0
	72.95	81.3	74.1	77.4	57.9	72.8	73.5	75.3	81.9	**75.7**	68.7	70.2	64.8	79.5	66.0	68.6	67.6	**80.5**	77.1	77.4	76.5	75.3	62.8
Ratio 7/8	−0.32	+0.0	+0.0	+0.4	−2.4	−1.3	+0.1	−0.6	−0.4	−0.6	−1.1	−3.9	+0.0	−0.1	+0.9	+0.6	−0.3	−0.6	+0.2	**+0.7**	+0.1	**+1.2**	−0.1
	72.94	81.3	73.8	77.3	55.3	74.0	74.2	75.3	81.3	74.3	68.6	68.2	65.2	79.6	67.3	69.3	67.4	79.4	77.4	**78.2**	76.9	**77.4**	62.7
Ratio 8/8	−0.13	−0.3	**+0.3**	**+0.5**	−0.4	−0.9	−0.7	+0.0	+0.2	+0.0	−0.5	−1.7	**+1.4**	−0.1	+0.5	−0.4	−0.9	−0.6	+0.5	+0.0	+0.2	−0.3	+0.1
	73.13	81.0	**74.1**	**77.5**	57.3	74.4	73.4	75.9	81.9	74.9	69.2	70.4	**66.6**	79.6	66.9	68.3	66.8	79.4	77.7	77.5	77.0	75.9	62.9
F) Various Sampling Ratios of umT5-xl Model																							
Ratio 4/8	75.66	83.4	76.3	80.2	62.7	77.2	76.9	79.1	84.0	78.8	71.8	75.6	65.5	80.6	68.1	71.8	68.3	80.3	79.5	79.4	**79.5**	78.6	66.8
Ratio 0/8	−0.15	**+0.4**	−0.4	−0.9	**+0.9**	−0.4	−0.7	+0.7	−0.2	−0.6	−0.2	**+1.6**	+0.3	+0.3	+0.0	−0.8	+0.2	−0.5	−0.4	−0.1	−0.7	+0.1	−1.6
	75.51	**83.8**	75.9	79.3	**63.6**	76.8	76.2	79.8	83.8	78.2	71.6	**77.2**	65.8	80.9	68.1	71.0	68.5	79.8	79.1	79.3	78.8	78.7	65.2
Ratio 1/8	−0.11	+0.3	−0.7	−0.4	+0.4	−0.8	−0.2	+0.9	−0.2	−0.8	+0.2	+0.3	−0.2	+0.3	+0.4	+0.2	−0.1	−0.5	−0.2	−0.3	−0.5	**+0.7**	−1.3
	75.55	83.7	75.6	79.8	63.1	76.4	76.7	80.0	83.8	78.0	72.0	75.9	65.3	80.9	68.5	72.0	68.2	79.8	79.3	79.1	79.0	**79.3**	65.5
Ratio 2/8	**+0.06**	+0.4	+0.1	+0.2	+0.8	−0.3	−0.4	**+1.1**	+0.0	−0.4	+0.5	+0.6	+0.1	+0.6	+0.9	−0.5	−0.2	−1.0	+0.0	+0.4	−0.3	+0.6	−1.6
	75.72	83.8	76.4	80.4	63.5	76.9	76.5	**80.2**	84.0	78.4	72.3	76.2	65.6	81.2	69.0	71.3	68.1	79.3	79.5	79.8	79.2	79.2	65.2
Ratio 3/8	−0.04	+0.2	−0.1	−0.1	+0.4	−0.3	−0.3	+0.5	+0.1	−0.3	+0.1	+0.2	+0.4	+0.4	**+1.2**	−0.5	**+0.4**	−0.9	−0.3	−0.2	−0.2	−0.1	−1.2
	75.62	83.6	76.2	80.1	63.1	76.9	76.6	79.6	84.1	78.5	71.9	75.8	65.9	81.0	**69.3**	71.3	**68.7**	79.4	79.2	79.2	79.3	78.5	65.6
Ratio 5/8	+0.00	+0.3	+0.2	**+0.4**	−0.3	−0.2	−0.3	+0.5	+0.0	+0.2	−0.5	+0.3	−0.1	+0.2	−0.2	+0.1	+0.1	−0.4	+0.4	+0.1	−0.1	+0.1	−0.7
	75.66	83.7	76.5	**80.6**	62.4	77.0	76.6	79.6	84.0	79.0	71.3	75.9	65.4	80.8	67.9	71.9	68.5	79.9	79.9	79.5	79.4	78.7	66.1
Ratio 6/8	−0.05	+0.1	+0.5	+0.3	−0.9	**+0.2**	−0.2	+0.5	−0.2	−0.1	+0.4	−1.7	+0.8	+0.6	+0.0	−0.1	−0.3	+0.1	+0.4	+0.3	−0.6	+0.2	−1.2
	75.61	83.5	76.8	80.5	61.8	**77.4**	76.7	79.6	83.8	78.7	72.2	73.9	66.3	81.2	68.1	71.7	68.0	80.4	79.9	79.7	78.9	78.8	65.6
Ratio 7/8	−0.12	+0.1	**+0.6**	+0.3	−0.9	−2.1	**+0.1**	+0.4	**+0.3**	−0.1	+0.6	−1.4	**+1.3**	+0.3	+0.0	+0.5	+0.1	−0.4	−0.2	−0.4	−0.2	−0.2	−1.0
	75.54	83.5	**76.9**	80.5	61.8	75.1	**77.0**	79.5	**84.3**	78.7	72.4	74.2	**66.8**	80.9	68.1	72.3	68.4	79.9	79.3	79.0	79.3	78.4	65.8
Ratio 8/8	−0.07	−0.1	+0.4	+0.3	−1.8	−3.2	−0.1	+0.0	+0.0	**+0.4**	**+0.8**	−3.1	+0.8	**+0.9**	−0.1	**+0.7**	+0.3	**+1.4**	**+0.5**	**+0.4**	−0.1	+0.0	**+0.2**
	75.59	83.3	76.7	80.5	60.9	74.0	76.8	79.1	84.0	**79.2**	**72.6**	72.5	66.3	**81.5**	68.0	**72.5**	68.6	**81.7**	**80.0**	**79.8**	79.4	78.6	**67.0**

Table 6: Ablations experiments on the CorefUD 1.3 minidev set (CoNLL score in %). The results are averages of 3 or more runs and for every run the epoch with best average score over the whole CorefUD is used.

Cross-Lingual Zero-Shot Evaluation Given that our model is multilingual, it can be used to perform coreference resolution in languages not exposed to during training. In order to evaluate the performance of our model in such a setting, we train several multilingual models on corpora from all but one language, and then evaluate their performance on the excluded corpora. The results are summarized in Table 6.A for the mT5-large model and in Table 6.B for the umT5-xl model. While

Model	TensorFlow			PyTorch				
	Compile time	Training throughput	Max batch	Cold-start compile	Warm-start compile	Eager throughput	Compiled throughput	Max batch
mT5 base	50s	7.1batch/s	39	55s	27s	8.0batch/s	10.3batch/s	58
mT5 large	91s	3.1batch/s	13	92s	50s	3.3batch/s	4.4batch/s	21
mT5 xl	95s	2.5batch/s	5	97s	51s	2.3batch/s	2.9batch/s	9

Table 7: Comparison of compilation and training times of CorPipe using the latest TensorFlow 2.19 and PyTorch 2.7 with the latest transformers 4.52.4 on a single A100 40GB GPU. The training throughput is measured using batch size of 4 for the xl model and 8 otherwise.

the cross-lingual zero-shot performance is substantially lower by roughly 14 percentage points, it is still higher than the baseline system of Pražák et al. (2021) and on par with the best LLM-track submission. Interestingly, the performance of umT5-xl is higher by more than 2 points, an increase consistent with the results in the supervised setting.

Segment Size The effect of context larger than the usual 512 tokens is quantified in Table 6.C for the mT5-large model and in Table 6.D for the umT5-xl model. The results show that the increase from 512 to 1 024 tokens leads to a significant performance increase of more than 2 percentage points, and the further increase to 2 560 tokens brings a smaller increase by less than 0.5 points.

Sampling Ratio During training, we sample sentences from the training corpora proportionally to the square root of their size, following for example van der Goot et al. (2021); Straka (2024); Straka et al. (2024). We quantify the impact of using different exponents (sampling ratios) in Table 6.E for the mT5-large model and in Table 6.F for the umT5-xl model. The results show that while the choice of 0.5 is reasonable, the sampling ratio has very little impact on the average performance. However, we can see a minor effect of the sampling ratio on the performance of the two largest corpora (the Czech ones), with the decrease of 0.5 to 1.5 percentage points for uniform sampling (sampling ratio 0) to the increase of 0.3 to 0.5 percentage points for proportional sampling (sampling ratio 1).

6 PyTorch vs TensorFlow

Having both PyTorch and TensorFlow implementations of CorPipe, we can compare the two variants in terms of training throughput and memory usage. To this end, we compare the CorPipe 23 using the latest TensorFlow 2.19 and CorPipe 25 utilizing the latest PyTorch 2.7, both with the latest transformers library 4.52.4, on a single A100 40GB GPU.

The results are presented in Table 7. For all the base, large, and xl sizes, the PyTorch implementation outperforms the TensorFlow implementation:

- The training throughput is higher by 16% for the xxl model up to 45% for the base model, when comparing compiled PyTorch models to compiled TensorFlow models.
- The PyTorch model cold-start compilation time is quite similar to TensorFlow; however, the warm-start compilation (reusing cached compilation files from preceding executions; happens automatically) is significantly shorter, being circa half of the TensorFlow time.
- The eager PyTorch model has comparable or slightly better performance than the compiled TensorFlow model.
- The PyTorch implementation has lower memory requirements, allowing batches larger by at least 50% to fit into the GPU memory.

Note that the difference might stem just from different mT5 implementations (FlashAttention, etc.), not necessarily from the frameworks themselves.

7 Conclusions

We introduced CorPipe 25, the winning submission to the CRAC 2025 Shared Task on Multilingual Coreference Resolution (Novák et al., 2025a). Our approach employs a three-stage pipeline architecture that first predicts empty nodes using a dedicated pretrained encoder model, then performs mention detection and coreference linking through a jointly trained system utilizing another pretrained encoder. This complete PyTorch reimplementation significantly outperforms all other submissions by substantial margins of 7 and 8 percentage points for our single model and ensemble variants, respectively. The source code and trained models are publicly available at `https://github.com/ufal/crac2025-corpipe`.

Acknowledgements

Our research has been supported by the OP JAK project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic and uses data provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

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

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and conjunction-aware word-level coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Miloslav Konopík, Anna Nedoluzhko, Martin Popel, Ondřej Pražák, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025a. Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches? In *Proceedings of The Sixth Workshop on Computational Approaches to Discourse and The Eight Workshop on Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2025)*, Suzhou, China. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, and 29 others. 2025b. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *Preprint*, arXiv:2504.06225.