ÚFAL CorPipe at CRAC 2022: Effectivity of Multilingual Models for Coreference Resolution

Milan Straka, Jana Straková

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

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks
    - works better than individual models

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks
    - works better than individual models
    - can be used even on unseen languages

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks
    - works better than individual models
    - can be used even on unseen languages
  - contrary to the baseline solution:
    - CorPipe first predicts mentions

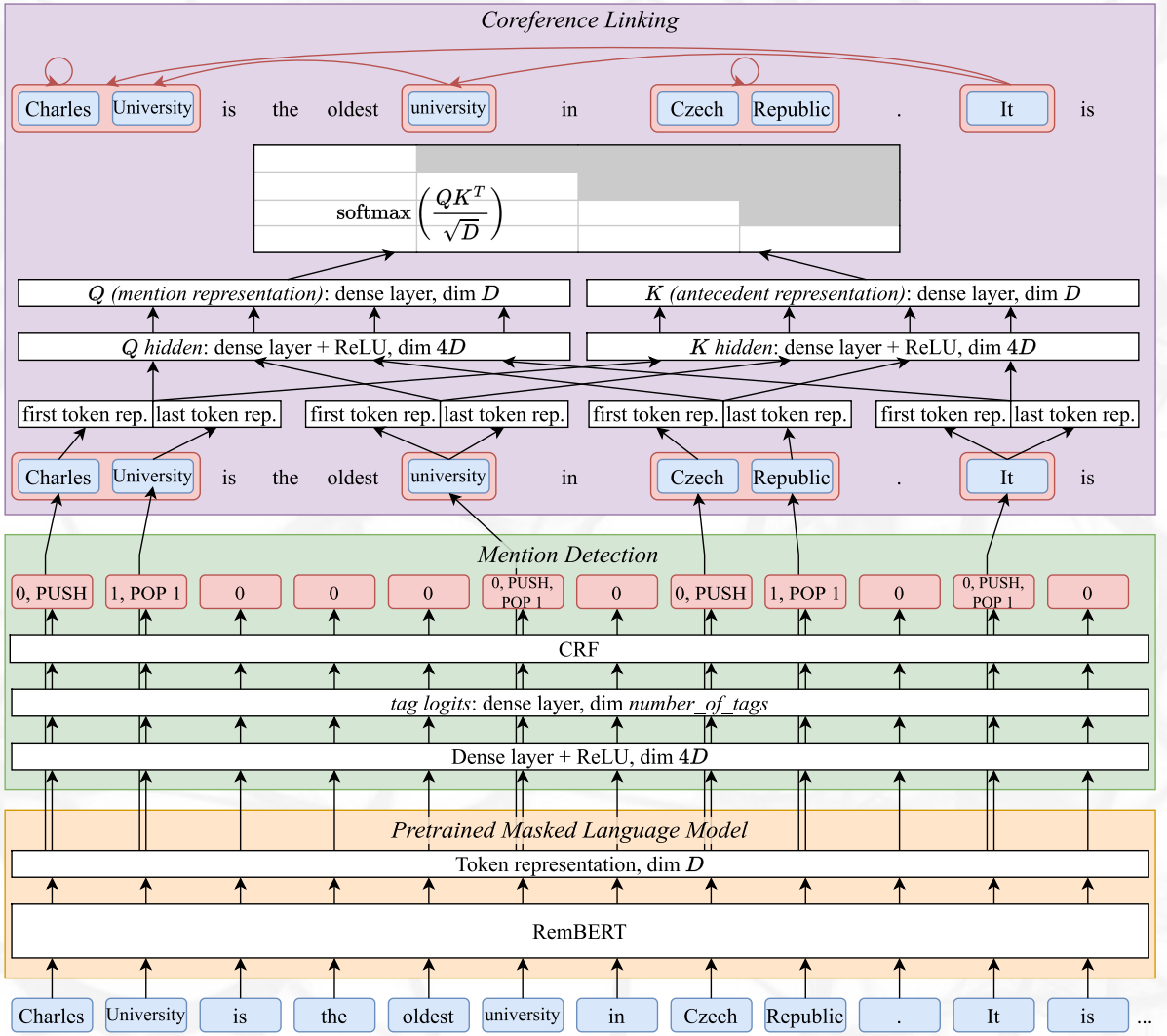
CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks
    - works better than individual models
    - can be used even on unseen languages
  - contrary to the baseline solution:
    - CorPipe first predicts mentions
    - only then it predicts links between the predicted mentions

CorPipe

- winning entry of the CRAC 2022 Shared Task on Multilingual Coreference Resolution
- distinguishing features:
  - a single multilingual model for all 13 treebanks
    - works better than individual models
    - can be used even on unseen languages
  - contrary to the baseline solution:
    - CorPipe first predicts mentions
    - only then it predicts links between the predicted mentions
    - both tasks performed by a single model

Corpipe Architecture



Official CRAC 2022 Results

Team/Submission	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
ÚFAL CorPipe	**70.72**	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
best dev	**1**	2	**1**	**1**	2	2	2	**1**	**1**	**1**	3	**1**	2	**1**
ÚFAL CorPipe	69.56	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
multilingual	2	**1**	2	2	3	**1**	**1**	3	2	3	2	3	**1**	2
UWB	67.64	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
ondfa†	3	4	4	4	**1**	3	4	4	4	4	**1**	2	4	3
ÚFAL CorPipe	64.30	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
individual	4	3	3	3	5	5	3	5	3	2	4	6	5	4
Barbora Dohnalová	59.72	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
berulasek	5	5	5	5	4	4	5	2	5	5	5	4	6	5
UWB	58.53	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.98	55.55	52.35	64.81	65.34	67.66
BASELINE‡	6	6	6	6	6	6	6	2	6	6	6	5	7	6
Matouš Moravec	55.05	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
moravec	7	7	7	7	7	7	7	2	7	7	7	7	8	7

Table 1: Official results of CRAC 2022 Shared Task on the test set (CoNLL score in %). The systems † and ‡ are described in Pražák and Konopik (2022) and Pražák et al. (2021), respectively; the rest in Žabokrtský et al. (2022).

Official CRAC 2022 Results

Team/Submission	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
ÚFAL CorPipe ***best dev***	**70.72** **1**	78.18 2	**78.59** **1**	**77.69** **1**	65.52 2	70.69 2	72.50 2	**39.00** **1**	**81.39** **1**	**65.27** **1**	63.15 3	**69.92** **1**	78.12 2	**79.34** **1**
ÚFAL CorPipe ***multilingual***	69.56 2	**78.49** **1**	78.49 2	77.57 2	59.94 3	**71.11** **1**	**73.20** **1**	33.55 3	80.80 2	64.35 3	63.38 2	67.38 3	**78.32** **1**	77.74 2
UWB *ondfa*†	67.64 3	70.55 4	74.07 4	72.42 4	**73.90** **1**	68.68 3	68.31 4	31.90 4	72.32 4	61.39 4	**65.01** **1**	68.05 2	75.20 4	77.50 3
ÚFAL CorPipe ***individual***	64.30 4	76.34 3	77.87 3	76.76 3	36.50 5	56.65 5	70.66 3	23.48 5	78.78 3	64.94 2	62.94 4	61.32 6	73.36 5	76.26 4
Barbora Dohnalová *berulasek*	59.72 5	64.67 5	70.56 5	67.95 5	38.50 4	57.70 4	63.07 5	36.44 2	66.61 5	56.04 5	55.02 5	65.67 4	65.99 6	68.17 5
UWB BASELINE‡	58.53 6	63.74 6	70.00 6	67.27 6	33.75 6	55.44 6	62.59 6	36.44 2	65.98 6	55.55 6	52.35 6	64.81 5	65.34 7	67.66 6
Matouš Moravec *moravec*	55.05 7	58.25 7	68.19 7	64.71 7	31.86 7	52.84 7	59.15 7	36.44 2	62.01 7	54.87 7	52.00 7	59.49 7	63.40 8	52.49 7

Table 1: Official results of CRAC 2022 Shared Task on the test set (CoNLL score in %). The systems † and ‡ are described in Pražák and Konopik (2022) and Pražák et al. (2021), respectively; the rest in Žabokrtský et al. (2022).

Team/Submission	Avg. with singletons
ÚFAL CorPipe, *best dev*	**72.98**
ÚFAL CorPipe, *multilingual*	71.81
ÚFAL CorPipe, *individual*	67.93
UWB, *ondfa*	58.06
Barbora Dohnalová, *berulasek*	50.84
UWB, BASELINE	49.69
Matouš Moravec, *moravec*	46.79

Test Results of RemBERT vs XLM-R and Multilinguality

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
XLM-R base, multilingual	67.8	77.1	75.8	74.3	54.7	66.9	70.1	38.5	77.6	64.2	62.3	69.4	73.3	76.6
Best base model, individual	-5.2	-4.0	+1.5	+2.2	-18.2	-9.8	-3.4	-15.0	-2.4	-2.4	-2.0	-8.1	+0.0	-5.6
Best base model, best dev	+0.4	-0.6	+1.5	+2.2	+2.0	+0.6	-1.0	-0.9	+1.2	-1.1	+0.6	+0.4	+0.0	+0.2
RemBERT, multilingual	+1.8	**+1.4**	+2.6	+3.3	+5.2	**+4.2**	**+3.1**	-4.9	+3.2	+0.1	**+1.1**	-2.0	**+5.0**	+1.1
RemBERT, individual	-3.5	-0.7	+2.0	+2.5	-18.2	-10.3	+0.6	-15.0	+1.2	+0.7	+0.7	-8.1	+0.0	-0.4
RemBERT, best dev	**+3.0**	+1.1	**+2.7**	**+3.4**	**+10.8**	+3.8	+2.4	**+0.5**	**+3.8**	**+1.0**	+0.9	**+0.5**	+4.8	**+2.7**

Comparing Pretrained Models on Dev

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
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 large multilingual	+1.5	**+1.7**	+1.8	+2.0	+0.3	**+4.1**	+2.1	-4.5	+2.2	+1.7	+3.1	-0.0	+2.9	+0.9
RemBERT multilingual	**+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**
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
XLM-R large individual	-0.6	+0.2	+2.8	+3.0	-7.7	-5.2	-0.9	-4.4	+1.0	+0.3	+3.7	-5.4	+3.5	-1.2
RemBERT individual	-4.7	+0.6	+2.8	+1.9	-23.0	-12.1	+0.7	-30.5	+1.1	+0.7	-0.4	-8.9	+2.7	-1.8
RemBERT 50% additional	+0.3	+1.0	+2.5	+2.4	-1.4	-0.5	+1.7	-8.3	+0.9	+1.3	+1.6	-3.5	+3.6	+1.7

Comparing Pretrained Models on Dev

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
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 large multilingual	+1.5	**+1.7**	+1.8	+2.0	+0.3	**+4.1**	+2.1	-4.5	+2.2	+1.7	+3.1	-0.0	+2.9	+0.9
RemBERT multilingual	**+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**
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
XLM-R large individual	-0.6	+0.2	+2.8	+3.0	-7.7	-5.2	-0.9	-4.4	+1.0	+0.3	+3.7	-5.4	+3.5	-1.2
RemBERT individual	-4.7	+0.6	+2.8	+1.9	-23.0	-12.1	+0.7	-30.5	+1.1	+0.7	-0.4	-8.9	+2.7	-1.8
RemBERT 50% additional	+0.3	+1.0	+2.5	+2.4	-1.4	-0.5	+1.7	-8.3	+0.9	+1.3	+1.6	-3.5	+3.6	+1.7

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**

Comparing Mixing Strategies on Dev

Experiment	Avg	ca	cs pcedt	cs pdt	de parc	de pots	en gum	en parc	es	fr	hu	lt	pl	ru
D) Effect of Mixing Ratios using XLM-R base Pretrained Model														
Logarithmic, w/o corpus id	**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**
Logarithmic, w/ corpus id	-0.4	-0.5	+0.1	+0.3	-0.8	-0.3	-0.6	-4.6	**+0.1**	**+0.6**	**+1.3**	-0.9	**+0.3**	-0.7
Uniform, w/o corpus id	-0.8	-0.5	-0.2	-0.9	-1.8	-3.5	-0.2	-1.9	+0.0	-0.0	+0.4	-1.1	+0.0	-1.5
Uniform, w/ corpus id	-1.6	-1.1	-0.5	-0.6	-6.4	-2.1	-0.4	-7.0	+0.1	+0.1	-0.5	-1.1	-0.6	-1.2
Linear, w/o corpus id	-0.3	**+0.1**	**+0.8**	**+1.1**	-1.1	-0.5	-0.5	-3.5	-0.1	+0.3	+1.0	**+0.3**	-0.1	-1.6
E) Effect of Mixing Ratios using RemBERT Pretrained Model														
Logarithmic, w/o corpus id	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**
Logarithmic, w/ corpus id	**+0.6**	+0.4	+0.1	+0.1	**+3.0**	**+1.2**	-0.1	**+5.8**	+0.3	**+0.9**	-2.4	-1.3	+0.1	-0.2
Uniform, w/o corpus id	+0.1	+1.2	-0.3	-0.1	+2.4	+0.5	**+0.0**	-0.9	-0.1	+0.7	-0.6	-0.2	**+0.1**	-1.2
Uniform, w/ corpus id	-0.1	-0.0	-0.2	-0.3	-4.2	+0.3	-0.1	+4.5	+0.4	+0.6	-1.0	-0.1	+0.1	-1.2
Linear, w/o corpus id	-0.1	**+1.3**	**+0.1**	**+0.2**	-2.3	-0.5	-1.5	+1.9	**+0.5**	+0.7	-1.0	**+0.4**	+0.0	-1.3

Zero-shot Evaluation on Dev

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

Thank You