

Coreference by Competition

Lessons from the CRAC Shared Task

Michal Novák

The team



Michal Novák



Martin Popel



Anna Nedoluzhko



Zdeněk Žabokrtský



Daniel Zeman



Miloslav Konopík



Ondřej Pražák



Jakub Sido



Milan Straka

Maciej Ogrodniczuk, Amir Zeldes, Barbora Dohnalová, Yilun Zhu, ...

Coreference by Competition

Lessons from the CRAC Shared Task

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution

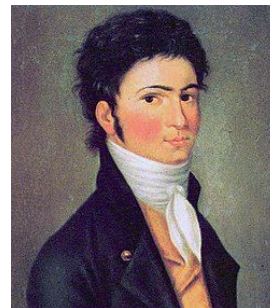
Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference

- two or more expressions in the text (mentions) refer to the same discourse entity

Beethoven was a musical genius. *The German composer* began going deaf in *his* late twenties, yet *he* continued to compose masterpieces.



- in some languages, mentions might not be even expressed on the surface (zero mentions)

Beethoven byl hudební génius. *Německý skladatel* začal ztrácet sluch ve *svých* dvaceti letech, přesto $\emptyset$ *subj* dál skládal mistrovská díla.

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference Resolution systems

Paper	Model	$\emptyset$ /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 _{mT5} ^{dev}	83.3 _{mT5}	$\mathcal{O}(n)$
Wu et al. (2020)	CorefQA	79.9 _{SpanB} ^{+QA}	83.1 _{SpanB} ^{+QA}			$\mathcal{O}(n)$
Straka (2023)	CorPipe		80.7 _{T5}	82.0 _{FT5}		1
Straka (2023)	CorPipe		77.2 _{mT5}	78.9 _{mT5}		1

from Straka (2023)

Coreference Resolution systems

evaluated
only on
English
OntoNotes
(and GAP in
some cases)

Paper	Model	Ø/ELMo/ base PLM	large PLM ~350M	xl PLM ~3B	xxl PLM ~11B	NN calls
Lee et al. (2017)	e2e	67.2 _Ø				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 _{mT5} ^{dev}	83.3 _{mT5}	$\mathcal{O}(n)$
Wu et al. (2020)	CorefQA	79.9 _{SpanB} ^{+QA}	83.1 _{SpanB} ^{+QA}			$\mathcal{O}(n)$
Straka (2023)	CorPipe		80.7 _{T5}	82.0 _{FT5}		1
Straka (2023)	CorPipe		77.2 _{mT5}	78.9 _{mT5}		1

from Straka (2023)

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Shared Tasks

- collaborative evaluation campaign organized within the research community
 - popular in the fields such as Natural Language Processing, Machine Learning
- provides a common problem, shared dataset and evaluation framework so that multiple teams can develop and compare their systems under the same conditions
- purposes:
 - to encourage research progress on a specific task
 - to increase reproducibility and comparability of the methods
 - to provide benchmark datasets, evaluation tools and baseline models
- examples:
 - CoNLL Shared Tasks
 - SemEval
 - WMT

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

- a multi-lingual collection of corpora annotated with coreference and anaphora
- harmonized using the same annotation scheme
- combines annotation of coreference/anaphora (always manual) with annotation of morphology and dependency syntax (manual if available, otherwise automatic)
- Universal Dependencies (Nivre et al., 2017)
 - source of inspiration
 - keeping the maximum compliance with its standards
- CorefUD 1.3 (Novák et al., 2025)
 - <http://hdl.handle.net/11234/1-5896>
 - 6th release since 2021
 - 28 datasets for 18 languages

Current languages and datasets

Public

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- English-GUM (Zeldes, 2017)
- English-LitBank (Bamman et al., 2019)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-PotsdamCC (Bourgonje and Stede, 2020)
- Norwegian-BokmaalNARC (Mæhlum et al., 2022)
- Norwegian-NynorskNARC (Mæhlum et al., 2022)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- Hindi-HDTB (Mujadia et al., 2016)

Non-public

- English-OntoNotes (Weischedel et al., 2011)
- English-ARRAU (Uryupina et al., 2020)

CorefUD 1.3

- French-Democrat (Landragin, 2021)
 - French-Ancor (Muzerelle et al., 2014)
 - Polish-PCC (Ogrodniczuk et al., 2013)
 - Lithuanian-LCC (Žitkus and Butkienė, 2018)
 - Russian-RuCor (Toldova et al., 2014)
 - Hungarian-SzegedKoref (Vincze et al., 2018)
 - Hungarian-KorKor (Vadász, 2022)
 - Turkish-ITCC (Pamay and Eryiğit, 2018)
 - Ancient Greek-PROIEL (Haug and Jøhndal, 2008)
 - Old Church Slavonic-PROIEL (Haug and Jøhndal, 2008)
 - Ancient Hebrew-PTNK (Swanson et al., 2024)
 - Korean-ECMT (Nam et al., 2020)
-
- English-PCEDT (Nedoluzhko et al., 2016)
 - Dutch-COREA (Hendrickx et al., 2008)

Current languages and datasets

CorefUD 1.3

Public

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- English-GUM (Zeldes, 2017)
- English-LitBank (Bamman et al., 2019)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-PotsdamCC (Bourgonje and Stede, 2020)
- Norwegian-BokmaalNARC (Mæhlum et al., 2022)
- Norwegian-NynorskNARC (Mæhlum et al., 2022)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- Hindi-HDTB (Mujadia et al., 2016)

Non-public

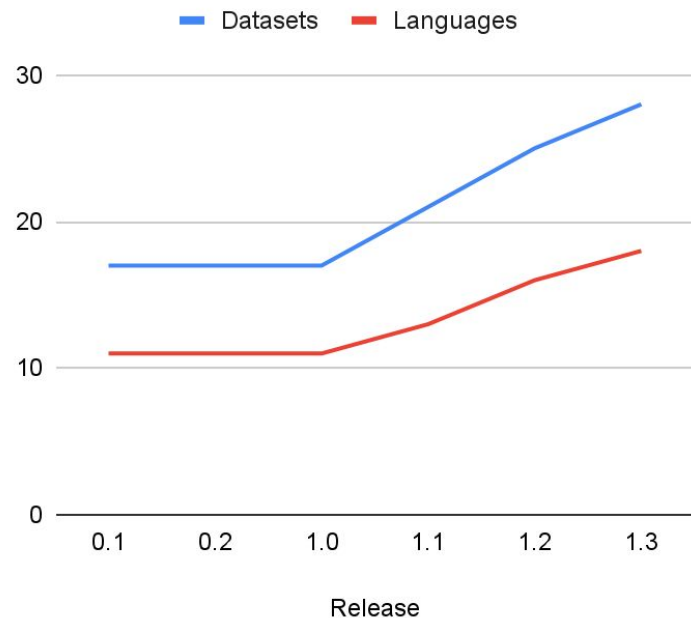
- English-OntoNotes (Weischedel et al., 2011)
- English-ARRAU (Uryupina et al., 2020)

- French-Democrat (Landragin, 2021)
- French-Ancor (Muzerelle et al., 2014)
- Polish-PCC (Ogrodniczuk et al., 2013)
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- Russian-RuCor (Toldova et al., 2014)
- Hungarian-SzegedKoref (Vincze et al., 2018)
- Hungarian-KorKor (Vadász, 2022)
- Turkish-ITCC (Pamay and Eryiğit, 2018)
- Ancient Greek-PROIEL (Haug and Jøhndal, 2008)
- Old Church Slavonic-PROIEL (Haug and Jøhndal, 2008)
- Ancient Hebrew-PTNK (Swanson et al., 2024)
- Korean-ECMT (Nam et al., 2020)

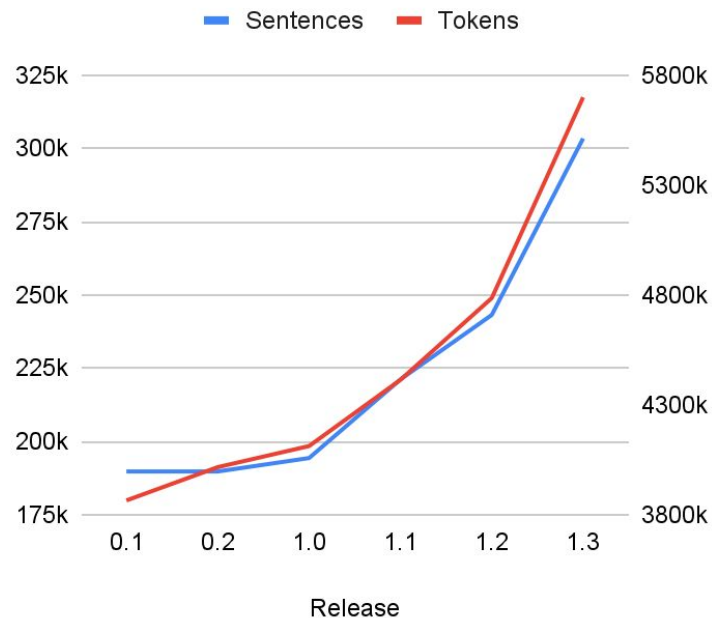
- English-PCEDT (Nedoluzhko et al., 2016)
- Dutch-COREA (Hendrickx et al., 2008)

Size over time

Datasets and Languages



Sentences and Tokens



- CorefUD 1.0 format
 - since CorefUD 1.0
 - fully compliant with the CoNLL-U format
 - may be included in UD releases
 - represented in the MISC field

```
# sent_id = ln94200-149-p2s3
# text = Pokud nebude jeho smlouva zrušena, je rozhodnut zanechat vůbec fotbalu, uvedl manažer brazilského hráče Ricardo Fuica.
# orig_file_sentence ln94200_149#5
1   Pokud      pokud      SCONJ  J,-----  _      5   mark      5:mark      _
2   nebude     být        AUX     VB-S---3F-NAI-- Aspect=... 5   aux:pass   5:aux:pass   _
3   jeho       jeho       DET     P9XXXZS3----- Gender[... 4   det        4:det        Entity=(e62601--1-gstype:spec) | Functor=4:APP
4   smlouva    smlouva    NOUN    NNFS1-----A---- Case=No... 5   nsubj:pass 5:nsubj:pass Functor=5:PAT
5   zrušena    zrušený    ADJ     VsQW----X-APP-- Aspect=... 8   advcl      8:advcl:pokud SpaceAfter=No | LDeriv=zrušit | Functor=8:COND
6   ,          ,          PUNCT   Z:-----  _      5   punct      5:punct      _
7   je         být        AUX     VB-S---3P-AAI-- Aspect=... 8   cop        8:cop        Functor=8:EFF
7.1 on        #PersPron PRON    _      Case=No... _      8:nsubj|9:nsubj:xsubj Entity=(e62601--1-gstype:spec) | Functor=8:ACT.cop, 9:ACT
8   rozhodnut  rozhodnutý ADJ     VsYS----X-APP-- Aspect=... 13  ccomp      13:ccomp     LDeriv=rozhodnout | Functor=13:PAT
9   zanechat   zanechat   VERB    Vf-----A-P--  Aspect=... 8   xcomp      8:xcomp      Functor=8:MANN
10  vůbec      vůbec      PART    TT-----  _      11  advmod:emph 11:advmod:emph Functor=11:EXT
11  fotbalu    fotbal     NOUN    NNIS2-----A---- Animacy... 9   obl:arg     9:obl:arg:gen SpaceAfter=No | Functor=9:PAT
12  ,          ,          PUNCT   Z:-----  _      8   punct      8:punct      _
13  uvedl      uvést      VERB    VpYS----R-AAP-- Aspect=... 0   root       0:root       Functor=0:PRED
14  manažer    manažer    NOUN    NNMS1-----A---- Animacy... 13  nsubj      13:nsubj     Entity=(e62605--1-gstype:spec | Functor=13:RSTR
15  brazilského brazilský  ADJ     AAMS2-----1A---- Animacy... 16  amod       16:amod      Entity=(e62601--2-gstype:spec | Functor=16:RSTR
16  hráče      hráč       NOUN    NNMS2-----A---- Animacy... 14  nmod       14:nmod:gen  Entity=e62601 | Functor=14:APP
17  Ricardo    Ricardo    PROPON NNMS1-----A---- Animacy... 14  flat       14:flat      Functor=14:RSTR
18  Fuica      Fuica      PROPON NNMS1-----A---- Animacy... 14  flat       14:flat      Entity=e62605 | Functor=14:ACT | SpaceAfter=No
19  .          .          PUNCT   Z:-----  _      13  punct      13:punct     _
```

- CorefUD 1.0 format
 - since CorefUD 1.0
 - fully compliant with the CoNLL-U format
 - may be included in UD releases
 - represented in the MISC field

```
# sent_id = ln94200-149-p2s3
# text = Pokud nebude jeho smlouva zrušena, je rozhodnut zanechat vůbec fotbalu, uvedl manažer brazilského hráče Ricardo Fuica.
# orig_file_sentence ln94200_149#5
1 Pokud pokud SCONJ J,----- 5 mark 5:mark -
2 nebude být AUX VB-S---3F-NAI-- Aspect=... 5 aux:pass 5:aux:pass -
3 jeho jeho DET P9XXXS3----- Gender[... 4 det 4:det Entity=(e62601--1-gstype:spec) | Functor=4:APP
4 smlouva smlouva NOUN NNFS1-----A---- Case=No... 5 nsubj:pass 5:nsubj:pass Functor=5:PAT
5 zrušena zrušený ADJ VsQW---X-APP-- Aspect=... 8 advcl 8:advcl:pokud SpaceAfter=No | LDeriv=zrušit | Functor=8:COND
6 , , PUNCT Z:----- - 5 punct 5:punct -
7 je být AUX VB-S---3P-AAI-- Aspect=... 8 cop 8:cop Functor=8:EFF
7.1 on #PersPron PRON - Case=No... - 8:nsubj | 9:nsubj:xsubj Entity=(e62601--1-gstype:spec) | Functor=8:ACT.cop, 9:ACT
8 rozhodnut rozhodnutý ADJ VsYS---X-APP-- Aspect=... 13 ccomp 13:ccomp LDeriv=rozhodnout | Functor=13:PAT
9 zanechat zanechat VERB Vf-----A-P-- Aspect=... 8 xcomp 8:xcomp Functor=8:MANN
10 vůbec vůbec PART TT----- - 11 advmod:emph 11:advmod:emph Functor=11:EXT
11 fotbalu fotbal NOUN NNIS2-----A---- Animacy... 9 obl:arg 9:obl:arg:gen SpaceAfter=No | Functor=9:PAT
12 , , PUNCT Z:----- - 8 punct 8:punct -
13 uvedl uvést VERB VpYS---R-AAP-- Aspect=... 0 root 0:root Functor=0:PRED
14 manažer manažer NOUN NNMS1-----A---- Animacy... 13 nsubj 13:nsubj Entity=(e62605--1-gstype:spec) | Functor=13:RSTR
15 brazilského brazilský ADJ AAMS2-----1A---- Animacy... 16 amod 16:amod Entity=(e62601--2-gstype:spec) | Functor=16:RSTR
16 hráče hráč NOUN NNMS2-----A---- Animacy... 14 nmod 14:nmod:gen Entity=e62601 | Functor=14:APP
17 Ricardo Ricardo PROPN NNMS1-----A---- Animacy... 14 flat 14:flat Functor=14:RSTR
18 Fuica Fuica PROPN NNMS1-----A---- Animacy... 14 flat 14:flat Entity=e62605 | Functor=14:ACT | SpaceAfter=No
19 . . PUNCT Z:----- - 13 punct 13:punct -
```

- Udapi (Popel et al., 2017)
 - toolkit for text and UD data manipulation
 - querying, statistics
 - format conversions
 - visualization
- coreference object model
 - mention
 - entity
 - bridging links

```
#!/usr/bin/env python3
import udapi

# Extract the words of the first sentence in the Spanish blind dev set.
doc = udapi.Document("es_ancora-corefud-dev.conllu")
trees = list(doc.trees)
words = trees[0].descendants
print([w.form for w in words])
# ['Los', 'jugadores', 'de', 'el', 'Espanyol', 'aseguraron', 'hoy', 'que',
# 'prefieren', 'enfrentar', 'se', 'a', 'el', 'Barcelona', 'en', 'la', 'final',
# 'de', 'la', 'Copa', 'de', 'el', 'Rey', 'en', 'lugar', 'de', 'en', 'las',
# 'semifinales', ',', 'tras', 'clasificar', 'se', 'ayer', 'ambos', 'equipos',
# 'catalanes', 'para', 'esta', 'ronda', '.']

# Create entity e1 with two mentions: "las semifinales" and "esta ronda"
e1 = doc.create_coref_entity()
e1.create_mention(words=words[27:29], head=words[28])
e1.create_mention(words=words[38:40], head=words[39])

# Create an empty node (zero) before the 9th word "prefieren".
zero = words[8].create_empty_child(deprel="nsubj", after=False, form="_")

# Make sure the input file es_ancora-corefud-dev.conllu is really
# the blind dev set without any empty nodes.
assert zero == trees[0].descendants_and_empty[8], "unexpected input file"

# Create entity e2 with two mentions:
# "Los jugadores de el Espanyol" and the newly created zero.
e2 = doc.create_coref_entity()
e2.create_mention(words=words[0:5], head=words[1])
e2.create_mention(words=[zero], head=zero)

# Print the newly created coreference entities.
udapi.create_block("corefud.PrintEntities").process_document(doc)

# Save the predictions into a CoNLL-U file
doc.store_conllu("output.conllu")
```

Visualization

Be it known then, that **Sir Walter**, like **a good father**, (having met with one or two private disappointments in very unreasonable applications), prided **himself** on remaining single for **his dear daughters'** sake.

For **one daughter**, **his eldest**, **he** would really have given up any thing, which **he** had not been very much tempted to do.

Elizabeth had succeeded, at sixteen, to all that was possible, of **her mother's** rights and consequence; and being very handsome, and very like **himself**, **her** influence had always been great, and **they** had gone on together most happily.

His two other children were of very inferior value.

Mary had acquired a little artificial importance, by becoming **Mrs Charles Musgrove**; but **Anne**, with an elegance of mind and sweetness of character, which must have placed **her** high with **any people of real understanding**, was **nobody** with either **father** or **sister**; **her** word had no weight, **her** convenience was always to give way--**she** was only **Anne**.

Visualization

Prof. Dr. Semih Koray'ın, dergimizde **bizim** ki eski yazılarından **onun** derlenerek kaleme alınan makalesini **onun** dosyamızın **bizim** sunuş yazısı **onun** yaptık **biz** .

Koray, **bilim etkinliğinin** **onun** **tarih** içinde **onun** ki serüvenini **onun** geleceğe de uzanarak analiz ederken **o** dosyamıza **bizim** güzel bir giriş de yapıyor **o** .

Dosyamız **bizim** esas olarak **iki bölümden** oluşuyor **o** .

1592년 (선조 25년) 임진왜란이 일어나자 세자(世子)로 책봉되었고, 함경도와 전라도 등지에서 군수품과 의병을 직접 모집하고 군량미를 모으는 데에도 힘썼다.

임진왜란 동안 세자의 몸으로 부왕을 돕고 전쟁 승리에 적지 않은 공을 세웠으나, 부왕의 인정과 칭찬은 커녕 견제와 냉대를 당해야만 했다.

1 תצא ריעה בחר ללאה אָשר ילדה לַ יעקב לַ ראות בִּבְנוֹת הָ אָרֶץ :

1 יָרָא אֱלֹהֵי הָ שָׁכַם בֶּן־חַמּוֹר הַ חַיִּי גִישָׁא הָ אָרֶץ וַיִּקַּח אִם־הָ וַיִּשְׁכַּב אִם־הָ וַיֵּגֶה הָ :

1 חֲדָקָה נִפְשׁוּ ׀ בִּ רְעִיָה בַח יַעֲקֹב וַיֵּאָהֶב אֶת־הָ פַעַר ׀ וַדְּבַר עַל־לֵב הָ פַעַר :

1 יֵאמֹר שָׁכַם אֶל־חַמּוֹר אָבִי ׀ לֵאמֹר קַח־לִי אֶת־הָ יְלִידָהּ הָ אֵאתָ לַ אִשָּׁה :

1 יַעֲקֹב שָׁמַע כִּי טָמְאָ אֶת־רְעִיָה בַח ׀ וַיִּכְגֵּי ׀ הִיוּ אֶת־מִקְנֵהוּ חַיִּי בִּהַ שְׂדֵה ׀ הַסִּטְשׁ יַעֲקֹב עַד־בֹּא ׀ :

Adding new datasets

- only datasets with free licenses
 - Chinese and Arabic OntoNotes
- datasets enhancing diversity preferred
 - language
 - ancient
 - non-European
 - Korean vs. Japanese?
 - domain
 - LitBank: English but fiction
 - Ancor: French but spoken

Adding new datasets

- implementing the conversion pipeline
 - by the core team members
 - majority of datasets
 - time-consuming
 - cooperation of the core team with external volunteers
 - outsourcing it with no or weak supervision
 - initialize and outsource
 - ideally by the authors of the datasets
 - to promote their dataset

Design decisions

- mention
 - no ID
 - full span specified
 - may be discontinuous
- mention head
 - determined from the dependency tree
 - using the Udapi block `corefud.MoveHead`
- coreference entities / clusters
 - grouping by co-indexing (not by links)
 - singletons allowed
- zero mentions
 - represented by empty nodes using enhanced UD graphs
 - empty nodes may have multiple parents
- bridging relations
 - cluster-to-cluster
 - mention-to-mention would require mention
- morphology and syntax
 - UD-based
 - gold if easy to convert
 - otherwise using UDPipe 2

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Coreference by Competition
Lessons from the CRAC Shared Task
... on Multilingual Coreference Resolution
... using CorefUD

Shared task details

- CRAC Shared Task on Multilingual Coreference Resolution
 - collocated with the Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)
 - based on CorefUD

Shared task details

- CRAC Shared Task on Multilingual Coreference Resolution
 - colocated with the Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)
 - based on CorefUD
- task
 - identify mentions in texts and predict which mentions belong to the same coreference cluster

Shared task	Languages	Zeros
SemEval 2010 (Recasens et al., 2010)	7	not stated
CoNLL 2012 (Pradhan et al., 2012)	3	removed

Shared task details

- CRAC Shared Task on Multilingual Coreference Resolution
 - colocated with the Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)
 - based on CorefUD
- task
 - identify mentions in texts and predict which mentions belong to the same coreference cluster

Shared task	Languages	Zeros
SemEval 2010 (Recasens et al., 2010)	7	not stated
CoNLL 2012 (Pradhan et al., 2012)	3	removed
CRAC 2022 (Žabokrtský et al., 2022)	10	included (pre-defined slots)
CRAC 2023 (Žabokrtský et al., 2023)	12	included (pre-defined slots)

Shared task details

- CRAC Shared Task on Multilingual Coreference Resolution
 - collocated with the Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)
 - based on CorefUD
- task
 - predict empty nodes
 - identify mentions in texts and predict which mentions belong to the same coreference cluster

Shared task	Languages	Zeros
SemEval 2010 (Recasens et al., 2010)	7	not stated
CoNLL 2012 (Pradhan et al., 2012)	3	removed
CRAC 2022 (Žabokrtský et al., 2022)	10	included (pre-defined slots)
CRAC 2023 (Žabokrtský et al., 2023)	12	included (pre-defined slots)
CRAC 2024 (Novák et al., 2024)	15	included

- scoring is complex
 - comparing clusters of mentions
 - MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998), CEAF-e, CEAF-m (Luo, 2005), BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
 - each one has P/R/F1

- scoring is complex
 - comparing clusters of mentions
 - **MUC** (Vilain et al., 1995), **B3** (Bagga and Baldwin, 1998), **CEAF-e**, CEAF-m (Luo, 2005), **BLANC** (Recasens and Hovy, 2011), **LEA** (Moosavi and Strube, 2016)
 - each one has P/R/**F1**
 - CoNLL F1 score
 - **average**
 - macro-averaged over all datasets

Evaluation I

- scoring is complex
 - comparing clusters of mentions
 - **MUC** (Vilain et al., 1995), **B3** (Bagga and Baldwin, 1998), **CEAF-e**, **CEAF-m** (Luo, 2005), **BLANC** (Recasens and Hovy, 2011), **LEA** (Moosavi and Strube, 2016)
 - each one has P/R/F1
 - CoNLL F1 score
 - **average**
 - macro-averaged over all datasets
- gold and predicted mentions must be matched
 - exact match: used traditionally for English
 - partial match: CRAC'22
 - head match: since CRAC'23

		Match		
		Exact	Partial	Head
Predicted mention	Gold mention			
		✓	✓	✓
		✓	✓	✗
		✗	✓	✓
		✗	✓	✗
		✗	✗	✓
		✗	✗	✗
		✗	✗	✗

Evaluation II

- zero matching
 - no special treatment until CRAC 2024
 - all empty nodes were already reconstructed in the input
 - predicted empty nodes no longer guaranteed to align 1:1 with the gold empty nodes
 - dependency-based matching
 - priority to the accurate assignment of both parents and dep. types, but parents are enough

1 

2 

3 

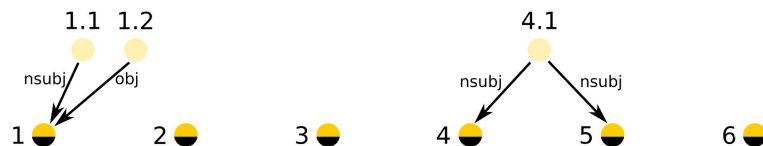
4 

5 

6 

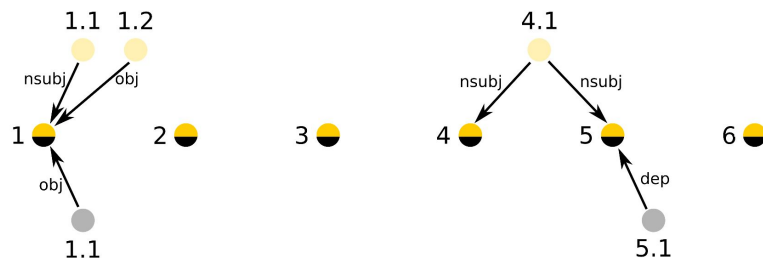
Evaluation II

- zero matching
 - no special treatment until CRAC 2024
 - all empty nodes were already reconstructed in the input
 - predicted empty nodes no longer guaranteed to align 1:1 with the gold empty nodes
 - dependency-based matching
 - priority to the accurate assignment of both parents and dep. types, but parents are enough



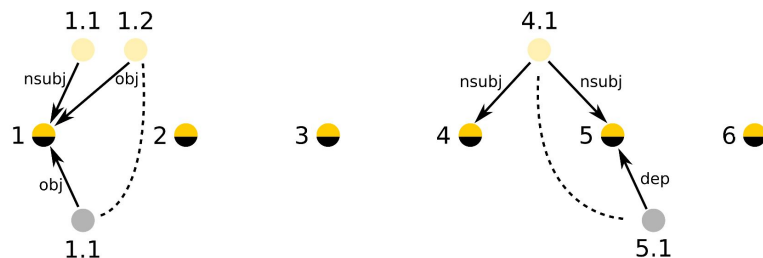
Evaluation II

- zero matching
 - no special treatment until CRAC 2024
 - all empty nodes were already reconstructed in the input
 - predicted empty nodes no longer guaranteed to align 1:1 with the gold empty nodes
 - dependency-based matching
 - priority to the accurate assignment of both parents and dep. types, but parents are enough



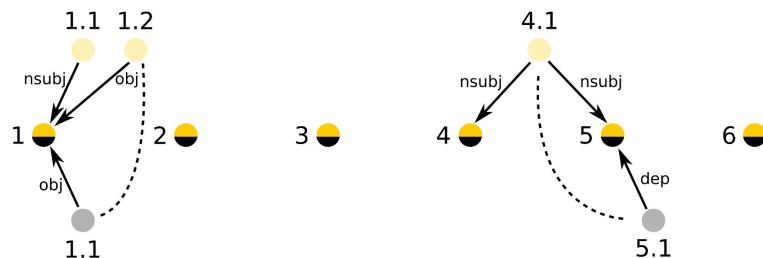
Evaluation II

- zero matching
 - no special treatment until CRAC 2024
 - all empty nodes were already reconstructed in the input
 - predicted empty nodes no longer guaranteed to align 1:1 with the gold empty nodes
 - dependency-based matching
 - priority to the accurate assignment of both parents and dep. types, but parents are enough



Evaluation II

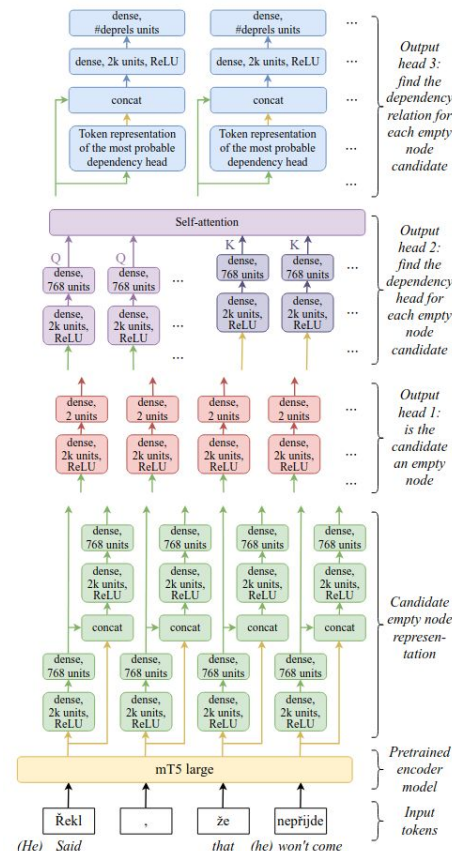
- zero matching
 - no special treatment until CRAC 2024
 - all empty nodes were already reconstructed in the input
 - predicted empty nodes no longer guaranteed to align 1:1 with the gold empty nodes
 - dependency-based matching
 - priority to the accurate assignment of both parents and dep. types, but parents are enough



- mention-decomposable score for zeros
 - difficult to compute traditional scores only on a specific type of mentions
 - allows it but due to oversimplification may lead to inaccuracies

Baseline systems

- empty nodes prediction (by Milan Straka)
 - based on XLM-RoBERTa large (Conneau et al., 2020)
 - two empty-node candidates for each word
 - its representation processed by three prediction heads:
 - empty node
 - word order
 - dependency relation
 - trained on a combination of all CorefUD datasets with zeros
 - macro-avg F1 = 82.9 (CRAC'24)
- coreference resolution (by Ondřej Pražák)
 - same each year
 - based on the system by (Pražák et al., 2021), originally proposed by (Lee et al., 2018)
 - built on multi-lingual BERT
 - same system for all languages



from Straka (2024)

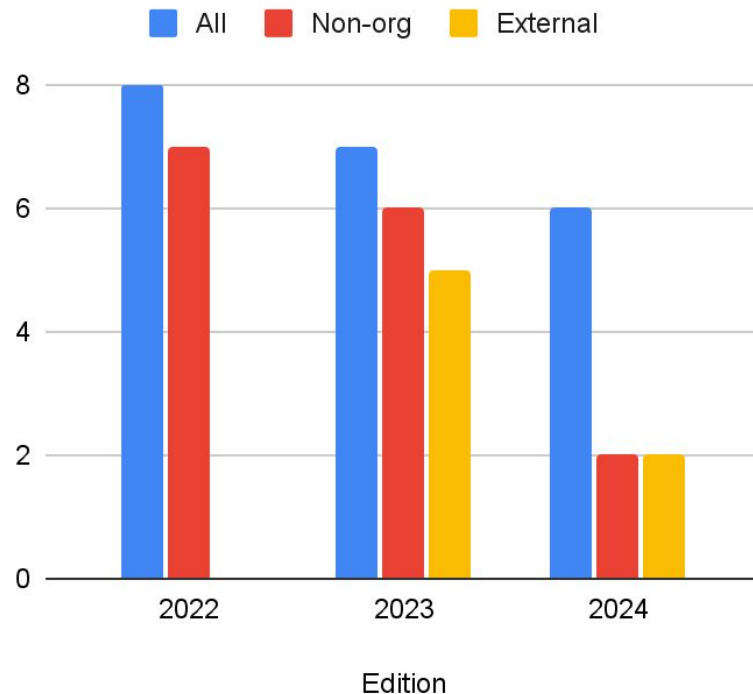
Participants

- advertising
 - among colleagues and our students
 - ACL Portal mailing list
 - direct emails to authors of coreference-related papers (by Anja Nedoluzhko)

Participants

- advertising
 - among colleagues and our students
 - ACL Portal mailing list
 - direct emails to authors of coreference-related papers (by Anja Nedoluzhko)

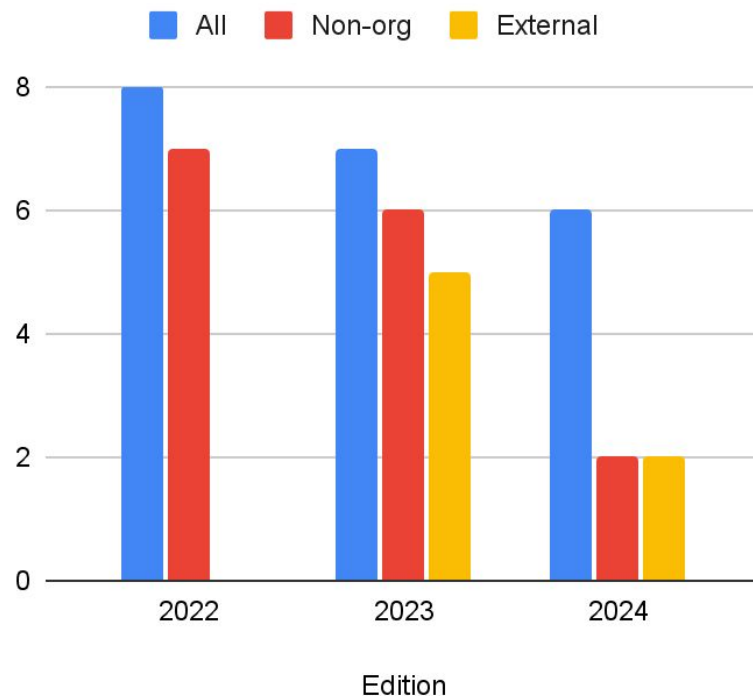
Submissions



Participants

- advertising
 - among colleagues and our students
 - ACL Portal mailing list
 - direct emails to authors of coreference-related papers (by Anja Nedoluzhko)
- difficult to attract
 - CR no longer attractive?
 - only workshop?
 - our organization and dissemination mistakes?

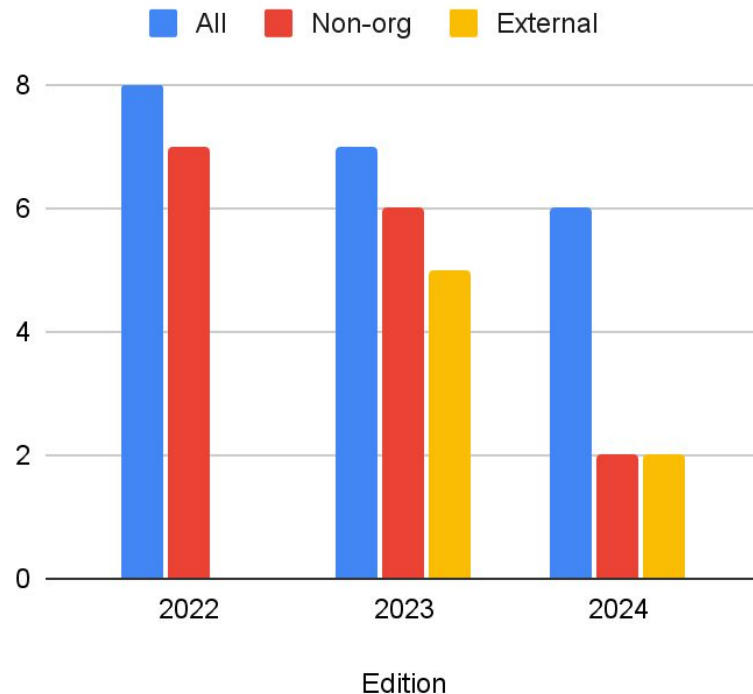
Submissions



Participants

- advertising
 - among colleagues and our students
 - ACL Portal mailing list
 - direct emails to authors of coreference-related papers (by Anja Nedoluzhko)
- difficult to attract
 - CR no longer attractive?
 - only workshop?
 - our organization and dissemination mistakes?
- still thankful to all the participants
 - Milan Straka (the winner of each edition)
 - Natalia Skachkova from DFKI
 - Ondřej Pražák from UWB

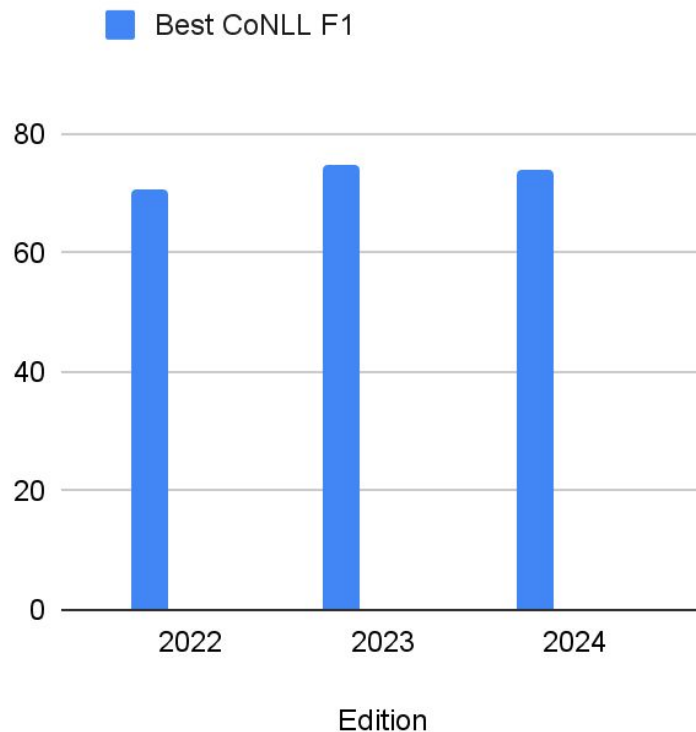
Submissions



Multilingual CR Performance

- winner's CoNLL F1
 - around 70-75
 - changes in data and evaluation setup across years
 - not directly comparable

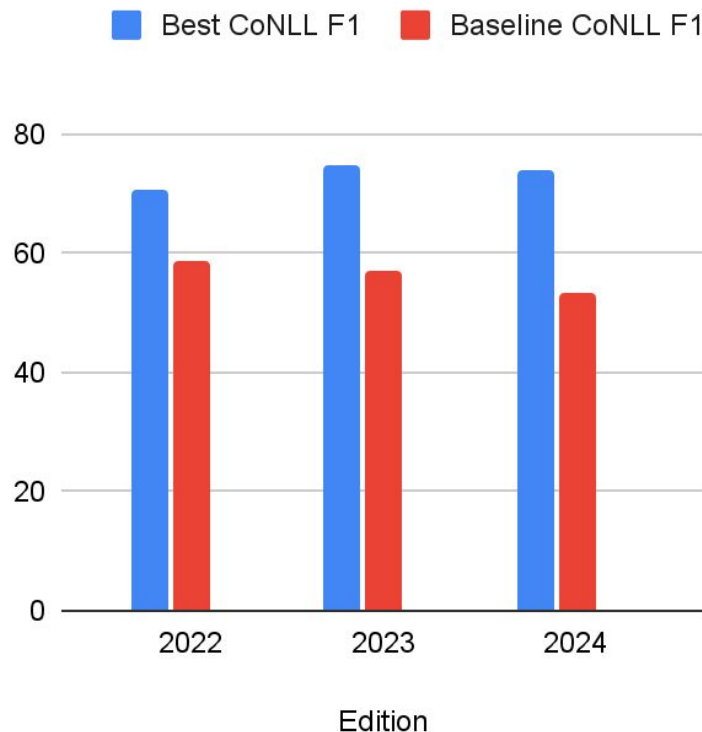
Performance



Multilingual CR Performance

- winner's CoNLL F1
 - around 70-75
 - changes in data and evaluation setup across years
 - not directly comparable
- baseline
 - still the same system, just retrained
 - the task gets more difficult

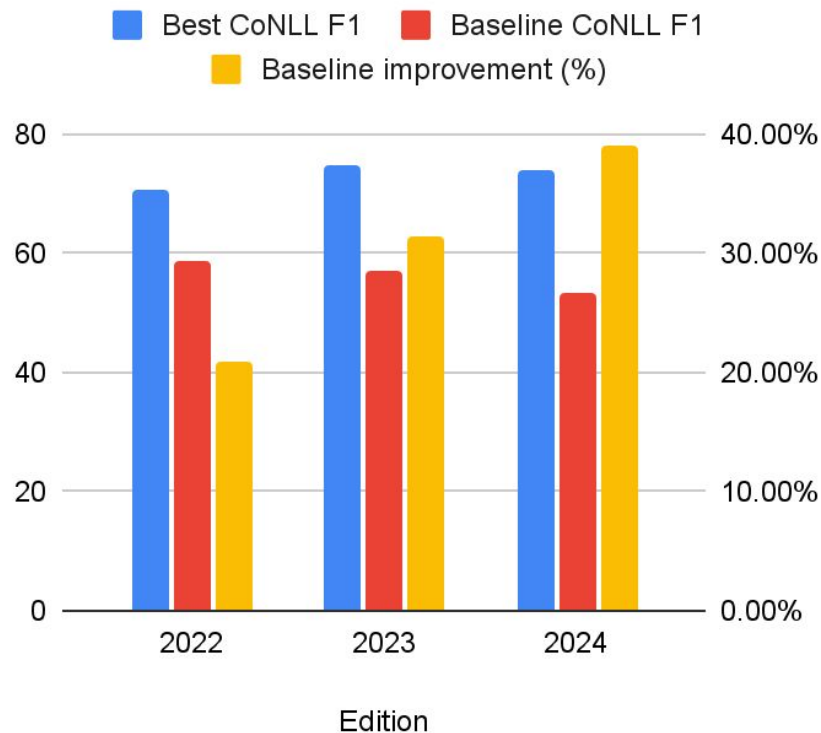
Performance



Multilingual CR Performance

- winner's CoNLL F1
 - around 70-75
 - changes in data and evaluation setup across years
 - not directly comparable
- baseline
 - still the same system, just retrained
 - the task gets more difficult
- improvement of the winner over the baseline (%)
 - fair comparison
 - 10 perc. points improvement each year

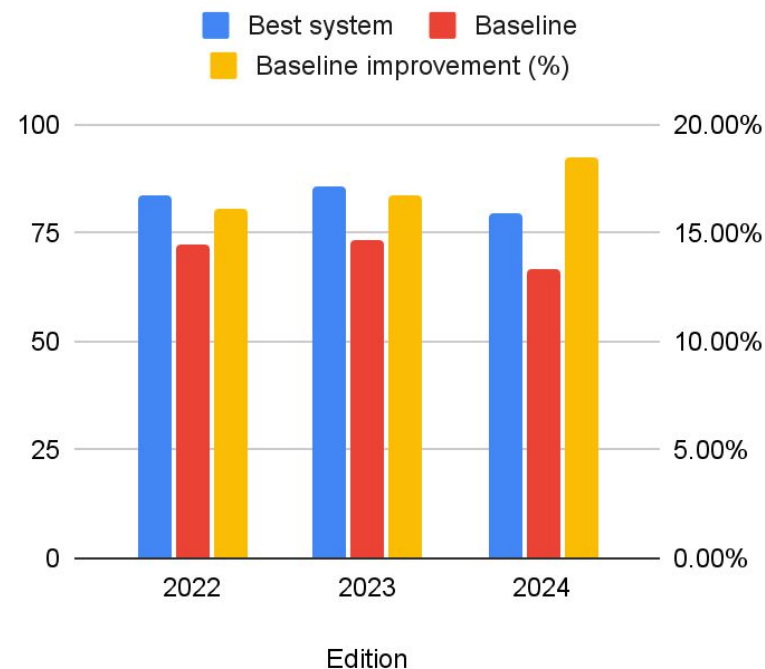
Performance



Performance on zeros

- mention-decomposable score for zeros
 - averaged over all datasets with zeros (2022: 5, 2023: 6, 2024: 10)
- moving to more realistic setup
 - 2023 -> 2024
 - drop in absolute performance
- improvement of the winner over the baseline

Performance on zeros



CRAC 2025 Shared Task

CRAC 2025 Shared Task

- focusing on Large Language Models (LLMs)
 - special track where only systems based on LLMs can be used
 - no special architectures allowed
- unconstrained track
 - the same as last years
- CorefUD 1.3
- participants
 - so far, 10+ registered
 - one of the projects in the ÚFAL's NPFL140 LLMs course (1 team)
- <https://ufal.mff.cuni.cz/corefud/crac25>

▶ Registration	as soon as possible
▶ Development phase	March 21 - June 3
▶ Evaluation phase	June 4-17
▶ Submission of system description papers	by TBA
▶ Workshop	TBA

The team



Michal Novák



Martin Popel



Anna Nedoluzhko



Zdeněk Žabokrtský



Daniel Zeman



Miloslav Konopík



Ondřej Pražák



Jakub Sido



Milan Straka

Maciej Ogrodniczuk, Amir Zeldes, Barbora Dohnalová, Yilun Zhu, ...

Summary

- standardizing multilingual coreference data
- standardizing evaluation framework
- turning attention to some more challenging aspects of coreference resolution

<https://ufal.cz/corefud>

<https://ufal.cz/corefud/crac25>