CRAC 2024, Miami

Findings of the Third Shared Task on Multilingual Coreference Resolution

Michal Novák[1], Barbora Dohnalová[1], Miloslav Konopík[2], Anna Nedoluzhko[1], Martin Popel[1], Ondřej Pražák[2], Jakub Sido[2], Milan Straka[1], Zdeněk Žabokrtský[1], Daniel Zeman[1]

November 15, 2024

ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

[1] Charles University, Prague, Czechia
[2] University of West Bohemia, Pilsen, Czechia

Outline

Introduction

Datasets

Evaluation Metrics

Participating Systems

Results and Comparison

Conclusion

Introduction

Motivation

- multilingual shared tasks: source of momentum in NLP subfields
    - e.g. CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006)
    - availability of the data is a limiting factor

Motivation

- multilingual shared tasks: source of momentum in NLP subfields
  - e.g. CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006)
  - availability of the data is a limiting factor
- CorefUD (Nedoluzhko et al., 2022a)
  - a multi-lingual collection of corpora annotated with coreference and anaphora
  - harmonized using the same annotation scheme

Motivation

- multilingual shared tasks: source of momentum in NLP subfields
  - e.g. CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006)
  - availability of the data is a limiting factor
- CorefUD (Nedoluzhko et al., 2022a)
  - a multi-lingual collection of corpora annotated with coreference and anaphora
  - harmonized using the same annotation scheme
- shared tasks on multilingual coreference resolution:

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	15	included

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.2
  - training (gold), dev (gold, no annot), test (no annot)

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.2
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.2
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline systems:
  - (1) for empty nodes prediction, (2) for coreference resolution
  - systems and their predictions on dev and test sets (3 starting points)

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.2
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline systems:
  - (1) for empty nodes prediction, (2) for coreference resolution
  - systems and their predictions on dev and test sets (3 starting points)
- Environment:
  - powered by CodaLab (`https://codalab.lisn.upsaclay.fr/competitions/19106`)
  - automatic validation, evaluation and ranking of the submissions

Shared Task

- Task:
  1. predict empty nodes
  2. identify mentions in texts and predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.2
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline systems:
  - (1) for empty nodes prediction, (2) for coreference resolution
  - systems and their predictions on dev and test sets (3 starting points)
- Environment:
  - powered by CodaLab (`https://codalab.lisn.upsaclay.fr/competitions/19106`)
  - automatic validation, evaluation and ranking of the submissions
- **`https://ufal.mff.cuni.cz/corefud/crac24`**

Changes to the 2023 edition

1. using a newer version of the collection: CorefUD 1.2
   - more low-resource and non-Latin-script languages: Ancient Greek, Ancient Hebrew, and Old Church Slavonic
   - new domain: novels with longer documents in LitBank

Changes to the 2023 edition

1. using a newer version of the collection: CorefUD 1.2
   - more low-resource and non-Latin-script languages: Ancient Greek, Ancient Hebrew, and Old Church Slavonic
   - new domain: novels with longer documents in LitBank
2. more focus on zeros
   - three new languages with zeros: Ancient Greek, Old Church Slavonic, and Turkish
   - slots for zeros (empty nodes) must be predicted

Datasets

CorefUD 1.2

- public edition of CorefUD 1.2 (Nedoluzhko et al., 2022b)
- 21 coreference datasets for 15 languages
- 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)
- the format is valid CoNLL-U; coreference information stored in the MISC column
- we followed the train/dev/test split of the collection

CorefUD 1.2: public datasets

- 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)
- French-Democrat (Landragin, 2021)
- 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)

CorefUD 1.2: new datasets

- 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)
- French-Democrat (Landragin, 2021)
- 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)

Data Statistics

CorefUD dataset	docs	sents	words	empty n.	entities	avg. len.	mentions
Ancient_Greek-PROIEL	19	6,475	64,111	6,283	3,215	6.6	21,354
Ancient_Hebrew-PTNK	40	1,161	28,485	0	870	7.2	6,247
Catalan-AnCora	1,298	13,613	429,313	6,377	17,558	3.6	62,417
Czech-PCEDT	2,312	49,208	1,155,755	35,654	49,225	3.4	168,055
Czech-PDT	3,165	49,428	834,720	21,808	46,628	3.3	154,905
English-GUM	217	12,147	211,920	115	8,270	4.4	36,733
English-LitBank	100	8,560	210,530	0	2,164	10.8	23,340
English-ParCorFull	19	543	10,798	0	188	4.4	835
French-Democrat	126	13,057	284,883	0	7,162	6.5	46,487
German-ParCorFull	19	543	10,602	0	243	3.7	896
German-PotsdamCC	176	2,238	33,222	0	880	2.9	2,519
Hungarian-KorKor	94	1,351	24,568	1,988	1,124	3.7	4,103
Hungarian-SzegedKoref	400	8,820	123,968	4,857	4,769	3.2	15,165
Lithuanian-LCC	100	1,714	37,014	0	1,087	4.0	4,337
Norwegian-BokmaalNARC	346	15,742	245,515	0	5,658	4.7	26,611
Norwegian-NynorskNARC	394	12,481	206,660	0	5,079	4.3	21,847
Old_Church_Slavonic-PROIEL	26	6,832	61,759	6,289	3,396	6.5	22,116
Polish-PCC	1,828	35,874	538,885	18,615	22,143	3.7	82,706
Russian-RuCor	181	9,035	156,636	0	3,515	4.6	16,193
Spanish-AnCora	1,356	14,159	458,418	8,112	19,445	3.6	70,663
Turkish-ITCC	24	4,732	55,358	11,584	4,019	5.4	21,569

Data Statistics

CorefUD dataset	docs	sents	words	empty n.	entities	avg. len.	mentions
Ancient_Greek-PROIEL	19	6,475	64,111	6,283	3,215	6.6	21,354
Ancient_Hebrew-PTNK	40	1,161	28,485	0	870	7.2	6,247
Catalan-AnCora	1,298	13,613	429,313	6,377	17,558	3.6	62,417
Czech-PCEDT	2,312	49,208	1,155,755	35,654	49,225	3.4	168,055
Czech-PDT	3,165	49,428	834,720	21,808	46,628	3.3	154,905
English-GUM	217	12,147	211,920	115	8,270	4.4	36,733
English-LitBank	100	8,560	210,530	0	2,164	10.8	23,340
English-ParCorFull	19	543	10,798	0	188	4.4	835
French-Democrat	126	13,057	284,883	0	7,162	6.5	46,487
German-ParCorFull	19	543	10,602	0	243	3.7	896
German-PotsdamCC	176	2,238	33,222	0	880	2.9	2,519
Hungarian-KorKor	94	1,351	24,568	1,988	1,124	3.7	4,103
Hungarian-SzegedKoref	400	8,820	123,968	4,857	4,769	3.2	15,165
Lithuanian-LCC	100	1,714	37,014	0	1,087	4.0	4,337
Norwegian-BokmaalNARC	346	15,742	245,515	0	5,658	4.7	26,611
Norwegian-NynorskNARC	394	12,481	206,660	0	5,079	4.3	21,847
Old_Church_Slavonic-PROIEL	26	6,832	61,759	6,289	3,396	6.5	22,116
Polish-PCC	1,828	35,874	538,885	18,615	22,143	3.7	82,706
Russian-RuCor	181	9,035	156,636	0	3,515	4.6	16,193
Spanish-AnCora	1,356	14,159	458,418	8,112	19,445	3.6	70,663
Turkish-ITCC	24	4,732	55,358	11,584	4,019	5.4	21,569

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD

Dataset	Empty nodes
grc_proiel	6,283
ca_ancora	6,377
cs_pcedt	35,654
cs_pdt	21,808
en_gum	115
hu_korkor	1,988
hu_szeged	4,857
cu_proiel	6,289
pl_pcc	18,615
es_ancora	8,112
tr_itcc	11,584

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD
- empty nodes newly left out from the test data
  - must be predicted by the systems
  - big shift towards the fully realistic setup
  - we provide baseline system to keep the task equally accessible to participants

Dataset	Empty nodes
grc_proiel	6,283
ca_ancora	6,377
cs_pcedt	35,654
cs_pdt	21,808
en_gum	115
hu_korkor	1,988
hu_szeged	4,857
cu_proiel	6,289
pl_pcc	18,615
es_ancora	8,112
tr_itcc	11,584

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD
- empty nodes newly left out from the test data
  - must be predicted by the systems
  - big shift towards the fully realistic setup
  - we provide baseline system to keep the task equally accessible to participants
- datasets with zeros extended
  - new
  - old, newly with zeros
  - old, better conversion of zeros

Dataset	Empty nodes
grc_proiel	6,283
ca_ancora	6,377
cs_pcedt	35,654
cs_pdt	21,808
en_gum	115
hu_korkor	1,988
hu_szeged	4,857
cu_proiel	6,289
pl_pcc	18,615
es_ancora	8,112
tr_itcc	11,584

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD
- empty nodes newly left out from the test data
  - must be predicted by the systems
  - big shift towards the fully realistic setup
  - we provide baseline system to keep the task equally accessible to participants
- datasets with zeros extended
  - new
  - old, newly with zeros
  - old, better conversion of zeros

Dataset	Empty nodes
grc_proiel	6,283
ca_ancora	6,377
cs_pcedt	35,654
cs_pdt	21,808
en_gum	115
hu_korkor	1,988
hu_szeged	4,857
cu_proiel	6,289
pl_pcc	18,615
es_ancora	8,112
tr_itcc	11,584

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD
- empty nodes newly left out from the test data
    - must be predicted by the systems
    - big shift towards the fully realistic setup
    - we provide baseline system to keep the task equally accessible to participants
- datasets with zeros extended
    - new
    - old, newly with zeros
    - old, better conversion of zeros

Dataset	Empty nodes
grc_proiel	6,283
ca_ancora	6,377
cs_pcedt	35,654
cs_pdt	21,808
en_gum	115
hu_korkor	1,988
hu_szeged	4,857
cu_proiel	6,289
pl_pcc	18,615
es_ancora	8,112
tr_itcc	11,584

Annotation Details: Format

- participants asked to predict coreference only (no bridging or other anaphoric relations)
- the `Entity` attribute
  - bracketing
  - entity/cluster ID
  - head
  - ~~other coreference-related attributes~~

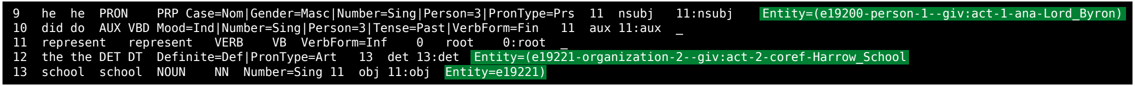
Annotation Details: Format

- participants asked to predict coreference only (no bridging or other anaphoric relations)
- the `Entity` attribute
  - bracketing
  - entity/cluster ID
  - head
  - ~~other coreference-related attributes~~

Annotation Details: Format

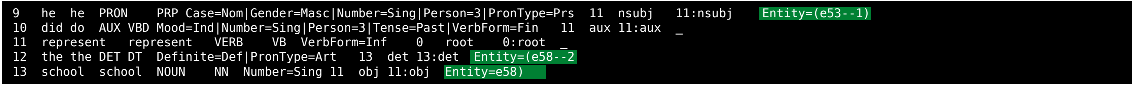
- participants asked to predict coreference only (no bridging or other anaphoric relations)
- the `Entity` attribute
  - bracketing
  - entity/cluster ID
  - head
  - ~~other coreference-related attributes~~

Gold file:



```
9   he  he  PRON    PRP Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  nsubj   11:nsubj    Entity=(e19200-person-1--giv:act-1-ana-Lord_Byron)
10  did do  AUX VBD Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   11  aux 11:aux  _
11  represent   represent   VERB    VB  VerbForm=Inf    0   root    0:root  _
12  the the DET DT  Definite=Def|PronType=Art   13  det 13:det  Entity=(e19221-organization-2--giv:act-2-coref-Harrow_School
13  school  school  NOUN    NN  Number=Sing 11  obj 11:obj  Entity=e19221)
```

Predicted file:



```
9   he  he  PRON    PRP Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11  nsubj   11:nsubj    Entity=(e53--1)
10  did do  AUX VBD Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   11  aux 11:aux  _
11  represent   represent   VERB    VB  VerbForm=Inf    0   root    0:root  _
12  the the DET DT  Definite=Def|PronType=Art   13  det 13:det  Entity=(e58--2
13  school  school  NOUN    NN  Number=Sing 11  obj 11:obj  Entity=e58)
```

Data preprocessing and starting points

- CorefUD data adjusted for the shared task
- *Gold data*
  - exactly the same, except for a minor technical modification
  - train and dev set
- *Input data*
  - much closer to the real-world scenario
  - morpho-syntactic features replaced with outputs of UDPipe 2 (Straka, 2018)
  - empty nodes removed
  - coreference annotation removed
- Starting points

	Baseline	
Starting point	**Empty nodes**	**Coreference**
Coreference and zeros from scratch	N	N
Coreference from scratch	Y	N
Refine the baseline	Y	Y

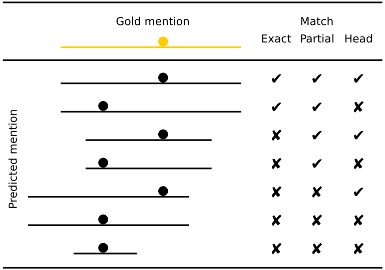
Evaluation Metrics

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching



Exact	Partial	Head
✓	✓	✓
✓	✓	✗
✗	✓	✓
✗	✓	✗
✗	✗	✓
✗	✗	✗
✗	✗	✗

- PM head is GM head (spans to disambiguate if multiple heads are matching)
- mention heads in CorefUD defined syntactically (Udapi block `corefud.MoveHead`)

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

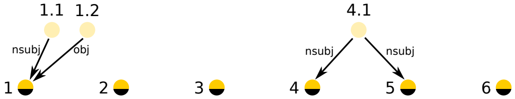
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough

1 ◒ 2 ◒ 3 ◒ 4 ◒ 5 ◒ 6 ◒

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

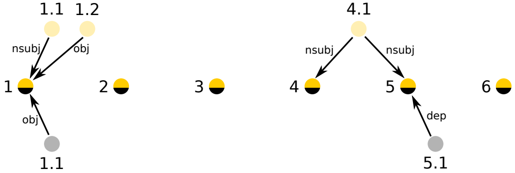
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

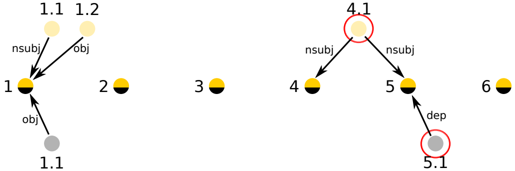
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

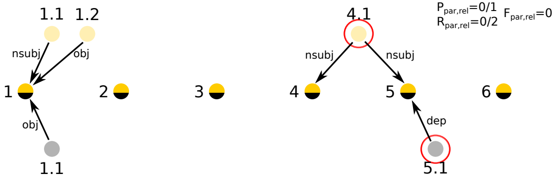
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

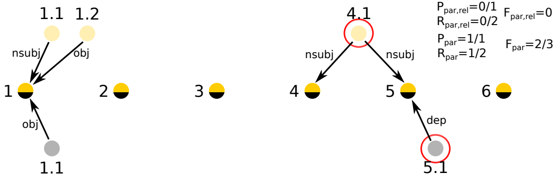
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

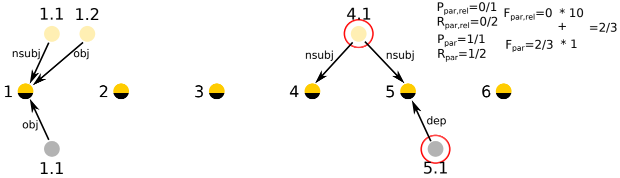
- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



$P_{par,rel}=0/1$ $R_{par,rel}=0/2$ $F_{par,rel}=0 * 10$

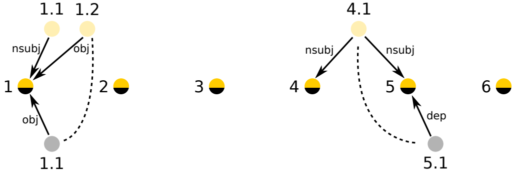
$P_{par}=1/1$ $R_{par}=1/2$ $F_{par}=2/3 * 1$

$+ \quad =2/3$

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

- predicted empty nodes not guaranteed to align 1:1 with the gold empty nodes
- searching for maximum 1:1 matching in a weighted bipartite graph of the empty nodes from the same sentence
- edge score: weighted sum of the F1 of predicting dependencies of zeros in the enhanced dependency graph
- priority to the accurate assignment of both parents and dep. types, but parents are enough



Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

- motivation: singletons not annotated in the majority of CorefUD datasets
- entities with a single mention deleted from both the GM and the PM

Primary Score

- CoNLL F1 score
- singletons excluded
- head match
- dependency-based zero matching

- unweighted average of the following F1 scores:
  - MUC (Vilain et al., 1995)
  - B^3 (Bagga and Baldwin, 1998)
  - CEAF-e (Luo, 2005)
- macro-averaged over all datasets

Supplementary Scores

- MUC, B^3, CEAF-e

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
- CoNLL F1 with exact or partial match

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
- CoNLL F1 with exact or partial match
- CoNLL F1 with singletons

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
- CoNLL F1 with exact or partial match
- CoNLL F1 with singletons
- Mention Overlap Ratio (MOR)
    - measures overlap of GMs and PMs, no matter to which entity they belong
    - Recall / Precision / F1

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
- CoNLL F1 with exact or partial match
- CoNLL F1 with singletons
- Mention Overlap Ratio (MOR)
  - measures overlap of GMs and PMs, no matter to which entity they belong
  - Recall / Precision / F1
- Anaphor-decomposable score for zeros
  - success rate of finding a correct antecedent for specified anaphor types
  - an application of the schema proposed by Tuggener (2014)
  - easy to interpret

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- builds on UA scorer 2.0 (Yu et al., 2023)

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- builds on UA scorer 2.0 (Yu et al., 2023)
- reuses its implementations of standard coreference measures

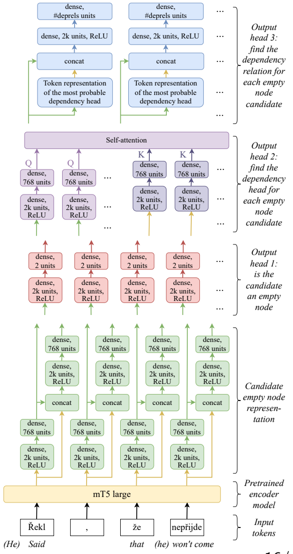
Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- builds on UA scorer 2.0 (Yu et al., 2023)
- reuses its implementations of standard coreference measures
- adds the following features:
  - head match
  - dependency-based matching of zero mentions

Participating Systems

Baselines

- **Empty nodes prediction**
  - 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
    - dependecy relation
  - trained on a combination of all CorefUD datasets with zeros
  - macro-avg F1 = 82.9
- **Coreference resolution**
  - same each year
  - based on the system by (Pražák et al., 2021), originally proposed by (Lee et al., 2017)
  - built on multi-lingual BERT
  - same system for all languages


dense, #deprels units
dense, 2k units, ReLU
concat
Token representation of the most probable dependency head
Output head 3: find the dependency relation for each empty node candidate
Self-attention
Q
K
dense, 768 units
dense, 2k units, ReLU
Output head 2: find the dependency head for each empty node candidate
dense, 2 units
dense, 2k units, ReLU
Output head 1: is the candidate an empty node
dense, 768 units
dense, 2k units, ReLU
concat
Candidate empty node represen-tation
mT5 large
Pretrained encoder model
Řekl
,
že
nepřijde
Input tokens
(He) Said
that
(he) won't come
...

Submissions

- 6 submissions by 4 teams

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
Baseline
Baseline-GZ
RitwikmishraFix

Submissions

- 6 submissions by 4 teams
  - all but one described in separate papers

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
Baseline
Baseline-GZ
RitwikmishraFix

Submissions

- 6 submissions by 4 teams
  - all but one described in separate papers
- submissions provided by the organizers

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
BASELINE
BASELINE-GZ
RitwikmishraFix

Submissions

- 6 submissions by 4 teams
  - all but one described in separate papers
- submissions provided by the organizers
  - automatic correction of non-valid files in Ritwikmishra submission

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
BASELINE
BASELINE-GZ
RitwikmishraFix

Submissions

- 6 submissions by 4 teams
  - all but one described in separate papers
- submissions provided by the organizers
  - automatic correction of non-valid files in Ritwikmishra submission
  - combination of both baseline systems

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
BASELINE
BASELINE-GZ
RitwikmishraFix

Submissions

- 6 submissions by 4 teams
  - all but one described in separate papers
- submissions provided by the organizers
  - automatic correction of non-valid files in Ritwikmishra submission
  - combination of both baseline systems
  - coreference resolution baseline applied on gold empty nodes

Submission
DFKI-CorefGen
CorPipe
CorPipe-single
CorPipe-2stage
Ondfa
Ritwikmishra
Baseline
Baseline-GZ
RitwikmishraFix

System Comparison: Basic Properties

Name	Starting point	Baseline	Official data	Pretrained model	Model size	Tuned per lang.
DFKI-CorefGen	From scratch	No	Yes	mT5-base	0.6B	No
CorPipe	From scratch	No	Yes	mT5-large, -xl, InfoXLM-large	3.7B	Yes
CorPipe-single	From scratch	No	Yes	mT5-large	0.5B	No
CorPipe-2stage	Coref from scratch	Empty node	Yes	mT5-large, -xl, InfoXLM-large	5.1B	Yes
Ondfa	Coref from scratch	Coref	Yes	mT5-xxl, XLM-R-large	6.3B	Yes
Ritwikmishra	Coref from scratch	No	No	XLM-R-base	0.3B	No

- either completely from scratch or use the empty nodes predictions

System Comparison: Basic Properties

Name	Starting point	Baseline	Official data	Pretrained model	Model size	Tuned per lang.
DFKI-CorefGen	From scratch	No	Yes	mT5-base	0.6B	No
CorPipe	From scratch	No	Yes	mT5-large, -xl, InfoXLM-large	3.7B	Yes
CorPipe-single	From scratch	No	Yes	mT5-large	0.5B	No
CorPipe-2stage	Coref from scratch	Empty node	Yes	mT5-large, -xl, InfoXLM-large	5.1B	Yes
Ondfa	Coref from scratch	Coref	Yes	mT5-xxl, XLM-R-large	6.3B	Yes
Ritwikmishra	Coref from scratch	No	No	XLM-R-base	0.3B	No

- either completely from scratch or use the empty nodes predictions
- one system does not even use the provided gold data

System Comparison: Basic Properties

Name	Starting point	Baseline	Official data	Pretrained model	Model size	Tuned per lang.
DFKI-CorefGen	From scratch	No	Yes	mT5-base	0.6B	No
CorPipe	From scratch	No	Yes	mT5-large, -xl, InfoXLM-large	3.7B	Yes
CorPipe-single	From scratch	No	Yes	mT5-large	0.5B	No
CorPipe-2stage	Coref from scratch	Empty node	Yes	mT5-large, -xl, InfoXLM-large	5.1B	Yes
Ondfa	Coref from scratch	Coref	Yes	mT5-xxl, XLM-R-large	6.3B	Yes
Ritwikmishra	Coref from scratch	No	No	XLM-R-base	0.3B	No

- either completely from scratch or use the empty nodes predictions
- one system does not even use the provided gold data
- increased interest in using mT5 as a base model

Results and Comparison

The Winner

CorPipe-2stage

Same team three times in a row. Congratulations!

Main Results: Primary Score

system	CoNLL F1
CorPipe-2stage	73.90
CorPipe	72.75
CorPipe-single	70.18
Ondfa	69.97
Baseline-GZ	54.60
Baseline	53.16
DFKI-CorefGen	33.38
RitwikmishraFix	30.63
Ritwikmishra	16.47

Main Results: Primary Score

system	CoNLL F1
CorPipe-2stage	73.90
CorPipe	72.75
CorPipe-single	70.18
Ondfa	69.97
Baseline-GZ	54.60
Baseline	53.16
DFKI-CorefGen	33.38
RitwikmishraFix	30.63
Ritwikmishra	16.47

- comparison to the Baseline
    - 2024: +21 points (+39%)

Main Results: Primary Score

system	CoNLL F1
CorPipe-2stage	73.90
CorPipe	72.75
CorPipe-single	70.18
Ondfa	69.97
Baseline-GZ	54.60
Baseline	53.16
DFKI-CorefGen	33.38
RitwikmishraFix	30.63
Ritwikmishra	16.47

- comparison to the Baseline
  - 2024: +21 points (+39%)
  - 2023: +18 points (+31%)
  - 2022: +12 points (+20%)

Main Results: Supplementary Scores

system	primary	MUC	B^3	CEAF-e	BLANC	LEA
CorPipe-2stage	**73.90**	**79** / **81** / **80**	**69** / **74** / **71**	**71** / **70** / **70**	**67** / **73** / **70**	**66** / **71** / **68**
CorPipe	72.75	79 / 80 / 79	69 / 72 / 70	71 / 68 / 69	67 / 72 / 69	65 / 69 / 67
CorPipe-single	70.18	77 / 76 / 77	68 / 67 / 67	69 / 66 / 67	66 / 66 / 66	64 / 63 / 64
Ondfa	69.97	75 / 81 / 78	64 / 72 / 67	64 / 67 / 65	62 / 71 / 65	61 / 69 / 64
Baseline-GZ	54.60	56 / 75 / 63	43 / 63 / 50	46 / 57 / 50	41 / 63 / 48	39 / 58 / 46
Baseline	53.16	54 / 73 / 62	41 / 62 / 49	44 / 56 / 49	39 / 62 / 46	37 / 57 / 44
DFKI-CorefGen	33.38	37 / 52 / 41	26 / 38 / 29	25 / 42 / 30	21 / 39 / 23	21 / 31 / 23
RitwikmishraFix	30.63	33 / 50 / 36	26 / 43 / 28	27 / 37 / 29	24 / 39 / 24	24 / 39 / 25
Ritwikmishra	16.47	18 / 31 / 18	15 / 27 / 15	15 / 22 / 16	13 / 23 / 12	13 / 25 / 13

* Recall / Precision / F1

Main Results: Supplementary Scores

system	primary	MUC	B^3	CEAF-e	BLANC	LEA
CorPipe-2stage	**73.90**	**79** / **81** / **80**	**69** / **74** / **71**	**71** / **70** / **70**	**67** / **73** / **70**	**66** / **71** / **68**
CorPipe	72.75	79 / 80 / 79	69 / 72 / 70	71 / 68 / 69	67 / 72 / 69	65 / 69 / 67
CorPipe-single	70.18	77 / 76 / 77	68 / 67 / 67	69 / 66 / 67	66 / 66 / 66	64 / 63 / 64
Ondfa	69.97	75 / 81 / 78	64 / 72 / 67	64 / 67 / 65	62 / 71 / 65	61 / 69 / 64
Baseline-GZ	54.60	56 / 75 / 63	43 / 63 / 50	46 / 57 / 50	41 / 63 / 48	39 / 58 / 46
Baseline	53.16	54 / 73 / 62	41 / 62 / 49	44 / 56 / 49	39 / 62 / 46	37 / 57 / 44
DFKI-CorefGen	33.38	37 / 52 / 41	26 / 38 / 29	25 / 42 / 30	21 / 39 / 23	21 / 31 / 23
RitwikmishraFix	30.63	33 / 50 / 36	26 / 43 / 28	27 / 37 / 29	24 / 39 / 24	24 / 39 / 25
Ritwikmishra	16.47	18 / 31 / 18	15 / 27 / 15	15 / 22 / 16	13 / 23 / 12	13 / 25 / 13

* Recall / Precision / F1

- *CorPipe-2stage* consistently best in all coreference scores

Primary Score Across Datasets

system	primary	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_parcorfull	de_potsdam	en_gum	en_litbank	en_parcorfull	es_ancora	fr_democrat	grc_proiel	hbo_ptnk	hu_korkor	hu_szeged	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipe-2stage	**73.90**	82.22	**74.85**	**77.18**	**61.58**	69.53	71.79	**75.66**	**79.60**	68.89	**82.46**	68.16	**71.34**	**72.02**	63.17	**69.97**	**75.79**	**79.81**	**78.01**	**78.50**	**83.22**	**68.18**
CorPipe	72.75	81.02	73.71	75.84	60.72	**71.68**	71.45	74.61	79.10	**69.75**	80.98	**68.77**	68.53	70.86	60.32	68.12	75.78	79.55	77.52	77.03	83.09	59.37
CorPipe-single	70.18	80.42	72.82	74.82	57.11	61.62	67.02	74.39	78.08	58.61	79.75	67.89	66.01	67.18	60.09	67.32	75.19	78.92	76.60	75.20	81.21	53.43
Ondfa	69.97	**82.46**	70.82	75.80	54.97	71.40	**71.91**	70.53	74.15	55.58	81.94	62.69	61.64	61.56	**64.86**	69.26	71.97	74.51	72.07	76.34	80.47	64.49
Baseline-GZ	54.60	69.59	68.93	66.15	27.56	47.21	55.65	63.18	63.54	33.08	70.64	53.62	31.87	24.60	41.65	54.64	62.00	64.96	63.70	67.00	65.83	51.16
Baseline	53.16	68.32	64.06	63.83	24.51	47.21	55.65	63.19	63.54	33.08	69.58	53.62	28.76	24.60	35.14	54.51	62.00	64.96	63.70	66.24	65.83	44.05
DFKI-CorefGen	33.38	34.77	32.89	30.88	22.52	23.07	45.85	35.49	46.59	32.69	37.76	36.34	25.87	37.96	23.53	33.85	42.73	37.92	35.69	27.19	47.79	9.65
RitwikmishraFix	30.63	27.05	0.00	0.00	6.79	25.35	48.90	48.64	61.47	53.12	30.04	43.63	5.60	0.12	33.40	30.28	44.31	56.41	53.17	0.00	53.89	20.97
Ritwikmishra	16.47	0.00	0.00	0.00	6.79	25.35	48.90	0.00	0.00	53.12	0.00	43.72	5.60	0.09	33.40	30.32	44.78	0.00	0.00	0.00	53.88	0.00

Primary Score Across Datasets

system	primary	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_parcorfull	de_potsdam	en_gum	en_litbank	en_parcorfull	es_ancora	fr_democrat	grc_proiel	hbo_ptnk	hu_korkor	hu_szeged	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipe-2stage	**73.90**	82.22	**74.85**	**77.18**	**61.58**	69.53	71.79	**75.66**	**79.60**	68.89	**82.46**	68.16	**71.34**	**72.02**	63.17	**69.97**	**75.79**	**79.81**	**78.01**	**78.50**	**83.22**	**68.18**
CorPipe	72.75	81.02	73.71	75.84	60.72	**71.68**	71.45	74.61	79.10	**69.75**	80.98	**68.77**	68.53	70.86	60.32	68.12	75.78	79.55	77.52	77.03	83.09	59.37
CorPipe-single	70.18	80.42	72.82	74.82	57.11	61.62	67.02	74.39	78.08	58.61	79.75	67.89	66.01	67.18	60.09	67.32	75.19	78.92	76.60	75.20	81.21	53.43
Ondfa	69.97	**82.46**	70.82	75.80	54.97	71.40	**71.91**	70.53	74.15	55.58	81.94	62.69	61.64	61.56	**64.86**	69.26	71.97	74.51	72.07	76.34	80.47	64.49
Baseline-GZ	54.60	69.59	68.93	66.15	27.56	47.21	55.65	63.18	63.54	33.08	70.64	53.62	31.87	24.60	41.65	54.64	62.00	64.96	63.70	67.00	65.83	51.16
Baseline	53.16	68.32	64.06	63.83	24.51	47.21	55.65	63.19	63.54	33.08	69.58	53.62	28.76	24.60	35.14	54.51	62.00	64.96	63.70	66.24	65.83	44.05
DFKI-CorefGen	33.38	34.77	32.89	30.88	22.52	23.07	45.85	35.49	46.59	32.69	37.76	36.34	25.87	37.96	23.53	33.85	42.73	37.92	35.69	27.19	47.79	9.65
RitwikmishraFix	30.63	27.05	0.00	0.00	6.79	25.35	48.90	48.64	61.47	53.12	30.04	43.63	5.60	0.12	33.40	30.28	44.31	56.41	53.17	0.00	53.89	20.97
Ritwikmishra	16.47	0.00	0.00	0.00	6.79	25.35	48.90	0.00	0.00	53.12	0.00	43.72	5.60	0.09	33.40	30.32	44.78	0.00	0.00	0.00	53.88	0.00

- results more diverse than last year

Primary Score Across Datasets

system	primary	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_parcorfull	de_potsdam	en_gum	en_litbank	en_parcorfull	es_ancora	fr_democrat	grc_proiel	hbo_ptnk	hu_korkor	hu_szeged	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipe-2stage	**73.90**	82.22	**74.85**	**77.18**	**61.58**	69.53	71.79	**75.66**	**79.60**	68.89	**82.46**	68.16	**71.34**	**72.02**	63.17	**69.97**	**75.79**	**79.81**	**78.01**	**78.50**	**83.22**	**68.18**
CorPipe	72.75	81.02	73.71	75.84	60.72	**71.68**	71.45	74.61	79.10	**69.75**	80.98	**68.77**	68.53	70.86	60.32	68.12	75.78	79.55	77.52	77.03	83.09	59.37
CorPipe-single	70.18	80.42	72.82	74.82	57.11	61.62	67.02	74.39	78.08	58.61	79.75	67.89	66.01	67.18	60.09	67.32	75.19	78.92	76.60	75.20	81.21	53.43
Ondfa	69.97	**82.46**	70.82	75.80	54.97	71.40	**71.91**	70.53	74.15	55.58	81.94	62.69	61.64	61.56	**64.86**	69.26	71.97	74.51	72.07	76.34	80.47	64.49
Baseline-GZ	54.60	69.59	68.93	66.15	27.56	47.21	55.65	63.18	63.54	33.08	70.64	53.62	31.87	24.60	41.65	54.64	62.00	64.96	63.70	67.00	65.83	51.16
Baseline	53.16	68.32	64.06	63.83	24.51	47.21	55.65	63.19	63.54	33.08	69.58	53.62	28.76	24.60	35.14	54.51	62.00	64.96	63.70	66.24	65.83	44.05
DFKI-CorefGen	33.38	34.77	32.89	30.88	22.52	23.07	45.85	35.49	46.59	32.69	37.76	36.34	25.87	37.96	23.53	33.85	42.73	37.92	35.69	27.19	47.79	9.65
RitwikmishraFix	30.63	27.05	0.00	0.00	6.79	25.35	48.90	48.64	61.47	53.12	30.04	43.63	5.60	0.12	33.40	30.28	44.31	56.41	53.17	0.00	53.89	20.97
Ritwikmishra	16.47	0.00	0.00	0.00	6.79	25.35	48.90	0.00	0.00	53.12	0.00	43.72	5.60	0.09	33.40	30.32	44.78	0.00	0.00	0.00	53.88	0.00

- results more diverse than last year
- tr_itcc fixed and newly with zeros
  - Baseline-2023: 22.75
  - Baseline-GZ: 51.16

Primary Score Across Datasets

system	primary	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_parcorfull	de_potsdam	en_gum	en_litbank	en_parcorfull	es_ancora	fr_democrat	grc_proiel	hbo_ptnk	hu_korkor	hu_szeged	lt_lcc	no_bokmaalnarc	no_nynorsknarc	pl_pcc	ru_rucor	tr_itcc
CorPipe-2stage	**73.90**	82.22	**74.85**	**77.18**	**61.58**	69.53	71.79	**75.66**	**79.60**	68.89	**82.46**	68.16	**71.34**	**72.02**	63.17	**69.97**	**75.79**	**79.81**	**78.01**	**78.50**	**83.22**	**68.18**
CorPipe	72.75	81.02	73.71	75.84	60.72	**71.68**	71.45	74.61	79.10	**69.75**	80.98	**68.77**	68.53	70.86	60.32	68.12	75.78	79.55	77.52	77.03	83.09	59.37
CorPipe-single	70.18	80.42	72.82	74.82	57.11	61.62	67.02	74.39	78.08	58.61	79.75	67.89	66.01	67.18	60.09	67.32	75.19	78.92	76.60	75.20	81.21	53.43
Ondfa	69.97	**82.46**	70.82	75.80	54.97	71.40	**71.91**	70.53	74.15	55.58	81.94	62.69	61.64	61.56	**64.86**	69.26	71.97	74.51	72.07	76.34	80.47	64.49
Baseline-GZ	54.60	69.59	68.93	66.15	27.56	47.21	55.65	63.18	63.54	33.08	70.64	53.62	31.87	24.60	41.65	54.64	62.00	64.96	63.70	67.00	65.83	51.16
Baseline	53.16	68.32	64.06	63.83	24.51	47.21	55.65	63.19	63.54	33.08	69.58	53.62	28.76	24.60	35.14	54.51	62.00	64.96	63.70	66.24	65.83	44.05
DFKI-CorefGen	33.38	34.77	32.89	30.88	22.52	23.07	45.85	35.49	46.59	32.69	37.76	36.34	25.87	37.96	23.53	33.85	42.73	37.92	35.69	27.19	47.79	9.65
RitwikmishraFix	30.63	27.05	0.00	0.00	6.79	25.35	48.90	48.64	61.47	53.12	30.04	43.63	5.60	0.12	33.40	30.28	44.31	56.41	53.17	0.00	53.89	20.97
Ritwikmishra	16.47	0.00	0.00	0.00	6.79	25.35	48.90	0.00	0.00	53.12	0.00	43.72	5.60	0.09	33.40	30.32	44.78	0.00	0.00	0.00	53.88	0.00

- results more diverse than last year
- tr_itcc fixed and newly with zeros
  - Baseline-2023: 22.75
  - Baseline-GZ: 51.16

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipe-2stage	86	80	66	**76**	**91**	**86**	64	75	87	82
CorPipe	81	74	62	75	84	81	63	70	82	69
CorPipe-single	79	72	60	73	83	78	60	68	79	63
Ondfa	**87**	79	66	72	90	81	**66**	**77**	86	**82**
Baseline-GZ	82	**83**	**80**	66	87	65	62	56	**87**	78
Baseline	77	72	61	56	83	66	49	53	82	70
DFKI-CorefGen	0	0	0	0	0	0	0	0	0	0
RitwikmishraFix	0	0	0	0	0	0	0	0	0	0
Ritwikmishra	0	0	0	0	0	0	0	0	0	0
Baseline-2023	82	82	79	–	87	–	64	59	62	–

* Recall / Precision / F1

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipe-2stage	86	80	66	**76**	**91**	**86**	64	75	87	82
CorPipe	81	74	62	75	84	81	63	70	82	69
CorPipe-single	79	72	60	73	83	78	60	68	79	63
Ondfa	**87**	79	66	72	90	81	**66**	**77**	86	**82**
Baseline-GZ	82	**83**	**80**	66	87	65	62	56	**87**	78
Baseline	77	72	61	56	83	66	49	53	82	70
DFKI-CorefGen	0	0	0	0	0	0	0	0	0	0
RitwikmishraFix	0	0	0	0	0	0	0	0	0	0
Ritwikmishra	0	0	0	0	0	0	0	0	0	0
Baseline-2023	82	82	79	–	87	–	64	59	62	–

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- best-performing systems aligned with overall scores across datasets

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipe-2stage	86	80	66	**76**	**91**	**86**	64	75	87	82
CorPipe	81	74	62	75	84	81	63	70	82	69
CorPipe-single	79	72	60	73	83	78	60	68	79	63
Ondfa	**87**	79	66	72	90	81	**66**	**77**	86	**82**
Baseline-GZ	82	**83**	**80**	66	87	65	62	56	**87**	78
Baseline	77	72	61	56	83	66	49	53	82	70
DFKI-CorefGen	0	0	0	0	0	0	0	0	0	0
RitwikmishraFix	0	0	0	0	0	0	0	0	0	0
Ritwikmishra	0	0	0	0	0	0	0	0	0	0
Baseline-2023	82	82	79	–	87	–	64	59	62	–

[*] Recall / Precision / F1

- anaphor-decomposable score on zeros
- best-performing systems aligned with overall scores across datasets
- predicting empty nodes, the task has become more challenging

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipe-2stage	86	80	66	**76**	**91**	**86**	64	75	87	82
CorPipe	81	74	62	75	84	81	63	70	82	69
CorPipe-single	79	72	60	73	83	78	60	68	79	63
Ondfa	**87**	79	66	72	90	81	**66**	**77**	86	**82**
Baseline-GZ	82	**83**	**80**	66	87	65	62	56	**87**	78
Baseline	77	72	61	56	83	66	49	53	82	70
DFKI-CorefGen	0	0	0	0	0	0	0	0	0	0
RitwikmishraFix	0	0	0	0	0	0	0	0	0	0
Ritwikmishra	0	0	0	0	0	0	0	0	0	0
Baseline-2023	82	82	79	–	87	–	64	59	62	–

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- best-performing systems aligned with overall scores across datasets
- predicting empty nodes, the task has become more challenging
- baseline performance jump on pl_pcc due to fixes in the conversion pipeline

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	cu_proiel	es_ancora	grc_proiel	hu_korkor	hu_szeged	pl_pcc	tr_itcc
CorPipe-2stage	86	80	66	**76**	**91**	**86**	64	75	87	82
CorPipe	81	74	62	75	84	81	63	70	82	69
CorPipe-single	79	72	60	73	83	78	60	68	79	63
Ondfa	**87**	79	66	72	90	81	**66**	**77**	86	**82**
Baseline-GZ	82	**83**	**80**	66	87	65	62	56	**87**	78
Baseline	77	72	61	56	83	66	49	53	82	70
DFKI-CorefGen	0	0	0	0	0	0	0	0	0	0
RitwikmishraFix	0	0	0	0	0	0	0	0	0	0
Ritwikmishra	0	0	0	0	0	0	0	0	0	0
Baseline-2023	82	82	79	–	87	–	64	59	62	–

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- best-performing systems aligned with overall scores across datasets
- predicting empty nodes, the task has become more challenging
- baseline performance jump on pl_pcc due to fixes in the conversion pipeline

Other Statistics

- see the paper

Conclusion

Summary

- summary of CRAC 2024 Multilingual Coreference Resolution Shared Task

Web

`https://ufal.mff.cuni.cz/corefud/crac24`

Summary

- summary of CRAC 2024 Multilingual Coreference Resolution Shared Task
- moving towards even more realistic setup
  - no pre-defined slots for zeros
  - more diverse languages

Web

`https://ufal.mff.cuni.cz/corefud/crac24`

Summary

- summary of CRAC 2024 Multilingual Coreference Resolution Shared Task
- moving towards even more realistic setup
  - no pre-defined slots for zeros
  - more diverse languages
- growing quality of the submissions

Web

`https://ufal.mff.cuni.cz/corefud/crac24`

Summary

- summary of CRAC 2024 Multilingual Coreference Resolution Shared Task
- moving towards even more realistic setup
    - no pre-defined slots for zeros
    - more diverse languages
- growing quality of the submissions
- we wish for more participants

Web

`https://ufal.mff.cuni.cz/corefud/crac24`

Future Editions

- we are organizing the shared task in 2025 again

Future Editions

- we are organizing the shared task in 2025 again
- planned extensions:
  - additional datasets (Japanese?)
  - push the shared task to the LLM era