CRAC 2023, Singapore

Findings of the Second Shared Task on Multilingual Coreference Resolution

Zdeněk Žabokrtský[1], Miloslav Konopík[2], Anna Nedoluzhko[1], **Michal Novák**[1], Maciej Ogrodniczuk[3], Martin Popel[1], Ondřej Pražák[2], Jakub Sido[2], Daniel Zeman[1]

December 7, 2023

[1] Charles University, Prague, Czechia
[2] University of West Bohemia, Pilsen, Czechia
[3] Polish Academy of Sciences, Warsaw, Poland

Outline

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*
CRAC 2023	12	included*

* already generated in the input

Shared Task

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

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.1
  - training (gold), dev (gold, no annot), test (no annot)

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.1
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.1
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline system:
  - based on (Pražák et al., 2021)
  - system and its predictions on dev and test sets

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.1
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline system:
  - based on (Pražák et al., 2021)
  - system and its predictions on dev and test sets
- Environment:
  - powered by CodaLab (`https://codalab.lisn.upsaclay.fr/competitions/11800`)
  - automatic validation, evaluation and ranking of the submissions

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.1
  - training (gold), dev (gold, no annot), test (no annot)
- Scorer:
  - CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- Baseline system:
  - based on (Pražák et al., 2021)
  - system and its predictions on dev and test sets
- Environment:
  - powered by CodaLab (`https://codalab.lisn.upsaclay.fr/competitions/11800`)
  - automatic validation, evaluation and ranking of the submissions
- **`https://ufal.mff.cuni.cz/corefud/crac23`**

Changes to the 2022 edition

1. using a newer version of the collection: CorefUD 1.1

Changes to the 2022 edition

1. using a newer version of the collection: CorefUD 1.1
2. automatic morpho-syntactic features in test files: more realistic evaluation scheme

Changes to the 2022 edition

1. using a newer version of the collection: CorefUD 1.1
2. automatic morpho-syntactic features in test files: more realistic evaluation scheme
3. head matching of mentions

Datasets

CorefUD 1.1

- public edition of CorefUD 1.1 (Nedoluzhko et al., 2022b)
- 17 coreference datasets for 12 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 1.1 datasets

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

CorefUD 1.1 datasets

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

Data Statistics

CorefUD dataset	docs	sents	words	zeros	entities	avg. len.	non-singletons
Catalan-AnCora	1298	13,613	429,313	6,377	18,030	3.5	62,417
Czech-PCEDT	2312	49,208	1,155,755	35,844	52,721	3.3	168,138
Czech-PDT	3165	49,428	834,720	22,389	78,747	2.4	154,983
English-GUM	195	10,761	187,416	99	27,757	1.9	32,323
English-ParCorFull	19	543	10,798	0	202	4.2	835
French-Democrat	126	13,057	284,883	0	39,023	2.0	46,487
German-ParCorFull	19	543	10,602	0	259	3.5	896
German-PotsdamCC	176	2,238	33,222	0	3,752	1.4	2,519
Hungarian-KorKor	94	1,351	24,568	1,988	1,134	3.6	4,103
Hungarian-SzegedKoref	400	8,820	123,968	4,857	5,182	3.0	15,165
Lithuanian-LCC	100	1,714	37,014	0	1,224	3.7	4,337
Norwegian-BokmaalNARC	346	15,742	245,515	0	53,357	1.4	26,611
Norwegian-NynorskNARC	394	12,481	206,660	0	44,847	1.4	21,847
Polish-PCC	1828	35,874	538,885	470	127,688	1.5	82,804
Russian-RuCor	181	9,035	156,636	0	3,636	4.5	16,193
Spanish-AnCora	1356	14,159	458,418	8,112	20,115	3.5	70,663
Turkish-ITCC	24	4,733	55,341	0	690	5.3	3,668

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)
```

Annotation Details: Morpho-Syntax

- CorefUD also comprises UD-like annotation of parts of speech, morphological features, and dependency syntax
  - manual annotation in original data kept also in CorefUD
  - otherwise parsed using UDPipe 2.0 (Straka, 2018)
- shared task data
  - train: as in CorefUD
  - dev, test: replaced with outputs of UDPipe for all datasets

Annotation Details: Morpho-Syntax

- CorefUD also comprises UD-like annotation of parts of speech, morphological features, and dependency syntax
  - manual annotation in original data kept also in CorefUD
  - otherwise parsed using UDPipe 2.0 (Straka, 2018)
- shared task data
  - train: as in CorefUD
  - dev, test: replaced with outputs of UDPipe for all datasets

Annotation Details: Zeros

- zeros are integral part of some of the datasets
- annotated using empty nodes from enhanced UD
- we keep the empty nodes in the test data
  - slightly unrealistic setup
  - presence of an empty node may indicate its anaphoricity
  - yet simpler and more accessible to participants

Dataset	Zeros
ca_ancora	6,377
cs_pcedt	35,844
cs_pdt	22,389
en_gum	99
hu_korkor	1,988
hu_szeged	4,857
pl_pcc	470
es_ancora	8,112

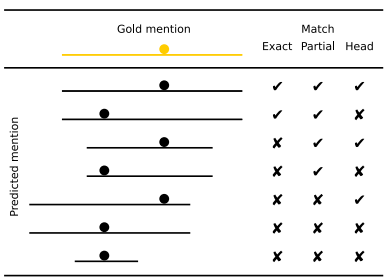
Evaluation Metrics

Primary Score

- CoNLL F1 score
- singletons excluded
- head match

Primary Score

- CoNLL F1 score
- singletons excluded
- head match



- **Exact**: the PM consists of the same words as the GM
- **Partial**: all PM words are included in the GM and GM head is one of PM words
- **Head**: PM head is GM head (spans to disambiguate if multiple heads are matching)

Primary Score

- CoNLL F1 score
- singletons excluded
- head match

- why not exact match?
  - some datasets (e.g. cs_pdt) do not specify mention spans, only heads
  - in general, mention boundaries may be difficult to specify
- why not partial match (used in 2022)?
  - it encouraged some participants to reduce their mentions to head words only
  - unfair comparison as not all participants did it
  - skewed exact match evaluation
- mention heads in CorefUD defined syntactically
  - Udapi block `corefud.MoveHead`
  - even for datasets with no heads originally annotated (e.g. de_potsdam)
  - participants could use the same

Primary Score

- CoNLL F1 score
- singletons excluded
- head match

- 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

- 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`)
- based on UA scorer 1.0 (Yu et al., 2022)

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- based on UA scorer 1.0 (Yu et al., 2022)
- reuses its implementations of standard coreference measures

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- based on UA scorer 1.0 (Yu et al., 2022)
- reuses its implementations of standard coreference measures
- adds the following features:
  - processing of CorefUD format
  - head match
  - handling of discontinuous mentions
  - allows for scoring zeros (they have to be already generated)
  - new scores: MOR and anaphor-decomposable score for zeros

Official scorer

- CorefUD scorer (`https://github.com/ufal/corefud-scorer`)
- based on UA scorer 1.0 (Yu et al., 2022)
- reuses its implementations of standard coreference measures
- adds the following features:
  - processing of CorefUD format
  - head match
  - handling of discontinuous mentions
  - allows for scoring zeros (they have to be already generated)
  - new scores: MOR and anaphor-decomposable score for zeros
- in the meantime, most of the new features integrated to UA scorer 2.0 (Yu et al., 2023)

Participating Systems

Baseline

- same as last year
- based on the coreference system by (Pražák et al., 2021)
- built on multi-lingual BERT
- going through all potential spans and maximizing gold antecedents
- same system for all languages

Submissions

- 8 submissions + baseline

Submission
Anonymous
BASELINE
CorPipe
DFKI-Adapt
DFKI-MPrompt
DeepBlueAI
McGill
Morfbase
Ondfa

Submissions

- 8 submissions + baseline
- 5 submissions described in separate papers

Submission
Anonymous
BASELINE
CorPipe
DFKI-Adapt
DFKI-MPrompt
DeepBlueAI
McGill
Morfbase
Ondfa

Submissions

- 8 submissions + baseline
- 5 submissions described in separate papers
- one team asked us to anonymize their submission

Submission
Anonymous
BASELINE
CorPipe
DFKI-Adapt
DFKI-MPrompt
DeepBlueAI
McGill
Morfbase
Ondfa

Submissions

- 8 submissions + baseline
- 5 submissions described in separate papers
- one team asked us to anonymize their submission
- one team have not provided any details about their system

Submission
Anonymous
BASELINE
CorPipe
DFKI-Adapt
DFKI-MPrompt
DeepBlueAI
McGill
Morfbase
Ondfa

System Comparison: Basic Properties

Name	Baseline	Pretrained model	Model size	Seq. length	Tuned per lang.	Batch size
Anonymous	No	xlm-roberta-base	1-20M (various)	512	Lang. families	16
BASELINE	Yes	bert-base	220M	512	No	1 doc
CorPipe	No	google/mt5-{large,xl}	567M, 1.7G	512, 2560	No	8, 12, 16, 32
DFKI-Adapt	Yes	bert-base	259M	512	Yes	1 doc
DFKI-MPrompt	Yes	bert-base + soft prompt	221M	512	No	1 sent + 1 doc
McGill	No	xlm-roberta-large	596M	512	No	1
Morfbase	Yes	bert-base	219M	512	No	256
Ondfa	Yes	xlm-roberta-large	600M	512	Yes	1 doc

System Comparison: Basic Properties

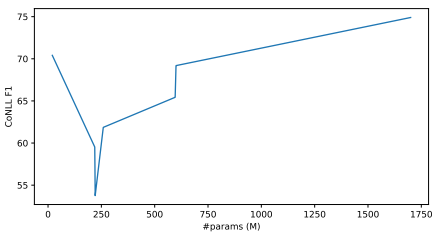
Name	Baseline	Pretrained model	Model size	Seq. length	Tuned per lang.	Batch size
Anonymous	No	xlm-roberta-base	1-20M (various)	512	Lang. families	16
Baseline	Yes	bert-base	220M	512	No	1 doc
CorPipe	No	google/mt5-{large,xl}	567M, 1.7G	512, 2560	No	8, 12, 16, 32
DFKI-Adapt	Yes	bert-base	259M	512	Yes	1 doc
DFKI-MPrompt	Yes	bert-base + soft prompt	221M	512	No	1 sent + 1 doc
McGill	No	xlm-roberta-large	596M	512	No	1
Morfbase	Yes	bert-base	219M	512	No	256
Ondfa	Yes	xlm-roberta-large	600M	512	Yes	1 doc

- more than a half built on the Baseline system

System Comparison: Basic Properties

Name	Baseline	Pretrained model	Model size	Seq. length	Tuned per lang.	Batch size
Anonymous	No	xlm-roberta-base	1-20M (various)	512	Lang. families	16
BASELINE	Yes	bert-base	220M	512	No	1 doc
CorPipe	No	google/mt5-{large,xl}	567M, 1.7G	512, 2560	No	8, 12, 16, 32
DFKI-Adapt	Yes	bert-base	259M	512	Yes	1 doc
DFKI-MPrompt	Yes	bert-base + soft prompt	221M	512	No	1 sent + 1 doc
McGill	No	xlm-roberta-large	596M	512	No	1
Morfbase	Yes	bert-base	219M	512	No	256
Ondfa	Yes	xlm-roberta-large	600M	512	Yes	1 doc

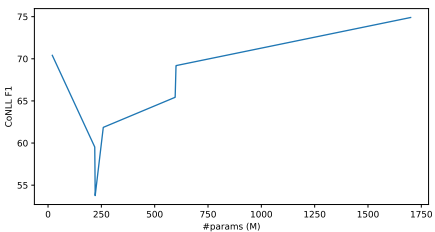
- more than a half built on the Baseline system
- scores improve with larger models (apart from some exceptions)



System Comparison: Basic Properties

Name	Baseline	Pretrained model	Model size	Seq. length	Tuned per lang.	Batch size
Anonymous	No	xlm-roberta-base	1-20M (various)	512	Lang. families	16
BASELINE	Yes	bert-base	220M	512	No	1 doc
CorPipe	No	google/mt5-{large,xl}	567M, 1.7G	512, 2560	No	8, 12, 16, 32
DFKI-Adapt	Yes	bert-base	259M	512	Yes	1 doc
DFKI-MPrompt	Yes	bert-base + soft prompt	221M	512	No	1 sent + 1 doc
McGill	No	xlm-roberta-large	596M	512	No	1
Morfbase	Yes	bert-base	219M	512	No	256
Ondfa	Yes	xlm-roberta-large	600M	512	Yes	1 doc

- more than a half built on the Baseline system
- scores improve with larger models (apart from some exceptions)



Results and Comparison

The Winner

CorPipe

Two years in a row. Congratulations!

Main Results: Primary Score

system	CoNLL F1
CorPipe	74.90
Anonymous	70.41
Ondfa	69.19
McGill	65.43
DeepBlueAI	62.29
DFKI-Adapt	61.86
Morfbase	59.53
BASELINE	56.96
DFKI-MPrompt	53.76

Main Results: Primary Score

system	CoNLL F1
CorPipe	74.90
Anonymous	70.41
Ondfa	69.19
McGill	65.43
DeepBlueAI	62.29
DFKI-Adapt	61.86
Morfbase	59.53
BASELINE	56.96
DFKI-MPrompt	53.76

- improvement of 18 points (31%) over the baseline
- 2022: improvement of 12 points (20%)

Main Results: Supplementary Scores

system	primary	MUC	B^3	CEAF-e	BLANC	LEA
CorPipe	**74.90**	**80** / **79** / **80**	**73** / **73** / **73**	**73** / **71** / **72**	**72** / **73** / **72**	**70** / **71** / **70**
Anonymous	70.41	74 / 78 / 76	65 / 72 / 68	67 / 68 / 67	63 / 71 / 66	62 / 69 / 65
Ondfa	69.19	74 / 78 / 75	64 / 71 / 67	64 / 67 / 66	62 / 70 / 65	61 / 68 / 64
McGill	65.43	69 / 76 / 71	60 / 69 / 63	58 / 68 / 62	58 / 68 / 61	57 / 66 / 60
DeepBlueAI	62.29	67 / 74 / 70	56 / 65 / 59	55 / 63 / 58	53 / 64 / 56	53 / 61 / 56
DFKI-Adapt	61.86	66 / 73 / 69	56 / 65 / 59	56 / 62 / 58	53 / 63 / 56	52 / 61 / 55
Morfbase	59.53	63 / 71 / 66	51 / 65 / 56	56 / 58 / 56	47 / 62 / 52	47 / 61 / 52
BASELINE	56.96	56 / 76 / 63	46 / 69 / 54	48 / 62 / 54	44 / 67 / 51	42 / 64 / 49
DFKI-MPrompt	53.76	57 / 67 / 61	45 / 60 / 50	49 / 56 / 51	41 / 57 / 45	40 / 55 / 45

* Recall / Precision / F1

Main Results: Supplementary Scores

system	primary	MUC	B^3	CEAF-e	BLANC	LEA
CorPipe	**74.90**	**80** / **79** / **80**	**73** / **73** / **73**	**73** / **71** / **72**	**72** / **73** / **72**	**70** / **71** / **70**
Anonymous	70.41	74 / 78 / 76	65 / 72 / 68	67 / 68 / 67	63 / 71 / 66	62 / 69 / 65
Ondfa	69.19	74 / 78 / 75	64 / 71 / 67	64 / 67 / 66	62 / 70 / 65	61 / 68 / 64
McGill	65.43	69 / 76 / 71	60 / 69 / 63	58 / 68 / 62	58 / 68 / 61	57 / 66 / 60
DeepBlueAI	62.29	67 / 74 / 70	56 / 65 / 59	55 / 63 / 58	53 / 64 / 56	53 / 61 / 56
DFKI-Adapt	61.86	66 / 73 / 69	56 / 65 / 59	56 / 62 / 58	53 / 63 / 56	52 / 61 / 55
Morfbase	59.53	63 / 71 / 66	51 / 65 / 56	56 / 58 / 56	47 / 62 / 52	47 / 61 / 52
BASELINE	56.96	56 / 76 / 63	46 / 69 / 54	48 / 62 / 54	44 / 67 / 51	42 / 64 / 49
DFKI-MPrompt	53.76	57 / 67 / 61	45 / 60 / 50	49 / 56 / 51	41 / 57 / 45	40 / 55 / 45

* Recall / Precision / F1

- *CorPipe* consistently best in all coreference scores

Primary Score Across Datasets

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

Primary Score Across Datasets

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

- ÚFAL CorPipe team dominant on all datasets

Primary Score Across Datasets

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

- ÚFAL CorPipe team dominant on all datasets
- low results on tr_itcc due to unfinished annotation

Primary Score Across Datasets

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

- ÚFAL CorPipe team dominant on all datasets
- low results on tr_itcc due to unfinished annotation
- de_parcorfull and en_parcorfull prone to inconsistencies likely due to their size

Singletons

system	primary	with singletons	
CorPipe	**74.90**	**76.82**	(+1.91)
Anonymous	70.41	73.20	(+2.79)
Ondfa	69.19	68.37	(-0.82)
McGill	65.43	68.23	(+2.80)
DeepBlueAI	62.29	54.51	(-7.78)
DFKI-Adapt	61.86	53.94	(-7.92)
Morfbase	59.53	52.07	(-7.47)
BASELINE	56.96	49.32	(-7.64)
DFKI-MPrompt	53.76	46.83	(-6.93)

Singletons

system	primary	with singletons	
CorPipe	**74.90**	**76.82**	(+1.91)
Anonymous	70.41	73.20	(+2.79)
Ondfa	69.19	68.37	(-0.82)
McGill	65.43	68.23	(+2.80)
DeepBlueAI	62.29	54.51	(-7.78)
DFKI-Adapt	61.86	53.94	(-7.92)
Morfbase	59.53	52.07	(-7.47)
BASELINE	56.96	49.32	(-7.64)
DFKI-MPrompt	53.76	46.83	(-6.93)

- *CorPipe* systems also best in evaluation with singletons

Singletons

system	primary	with singletons	
CorPipe	**74.90**	**76.82**	(+1.91)
Anonymous	70.41	73.20	(+2.79)
Ondfa	69.19	68.37	(-0.82)
McGill	65.43	68.23	(+2.80)
DeepBlueAI	62.29	54.51	(-7.78)
DFKI-Adapt	61.86	53.94	(-7.92)
Morfbase	59.53	52.07	(-7.47)
Baseline	56.96	49.32	(-7.64)
DFKI-MPrompt	53.76	46.83	(-6.93)

- *CorPipe* systems also best in evaluation with singletons
- suggests that the submissions with improvements were optimized also for singletons (unlike the others)

Exact Match

system	primary	partial-match	exact-match	*MOR
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**79** / 80 / **79**
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	74 / 78 / 76
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	52 / 83 / 63
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	59 / 82 / 67
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	61 / 81 / 67
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	58 / 80 / 66
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	59 / 78 / 66
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49 / **87** / 61
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	57 / 71 / 62

* Recall / Precision / F1

Exact Match

system	primary	partial-match	exact-match	*MOR
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**79** / 80 / **79**
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	74 / 78 / 76
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	52 / 83 / 63
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	59 / 82 / 67
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	61 / 81 / 67
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	58 / 80 / 66
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	59 / 78 / 66
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49 / **87** / 61
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	57 / 71 / 62

* Recall / Precision / F1

- *CorPipe* performs the best even in terms of partial and exact matching

Exact Match

system	primary	partial-match	exact-match	*MOR
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**79** / 80 / **79**
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	74 / 78 / 76
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	52 / 83 / 63
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	59 / 82 / 67
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	61 / 81 / 67
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	58 / 80 / 66
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	59 / 78 / 66
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49 / **87** / 61
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	57 / 71 / 62

* Recall / Precision / F1

- *CorPipe* performs the best even in terms of partial and exact matching
- for some datasets, the Ondfa system predicted only the head word

Exact Match

system	primary	partial-match	exact-match	*MOR
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**79** / 80 / **79**
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	74 / 78 / 76
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	52 / 83 / 63
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	59 / 82 / 67
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	61 / 81 / 67
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	58 / 80 / 66
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	59 / 78 / 66
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49 / **87** / 61
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	57 / 71 / 62

* Recall / Precision / F1

Exact Match

system	primary	partial-match	exact-match	*MOR
CorPipe	**74.90**	**73.33** (-1.57)	**71.46** (-3.44)	**79** / 80 / **79**
Anonymous	70.41	69.23 (-1.18)	67.09 (-3.32)	74 / 78 / 76
Ondfa	69.19	68.93 (-0.26)	53.01 (-16.18)	52 / 83 / 63
McGill	65.43	64.56 (-0.88)	63.13 (-2.30)	59 / 82 / 67
DeepBlueAI	62.29	61.32 (-0.98)	59.95 (-2.34)	61 / 81 / 67
DFKI-Adapt	61.86	60.83 (-1.03)	59.18 (-2.69)	58 / 80 / 66
Morfbase	59.53	58.49 (-1.05)	56.89 (-2.64)	59 / 78 / 66
BASELINE	56.96	56.28 (-0.68)	54.75 (-2.21)	49 / **87** / 61
DFKI-MPrompt	53.76	51.62 (-2.15)	50.42 (-3.35)	57 / 71 / 62

* Recall / Precision / F1

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

* Recall / Precision / F1

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

* Recall / Precision / F1

- anaphor-decomposable score on zeros

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- over 90 F1 for best-performing systems on some of the datasets

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- over 90 F1 for best-performing systems on some of the datasets
- results on pl_pcc unreliable due to a small number of converted zeros

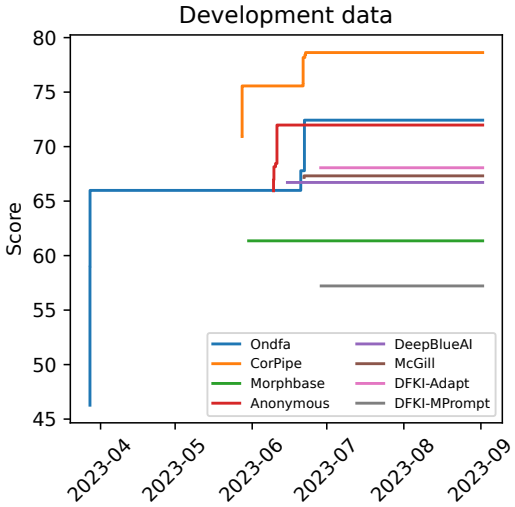
Performance on Zeros

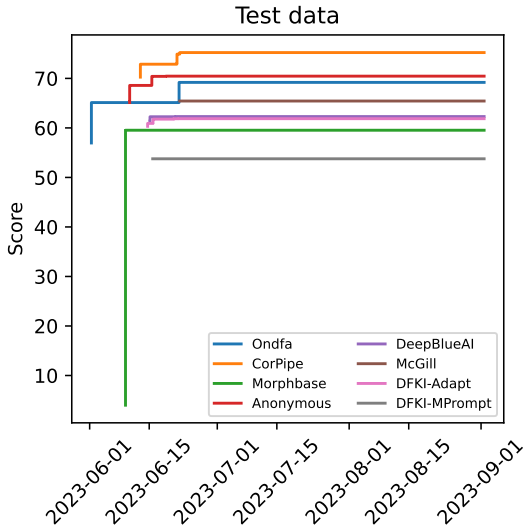
system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_korkor	hu_szeged	pl_pcc
CorPipe	**93** / **92** / **92**	**91** / **92** / **92**	**87** / **88** / **87**	**94** / 95 / **95**	**82** / 89 / **85**	**88** / 70 / 78	75 / 69 / 72
Anonymous	91 / 90 / 91	90 / 91 / 90	86 / 86 / 86	94 / 95 / 94	79 / **89** / 84	83 / **74** / 78	71 / 63 / 67
Ondfa	91 / 90 / 91	90 / 92 / 91	86 / 87 / 87	**94** / 94 / 94	77 / 87 / 82	86 / 74 / **79**	**79** / 73 / **76**
McGill	89 / 90 / 89	88 / 89 / 89	82 / 87 / 84	92 / **95** / 94	81 / 85 / 83	81 / 73 / 77	71 / 65 / 68
DeepBlueAI	85 / 89 / 87	86 / 90 / 88	83 / 86 / 85	91 / 94 / 93	75 / 79 / 77	78 / 70 / 74	**79** / 68 / 73
DFKI-Adapt	85 / 84 / 84	84 / 85 / 84	78 / 81 / 80	89 / 89 / 89	67 / 77 / 72	67 / 61 / 64	62 / 68 / 65
Morfbase	84 / 85 / 85	81 / 84 / 83	78 / 81 / 80	88 / 89 / 88	57 / 73 / 64	61 / 57 / 59	33 / 40 / 36
BASELINE	82 / 82 / 82	81 / 84 / 82	77 / 81 / 79	87 / 88 / 87	60 / 68 / 64	61 / 57 / 59	50 / **80** / 62
DFKI-MPrompt	78 / 83 / 80	78 / 85 / 81	72 / 79 / 75	78 / 87 / 82	69 / 70 / 69	59 / 45 / 51	46 / 55 / 50

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- over 90 F1 for best-performing systems on some of the datasets
- results on pl_pcc unreliable due to a small number of converted zeros
- however, zeros were already generated in the input

Evolution of Competition





Other Statistics

- see the paper

Conclusion

Summary

- summary of CRAC 2023 Multilingual Coreference Resolution Shared Task

Web

`https://ufal.mff.cuni.cz/corefud/crac23`

Summary

- summary of CRAC 2023 Multilingual Coreference Resolution Shared Task
- (slowly) growing number of participants

Web

`https://ufal.mff.cuni.cz/corefud/crac23`

Summary

- summary of CRAC 2023 Multilingual Coreference Resolution Shared Task
- (slowly) growing number of participants
- growing quality of the submissions
- we want to keep evaluation available even after the shared task

Web

`https://ufal.mff.cuni.cz/corefud/crac23`

Future Editions

- we are organizing the shared task in 2024 again

Future Editions

- we are organizing the shared task in 2024 again
- possible extensions:
  - fixing minor errors in CorefUD harmonization procedure
  - additional datasets (non-European?)
  - more realistic setup for zeros