CRAC 2022, Gyeongju, South Korea

Findings of the 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], Yilun Zhu[4]

October 17, 2022

[1] Charles University, Prague, Czechia
[2] University of West Bohemia, Pilsen, Czechia
[3] Polish Academy of Sciences, Warsaw, Poland
[4] Georgetown University, Washington, DC, USA

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 1.0 (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 1.0 (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	10	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 1.0
  - training (gold), dev (gold, no annot), eval (no annot)

Shared Task

- Task:
  1. identify mentions in texts
  2. predict which mentions belong to the same coreference cluster
- Data:
  - CorefUD 1.0
  - training (gold), dev (gold, no annot), eval (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 1.0
  - training (gold), dev (gold, no annot), eval (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 1.0
  - training (gold), dev (gold, no annot), eval (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/4891`)
  - 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 1.0
  - training (gold), dev (gold, no annot), eval (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/4891`)
  - automatic validation, evaluation and ranking of the submissions
- **`https://ufal.mff.cuni.cz/corefud/crac22`**

Datasets

CorefUD 1.0

- public edition of CorefUD 1.0 (Nedoluzhko et al., 2022b)
- 13 coreference datasets for 10 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 1.0 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)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- Polish-PCC (Ogrodniczuk et al., 2013)
- Hungarian-SzegedKoref (Vincze et al., 2018)
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- Russian-RuCor (Toldova et al., 2014)

Annotation Details: Format

Key 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)
```

Response 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--1
13  school  school  NOUN    NN  Number=Sing 11  obj 11:obj  Entity=e58)
```

- 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: 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	43,054
cs_pdt	32,617
en_gum	92
hu_szeged	4,857
pl_pcc	470
es_ancora	8,112

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)
- performance of systems exploiting morpho-syntax may be overestimated

Evaluation Metrics

Primary Score

- CoNLL F1 score
- singletons excluded
- partial matching

Primary Score

- CoNLL F1 score
- singletons excluded
- partial matching

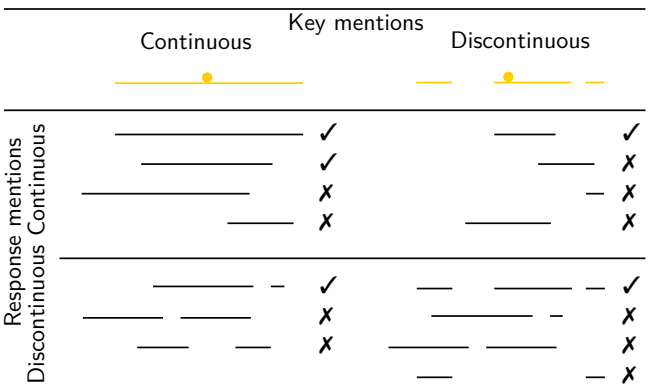
- motivations:
  - some datasets (e.g. cs_pdt) do not specify mention spans, only heads
  - in general, mention boundaries may be difficult to specify
  - some corpora thus define a unit carrying the most important information (head or minimal span)
- mention heads in CorefUD defined syntactically
  - coreference heads often correspond to syntactic heads

    (Nedoluzhko et al., 2021)
- allows for matching discontinuous mentions

Primary Score

- CoNLL F1 score
- singletons excluded
- partial matching

- a response mention matches a key mention if:
  - all its words are included in the key mention
  - the key head is one of the response mention words



Primary Score

- CoNLL F1 score
- singletons excluded
- partial matching

- motivation: singletons not annotated in the majority of CorefUD datasets
- entities with a single mentions deleted from both the key and the response

Primary Score

- CoNLL F1 score
- singletons excluded
- partial 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 matching

Supplementary Scores

- MUC, B^3, CEAF-e
- BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016)
- CoNLL F1 with exact matching
- 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 matching
- CoNLL F1 with singletons
- Mention Overlap Ratio (MOR)
  - measures overlap of key and response mentions, 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 matching
- CoNLL F1 with singletons
- Mention Overlap Ratio (MOR)
  - measures overlap of key and response mentions, 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 Universal Anaphora scorer (Yu et al., 2022)

Official scorer

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

Official scorer

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

Participating Systems

Baseline

- 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 TODO: understand the baseline system better

Teams

- 5 teams
- 9 submissions (including the baseline system)
  - each team was allowed to submit at most 3 systems

Teams

- 5 teams
- 9 submissions (including the baseline system)
  - each team was allowed to submit at most 3 systems

Team	Submission
ÚFAL CorPipe	straka
	straka-single-multilingual-model
	straka-only-single-treebank-data
UWB	ondfa
	BASELINE
Matouš Moravec	Moravec
Barbora Dohnalová	berulasek
	simple-rule-based
Karol Saputa	k-sap

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

- based on baseline
  - system
  - predictions

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	Baseline	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

- rule-based approach vs. deep learning

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

- **Moravec**
  - rule-based post-processing of baseline predictions
  - exploits the output of named entity recognition using NameTag (Straková et al., 2019)

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

- **berulasek**
  - rule-based post-processing of baseline predictions
  - reduces mention spans to heads
  - links proper nouns with the same lemma

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

- **simple-rule-based**
  - links each pronoun to the nearest previous pronoun of the same gender
  - applies *berulasek* post-processing

System Comparison: Basic Properties

Team	Submission	Baseline based	Approach
ÚFAL CorPipe	straka	No	DL
	straka-single-multilingual-model	No	DL
	straka-only-single-treebank-data	No	DL
UWB	ondfa	Yes	DL
	BASELINE	–	DL
Matouš Moravec	Moravec	Yes – files only	rule-based postprocess of DL
Barbora Dohnalová	berulasek	Yes – files only	rule-based postprocess of DL
	simple-rule-based	No	rules
Karol Saputa	k-sap	No	DL

System Comparison: DL-based

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	∼10

System Comparison: DL-based

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	∼10

- large pre-trained models
- hundreds of millions parameters

System Comparison: DL-based

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	∼10

- large pre-trained models
- hundreds of millions parameters
- mostly multi-lingual

System Comparison: DL-based

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	∼10

- large pre-trained models
- hundreds of millions parameters
- mostly multi-lingual
- *k-sap*: for Polish only

System Comparison: DL-based

Team	Submission	Model	SL	Size	Batch size	Updates	HParams
ÚFAL CorPipe	straka	google/rembert	512	614M	8	960k	4
	straka-single...	google/rembert	512	614M	8	960k	4
	straka-only...	google/rembert	512	614M	8	960k	4
UWB	ondfa	xlm-roberta-large	512	600M	1	800k	4
	BASELINE	multiling. BERT	512	220M	1	800k	0
Karol Saputa	k-sap	allegro/herbert-base-cased	512	415M	Dynamic	27k	∼10

- large pre-trained models
- hundreds of millions parameters
- mostly multi-lingual
- *k-sap*: for Polish only
- maximum sequence length: 512 sub-words

Results and Comparison

ÚFAL CorPipe: *straka*

Congratulations!

Main Results: Primary Score

system	CoNLL F1
straka	**70.72**
straka-single...	69.56
ondfa	67.64
straka-only...	64.30
berulasek	59.72
BASELINE	58.53
moravec	55.05
simple-rule-based	18.14
k-sap	5.90

Main Results: Primary Score

system	CoNLL F1
straka	**70.72**
straka-single...	69.56
ondfa	67.64
straka-only...	64.30
berulasek	59.72
BASELINE	58.53
moravec	55.05
simple-rule-based	18.14
k-sap	5.90

- improvement of 12 points (20%) over the baseline

Main Results: Supplementary Scores

system	primary	*MUC	B^3	CEAF-e	BLANC	LEA
straka	**70.72**	**74** / 76 / **74**	**67** / **72** / **68**	**71** / **70** / **70**	**63** / 70 / **65**	**63** / **69** / **65**
straka-single...	69.56	72 / 76 / 73	65 / 72 / 67	67 / 70 / 68	61 / **71** / 64	62 / 68 / 64
ondfa	67.64	69 / **76** / 72	61 / 71 / 65	62 / 69 / 65	59 / 69 / 63	58 / 67 / 62
straka-only...	64.30	65 / 71 / 68	58 / 68 / 62	61 / 67 / 63	55 / 66 / 59	54 / 63 / 58
berulasek	59.72	58 / 76 / 64	50 / 70 / 57	52 / 67 / 58	46 / 70 / 53	45 / 66 / 53
BASELINE	58.53	56 / 74 / 63	48 / 69 / 56	51 / 66 / 57	45 / 68 / 51	44 / 64 / 51
moravec	55.05	53 / 70 / 60	45 / 65 / 52	50 / 59 / 53	41 / 59 / 46	41 / 60 / 48
simple-rule-based	18.14	14 / 22 / 16	14 / 26 / 17	23 / 27 / 22	10 / 20 / 11	7 / 17 / 9
k-sap	5.90	6 / 7 / 6	5 / 7 / 6	5 / 6 / 6	5 / 7 / 6	5 / 6 / 6

* Recall / Precision / F1

Main Results: Supplementary Scores

system	primary	*MUC	B^3	CEAF-e	BLANC	LEA
straka	**70.72**	**74** / 76 / **74**	**67** / **72** / **68**	**71** / **70** / **70**	**63** / 70 / **65**	**63** / **69** / **65**
straka-single...	69.56	72 / 76 / 73	65 / 72 / 67	67 / 70 / 68	61 / **71** / 64	62 / 68 / 64
ondfa	67.64	69 / **76** / 72	61 / 71 / 65	62 / 69 / 65	59 / 69 / 63	58 / 67 / 62
straka-only...	64.30	65 / 71 / 68	58 / 68 / 62	61 / 67 / 63	55 / 66 / 59	54 / 63 / 58
berulasek	59.72	58 / 76 / 64	50 / 70 / 57	52 / 67 / 58	46 / 70 / 53	45 / 66 / 53
BASELINE	58.53	56 / 74 / 63	48 / 69 / 56	51 / 66 / 57	45 / 68 / 51	44 / 64 / 51
moravec	55.05	53 / 70 / 60	45 / 65 / 52	50 / 59 / 53	41 / 59 / 46	41 / 60 / 48
simple-rule-based	18.14	14 / 22 / 16	14 / 26 / 17	23 / 27 / 22	10 / 20 / 11	7 / 17 / 9
k-sap	5.90	6 / 7 / 6	5 / 7 / 6	5 / 6 / 6	5 / 7 / 6	5 / 6 / 6

* Recall / Precision / F1

- *straka* 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_szeged	lt_lcc	pl_pcc	ru_rucor
straka	**70.72**	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
straka-single...	69.56	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
ondfa	67.64	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
straka-only...	64.30	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
berulasek	59.72	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
BASELINE	58.53	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.99	55.55	52.35	64.81	65.34	67.66
moravec	55.05	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
simple-rule-based	18.14	15.58	5.51	9.48	29.81	19.41	21.99	11.37	16.64	21.74	17.00	27.53	15.69	24.06
k-sap	5.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.67	0.00

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_szeged	lt_lcc	pl_pcc	ru_rucor
straka	**70.72**	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
straka-single...	69.56	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
ondfa	67.64	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
straka-only...	64.30	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
berulasek	59.72	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
BASELINE	58.53	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.99	55.55	52.35	64.81	65.34	67.66
moravec	55.05	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
simple-rule-based	18.14	15.58	5.51	9.48	29.81	19.41	21.99	11.37	16.64	21.74	17.00	27.53	15.69	24.06
k-sap	5.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.67	0.00

- ÚFAL CorPipe team dominant on most 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_szeged	lt_lcc	pl_pcc	ru_rucor
straka	**70.72**	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
straka-single...	69.56	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
ondfa	67.64	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
straka-only...	64.30	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
berulasek	59.72	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
BASELINE	58.53	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.99	55.55	52.35	64.81	65.34	67.66
moravec	55.05	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
simple-rule-based	18.14	15.58	5.51	9.48	29.81	19.41	21.99	11.37	16.64	21.74	17.00	27.53	15.69	24.06
k-sap	5.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.67	0.00

- ÚFAL CorPipe team dominant on most datasets
- outperformed by *ondfa* only on de_parcorfull and hu_szeged

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_szeged	lt_lcc	pl_pcc	ru_rucor
straka	**70.72**	78.18	**78.59**	**77.69**	65.52	70.69	72.50	**39.00**	**81.39**	**65.27**	63.15	**69.92**	78.12	**79.34**
straka-single...	69.56	**78.49**	78.49	77.57	59.94	**71.11**	**73.20**	33.55	80.80	64.35	63.38	67.38	**78.32**	77.74
ondfa	67.64	70.55	74.07	72.42	**73.90**	68.68	68.31	31.90	72.32	61.39	**65.01**	68.05	75.20	77.50
straka-only...	64.30	76.34	77.87	76.76	36.50	56.65	70.66	23.48	78.78	64.94	62.94	61.32	73.36	76.26
berulasek	59.72	64.67	70.56	67.95	38.50	57.70	63.07	36.44	66.61	56.04	55.02	65.67	65.99	68.17
BASELINE	58.53	63.74	70.00	67.27	33.75	55.44	62.59	36.44	65.99	55.55	52.35	64.81	65.34	67.66
moravec	55.05	58.25	68.19	64.71	31.86	52.84	59.15	36.44	62.01	54.87	52.00	59.49	63.40	52.49
simple-rule-based	18.14	15.58	5.51	9.48	29.81	19.41	21.99	11.37	16.64	21.74	17.00	27.53	15.69	24.06
k-sap	5.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.67	0.00

- ÚFAL CorPipe team dominant on most datasets
- outperformed by *ondfa* only on de_parcorfull and hu_szeged
- *k-sap* ranks 3rd on pl_pcc, ignoring the other datasets

Singletons

system	primary	with-singletons	
straka	**70.72**	**72.98**	(+2.26)
straka-single...	69.56	71.81	(+2.25)
ondfa	67.64	58.06	(-9.58)
straka-only...	64.30	67.93	(+3.63)
berulasek	59.72	50.84	(-8.88)
BASELINE	58.53	49.69	(-8.84)
moravec	55.05	46.79	(-8.27)
simple-rule-based	18.14	17.13	(-1.00)
k-sap	5.90	3.83	(-2.07)

Singletons

system	primary	with-singletons	
straka	**70.72**	**72.98**	(+2.26)
straka-single...	69.56	71.81	(+2.25)
ondfa	67.64	58.06	(-9.58)
straka-only...	64.30	67.93	(+3.63)
berulasek	59.72	50.84	(-8.88)
BASELINE	58.53	49.69	(-8.84)
moravec	55.05	46.79	(-8.27)
simple-rule-based	18.14	17.13	(-1.00)
k-sap	5.90	3.83	(-2.07)

- *straka** systems also best in evaluation with singletons

Singletons

system	primary	with-singletons	
straka	**70.72**	**72.98**	(+2.26)
straka-single...	69.56	71.81	(+2.25)
ondfa	67.64	58.06	(-9.58)
straka-only...	64.30	67.93	(+3.63)
berulasek	59.72	50.84	(-8.88)
BASELINE	58.53	49.69	(-8.84)
moravec	55.05	46.79	(-8.27)
simple-rule-based	18.14	17.13	(-1.00)
k-sap	5.90	3.83	(-2.07)

- *straka** systems also best in evaluation with singletons
- the only ones that are positively affected

Singletons

system	primary	with-singletons	
straka	**70.72**	**72.98**	(+2.26)
straka-single...	69.56	71.81	(+2.25)
ondfa	67.64	58.06	(-9.58)
straka-only...	64.30	67.93	(+3.63)
berulasek	59.72	50.84	(-8.88)
BASELINE	58.53	49.69	(-8.84)
moravec	55.05	46.79	(-8.27)
simple-rule-based	18.14	17.13	(-1.00)
k-sap	5.90	3.83	(-2.07)

- *straka** systems also best in evaluation with singletons
- the only ones that are positively affected
- suggests that ÚFAL CorPipe optimized also for singletons (unlike the other teams)

Exact Match

system	primary	exact-match	*MOR
straka	**70.72**	33.18 (-37.54)	32 / 83 / 45
straka-single...	69.56	33.06 (-36.51)	32 / 84 / 45
ondfa	67.64	54.73 (-12.91)	**52** / 84 / **62**
straka-only...	64.30	32.28 (-32.02)	30 / 83 / 43
berulasek	59.72	31.50 (-28.22)	27 / **88** / 40
BASELINE	58.53	**56.72** (-1.82)	49 / 86 / 61
moravec	55.05	52.68 (-2.37)	49 / 81 / 60
simple-rule-based	18.14	12.60 (-5.54)	16 / 55 / 23
k-sap	5.90	5.84 (-0.05)	5 / 7 / 6

* Recall / Precision / F1

Exact Match

system	primary	exact-match	*MOR
straka	**70.72**	33.18 (-37.54)	32 / 83 / 45
straka-single...	69.56	33.06 (-36.51)	32 / 84 / 45
ondfa	67.64	54.73 (-12.91)	**52** / 84 / **62**
straka-only...	64.30	32.28 (-32.02)	30 / 83 / 43
berulasek	59.72	31.50 (-28.22)	27 / **88** / 40
BASELINE	58.53	**56.72** (-1.82)	49 / 86 / 61
moravec	55.05	52.68 (-2.37)	49 / 81 / 60
simple-rule-based	18.14	12.60 (-5.54)	16 / 55 / 23
k-sap	5.90	5.84 (-0.05)	5 / 7 / 6

* Recall / Precision / F1

- BASELINE system performs the best in terms of exact matching

Exact Match

system	primary	exact-match	*MOR
straka	**70.72**	33.18 (-37.54)	32 / 83 / 45
straka-single...	69.56	33.06 (-36.51)	32 / 84 / 45
ondfa	67.64	54.73 (-12.91)	**52** / 84 / **62**
straka-only...	64.30	32.28 (-32.02)	30 / 83 / 43
berulasek	59.72	31.50 (-28.22)	27 / **88** / 40
BASELINE	58.53	**56.72** (-1.82)	49 / 86 / 61
moravec	55.05	52.68 (-2.37)	49 / 81 / 60
simple-rule-based	18.14	12.60 (-5.54)	16 / 55 / 23
k-sap	5.90	5.84 (-0.05)	5 / 7 / 6

* Recall / Precision / F1

- BASELINE system performs the best in terms of exact matching
- the teams optimized for the primary score, which is based on partial matching

Exact Match

system	primary	exact-match	*MOR
straka	**70.72**	33.18 (-37.54)	32 / 83 / 45
straka-single...	69.56	33.06 (-36.51)	32 / 84 / 45
ondfa	67.64	54.73 (-12.91)	**52** / 84 / **62**
straka-only...	64.30	32.28 (-32.02)	30 / 83 / 43
berulasek	59.72	31.50 (-28.22)	27 / **88** / 40
BASELINE	58.53	**56.72** (-1.82)	49 / 86 / 61
moravec	55.05	52.68 (-2.37)	49 / 81 / 60
simple-rule-based	18.14	12.60 (-5.54)	16 / 55 / 23
k-sap	5.90	5.84 (-0.05)	5 / 7 / 6

* Recall / Precision / F1

- BASELINE system performs the best in terms of exact matching
- the teams optimized for the primary score, which is based on partial matching
- some teams even reduced the mention spans to heads in post-processing

Exact Match

system	primary	exact-match	*MOR
straka	**70.72**	33.18 (-37.54)	32 / 83 / 45
straka-single...	69.56	33.06 (-36.51)	32 / 84 / 45
ondfa	67.64	54.73 (-12.91)	**52** / 84 / **62**
straka-only...	64.30	32.28 (-32.02)	30 / 83 / 43
berulasek	59.72	31.50 (-28.22)	27 / **88** / 40
BASELINE	58.53	**56.72** (-1.82)	49 / 86 / 61
moravec	55.05	52.68 (-2.37)	49 / 81 / 60
simple-rule-based	18.14	12.60 (-5.54)	16 / 55 / 23
k-sap	5.90	5.84 (-0.05)	5 / 7 / 6

* Recall / Precision / F1

- BASELINE system performs the best in terms of exact matching
- the teams optimized for the primary score, which is based on partial matching
- some teams even reduced the mention spans to heads in post-processing
- confirmed by low MOR recall scores

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_szeged	pl_pcc
straka	**91** / 91 / **91**	**91** / **92** / **92**	87 / **90** / **89**	**94** / **95** / **95**	79 / 71 / 75	62 / 60 / 61
straka-single...	91 / **92** / 91	91 / 92 / 92	**88** / 90 / 89	94 / 95 / 95	76 / **76** / 76	**79** / 83 / **81**
ondfa	88 / 88 / 88	88 / 92 / 90	85 / 89 / 87	92 / 94 / 93	**81** / 74 / **77**	62 / 60 / 61
straka-only...	89 / 88 / 88	90 / 92 / 91	87 / 89 / 88	92 / 92 / 92	74 / 70 / 72	71 / 63 / 67
berulasek	82 / 83 / 82	84 / 86 / 85	80 / 84 / 82	87 / 89 / 88	55 / 54 / 54	42 / 50 / 45
BASELINE	82 / 82 / 82	84 / 86 / 85	80 / 83 / 82	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
moravec	81 / 82 / 82	84 / 85 / 84	80 / 83 / 81	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
simple-rule-based	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
k-sap	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	4 / **100** / 8

* Recall / Precision / F1

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_szeged	pl_pcc
straka	**91** / 91 / **91**	**91** / **92** / **92**	87 / **90** / **89**	**94** / **95** / **95**	79 / 71 / 75	62 / 60 / 61
straka-single...	91 / **92** / 91	91 / 92 / 92	**88** / 90 / 89	94 / 95 / 95	76 / **76** / 76	**79** / 83 / **81**
ondfa	88 / 88 / 88	88 / 92 / 90	85 / 89 / 87	92 / 94 / 93	**81** / 74 / **77**	62 / 60 / 61
straka-only...	89 / 88 / 88	90 / 92 / 91	87 / 89 / 88	92 / 92 / 92	74 / 70 / 72	71 / 63 / 67
berulasek	82 / 83 / 82	84 / 86 / 85	80 / 84 / 82	87 / 89 / 88	55 / 54 / 54	42 / 50 / 45
BASELINE	82 / 82 / 82	84 / 86 / 85	80 / 83 / 82	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
moravec	81 / 82 / 82	84 / 85 / 84	80 / 83 / 81	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
simple-rule-based	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
k-sap	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	4 / **100** / 8

* Recall / Precision / F1

- anaphor-decomposable score on zeros

Performance on Zeros

system	ca_ancora	cs_pdt	cs_pcedt	es_ancora	hu_szeged	pl_pcc
straka	**91** / 91 / **91**	**91** / **92** / **92**	87 / **90** / **89**	**94** / **95** / **95**	79 / 71 / 75	62 / 60 / 61
straka-single...	91 / **92** / 91	91 / 92 / 92	**88** / 90 / 89	94 / 95 / 95	76 / **76** / 76	**79** / 83 / **81**
ondfa	88 / 88 / 88	88 / 92 / 90	85 / 89 / 87	92 / 94 / 93	**81** / 74 / **77**	62 / 60 / 61
straka-only...	89 / 88 / 88	90 / 92 / 91	87 / 89 / 88	92 / 92 / 92	74 / 70 / 72	71 / 63 / 67
berulasek	82 / 83 / 82	84 / 86 / 85	80 / 84 / 82	87 / 89 / 88	55 / 54 / 54	42 / 50 / 45
BASELINE	82 / 82 / 82	84 / 86 / 85	80 / 83 / 82	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
moravec	81 / 82 / 82	84 / 85 / 84	80 / 83 / 81	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
simple-rule-based	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
k-sap	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	4 / **100** / 8

* 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_szeged	pl_pcc
straka	**91** / 91 / **91**	**91** / **92** / **92**	87 / **90** / **89**	**94** / **95** / **95**	79 / 71 / 75	62 / 60 / 61
straka-single...	91 / **92** / 91	91 / 92 / 92	**88** / 90 / 89	94 / 95 / 95	76 / **76** / 76	**79** / 83 / **81**
ondfa	88 / 88 / 88	88 / 92 / 90	85 / 89 / 87	92 / 94 / 93	**81** / 74 / **77**	62 / 60 / 61
straka-only...	89 / 88 / 88	90 / 92 / 91	87 / 89 / 88	92 / 92 / 92	74 / 70 / 72	71 / 63 / 67
berulasek	82 / 83 / 82	84 / 86 / 85	80 / 84 / 82	87 / 89 / 88	55 / 54 / 54	42 / 50 / 45
BASELINE	82 / 82 / 82	84 / 86 / 85	80 / 83 / 82	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
moravec	81 / 82 / 82	84 / 85 / 84	80 / 83 / 81	87 / 88 / 87	52 / 51 / 52	42 / 50 / 45
simple-rule-based	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
k-sap	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	4 / **100** / 8

* Recall / Precision / F1

- anaphor-decomposable score on zeros
- over 90 F1 for best-performing systems on some of the datasets
- however, zeros were already generated in the input

Other Analyses

- both automatic and manual
- see the paper

Conclusion