CRAC'21 Panel Discussion

The Universal Anaphora Extension of the CONLL-U Markup Scheme

Anna Nedoluzhko, Michal Novák, Martin Popel,
Zdeněk Žabokrtský and Dan Zeman (Charles University, Prague)
in cooperation with Amir Zeldes (Georgetown University)

📅 November 11, 2021

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

Outline

1. Existing **diversity** in coreference resources
2. The **CorefUD** collection in a nutshell: already harmonized resources for 11 languages
3. A CRAC 2022 **shared task proposal** on multilingual coreference resolution

Diversity of existing coreference resources

Diversity of content

Observed differences along several dimensions:

- **mention span** – a linearly delimited sequence of tokens, or a syntactically delimited element (in a constituency or dependency tree)?
- **classification** of mentions?
- coreference **grouping** – chain-based, or cluster-based?
- **non-identity** anaphora relations included too?
- handling of **specific relations**: apposition, predication, split antedent ...
- presence of annotated **zeros** (e.g. pro-drops)?
- **other NLP annotations** present in the data: lemmatization, POS tagging, syntactic trees, named entities
- and many others differences ...

Diversity of file formats (selected examples)

- **CoNLL 2011 / CoNLL 2012 / SemEval 2010** (Pradhan et al., 2012, 2011, Recasens et al., 2010)
    - plain-text based, column-based
    - identity coreference only
    - coreference in the last column in open-close notation
    - CoNLL 2011 and 2012 Shared tasks set the standard for its representation and evaluation
- **MMAX / MMAX2** (Müller and Strube, 2001, 2006)
    - XML-based
    - broad variety of linguistic phenomena, including anaphora
    - ARRAU, Polish Coreference Corpus, COREA, Potsdam Commentary Corpus, ParCorFull
    - numerous variations of the format
- **Prague Markup Language** (Pajas and Štěpánek, 2006)
    - XML-based
    - broad variety of linguistic phenomena, including anaphora
    - Prague Dependency Treebank, Prague Czech-English Dependency Treebank
    - rarely used outside UFAL

Diversity of file formats – a generalization

- we cannot escape from the trade-off between:
  - **simplicity and robustness** (but then limited expressive power),
  - versus **flexibility and extensibility** (but then difficult maintainability and danger of divergence)
- lessons taken from UD
  - extremely **simplified scheme is beneficial** for community growth
  - it is crucial to have a **single format** already in **early stages**
  - **automatic validators** are extremely valuable

Universal Anaphora developments 2020-2021 (our view!)

1. Universal Anaphora theme opened on CRAC 2020; UA Initiative announced then
2. three file formats discussed extensively:
   - an **XML-based** format, versatile, easy to extend with additional layers of annotation,
   - an **extension of the CoNLL-U** file format, with added columns,
   - a file format **strictly compliant with the CoNLL-U** standard
3. we (both in Prague and in Georgetown) prefer strongly the third option

   Note: technically, it is not an extension, just an additional convention within the CoNLL-U's MISC column

Universal Anaphora developments 2020-2021 (our view!), cont.

6. Prague's proof of concept: the CorefUD collection, 17 coreference datasets converted to CoNLL-U, completed in March 2021, released on Lindat
7. negotiation with Amir Zeldes in April 2021: agreed to accept Amir's convention used in GUM for the MISC column (details)
8. a new CorefUD release planned for January 2022, based on the GUM style
9. CorefUD Python API will be modified accordingly

Hence, from our perspective, the file format question is basically solved :-)

CorefUD in a nutshell

17 coreference datasets harmonized in CorefUD 0.1

free licenses

- 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)

non-free licenses

- English-OntoNotes (Weischedel et al., 2011)
- English-ARRAU (Uryupina et al., 2020)
- Dutch-COREA (Hendrickx et al., 2008)
- English-PCEDT (Nedoluzhko et al., 2016)

Diversity in existing resources: relations

	Coref. grouping		Relations among mentions					
CorefUD dataset	cluster-based	link-based	singletons	appos.	pred.	split antec.	disc. deixis	bridg.
Catalan-AnCora	✓	×	✓	✓	✓	✓	✓	×
Czech-PCEDT	×	✓	(✓)	(✓)	(✓)	✓	✓	×
Czech-PDT	×	✓	(✓)	(✓)	(✓)	✓	✓	✓
English-GUM	✓	×	✓	✓	✓	✓	✓	✓
English-ParCorFull	✓	×	×	✓	(✓)	✓	✓	×
French-Democrat	✓	×	✓	×	×	×	×	×
German-ParCorFull	✓	×	×	✓	(✓)	✓	✓	×
German-PotsdamCC	×	✓	✓	✓	✓ ?	×	✓	×
Hungarian-SzegedKoref	✓	×	×	✓	?	×	✓	✓
Lithuanian-LCC	×	✓	×	×	×	✓	×	×
Polish-PCC	✓	×	✓	✓	✓	×	✓	✓
Russian-RuCor	✓	×	×	✓	✓	×	×	×
Spanish-AnCora	✓	×	✓	✓	✓	✓	✓	×
Dutch-COREA	×	✓	✓	✓	✓	×	✓	✓
English-ARRAU	✓	✓	✓	✓	✓	✓	✓	✓
English-OntoNotes	✓	×	×	✓	×	×	(✓)	×
English-PCEDT	×	✓	(✓)	(✓)	(✓)	✓	✓	×

Example of extracted statistics: non-singleton mentions

	mentions				distribution of lengths					
CorefUD dataset	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	62,417	128	134	4.2	10.2	34.6	19.6	7.5	4.5	23.7
Czech-PCEDT	178,475	154	79	3.4	23.0	28.5	16.1	8.3	4.1	20.0
Czech-PDT	169,644	203	99	2.9	17.2	36.4	18.7	8.5	4.1	15.1
English-GUM	22,896	170	95	2.6	0.0	54.8	20.6	8.4	3.9	12.3
English-ParCorFull	720	67	37	2.1	0.0	59.0	24.4	6.0	2.9	7.6
French-Democrat	47,172	166	71	1.7	0.0	64.2	21.7	6.4	2.5	5.3
German-ParCorFull	900	85	30	2.0	0.0	65.0	17.4	6.2	4.0	7.3
German-PotsdamCC	2,523	76	34	2.6	0.0	34.8	32.4	15.5	6.4	10.9
Hungarian-SzegedKoref	15,182	122	36	1.6	15.1	37.4	32.5	10.2	2.6	2.2
Lithuanian-LCC	4,337	117	19	1.5	0.0	69.1	16.6	11.1	1.2	2.0
Polish-PCC	82,865	154	108	2.1	0.3	68.7	14.9	5.2	2.7	8.2
Russian-RuCor	16,254	104	18	1.7	0.0	68.9	16.3	6.7	3.5	4.6
Spanish-AnCora	70,675	137	90	4.4	11.4	35.3	17.6	7.6	4.0	24.1
Dutch-COREA	8,663	62	60	2.6	0.0	42.5	33.1	8.6	4.0	11.7
English-ARRAU	31,906	139	75	2.9	0.0	45.4	26.9	10.7	4.2	12.8
English-OntoNotes	209,435	128	94	2.5	0.0	56.3	19.8	8.1	4.2	11.7
English-PCEDT	183,984	157	88	3.6	19.3	28.0	17.0	10.6	4.8	20.3

CRAC 2022 shared task proposal

Motivation for a coreference shared task proposal

- inspiration: the immense effect of the CoNLL-X Shared Task on Multilingual Dependency Parsing (2006) on the parsing community

- a similar number of languages
  - CoNLL-X in 2006: 12 languages
  - CorefUD in 2021: 11 languages

- ⟹ now is the right time! :-)

Data for the shared task

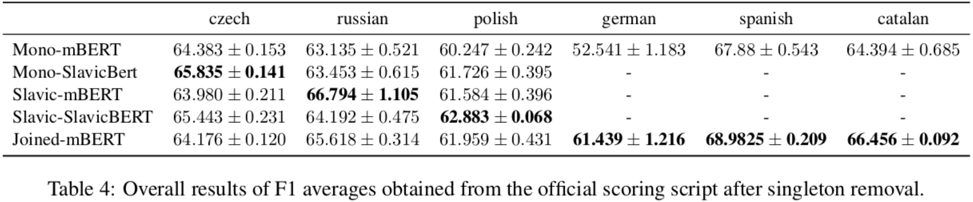
- CorefUD **public edition** – sufficiently free licenses
  - 13 datasets for 10 languages (1 dataset for Catalan, 2 for Czech, 2 for English, 1 for French, 2 for German, 1 for Hungarian, 1 for Lithuanian, 1 for Polish, 1 for Russian, and 1 for Spanish)
- CorefUD **non-public edition** – converted, but undistributable
  - 4 more datasets for 2 languages (1 dataset for Dutch, and 3 for English)
  - inclusion into the shared task up to copyright holders' decisions
  - or, if legally possible, replacing the text with underscores?
- train/dev/test split already defined in CorefUD (preserved from original resources)
- all test portions are kept unpublished

Evaluation measure?

- no straightforward natural measure for coreference resolution (nothing comparable e.g. to UAS for dependency parsing)
- a common solution: an average of MUC, B3 and CEAF scores (or BLANC)
- existing scorers
  - Perl: `https://github.com/conll/reference-coreference-scorers`
  - Python: `https://github.com/juntaoy/universal-anaphora-scorer`
- perhaps a Python reimplementation tailored for the CoNLL-U format would be useful

A baseline system?

- experimental results available already now for a subset of the CorefUD datasets: Pražák, Ondřej, Miloslav Konopík, and Jakub Sido. "Multilingual Coreference Resolution with Harmonized Annotations." arXiv:2107.12088 (2021):



	czech	russian	polish	german	spanish	catalan
Mono-mBERT	64.383 ± 0.153	63.135 ± 0.521	60.247 ± 0.242	52.541 ± 1.183	67.88 ± 0.543	64.394 ± 0.685
Mono-SlavicBert	**65.835 ± 0.141**	63.453 ± 0.615	61.726 ± 0.395	-	-	-
Slavic-mBERT	63.980 ± 0.211	**66.794 ± 1.105**	61.584 ± 0.396	-	-	-
Slavic-SlavicBERT	65.443 ± 0.231	64.192 ± 0.475	**62.883 ± 0.068**	-	-	-
Joined-mBERT	64.176 ± 0.120	65.618 ± 0.314	61.959 ± 0.431	**61.439 ± 1.216**	**68.9825 ± 0.209**	**66.456 ± 0.092**

Table 4: Overall results of F1 averages obtained from the official scoring script after singleton removal.

Multiple tracks?

- coreference track alone?
- a bridging track?
- a surprise language track? (a few not-yet-harmonized resources waiting in a queue)

Possible co-organizers

- the team in **Charles University** (Prague)
  - Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský and Dan Zeman
  - CorefUD data providers
  - Python API providers
- the team in **University of West Bohemia** (Pilsen)
  - Ondřej Pražák, Miloslav Konopík, Jakub Sido
  - providers of a baseline system
- possibly a student of Amir Zeldes in **Georgetown University**
- and hopefully some more volunteers :-)

Conclusions

Conclusions

- We believe CorefUD is mature enough to provide data for a shared task on multilingual coreference resolution
- QUESTION 1: Is there a space for the proposed shared task within CRAC 2022?
- QUESTION 2: If not, can we find some other opportunity in 2022?
- QUESTION 3: Anyone potentially interested in participating in such a shared task?

More about CorefUD: `https://ufal.mff.cuni.cz/corefud`

Thank you!