Recent Developments in the Universal Anaphora Scorer

Juntao Yu[1] and **Michal Novak**[2]

[1]University of Essex, UK; [2]Charles University, Czech Republic

(Joint work with **Sopan Khosla, Nafise Moosavi, Silviu Paun, Martin Popel, Sameer Pradhan, Zdeněk Žabokrtský, Yilun Zhu,** and **Massimo Poesio**)

CRAC@COLING 2022

Outline of the talk

- **UA scorer v1.0** and its use in the CODI/CRAC 2021 and 2022 ST
- **CorefUD scorer** and its use in the CRAC 2022 ST
- **UA scorer v2.0**
- Q&A

The Universal Anaphora Initiative

- Beyond Identity Anaphora:
  - **Split-antecedent anaphora**
    - E.g. [John]$_1$ met [Mary]$_2$. [He]$_1$ greeted [her]$_2$. [They]$_{1,2}$ went to the movies.
  - **Discourse deixis**
    - E.g. ... when suddenly a White Rabbit with pink eyes ran close by her. There was nothing so VERY remarkable in **[that];** nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at **[this],** but at the time it all seemed quite natural); ....
  - **Bridging references**
    - E.g. she found herself in [a long, low hall, which was lit up by a row of lamps hanging from [the roof]].

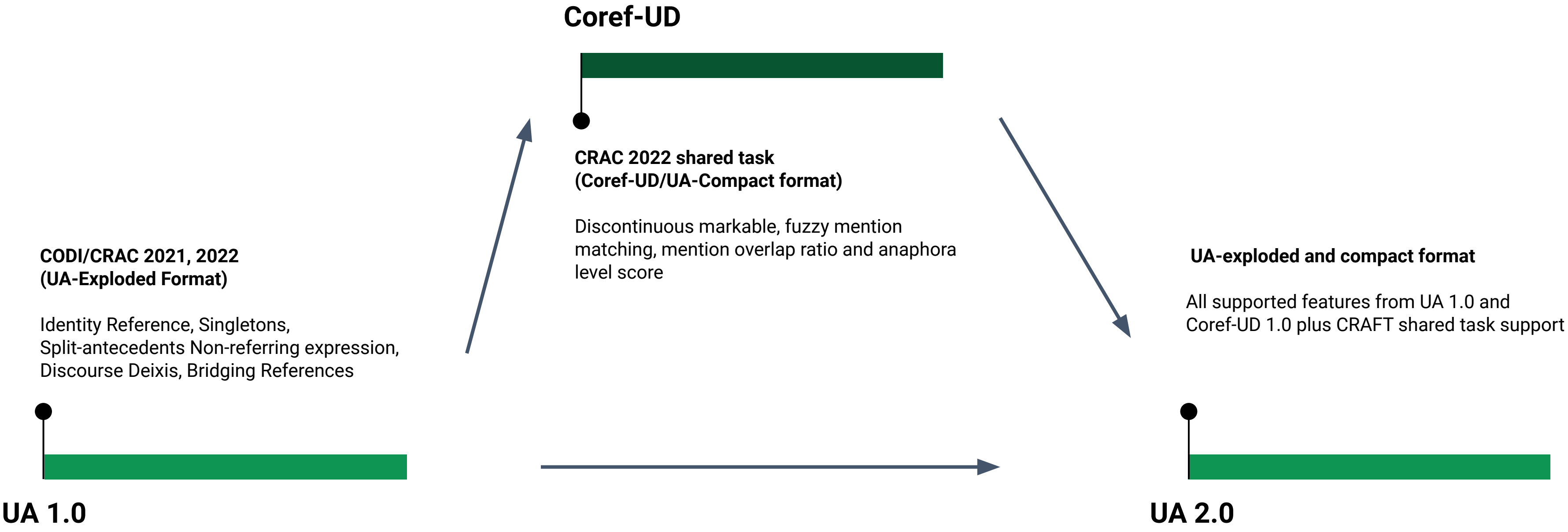
[1] http://www.universalanaphora.org

The Universal Anaphora Scorer

An extension to the Reference Coreference Scorer (Pradhan et al, 2014) scoring the interpretation of

- Identity Reference
- Singletons
- Split-antecedent anaphora
- Non-referring expression
- Discourse Deixis
- Bridging References
- Discontinuous markable
- Partial/Fuzzy Mention matching

The UA scorer timeline



Coref-UD

CRAC 2022 shared task
(Coref-UD/UA-Compact format)

Discontinuous markable, fuzzy mention matching, mention overlap ratio and anaphora level score

CODI/CRAC 2021, 2022
(UA-Exploded Format)

Identity Reference, Singletons, Split-antecedents Non-referring expression, Discourse Deixis, Bridging References

UA-exploded and compact format

All supported features from UA 1.0 and Coref-UD 1.0 plus CRAFT shared task support

UA 1.0

UA 2.0

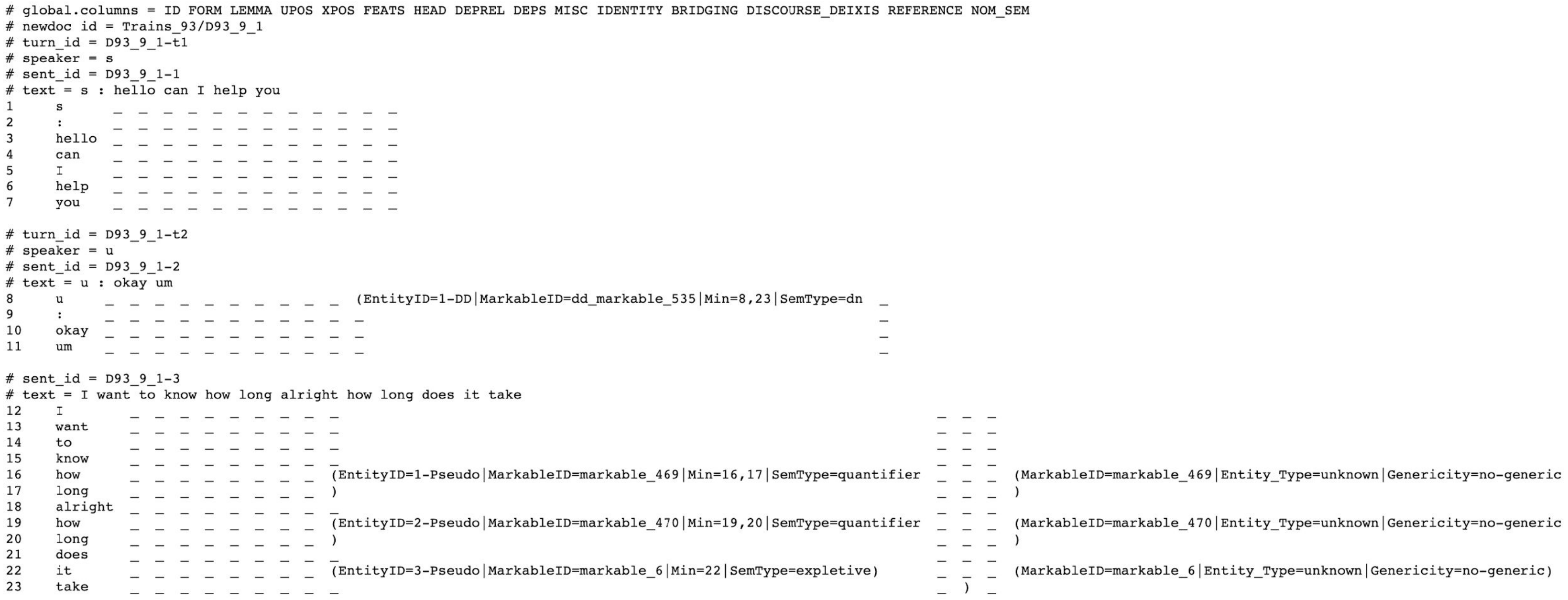
The Universal Anaphora Scorer, 1.0

The input: CONLL-UA Exploded Format

- an extension of the **CONLL-U** tabular format defined for Universal Dependencies broadly compatible with the **standard CONLL format** used by the Coreference Reference Scorer

- **Identity column**:
  - specifying the entity a markable refers to in the case of a referring markable and, optionally, whether the markable is referring or not, what its head is, and, for split antecedents, the set they belong to;
- **Bridging column**:
  - specifying the anchor, its most recent mention, and, optionally, the associative relation;
- **Discourse Deixis column**:
  - whose markables specify the non-nominal antecedents of discourse deixis, represented exactly as in the Identity layer. This makes it possible to adopt for discourse deixis the same metrics used for identity anaphora.

The input: CONLL-UA Exploded Format



```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC IDENTITY BRIDGING DISCOURSE_DEIXIS REFERENCE NOM_SEM
# newdoc id = Trains_93/D93_9_1
# turn_id = D93_9_1-t1
# speaker = s
# sent_id = D93_9_1-1
# text = s : hello can I help you
1     s      _  _  _  _  _  _  _  _  _  _  _  _
2     :      _  _  _  _  _  _  _  _  _  _  _  _
3     hello  _  _  _  _  _  _  _  _  _  _  _  _
4     can    _  _  _  _  _  _  _  _  _  _  _  _
5     I      _  _  _  _  _  _  _  _  _  _  _  _
6     help   _  _  _  _  _  _  _  _  _  _  _  _
7     you    _  _  _  _  _  _  _  _  _  _  _  _

# turn_id = D93_9_1-t2
# speaker = u
# sent_id = D93_9_1-2
# text = u : okay um
8     u     _  _  _  _  _  _  _  _  _  _  (EntityID=1-DD|MarkableID=dd_markable_535|Min=8,23|SemType=dn   _
9     :     _  _  _  _  _  _  _  _  _  _  _                                                               _
10    okay  _  _  _  _  _  _  _  _  _  _  _                                                               _
11    um    _  _  _  _  _  _  _  _  _  _  _                                                               _

# sent_id = D93_9_1-3
# text = I want to know how long alright how long does it take
12    I        _  _  _  _  _  _  _  _  _                                                                         _  _  _
13    want     _  _  _  _  _  _  _  _  _                                                                         _  _  _
14    to       _  _  _  _  _  _  _  _  _                                                                         _  _  _
15    know     _  _  _  _  _  _  _  _  _                                                                         _  _  _
16    how      _  _  _  _  _  _  _  _  (EntityID=1-Pseudo|MarkableID=markable_469|Min=16,17|SemType=quantifier   _  _  _  (MarkableID=markable_469|Entity_Type=unknown|Genericity=no-generic
17    long     _  _  _  _  _  _  _  _  )                                                                         _  _  _  )
18    alright  _  _  _  _  _  _  _  _  _                                                                         _  _  _
19    how      _  _  _  _  _  _  _  _  (EntityID=2-Pseudo|MarkableID=markable_470|Min=19,20|SemType=quantifier   _  _  _  (MarkableID=markable_470|Entity_Type=unknown|Genericity=no-generic
20    long     _  _  _  _  _  _  _  _  )                                                                         _  _  _  )
21    does     _  _  _  _  _  _  _  _  _                                                                         _  _  _
22    it       _  _  _  _  _  _  _  _  (EntityID=3-Pseudo|MarkableID=markable_6|Min=22|SemType=expletive)        _  _  _  (MarkableID=markable_6|Entity_Type=unknown|Genericity=no-generic)
23    take     _  _  _  _  _  _  _  _  _                                                                         _  )  _
```

https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/documents/UA_CONLL_U_Plus_proposal_v1.0.md

Identity Anaphora

Identity Column

(EntityID=10|MarkableID=markable_11|Min=5|SemType=do|ElementOf=23)

- Evaluating coreference relations only (e.g. in CoNLL 2012)
- Evaluating coreference relations and singletons (e.g. in CRAC 2018)
- Evaluating coreference relations (include split-antecedents) and singletons (e.g. in CODI-CRAC 2021,2022)

Supported metrics: MUC, B^3 , CEAF, BLANC, LEA

Split-antecedent anaphora

Identity Column

- Implements a new method proposed by Paun et al. (2022).
- Treating the antecedents of split-antecedent anaphors as a new type of mention: **accommodated sets**.

- E.g. [John]$_1$ met [Mary]$_2$. [He]$_1$ greeted [her]$_2$. [They]$_{1,2}$ went to the movies.

  - [He]1 ∈ Coref Chain 1 (John) = { [John], [He] }
  - [her]2 ∈ Coref Chain 2 (Mary) = { [Mary], [her] }
  - [They]1,2 ∈ Coref Chain 3 ({John,Mary}) = { {1,2}, [they] }

Split-antecedent anaphora

Identity Column

- Evaluation Steps
  - Identifies all accommodated sets in the key and response.
  - The relevant F1 scores are then calculated for pairs (key-response) of accommodated sets to create a similarity matrix between all accommodated sets in the key and response.
  - The KuhnMunkres algorithm (Kuhn, 1955; Munkres, 1957) is used to search for the best alignments.
  - The standard metrics are adjusted to allow partially matched mentions (i.e. the accommodated sets).

- The procedures for the standard mentions are unchanged, while for computation associated with accommodated sets, **partial credit** is awarded on how well the accommodated sets are resolved.

- Evaluating only split-antecedents
  - The micro-average F1 of all the split-antecedents in the key and response.

Non-referring expressions

Identity Column

(EntityID=4-**Pseudo**|MarkableID=markable_6|Min=17|SemType=predicate)

- E.g. [It] was late at night

- Follow CRAC 2018, non-referring expressions were scored separately.
- An F1 score is computed between the collection of non-referring expressions in the key and the response.

Discourse Deixis

Discourse Deixis Column

- Discourse deixis is similar to coreference
  - Both form clusters by linking the anaphors to their antecedents.
  - Both have split-antecedent anaphors that refer to multiple antecedents

- The main difference:
  - In coreference, antecedents are introduced using nominal phrases.
  - In discourse deixis they are introduced using non-nominal phrases (segments).

Discourse Deixis

Discourse Deixis Column

(EntityID=1-DD|MarkableID=dd_markable_2|Min=19,32|SemType=dn|ElementOf=6-DD)

- Discourse deixis is evaluated in the same way as entity anaphora;
  - Discourse deixis evaluation now works with predicted mentions/segments
  - By adopting the generalization of the standard identity reference metrics to split antecedents, we can use the scorer for the very common case of discourse deixis with more than one segment antecedent.

Bridging References

Bridging Column

(MarkableID=markable_9|Rel=subset|MentionAnchor=markable_3|EntityAnchor=3)

- [All doors in the town]_markable_1 ∈ Coref Chain 3 (All doors in the town) = { [All doors in the town]_markable_1, [those doors]_markable_3 }
- [The red doors]_markable_9

- Mention-based F1
- Entity-based F1
- Anaphora recognition F1

The CODI/CRAC 2021,2022 Shared Task

- The new UA scorer was used as the official scorer for the CODI-CRAC 2021,2022 shared task.
- Identity (Task 1)
  - coreference relations (including split-antecedents) and singletons
- Bridging (Task 2)
- Discourse Deixis (Task 3)

- The CODI/CRAC 2022 Shared Task with new datasets and additional gold anaphora phrases.
- https://codalab.lisn.upsaclay.fr/competitions/614
- https://aclanthology.org/2022.codi-crac.1/

CODI-CRAC 2021, 2022: Scoring

- **Identity Anaphora:** CONLL (the unweighted average of MUC, B3, CEAF)
  - `python ua-scorer.py key system`
- **Discourse Deixis:** Same as above
  - `python ua-scorer.py key system evaluate_discourse_deixis`
- **Bridging:** Entity F1
  - `python ua-scorer.py key system keep_bridging`

CODI-CRAC 2021, 2022: Baselines

- Baseline Systems
  - **Identity Anaphora** (Xu and Choi, 2020)
  - **Bridging** (Yu and Poesio 2020)
  - **Discourse Deixis (**Previous utterance [‘this’, ‘that’])

https://github.com/juntaoy/codi-crac2022_scripts

The CorefUD Scorer

The CorefUD Scorer

- used in the CRAC 2022 Shared Task on Multilingual CR (Žabokrtský et al., 2022)
- based on the UA 1.0 scorer
- parses the CorefUD 1.0 format
- supports discontinuous mentions
- supports zeros (must be pre-generated)
- adds two auxiliary measures (work in progress)

Discontinuous mentions in CorefUD

- present in several CorefUD datasets (Nedoluzhko et al., 2022)
  - **Společnost Nomura** rozjela **se společností American Express Co.** projekt …
    *[**The Nomura company** started a project **with the American Express company**]*
  - Aber **in der Mönchen-** und in der Zinnaer **Straße** …
    *[But **in the Mönchen-** and the Zinnaer **Street** …]*
  - … i **niestworzone** opowiadał **historie**.
    *[… and he was telling **unbelievable stories**.]*

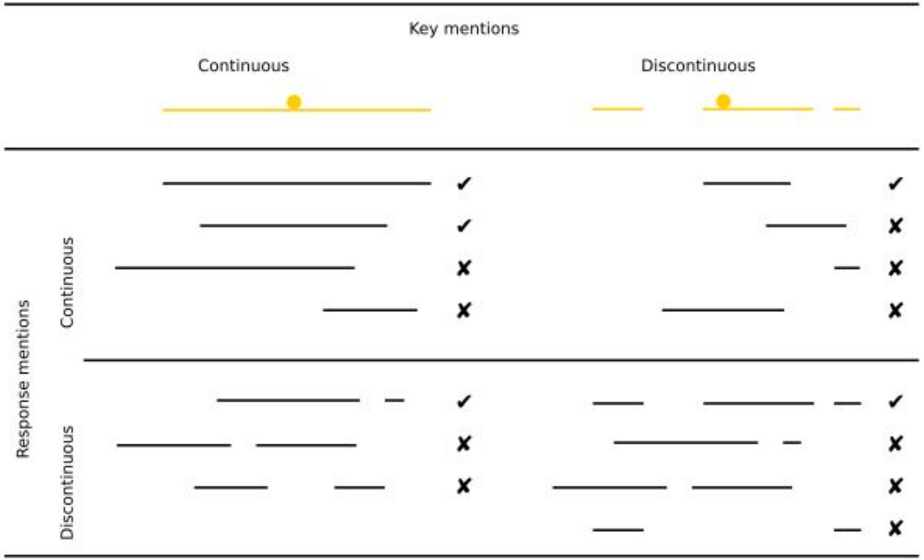
Dataset	**% disc**	**Dataset**	**% disc**	**Dataset**	**% disc**
ca-AnCora	0	de-ParCorFull	0.3	es-AnCora	0
cs-PCEDT	1.2	de-PotsdamCC	6.3	nl-COREA	0.3
cs-PDT	1.4	hu-SzegedKoref	0.4	en-ARRAU	1.2
en-GUM	0	lt-LCC	0	en-OntoNotes	0
en-ParCorFull	0.7	pl-PCC	1.1	en-PCEDT	1.45
fr-Democrat	0	ru-RuCor	0.5		

Discontinuous mentions in other scorers

- UA scorer 1.0 (and originally in the CoNLL scorer)
  - discontinuities not supported
  - partial matching based on head/MIN
- CRAFT shared task on CR in biomedical domain (Baumgartner et al., 2019)
  - discontinuities supported
  - partial matching based on overlap with the first part

Discontinuous mentions in CorefUD scorer

- discontinuities supported: matching defined in terms of set-subset relations
- partial matching based on head/MIN
- 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

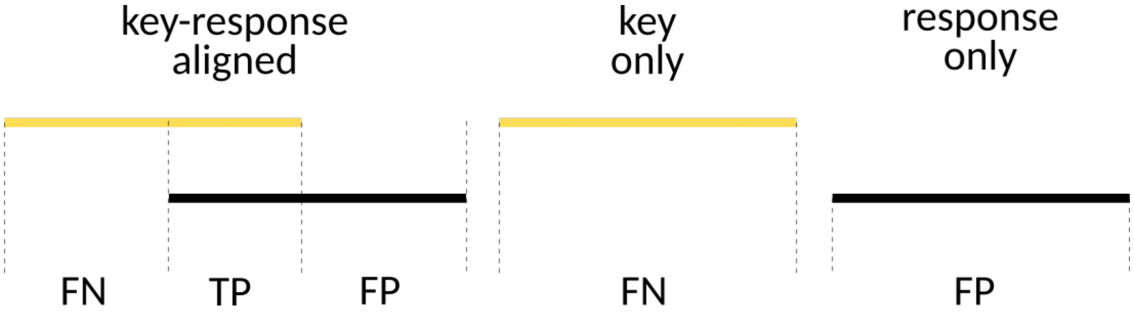


Mention Overlap Ratio

- score to calculate the ratio of mention overlapping, no matter to which entity a mention belongs

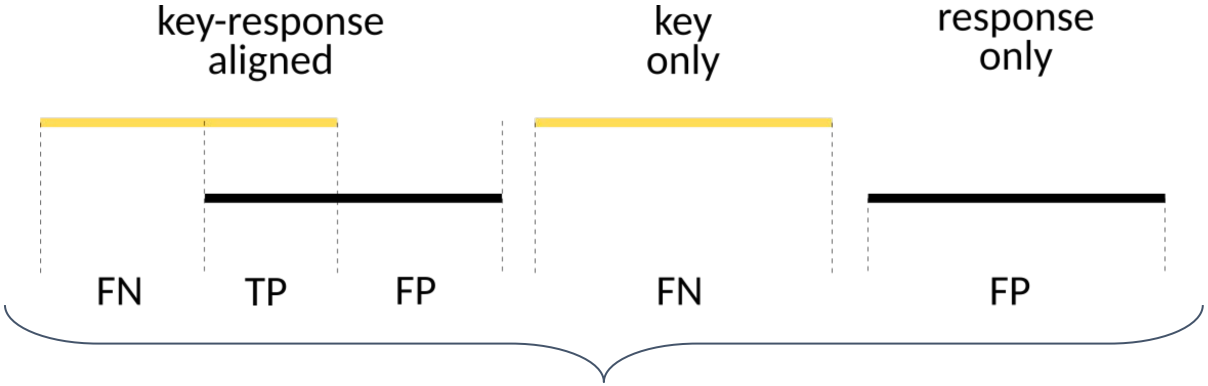
Mention Overlap Ratio

- score to calculate the ratio of mention overlapping, no matter to which entity a mention belongs



Mention Overlap Ratio

- score to calculate the ratio of mention overlapping, no matter to which entity a mention belongs



∑TP ∑FP ∑FN

Precision
Recall
F-score

Anaphor-decomposable scores

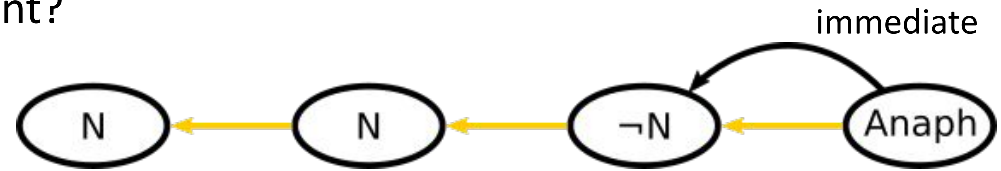
- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative

Anaphor-decomposable scores

- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?

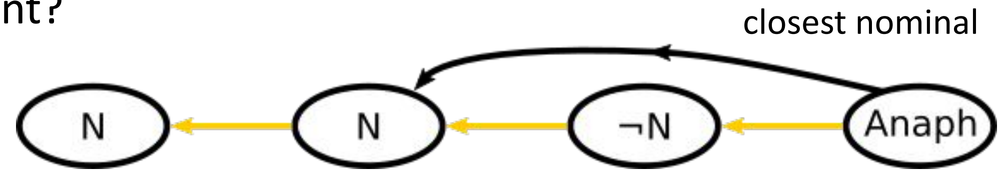
Anaphor-decomposable scores

- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?



Anaphor-decomposable scores

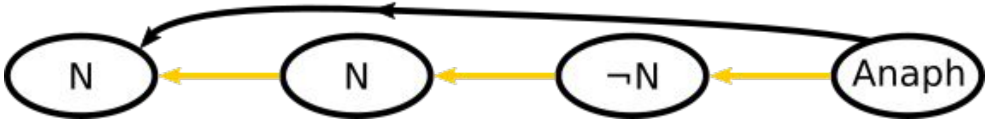
- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?



Anaphor-decomposable scores

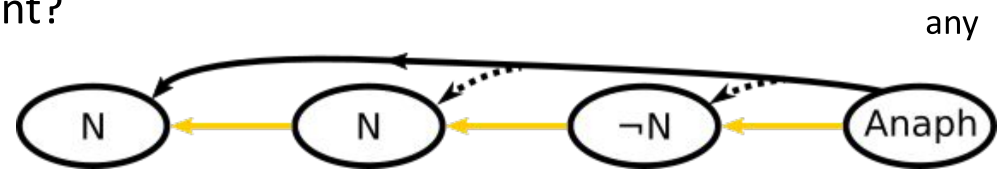
- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?

first nominal



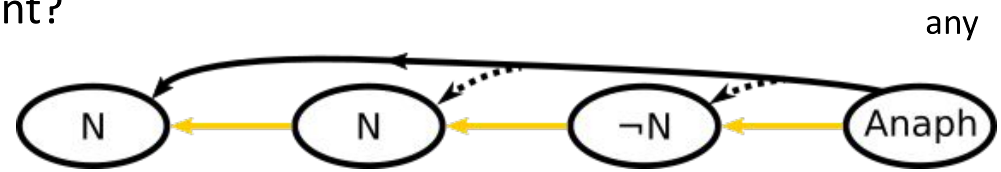
Anaphor-decomposable scores

- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?



Anaphor-decomposable scores

- standard coreference resolution measures
  - treat coreference as clustering
  - all mentions equally ⇒ hard to be calculated only for specific mention types
  - hard to interpret and insufficiently informative
- Application-Related Coreference Scores (ARCS; Tuggener 2014)
  - for each anaphor (no first mentions), we aggregate the following counts:
    - FP: anaphoric in the response, non-anaphoric in the key
    - FN: non-anaphoric in the response, anaphoric in the key
    - WL: anaphoric in both, but antecedent not correctly resolved
    - TP: correctly resolved antecedent
  - correctly resolved antecedent?
- in CorefUD scorer
  - any antecedent
  - only for zeros so far



The Universal Anaphora Scorer 2.0

UA-scorer 2.0

Merge features of UA-scorer 1.0 and CorefUD scorer and more

- Support both UA-exploded and UA-compact (CorefUD) format
- Identity Reference
- Singletons
- Split-antecedent anaphora
- Non-referring expression
- Discourse Deixis
- Bridging References
- Discontinuous markable
- Partial/Fuzzy Mention matching
- Mention matching score
- Anaphora level score
- CRAFT shared task support

Discontinuous mention in UA-Exploded format

- MarkableID

```
1    …   (EntityID=10|MarkableID=markable_11     …
2    …   )
3    …                                               …
4    …   (EntityID=10|MarkableID=markable_11)    …
```

→

Markable_11 = [[1,2], [4,4]]

Partial Boundary Overlap in Exploded format

```
1    …   (EntityID=10|MarkableID=markable_11      …
2    …   (EntityID=11|MarkableID=markable_12      …
3    …   )                                        …
4    …   )                                        …
```

markable_11 = [1,3] ⇒ [1,4]

markable_12 = [2,4] ⇒ [2,3]

Partial Boundary Overlap in Exploded format

```
1    …   (EntityID=10|MarkableID=markable_11     …
2    …   (EntityID=11|MarkableID=markable_12     …
3    …   markable_11)                            …
4    …   markable_12)                            …
```

markable_11 = [1,3]

markable_12 = [2,4]

```
python ua-scorer.py key system allow_boundary_crossing
```

Fuzzy mention matching

- CorefUD (default)

  -
    ```
    python ua-scorer.py key system min
    ```

- CRAFT

  -
    ```
    python ua-scorer.py key system craft
    ```

CRAFT 2019 shared task: https://aclanthology.org/D19-5725.pdf

Code

- The code is available from our GitHub pages:
- https://github.com/juntaoy/universal-anaphora-scorer
- https://github.com/ufal/corefud-scorer