Universal Anaphora 1.0 - Proposal for Discussion

Massimo Poesio Amir Zeldes Anna Nedoluzhko
Sopan Khosla Ramesh Manuvinakurike Nafise Moosavi
Vincent Ng Maciej Ogrodniczuk Sameer Pradhan
Carolyn Rose Michael Strube Juntao Yu Yulia Grishina
Yufang Hou Fred Landragin

8th March 2021

1 Introduction

Thanks to recent advances in NLP and representation learning, interest in coreference resolution and related tasks is at a high point. At the same time, large scale resources are still only available for few languages, and languages with multiple resources (notably English) do not employ a uniform standard, neither in terms of annotation formats, nor in terms of guidelines. These limitations need addressing as we move towards broader coverage of languages and text types, as well as work on multilingual and reusable tools for different types of anaphora.

The Universal Dependencies initiative has been very successful at progressively developing agreed upon standards concerning the annotation and markup of syntactic dependency information across multiple languages and domains, showing a blueprint for a way forward. The long-term objective of the UD-inspired **Universal Anaphora** initiative proposed here is to do the same for anaphoric information.

The 2011 and 2012 CONLL shared tasks (Pradhan et al., 2012) provided a solid foundation for further research in the field by developing a markup scheme that crystallized the aspects of the anaphora resolution task on which there was most agreement – identity coreference – while providing a reference scorer (Pradhan et al., 2014) that has also greatly contributed to the explosion of research in this area. The objective of the UA initiative is to enable further progress in the empirical study of anaphora by covering not just coreference, but all aspects of anaphoric interpretation from identity of sense anaphora to bridging to discourse deixis (although not all anaphori-

cally annotated corpora cover all of these phenomena); and not just English, but all languages.

The more modest objective of this stage 1.0 of the initiative is to come up with an agreed-upon **markup scheme** that can be used to encode the information in the existing corpora, which will enable us to create a collection of corpora all encoded using the same scheme. But we hope that this exercise will prove a useful starting point for further discussion on the annotation schemes, as well. A unified markup scheme hopefully will also allow us to develop an extension of the CONLL scorer able to score not just identity anaphora, but also other aspects of anaphoric interpretation such as the identification of non-referring expressions, as done in the 2018 CRAC Shared Task, as well as bridging reference and discourse deixis resolution.

We are fully aware that there is at the moment only partial agreement on the anaphoric phenomena that should be covered by such a scheme, and on the details of how they should be annotated. We will therefore adopt a similar strategy to that adopted in the MATE proposal (Poesio et al., 1999)– namely, identify those aspects of the proposal which are **core** and those that are **optional**. We will also (attempt to) distinguish between **mandatory** aspects of the proposal which all datasets consistent with the initiative are expected to agree to, and **optional** ones.

Comment [AMIR1]: Isn't discourse deix. a subtype of identity coref? if a group has multiple NP and non-NP mentions, we'd need to have both types in one set, no?

2 Key Ingredients of the UA 1.0 Proposal

The key ingredients of the proposal are as follows:

1. A specification of the phenomena to be covered
2. The markup format
3. The definition of markable
4. The Core Anaphoric Layers
5. The Additional Anaphoric and Reference Layers
6. The Non-Anaphoric Layers
7. The Universal Anaphora Scorer

Comment [AMIR2]: Other possibilities: search across corpora like `http://match.grew.fr/` for UD, format converters

3 The anaphoric phenomena to be covered

This proposal is designed (i) to cover all aspects of anaphoric information currently annotated in existing projects (Poesio et al., 2016) (ii) identifying some of these aspects as required, but (iii) without requiring all projects to annotate all of this information, and (iv) leaving room for future extensions (e.g., to cover ellipsis, or domain restriction).

Core anaphoric phenomena Most modern anaphoric annotation projects (i) cover basic identity anaphora as in (3), and (ii) distinguish identity anaphora from predication as in (3).

(1) [Mary]$_i$ bought [a new dress]$_j$ but [it]$_j$ didn't fit [her]$_i$.

(2) [Mary]$_i$ is [a teacher].

We would say that marking relations such as (i) is a minimum requirement for a project to qualify as an anaphoric annotation project. One question we need to tackle in the discussion is whether we also want to require 'UA compatible' datasets to satisfy (ii) (in their current forms, this would exclude, for instance, the English PRECO corpus (Chen et al., 2018), GUM (Zeldes, 2017), and several other language corpora, such as PCC (Stede, 2004) for German).

amir: I have some doubts about distinguishing pred. coref – some corpora include it without distinction (GUM), other don't at all (ON), and a few distinguish the two – I added a few to the list above. There are rather ambiguous cases, including "as" PPs, adverbial NPs without copula: "A teacher since 2000 herself, these events challenged Mary's resolve" and more. After seeing a bunch of bad cases like these in GUM, we decided to collapse i and ii based on set-theoretical identity as the coref criterion.

massimo: Oh right, I didn't know that GUM didn't distinguish between identity and predication (NB: ONTONOTES does distinguish between identity and predication altough it marks both, same as *Phrase Detectives*). I agree that distinguishing between identity and predication is not always easy, and I'm certainly not advocating a uniform treatment of, e.g., copular clauses as predicative–e.g., in ARRAU we would treat *The Morning Star is Venus* as a case of identity–but semantically those are really distinct relations, and merging them leads to the problems discussed by van Deemter and Kibble (2000)–how do you handle those in GUM?

Most modern annotation projects also do not impose *semantic* restrictions on the type of references to be marked–i.e., references to all objects are considered, unlike earlier efforts such as ACE, or the limited restriction in ONTONOTES on coref for noun modifiers which are not named entities. Again, a question to be discussed is whether we want to require all UA-included efforts to cover identity relations between all objects.

(In ONTONOTES, ARRAU, GUM and other projects, the **semantic category** of every markable is specified. We believe that an optional layer for this category should be included in the markup format, though it is not a core requirement, see Section 8.)

Zero anaphora as in (3) is one of the most common forms of anaphoric reference in languages with unrealized argument such as Arabic, Chinese, Italian, or Japanese.

(3) [IT] Giovanni è in ritardo, così mi ha chiesto se posso incontrarlo al cinema.
[EN] John is late so he asked me if I can meet him at the movies.
(Poesio et al., 2016, ex. 9, p. 29)

Zero anaphora is annotated in Arabic and Chinese ONTONOTES as in (3), using an asterisk * to indicate the position of the empty category:

(4) TRT . . . * . . .
Alhariri's statement included more details ...in which (he) emphasized that the council of ministers of Lebanon is the only representative ...

Zeros are annotated in other corpora for Chinese as well as the corpus for Catalan and Spanish, the LIVEMEMORIES corpus for Italian (Rodriguez et al., 2010), and the NAIST corpus for Japanese, among others, but using different markup formats–for instance, in LIVEMEMORIES, instead of adding * to the text, verbs are used as markables. One of the issues to be discussed is whether we can come up with a uniform way of representing these markables.

Minimally beyond the core In ONTONOTES, event anaphora, a subtype of **discourse deixis** Webber (1991); Kolhatkar et al. (2018) is marked. We propose that there should be an (optional) discourse deixis layer (see Section 7), and that this layer should be used for event anaphora.

amir: Why not just make this part of regular coref? What would you do if you have two NPs and 1 VP in a group? You could use an added feature to indicate 'discourse deixis'

Comment [AMIR3]: I added the ON case since I think it is also semantic.

Comment [AMIR4]: an important consideration is alignability to existing data when UA is added to a corpus, which biases me to prefer pro-dropped verbs as markables, rather than adding empty tokens

Comment [MASSIMO5]: I com-

massimo: Basically, it's a pragmatic decision - yes, discourse deixis can be viewed as a type of identity reference, but (i) linguistically, it requires the introduction of a new object in the discourse model, so it also resembles normal deixis (ii) computationally, very few systems have the ability to resolve such cases, so keeping this type of references in a separate layer (or mark them using an extra feature) in such a way that systems can be evaluated on the ability to resolve this type of anaphors separately seems like a good idea–this is what ONTONOTES does I believe?

One often critized aspect of ONTONOTES is that non-referring nominals are not marked except for predicative NPs.

amir: AFAIK ON doesn't include predicative NP coref either

massimo: The appositions and copulas are predicative NPs, and they are marked in a special way

We propose the scheme should include a layer storing the information contained in the `reference` attribute in ARRAU–i.e., specifying whether a nominal is referring, or an expletive, or predicative, or a quantifier (see Section 6). I would suggest to call this layer `Sem_Type`. It has to be decided whether this layer should be mandatory or optional.

amir: I think it has to be optional since most corpora don't have it

massimo: OK with me

Most anaphorically annotated corpora are focused on written text, although they typically also include spoken monologue. But more and more interest is being paid to coreference in **spoken dialogue**, for example in GUM. A desirable option for nominal annotation in dialogue is some way for handling deixis (*you*, *we*) and references to the visual situation. A more thorough treatment would also include some provision for dealing with discontinuous markables, which are present both in TRAINS and SWITCHBOARD (see Section 4).

In ONTONOTES, **identity of sense**, as in *one*-anaphora, is not marked (see example (7)). In ARRAU and in GUM, *one*-anaphora is marked as a type of bridging reference. These references could be specified in the (optional) bridging reference layer.

(5) Five and Seven said nothing, but looked at Two. Two began, in a low voice, "Why, the fact is, you see, Miss, this here ought to have been [a *red* rose-tree], and we put [a white one] in by mistake; and, if the Queen was to find it out, we should all have our heads cut off, you know. So you see, Miss, we're doing our best, afore she comes,

Comment [AMIR6]: there are a few types of these, incl. "another", "the other"...

Comment [MASSIMO7]: Indeed ....

to–" At this moment, Five, who had been anxiously looking across the garden, called out, "The Queen! The Queen!" and the three gardeners instantly threw themselves flat upon their faces. There was a sound of many footsteps and Alice looked 'round, eager to see the Queen. (From *Alice in Wonderland*, *Phrase Detectives* corpus.)

In ONTONOTES, **plural reference** is annotated when it's a reference to an antecedent introduced via a plural (as in (6a)) or via a conjunction (as in (6b)) but not in cases of **split-antecedent** reference as in (6c) and (6d). These references are annotated in ARRAU, in GUM and the *Phrase Detectives* corpus.

(6) a. 'And who are THESE?' said the Queen, pointing to [the three gardeners who were lying round the rose tree]$_i$; for, you see, as [they]$_i$ were lying on their faces, and the pattern on their backs was the same as the rest of the pack, she could not tell whether they were gardeners, or soldiers, or courtiers, or three of her own children.;

b. She could hear the rattle of the teacups as [[the March Hare] and [his friends]]$_i$ shared [their]$_i$ never-ending meal.

c. 'In THAT direction,' the Cat said, waving its right paw round, 'lives [a Hatter]$_i$: and in THAT direction,' waving the other paw, 'lives [a March Hare]$_j$. Visit either you like: [they]$_{i,j}$'re both mad.' (From *Alice in Wonderland*, *Phrase Detectives* corpus.)

d. Alice had no idea what [Latitude]$_i$ was, or [Longitude]$_j$ either, but thought [they]$_{i,j}$ were nice grand words to say.

A mixed approach to **reference to generic expressions** is adopted in ONTONOTES. Generic reference using pronouns and nominals to generic antecedents, as in (7a) or (7b), is annotated; but generic reference via bare plurals, as in (7c), is not.

(7) a. (12) [Officials]$_i$ said [they]$_i$ are tired of making the same statements.

b. (13) [Meetings]$_i$ are most productive when [they]$_i$ are held in the morning. [Those meetings]$_i$, however, generally have the worst attendance.

c. (14) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for [*cataract surgery]. The lens' foldability enables it to be inserted in smaller incisions than are now possible for [*cataract surgery].

In ARRAU and in GUM, all of these types of coreference are uniformly annotated; in addition, identity between references to kinds expressed as bare

nominals in premodifier position, as in (7), is also allowed in both corpora, whereas in ONTONOTES it is only allowed for proper nouns (e.g. ***Hong Kong*** *government* can contain a reference to ***Hong Kong***).

(8) Even the volatility created by [stock$_i$ index arbitrage and other computer-driven trading strategies isn't entirely bad, in Mr. Connolly's view. For the long-term investor who picks [stocks$_i$ carefully, the price volatility can provide welcome buying opportunities as short-term players scramble frantically to sell [stocks$_i$ in a matter of minutes.

One issue to be discussed is the extent to which such discrepancies can / should be reconciled.

Further beyond the core There are other forms of anaphoric reference besides identity, and there are now a number of corpora annotating (a subset of) these forms.

Possibly the most studied of these types of anaphora is **bridging reference** or **associative anaphora** (7) (Clark, 1977; Hawkins, 1978; Prince, 1981), which we also take to cover *other* anaphora as in (7).

(9) There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in [a long, low hall, which was lit up by a row of lamps hanging from [the roof]].

There were doors all round the hall, but they were all locked; and when Alice had been all the way down one side and up the other, trying every door, she walked sadly down [the middle], wondering how she was ever to get out again.

(10) There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in a long, low hall, which was lit up by a row of lamps hanging from the roof.

There were doors all round the hall, but they were all locked; and when Alice had been all the way down [one side] and up [the other], trying every door, she walked sadly down the middle, wondering how she was ever to get out again.

There are discrepancies between the various annotation projects not only on the subset of associative relations to mark, but also on the definition of the phenomenon - e.g. whether to annotate references via non-anaphoric expressions, as in (7), which is annotated according to a cohesion-based view of bridging (GNOME, ARRAU, GUM) but not in a lexically-based view according to which bridging is licensed by the existence of an implicit anaphor in the representation of the nominal, as in ISNOTES (see, e.g., Roesiger et al. (2018); Yu and Poesio (2020) for discussion).

(11) I visited [Spain] this summer. I particularly enjoyed [Madrid].

In ONTONOTES **event anaphora** is annotated, as in (7), but not full **discourse deixis**, as in (7).

(12) So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, 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); ...

(13) 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); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

Discourse deixis has been annotated in ARRAU, but doing so requires minimally to include a clausal layer. In GUM it is annotated in the same way as identity coreference, and is identified by markables having a non-nominal head.

amir: see note above: I'm not sure why this needs a separate layer, and don't know what you'd do if you have a longer mixed nominal/verbal chain

massimo: see above: I would separate evaluation on discourse deixis from evaluation on identity anaphora to nominally introduced antecedents in part for pragmatic reasons, in part because linguistically DD is in between IA and normal ('situational') deixis

Finally–in ARRAU and other corpora such as *Phrase Detectives* it is recognized that certain anaphoric expressions are **ambiguous**. For example, Poesio et al. (2007) discussed the cases of so-called **justified sloppiness** in anaphoric reference illustrated in example (7), encountered during the annotation of the TRAINS corpus of task-oriented dialogues collected at the University of Rochester (Heeman and Allen, 1995):

(14)
```
 3.1    M: can we .. kindly hook up
3.2     : uh
3.3     : engine E2 to the boxcar at ..
          Elmira
4.1    S: ok
5.1    M: +and+ send it to Corning
5.2     : as soon as possible please
6.1    S: okay
        [2sec]
7.1    M: do let me know when it gets
          there
8.1    S: okay it'll /
8.2     : it should get there at 2 AM
9.1    M: great
9.2     : uh can you give the
9.3     : manager at Corning instructions
          that
9.4     : as soon as it arrives
9.5     : it should be filled with
          oranges
10.1   S: okay
10.2    : then we can get that filled
```

In this example, it is not clear whether the pronoun *it* in 5.1 refers to *the engine E2* which has been hooked up to *the boxcar at Elmira*, to the boxcar itself, or indeed whether that matters. It's only at utterance 9.5 that we get evidence that *it* probably referred to *the boxcar at Elmira*, since it is only boxcars that can be filled with oranges. Evidence that subjects disagree on such cases was discussed, e.g., in (Poesio and Artstein, 2005; Poesio et al., 2006), and similar cases of disagreements have been found in

all large-scale anaphoric annotation projects (Versley, 2008; Recasens et al., 2011; Pradhan et al., 2012). The question is how to encode these multiple interpretations. One option would be to do as done in ARRAU and just have one duplicate layer to store the second interpretation in case of ambiguous anaphoric expressions. (This would presumably make it simple for systems who can't use multiple interpretations to ignore the second interpretation.) A second possibility would be to encode alternative interpretations directly in the layers–e.g., use semicolon-separated values in the identity layer to store alternative interpretations. This would be more flexible but could get messy quickly.

4 Markable definition

We mostly seem to agree on considering as markables for English:

- All NPs, whether referring or non-referring, and whether singletons or not;
- Possessive pronouns / determiners

This definition would need to be extended for languages other than English to cover

- Some markables for annotating zeros (e.g., verbs);
- Some markables for clitics (e.g., work off lemma layer / the dependency layer)

In addition, we may want to allow

- Some nominal modifiers for genericity
- Discontinuous markables for dialogue

amir: I'd leave this for a V2.0, makes integration with UD harder

massimo: We will need to do something about this for the CODI shared task, both TRAINS and Switchboard have these

Comment [AMIR8]: in GUM non-ref. NPs are unannotated, though findable via syntax trees

Comment [MASSIMO9]: would it be possible to automatically produce an 'extended version' of GUM in which the additional markables are extracted from the syntax tree, and then the `Sem_Type` attribute is used to mark these additional NPs as expletives? (Is it just expletives?)

Comment [AMIR10]: I'd prefer to handle it like

For further discussion

- Which layer to use to identify markables: a Universal Dependency layer (which makes it necessary to have one), or the lemma

  amir: you can use the UD conllu format and leave deprel cols empty!

  massimo: I am happy to agree on using the UD conllu format, but we still need to specify which layer we attach the anaphoric annotations to?

- How to mark empty categories
- Whether to allow discontinuous markables
- Whether to annotate MIN and how to define it

5 Markup format

The markup format will be an extension of the CONLL-U tabular format defined for Universal Dependencies.[1] Two versions of the format will be defined:

- A 'compact' version (UA-COMPACT) with all anaphoric information stored in the `MISC` column of the basic CONLL-U format. This format will be used for documents to be included in the UD distribution. Two proposals in this direction have been put forward, by Amir Zeldes and by the Prague group.
- An 'exploded' version (UA-EXPLODED) in 'CONLL-U Plus' format, i.e, with additional columns for the separate layers discussed in the following Section. The format for these columns will be based on the format defined for CONLL and extended in CRAC. This is the format used by the current version of the Universal Anaphora scorer and for the 2021 CODI/CRAC shared task.

The intention is for both formats to contain the same information specified by the layers, and to be mutually convertible.

[1] `https://universaldependencies.org/format.html`

6 The Core Anaphoric Layers

These layers are used to encode the core anaphoric information. They will be stored in the `Misc` layer in UA-COMPACT, and in separate columns in UA-EXPLODED.

Mandatory layers:

- **Identity**: identity coreference relations, encoded using a CONLL-style format according to which the entity (coreference chain) to which referring NPs refer is specified.

Optional layers:

- **MIN** (= head +)
- **Sem_Type** (using a variant of the ARRAU scheme, with values **expletive**, **dn**, **do**, **predicate**, **quantifier**, **coord**, **other-nonref**)

For further discussion:

1. Until we standardize markable definition and guidelines regarding which types of identity anaphora to annotate, very different results are going to be obtained when analyzing / training with the different corpora, but this will have to be left for UA 2.0
2. How to encode the reference to entities - CONLL style or SEMEVAL / EVALITA / CRAC style?
3. The **MIN** layer is likely to mostly contain the same information as the **HEAD** layer in CONLL-U (see below), but for the moment we keep it as a separate layer as there are cases in which the head specification adopted in UD doesn't seem to be appropriate for anaphoric reference (e.g., in the case of coordination). Also, **MIN** is specified as optional following Amir's suggestion.

7 The Additional Anaphoric and Reference Layers

These layers store information about additional anaphoric phenomena. Note that some of these layers provide information that arguably could be considered to be about coreference–clearly so in the case of split-antecedent

anaphora, whereas discourse deixis could also be considered a case of deictic reference. All of these layers are optional as very few corpora provide all of this information.

Mandatory layers: none

Optional layers:

- **split**, for split antecedent plurals, specifying for each split antecedent the set they belong to.
- **bridging** (Associative Anaphora)
- **discourse_deixis**
- **deixis** This layer would be used to store information about deictic references to objects in the visual situation for multimodal reference domains, as well as for entity links to outside knowledge bases (e.g., Wikipedia)

For further discussion:

1. For bridging we can certainly find a syntactic format which can be used to encode annotations obtained through different guidelines, but can we find a way to encode the different notions of bridging used e.g., in ARRAU and in ISNOTES / BASHI (Roesiger et al., 2018)?
2. Are we all happy to use a single layer to encode both deictic reference and entity linking?

8 The Non-Anaphoric Layers

The proposed UA format builds on the CONLL-U format from UD, which includes ways to specify all the required non-anaphoric information from the CONLL format.

Mandatory layers: (these are mostly from CONLL):

- **newdoc** (corresponding to the **Doc-ID** layer in the CONLL format)
- **sent_id** (corresponding to the **Sentence-ID** layer in the CONLL format)

- **ID** for token identifier (corresponding to the **Token-ID** layer in CONLL but with provision for multi-word tokens)
- **FORM** (corresponding to the **Token** layer in CONLL format)

Optional layers: (many of these also from CONLL)

- Additional information from CONLL-U:
  - **newpar** for paragraph beginning
  - **text** and **text_en** containing untokenized version of a sentence text
- Linguistic layers from CONLL-U:
  - **LEMMA**
  - **UPOS**
  - **XPOS**
  - **FEATS**, specifying morpho syntactic features such as agreement features–gender, number, etc - corresponding to the **MORPHOSYN** layer in CONLL but using the Universal Features format.[2]
  - **HEAD** the head in the dependency sense.
  - **DEPREL** universal dependency relation to the head.
- Additional linguistic layers not in UD:
  - **CONSTITUENCY** following the Penn Treebank format adopted in CONLL
  - **WORDSENSE** following the CONLL format
  - **PROPOSITION**, following the PROPBANK format adopted in CONLL
  - **Nom_Sem** layer providing additional semantic information about nominals not contained in either **FEATS** or the other layers, including
    - **entity type** information (type of object–an extended version of the set of categories used for named entities)
    - **genericity**
    - **functionality** in the sense of Loebner as in GNOME
  - **rhetorical structure** layers (e.g., RST, PDTB) following e.g., the format in GUM

[2] `https://universaldependencies.org/u/feat/index.html`

For further discussion:

1. Should the lemma layer be used as base layer for the anaphoric layers instead of the token layer? Or perhaps we should just use the dependency layer? (Whichever we use should be made mandatory.)

   amir: I would follow conllu here again, define markables using the (sub)tokens, and use the conllu lemma col, which can be separate from orthographic supertokens, see `https://universaldependencies.org/format.html`

9 The Universal Anaphora Scorer

The Universal Anaphora (UA) scorer is a Python scorer for the varieties of anaphoric reference covered by the Universal Anaphora guidelines, which include identity reference, split antecedent plurals, identification of non-referring expressions, bridging reference, and discourse deixis. The scorer builds on the original Reference Coreference scorer[3] Pradhan et al. (2014) developed for scoring the CONLL 2011 and 2012 shared tasks using the ONTONOTES corpus (Pradhan et al., 2011, 2012) and its reimplementation in Python by Moosavi,[4] also extended to compute the LEA score Moosavi and Strube (2016) and to evaluate non-referring expressions evaluation and cover singletons for the 2018 CRAC shared task Poesio et al. (2018). The scorer reports scores for Identity reference (with and without singletons and non-referring expressions–in the modality without singletons and non referring expressions the scorer is compatible with the original Coreference Reference scorer–split antecedents, bridging reference, and discourse deixis. For Identity reference it reports the MUC Vilain (1995), B-cubed Bagga and Baldwin (1998), CEAF Luo (2005), averaged CONLL Denis and Baldridge (2009); Pradhan et al. (2014), BLANC Recasens and Hovy (2011), and LEA Moosavi and Strube (2016) scores. The same scores are also computed for discourse deixis, that is treated as a generalized case of event coreference. For split antecedents, a generalization of these metrics to split antecedents due to Paun, Yu *et al* was developed. Accuracy is computed for bridging.

[3] `https://github.com/conll/reference-coreference-scorers`
[4] `https://github.com/ns-moosavi/LEA-coreference-scorer`

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proc. of the LREC workshop on Linguistic Coreference*, pages 563–566, Granada.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. http://www.aclweb.org/anthology/D18-1016 PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of EMNLP*, pages 172–181, Brussels, Belgium.

Herbert H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, London and New York.

K. van Deemter and R. Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637. Squib.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96.

J. A. Hawkins. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

Peter A. Heeman and James F. Allen. 1995. The TRAINS-93 dialogues. TRAINS Technical Note TN 94-2, University of Rochester, Dept. of Computer Science, Rochester, NY.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. https://doi.org/10.1162/coli_a_00327 Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. NAACL / EMNLP*, Vancouver.

N. S. Moosavi and M. Strube. 2016. https://doi.org/10.18653/v1/P16-1060 A proposal for a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin.

Massimo Poesio and Ron Artstein. 2005. http://www.aclweb.org/anthology/W05-0311 The reliability of anaphoric

annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabien Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. https://doi.org/10.18653/v1/W18-0702 Anaphora resolution with the ARRAU corpus. In *Proc. of the NAACL Worskhop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 11–22, New Orleans.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 4. Springer.

Massimo Poesio, Uwe Reyle, and Rosemary Stevenson. 2007. Justified sloppiness in anaphoric reference. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3, pages 11–34. Kluwer.

Massimo Poesio, Patrick Sturt, Ron Arstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42(2):157–175.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. https://doi.org/10.3115/v1/P14-2006 Scoring coreference partitions of predicted mentions: A reference implementation. In *Proc. of the ACL*, pages 30–35, Baltimore.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proc. of the CONLL Shared Task*, Jeju, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. http://www.aclweb.org/anthology/W11-1901 Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proc. of the 15th*

CONLL: Shared Task, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Marta Recasens and Ed Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

Marta Recasens, Ed Hovy, and M. Antonia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

K-J. Rodriguez, F. Delogu, Y. Versley, E. Stemle, and M. Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proc. LREC (poster)*.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. https://www.aclweb.org/anthology/C18-1298 Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manfred Stede. 2004. The potsdam commentary corpus. In *Proceeding of the ACL 2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

John; Aberdeen John; Connolly Dennis; Hirschman Lynette Vilain, Marc; Burger. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the Sixth Message Understanding Conference*, pages 45–52.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proc. of COLING*.

Amir Zeldes. 2017. https://doi.org/http://dx.doi.org/10.1007/s10579-016-9343-x The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.