Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4873–4883
Marseille, 20-25 June 2022

The Universal Anaphora Scorer

Juntao Yu[1], Sopan Khosla[2,3,*], Nafise Sadat Moosavi[4], Silviu Paun[5], Sameer Pradhan[6,7] and **Massimo Poesio[5]**
[1]Univ. of Essex, UK; [2]Carnegie Mellon Univ., USA; [3]AWS AI, Amazon, USA; [4]Univ. of Sheffield, UK;
[5]Queen Mary Univ., UK; [6]LDC, Univ. of Pennsylvania, USA; [7]cemantix.org
j.yu@essex.ac.uk; sopankh@amazon.com;
m.poesio@qmul.ac.uk

Abstract

The aim of the Universal Anaphora initiative is to push forward the state of the art in anaphora and anaphora resolution by expanding the aspects of anaphoric interpretation which are or can be reliably annotated in anaphoric corpora, producing unified standards to annotate and encode these annotations, deliver datasets encoded according to these standards, and developing methods for evaluating models carrying out this type of interpretation. Such expansion of the scope of anaphora resolution requires a comparable expansion of the scope of the scorers used to evaluate this work. In this paper, we introduce an extended version of the Reference Coreference Scorer (Pradhan et al., 2014) that can be used to evaluate the extended range of anaphoric interpretation included in the current Universal Anaphora proposal. The UA scorer supports the evaluation of identity anaphora resolution and of bridging reference resolution, for which scorers already existed but not integrated in a single package. It also supports the evaluation of split antecedent anaphora and discourse deixis, for which no tools existed. The proposed approach to the evaluation of split antecedent anaphora is entirely novel; the proposed approach to the evaluation of discourse deixis leverages the encoding of discourse deixis proposed in Universal Anaphora to enable the use for discourse deixis of the same metrics already used for identity anaphora. The scorer was tested in the recent CODI-CRAC 2021 Shared Task on Anaphora Resolution in Dialogues.

Keywords: Anaphora Resolution, Evaluation, Univeral Anaphora

1. Introduction

The performance of models for single-antecedent anaphora resolution on the aspects of anaphoric interpretation annotated in the reference ONTONOTES dataset (Pradhan et al., 2012) has greatly improved in recent years (Wiseman et al., 2015; Lee et al., 2017; Lee et al., 2018; Kantor and Globerson, 2019; Joshi et al., 2020). So the attention of the community has started to turn to more complex cases of anaphora not represented in ONTONOTES.

Well-known examples of this trend are work on the cases of anaphora whose interpretation requires some form of commonsense knowledge tested by benchmarks for the Winograd Schema Challenge (Rahman and Ng, 2012; Liu et al., 2017; Sakaguchi et al., 2020), or the pronominal anaphors that cannot be resolved purely using gender, for which benchmarks such as GAP have been developed (Webster et al., 2018). Another fruitful line of research has been devoted to creating datasets covering genres other than news, such as conversation (Muzerelle et al., 2014; Uryupina et al., 2020; Khosla et al., 2021), fiction (Bamman et al., 2020) or scientific articles (Cohen et al., 2017).

Further research has been carried out on aspects of anaphoric interpretation that go beyond identity anaphora but are covered by datasets such as ARRAU (Poesio et al., 2018; Uryupina et al., 2020) and GUM (Zeldes, 2017) for English, the Prague Dependency Treebank (Nedoluzhko, 2013) for Czech, and ANCORA for Catalan and Spanish (Recasens and Martí, 2010). These include, e.g., bridging reference (Clark, 1977; Hou et al., 2018; Hou, 2020; Yu and Poesio, 2020; Kobayashi and Ng, 2021), discourse deixis (Webber, 1991; Marasović et al., 2017; Kolhatkar et al., 2018) or split-antecedent anaphora (Eschenbach et al., 1989; Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2020; Yu et al., 2021).

The objective of the Universal Anaphora initiative, or UA,[1] is to coordinate these efforts to push forward the state of the art in anaphora research. The initiative, modelled on Universal Dependencies,[2] aims to achieve this by expanding the aspects of anaphoric interpretation which are or can be reliably annotated in anaphoric corpora, producing unified standards to annotate and encode these annotations, delivering datasets encoded according to these standards, and developing methods for evaluating this type of interpretation.

Like Universal Dependencies, Universal Anaphora is meant to push forward the state of the art in anaphora both from the linguistic and from the NLP perspective. One key issue in this last regard is how to assess a system's ability to carry out these more advanced types of anaphoric interpretation. This requires a scorer that can evaluate the interpretation produced by a system for, e.g., bridging reference, discourse deixis, or split-antecedent plurals. Partial scorers exist and have been used, e.g., in the 2018 CRAC Shared Task (Poesio et al., 2018). However, no standardized scorer exists for many of these aspects of anaphoric interpreta-

* Work done when the author was a student at CMU

[1]http://www.universalanaphora.org
[2]https://universaldependencies.org/

tion, and the partial scorers suffer from a number of limitations. Also, the solutions proposed for some of these aspects of interpretation, such as split-antecedent anaphors (Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2020) are not entirely satisfactory, even though these are cases of identity anaphora after all.

In this paper we present the new Universal Anaphora scorer for anaphoric interpretation, a Python extension of the Reference Coreference Scorer (Pradhan et al., 2014) and of the Generalized Coreference Scorer developed by Moosavi for the CRAC 2018 Shared Task.[3] The UA scorer is the first scorer able to evaluate system performance in all aspects of anaphoric interpretation covered by the current version of the Universal Anaphora proposal. This scorer was used in the CODI-CRAC 2021 Shared Task in Anaphora Resolution in Dialogue[4] (Khosla et al., 2021) and will be used in the forthcoming 2022 edition of the shared task.[5]

We begin with some background on the Universal Anaphora initiative in Section 2. Next, we discuss research on scoring anaphora in Section 3. The new scorer is presented in Section 4. In Section 5, we discuss the CODI-CRAC 2021 shared task. In Section 6, we analyze its results, with a focus on issues related to the scoring.

2. The Universal Anaphora Initiative

The **Universal Anaphora** (UA) initiative was launched in 2020 in order to enable further progress in the empirical study of anaphora by coordinating the many existing efforts to annotate not just identity coreference, but all aspects of anaphoric interpretation from identity of sense anaphora to bridging to discourse deixis; and not just for English, but all languages. Progress so far includes a first proposal concerning the range of phenomena to be covered, as well as a survey of the range of existing anaphoric annotations and a proposal for a markup format extending the CONLL-U format developed by the **Universal Dependencies** initiative with mechanisms for marking up the range of anaphoric information covered by UA.

2.1. Scope: Beyond Identity Anaphora

Most modern anaphoric annotation projects cover basic identity anaphora as in (1).

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

However, many other types of identity anaphora exist, as well as other types of anaphoric relations that are not annotated in ONTONOTES but are annotated in other corpora.

Split-antecedent anaphora In ONTONOTES, plural reference is only marked when the antecedent is mentioned by a single noun phrase. However, **split-antecedent anaphors** are also possible (Eschenbach et al., 1989; Kamp and Reyle, 1993), as in (2). These are also cases of plural identity coreference, but to sets composed of two or more entities introduced by separate noun phrases. Such references are annotated in, e.g., ARRAU (Uryupina et al., 2020), GUM (Zeldes, 2017) and *Phrase Detectives* (Poesio et al., 2019).

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

Discourse deixis In ONTONOTES, **event anaphora**, a subtype of **discourse deixis** (Webber, 1991; Kolhatkar et al., 2018) is marked, as exemplified by *that* in (3), which refers to the event of a white rabbit with pink ears running past Alice; but not the whole range of abstract anaphora, illustrated by, e.g., *this*, which refers to the fact that the Rabbit was able to talk. A more extensive annotation of event anaphora is found in corpora such as the multi-sentence AMR corpus (O'Gorman et al., 2018) and more complex discourse deictic references are marked in, e.g., ANCORA and ARRAU.

(3) ... 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 Possibly the most studied of non-identity anaphora is **bridging reference** or **associative anaphora** (Clark, 1977; Hawkins, 1978; Prince, 1981) as in (4), where bridging reference / associative anaphora *the roof* refers to an object which is related to / associated with, but not identical to, the *hall*. In UA, non-identity anaphora is also taken to cover *other* anaphora as well as other cases of association such as identity of sense anaphora, etc. (Poesio, 2016).

(4) 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]].

2.2. CONLL-UA

The markup format proposed in UA, called CONLL-UA,[6] is an extension of the CONLL-U tabular format

[3] https://github.com/ns-moosavi/coval

[4] https://competitions.codalab.org/competitions/30312

[5] https://codalab.lisn.upsaclay.fr/competitions/614

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

defined for Universal Dependencies. The format specifies the following layers in addition to those defined in UD:

- an `Identity` layer, 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;
- a `Bridging` layer, specifying the anchor, its most recent mention, and, optionally, the associative relation;
- a `Discourse_Deixis` layer, 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.

Two inter-convertible versions of the UA format have been defined: the 'compact' and the 'exploded' formats. The 'compact' format encodes all the anaphoric interpretations in the '`Misc`' column of the CONLL-U format; this makes it fully compatible with the Universal Dependencies format. As a result, the resources collected for the UA can also be used by the Universal Dependencies community. The 'exploded' format is also based on the CONLL-U format, but instead of putting all information in the `Misc` column, it uses separate columns to accommodate different types of anaphoric information. The format has a focus on the NLP community, to make it easy to interpret by humans and systems for anaphora resolution. The Universal Anaphora scorer currently supports the 'exploded' format, but an extension supporting the 'compact' format has been developed for the 2022 CRAC-CorefUD shared task on Multilingual Coreference Resolution.[7]

3. Scoring Anaphoric Reference

3.1. Scoring Identity Anaphora

Evaluation is an issue with most areas of NLP, but it has proven a particularly difficult issue with identity anaphora, or coreference (Luo and Pradhan, 2016). There are various scoring methods for evaluating coreference resolution. MUC (Vilain et al., 1995) is a link-based metric that computes a score based on the minimum number of additional or missing coreference links in the system output compared to the gold clusters. B^3(Bagga and Baldwin, 1998) is a mention-based metric that performs the evaluation based on the number of common mentions between system and gold coreference chains. CEAF (Luo, 2005) is an entity-based metric that first finds the best alignment between the system and gold coreference chains and then computes the evaluation score based on the number of common mentions or the number of common links between the aligned chains. BLANC (Recasens and Hovy, 2011) is another linked-based metrics that considers both coreference and non-coreferring links for performing the evaluation. Finally, LEA (Moosavi and Strube, 2016) is a linked-based entity-aware metric that performs the evaluations based on the number of common coreference links in each system and gold coreference chains. Since the two CONLL shared tasks (Pradhan et al., 2012) it has become customary to score systems using the average F1 value of MUC, B^3 and CEAF, as originally proposed by (Denis and Baldridge, 2009)–this average, originally known as MELA, has since become known as the CONLL metric.

3.2. The Reference Coreference Scorer

During the period between the MUC evaluations and the CONLL 2011 and CONLL 2012 shared tasks on coreference resolution, the reference implementation for only one evaluation metric—MUC—was made available by the proposers of the metric itself. Neither of the three intervening metrics came with reference implementations. To further complicate matters, only the article introducing the MUC metric explicitly provided steps for computing the MUC score for predicted mentions. This resulted in two fundamental misunderstandings across the research community in regards to the other three metrics:

1. An assumption that both B^3 and CEAF metrics could not handle predicted mentions and needed to be modified in some way.
2. Not realizing that the BLANC metric was defined to handle only gold mentions.

The result were multiple implementations of the metrics with variations for scoring predicted mentions, and the inaccurate computation of the BLANC metric, for handling predicted mentions (Pradhan et al., 2014).
The CONLL 2012 shared task used the first open source implementation of all scoring metrics created for the SEMEVAL 2010 shared task (Recasens et al., 2010). This was a significant step forward for the community in getting consistent evaluation scores across institutions (and therefore published articles). However, it was also built on top of the above two misunderstandings.
Soon after the conclusion of the CONLL 2012 shared task, these issues were uncovered. A committee of coreference researchers—including almost all the original proposers of the existing metrics MUC, B^3, CEAF, and BLANC— created a open source, reference implementation [8] for the research community (Pradhan et al., 2014), which also included an extension of BLANC to handle predicted mentions (Luo et al., 2014). This became generally known as the Reference Coreference Scorer.

[7] https://ufal.mff.cuni.cz/corefud/crac22

[8] http://github.com/conll/reference-coreference-scorers

This implementation is however *only limited to evaluating coreference chains* and does not evaluate other types of relations including non-identity anaphora, split-antecedent anaphora, bridging references, and discourse deixis.

3.3. Scoring Non-Identity Anaphora

Unlike with identity anaphora, for which a generally accepted if not entirely satisfactory scoring mechanism has emerged, no standards exist to evaluate system performance at the other aspects of anaphoric interpretation, and most proposals are only concerned with evaluation on gold mentions.

Split-antecedent anaphora There is limited work on split-antecedent anaphora resolution and its evaluation. (Vala et al., 2016; Yu et al., 2020) only evaluate their models on split-antecedent anaphors, and on gold mentions only. These methods compute precision, recall, and F1 measures based on the links between split-antecedent gold anaphors and their antecedent. Zhou and Choi (2018) propose to evaluate split-antecedent resolution using the standard CONLL scorer. They do this by adding the plural mention to each of the clusters for its atomic elements: for example, they represent the {{*John, Mary*}, *They*} entity as two gold clusters–{John, They} and {Mary, They}. This representation however violates the fundamental assumption behind the notion of coreference chain–that all mentions in a chain refer to the same entity.

Bridging references With bridging references we have a fairly clear idea on how to evaluate systems; the main problem is that there is no complete agreement on the definition of bridging reference, and that many bridging references associate on more than one entity in context.

Corpora annotated with bridging information may provide two types of information about a bridging reference: the entity the bridging reference is associated with, or **anchor**, and the most recent mention of this entity. Many systems only carry out the second of these steps; some systems perform both. In early work (Vieira and Poesio, 2000) bridging descriptions were evaluated by hand in order to take disagreements into account. Poesio et al. (2004) introduced **entity evaluation**: a system's output is considered correct as long as the *anchor* (i.e., the entity) is identified correctly, whether or not its most recent mention is. Unlike Vieira and Poesio (2000), Poesio et al. (2004) required a system to identify the same anchor as in the gold even if the response was plausible. Hou et al. (2018) introduced the more stringent **mention evaluation**, which also requires a system to identify the exact mention of the anchor that is annotated in the corpus.

Discourse deixis There has been limited work on resolving discourse deixis and evaluating systems carrying out such task (Kolhatkar et al., 2018). In most annotations of discourse deixis either a clause or a verb is marked as the antecedent of the discourse deixis, and most systems use some version of accuracy to evaluate whether the system identified the correct antecedent. A more lenient metric, **Success@N**, was proposed by Kolhatkar (e.g., (Kolhatkar and Hirst, 2014)) and also used by Marasović et al. (2017) and in the CRAC 2018 Shared Task (Poesio et al., 2018). SUCCESS@N is the proportion of instances where the gold answer–the unit label–occurs within a system's first n choices. (S@1 is standard precision.)

The approach to representing discourse deixis adopted in Universal Anaphora however makes it possible however to adopt for discourse deixis the same approach to evaluation adopted in **event coreference** (Lu and Ng, 2018)–namely, evaluate discourse deixis using the same metrics as entity coherence, thus assessing e.g., a system's ability to evaluate whole coreference chains started with a discourse deictic reference, but with a non-nominal first mention. The approach to representing discourse deixis adopted in UA makes it possible to adopt this approach adopted in the UA scorer; as far as we know, this is the first time this approach has been used for discourse deixis.

3.4. The CRAC 2018 Shared Task

The one previous shared task focused on evaluating system performance at tasks other than identity anaphora was the Shared Task on Anaphora Resolution with ARRAU at CRAC 2018 (Poesio et al., 2018), which employed the ARRAU corpus. That shared task was articulated around three tasks: identity coreference (including identification of non-referring expressions), bridging references, and discourse deixis. The organization of the shared task resulted in the development of an extended version of the Reference Coreference Scorer which also scores non-referring expressions. Separate scorers were developed for bridging reference resolution, carrying out both mention-based evaluation and entity-based evaluation of bridging references, as done by Hou et al. (2018), and for discourse deixis, based on Kolhatkar and Hirst (2014). The scorer presented in this paper integrates all these evaluations in a single scorer, adding the new ability to score split antecedent anaphora and an entirely new approach to discourse deixis evaluation.

4. The Universal Anaphora Scorer

The new 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 [9] (Pradhan et al., 2014) developed for use in the CONLL 2011 and 2012 shared tasks on the

[9] https://github.com/conll/reference-coreference-scorers

ONTONOTES corpus (Pradhan et al., 2012) and its reimplementation in Python by Moosavi,[10] developed for the CRAC 2018 shared task (Poesio et al., 2018).

4.1. Identity Reference

In 'exploded' format, identity reference (cluster id) is specified in the `Identity` column, which includes both singular clusters (including singletons) and split-antecedents. Parentheses are used to specify the boundaries of the mention, as in the CONLL format, and a set of attributes is attached to the opening parentheses to specify the annotations. This includes the cluster id (`EntityID`), markable id (`MarkableID`), the minimum span (`Min`) and the semantic type (`SemType`) (non-referring, new, old) of the mention. Split-antecedent information is annotated on the antecedents's row using an '`ElementOf`' attribute that specifies the cluster id of the split antecedent plural anaphor. The following is an example of the IDENTITY column:

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

The scorer computes all major metrics for identity reference including MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), CONLL (the unweighted average of MUC, B^3, and CEAF) (Pradhan et al., 2014), BLANC (Luo et al., 2014; Recasens and Hovy, 2011), and LEA (Moosavi and Strube, 2016) scores. The scorer preserves the settings used in the Reference Coreference scorer for the CONLL shared tasks, and its scores are consistent with those of that scorer.

Three score-reporting options are available: The first option mirrors the evaluation used in the CONLL shared tasks (Pradhan et al., 2012) which excludes singletons and split-antecedents from evaluation. In this setting, split-antecedents are ignored when constructing the clusters, and after the clusters are constructed singletons are filtered out. I.e., only clusters with multiple mentions are evaluated.

The second option is the one used in the identity anaphora sub-task of the CRAC shared task (Poesio et al., 2018). This evaluation includes singletons, but not split-antecedents. For this setting, split-antecedents are ignored when constructing the clusters but singletons are kept for the evaluation.

Finally, the scorer can include both singletons and split-antecedent anaphors; this is the format used in CODI-CRAC 2021 (Khosla et al., 2021). Clusters include both split-antecedents and singletons. For split antecedents, a generalization of the existing coreference metrics was developed; this is briefly reviewed in the next subsection.[11]

4.2. Split Antecedent Anaphora

As discussed in Section 3, the evaluation metrics for split antecedent anaphora proposed in previous work (e.g. Vala et al. (2016; Zhou and Choi (2018)) are not entirely satisfactory. The UA scorer implements a new method proposed by Paun et al. (2021), for scoring split-antecedent anaphora based on the idea of treating the antecedents of split-antecedent anaphors as a new type of mention, **accommodated sets**–set denoting entities which have the split antecedents as elements. So for instance, in example (2), split-antecedent anaphor $[They]_{1,2}$ is encoded as belonging to a coreference chain whose first element is the accommodated set $\{1,2\}$ with the coreference chains for *John* and *Mary* as elements. Schematically,

$[\text{He}]_1 \in$ Coref Chain 1 (**John**) = { [John], [He] }
$[\text{her}]_2 \in$ Coref Chain 2 (**Mary**) = { [Mary], [her] }
$[\text{They}]_{1,2} \in$ Coref Chain 3 (**{John,Mary}**) = { {1,2}, [they] }

The proposed generalization also gives partial credit to interpretations of split-antecedent anaphors which do not identity all split antecedents.

More specifically, to include split-antecedents in the evaluation, the scorer first 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 Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) is used to search for the best alignments. To preserve the uniqueness of the different metrics, the relevant F1 scores are computed by applying the metric in question to the singular clusters that form the split-antecedents. For example, when computing the MUC score, the MUC F1 is used to find the optimal alignment between accommodated sets in the key and the response. Once the accommodated sets are aligned, 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 credits are rewarded on how well the accommodated sets are resolved. The treatment for individual metrics are slightly different depending on the nature of the metrics. We refer the reader to Paun et al. (2021) for detailed discussion.

The scorer also provides an additional option to allow computing scores for split-antecedent plurals only. This option is useful to assess a system's performance on resolving split-antecedent references, which are not

[10] https://github.com/ns-moosavi/coval

[11] Due to the complexity of the proposed method, we provide a full description in a separate paper (Paun et al., 2021) which focuses on the evaluation of split-antecedent anaphora.

very frequent. The score is defined as the micro-average F1 of all the split-antecedent anaphors in the key and response for all the supported metrics.

4.3. Non-referring expressions

A key aspect of anaphoric interpretation is correctly determining whether nominal phrases like markable *it* in (5) are referring or not, and to distinguish such noun phrases from singleton mentions.

(5) [It] was late at night.

In the CONLL-UA exploded format, non-referring expressions are associated with pseudo entities in the `Identity` column– i.e., a '`-Pseudo`' suffix is appended to the cluster ids (`EntityID`) to distinguish them from singleton mentions. The semantic type (`SemType`) attribute is used to specify the non-referring type in detail for corpora such as ARRAU or CODI-CRAC 2021 in which such distinctions are made (e.g. predicate, idiom). The following is an example of how a non-referring (predicative) NP is specified in the `Identity` column:

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

The new UA scorer follows the scorer developed for the CRAC 2018 Shared Task in that non-referring expressions are not treated as singletons in the evaluation of identity reference. Instead, non-referring expressions are separated from identity references when inputted to the scorer. More specifically, the collection of non-referring expressions in both the key and the response is identified and the scorer computes an F1 score for non-referring expressions only. The F1 score for non-referring expression is reported separately from the F1 scores for identity reference.

4.4. Discourse Deixis

The UA scorer supports the extension to discourse deixis proposed in version 1.0 of the Universal Anaphora specification of anaphoric phenomena by implementing an entirely new approach to evaluation of discourse deixis supporting the evaluation. This new approach is enabled by the way discourse deixis is encoded in the UA markup.

As mentioned in Section 3, in the most recent previous work discourse deixis is evaluated using the 'Success@N' metric (Kolhatkar and Hirst, 2014), which is based on the assumption that gold anaphors are given. The metric gives credit to a system if the gold segments are retrieved within the top N interpretations of the system. This approach is limited, both in that it requires gold anaphors, and because as it treats individual anaphors separately without assessing the quality of the clusters they formed.

But discourse deixis is similar to coreference, in that both form clusters by linking the anaphors to their antecedents. Another important similarity is that in both cases we can have split-antecedent anaphors that refer to multiple antecedents–in fact, split antecedent reference is the norm for discourse deixis. The main difference is that, in coreference, antecedents are introduced using nominal phrases, whereas in discourse deixis they are introduced using non-nominal phrases (segments).

In the UA markup, discourse deixis is specified in the `Discourse_deixis` column of the 'exploded' format, and the same attributes are used as for the `Identity` column. The only difference is that the cluster id (`EntityID`) and the markable id (`MarkableID`) of the segments are highlighted with a '`-DD`' suffix and '`dd_`' prefix respectively, to avoid confusion in visual inspection. An example `Discourse_deixis` row is:

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

This representation enables the application of coreference metrics to evaluate discourse deixis. Particularly given that our new scorer provides a way to incorporate split-antecedents into the standard metrics, which therefore are discourse deixis-ready. This is exactly how the UA scorer evaluates discourse deixis: it computes the same MUC, B^3, CEAF, CONLL, BLANC and LEA metrics as for identity anaphora.

In other words, the UA scorer introduces two novelties in the scoring of discourse deixis. First, discourse deixis is evaluated in the same way as entity anaphora; this makes it possible to use the same metrics used for identity reference for evaluating discourse deixis as well. One of the advantages of this approach is that discourse deixis evaluation now works with predicted mentions/segments, which we hope will encourage research on discourse deixis to move to a more realistic setting. Second, 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.

4.5. Bridging References

In UA format, bridging references are specified in the `Bridging` column of the 'exploded' format. The attributes for bridging include the markable ID (`MarkableID`), a mention of anchor entity (`MentionAnchor`), the cluster id of the antecedent (`EntityAnchor`) and the bridging relationship (`Rel`). For example:

```
(MarkableID=markable_9|\
Rel=subset-inv|\
```

```
MentionAnchor=markable_1|\
EntityAnchor=3)
```

For bridging references, the scorer reports three scores: the two metrics computed by the scorer used for CRAC 2018 shared task–mention-based F1 and entity-based F1–and, in addition, anaphora recognition F1.
Mention-based F1 for bridging evaluates a system's ability to predict the correct anaphora and the mention of the anchor specified in the annotation (this is usually the closest or most suitable mention). `MentionAnchor` is used for this type of evaluation.
Entity-based F1 is more relaxed than mention-based F1, and does not require the system to predict exactly the same mention as the gold annotation. Instead, a system's interpretation is deemed correct as long as any mention of the correct anchor (`EntityAnchor`) is found, as done e.g., in Poesio et al. (2018).
Finally, anaphora recognition F1 is used to assess the system's ability to identify bridging anaphors.

5. The CODI/CRAC 2021 Shared Task

The new UA scorer was used as the official scorer for the CODI-CRAC 2021 shared task. A brief description of the task is provided here to give some context for interpreting the scorer's results; for more details please see (Khosla et al., 2021).

5.1. The tasks

Following the structure of the CRAC 2018 Shared Task, CODI-CRAC 2021 was articulated around three tasks covering identity anaphora, bridging anaphora, and discourse deixis. Participants could submit to one or more tasks.

5.2. Gold and Predicted Settings

Bridging reference resolution and discourse deixis are very difficult tasks. In consideration of this, the Bridging (Task 2) and Discourse Deixis (Task 3) tasks were further divided into system and gold settings, according to whether the markables would be predicted by the system or provided by the organizers. The two settings were run in order; the gold setting only became available after the runs under the system setting had been submitted. The two settings were scored separately.

5.3. Settings of the UA Scorer used

The following settings of the UA scorer were used for the individual tasks.[12]

Task 1 the evaluating coreference relations (including split-antecedents) and singletons modality was used. Non-referring expressions identification was not scored.

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

[12]For a full description of the task(s), see `https://github.com/sopankhosla/codi2021_scripts/blob/main/2021_CODI_CRAC_Introduction.md`

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval AR					
Emory	**80.33**	**63.98**	**78.41**	**74.49**	**74.3**
UTD_NLP	79.56	57.38	77.50	72.64	71.8
KU_NLP	69.16	57.59	71.09	65.67	65.9
DFKI	64.99	43.93	59.93	53.55	55.6
SCIR	55.92	39.46	52.25	51.63	49.8
Baseline	52.45	36.11	51.97	45.80	46.6
DFKI	61.26	00.00	59.20	51.24	42.9

Table 1: Performance on Task 1 (Evaluation Phase) – Identity Anaphora (CoNLL Avg. F1)

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval Br (Gold)					
UTD_NLP	**19.73**	**19.65**	**31.40**	**21.10**	**23.0**
KU_NLP	16.67	15.30	18.79	18.33	17.3
INRIA	9.35	6.00	16.28	7.79	9.9
Baseline	6.35	6.21	13.77	5.39	7.9
Eval Br (Pred)					
UTD_NLP	**13.98**	**13.33**	**21.92**	**15.26**	**16.1**
KU_NLP	13.46	10.25	12.32	10.99	11.8
Baseline	6.01	4.94	9.34	3.78	6.0

Table 2: Performance on Task 2 (Evaluation Phase) – Bridging Anaphora (Entity F1)

Task 2 the scorer was called using the command:

```
python ua-scorer.py key system \
    keep_bridging
```

Task 3 the scorer was called using the command:

```
python ua-scorer.py key system \
     evaluate_discourse_deixis
```

5.4. The CODI-CRAC 2021 Corpus

The only existing dataset covering the full range of phenomena and with some coverage of dialogue, the ARRAU data used for the CRAC 2018 Shared Task, was used as training material and as one of the development sets. In addition, new data from four dialogue corpora–AMI, LIGHT, PERSUASION and SWITCHBOARD–were annotated for development and testing using the same annotation scheme used in ARRAU. The dataset was annotated using the MMAX2 tool (Müller and Strube, 2006). After annotation, the documents were converted into the CONLL-UA 'exploded' format. All the publically distributable data (AMI, LIGHT and PERSUASION) are available from the Codalab shared task site and the Universal Anaphora site. For more details see (Khosla et al., 2021).

6. Results and Discussion

In this section, we discuss the scorer's use in each of the tasks.

Team	LIGHT	AMI	PERS.	SWBD.	Avg.
Eval DD (Gold)					
UTD_NLP	**43.44**	**36.91**	**52.09**	**40.44**	**43.2**
Eval DD (Pred)					
UTD_NLP	**42.70**	**35.35**	**39.64**	**35.43**	**38.3**
DFKI	20.97	17.43	23.76	23.86	21.5
Baseline	12.12	15.75	18.27	13.55	14.9

Table 3: Performance on Task 3 (Evaluation Phase) – Discourse Deixis (CoNLL Avg. F1)

6.1. Task 1 – Identity Anaphora

Task 1 saw the highest interest as five teams submitted a total of 36 runs to the official leaderboard. The results on different sub-corpora are reported in Table 1. Although only the CoNLL Avg. F1 scores were reported on the leaderboard, the scorer's setting utilized in this task allowed the organizers to provide additional details to participants about their systems' performances (Precision, Recall, and F1 scores) on multiple state-of-the-art metrics like B^3, CEAF, BLANC, MUC, and LEA. After the culmination of the eval-phase of the shared-task,

```
python ua-scorer.py key system \
    remove_singletons \
    remove_split_antecedent
```

a mode that is compatible with the Reference Coreference scorer was used by the organizers to evaluate systems only on coreferring markables. The analysis showed that every system lost about 5-8 CoNLL Avg. F1 points against their performance on the competition setting, across all four datasets. This reveals a slight bias the systems might have towards creating singleton clusters. Using the mode:

```
python ua-scorer.py key system \
    only_split_antecedent
```

to isolate systems' performances on split-antecedents, the organizers found that even though the participating systems achieved high overall scores on Task 1, none of them were able to handle split-antecedents correctly, thus highlighting the need for further research in this direction.

6.2. Task 2 – Bridging Anaphora

Three teams participated in Task 2 with *INRIA* only participating in the gold mention setting. Entity F1 scores were reported for each sub-corpora. Precision/recall/F1 scores for other two metrics – mention-based and anaphora recognition, were also output by the scorer to aid teams in evaluating different modules of their systems. Table 2 summarizes the performance of each team on this task.

6.3. Task 3 – Discourse Deixis

25 runs were received for Task 3. Two teams submitted to the predicted mention setting (Eval DD (Pred)) with *UTD_NLP* achieving performance around 35–42 CoNLL Avg. F1 percentage points on the different sub-corpora, almost doubling the score of the second team. When the gold markables were also released (Eval DD (Gold)), the system submitted by *UTD_NLP* managed a jump of more than 12 points on PERSUASION (Table 3). As discussed earlier, the scoring and metrics used for Task 3 were similar to that of Task 1. The command included an additional argument `evaluate_discourse_deixis` to only evaluate discourse deixis instances. The scorer also reported other state-of-the-art metrics like B^3, CEAF, BLANC, MUC, and LEA for participants' reference.

Across all three tasks, the availability of different (scoring) modes and settings in the scorer allowed for a deeper understanding of the performance of different participating teams.

7. Conclusion and Future Work

NLP research is driven not just by the availability of resources providing the desired information, but also by the existence of standardized scorers allowing researchers to evaluate their systems in a reliable way. In this paper we presented the new Universal Anaphora scorer designed to evaluate models carrying out a fuller form of anaphoric interpretation. The scorer was tested in the CODI-CRAC 2021 shared task proving reliable and able to provide insight in the performance of participating systems; we hope that it will encourage research in so-far under-researched aspects of anaphora. Future plans include extending the scorer to cover discontinuous mentions–mentions broken into separate segments of text, as founds e.g., in completions which are common in dialogue (Poesio and Rieser, 2010)–and then to cover cases of disagreement on anaphoric interpretation, which are particularly common in dialogue (Poesio and Artstein, 2005) but can be found in all genres and are very numerous in the *Phrase Detectives* corpus (Poesio et al., 2019).

8. Acknowledgements

The work of Juntao Yu, Silviu Paun and Massimo Poesio was funded by the DALI project, ERC Grant 695662.

9. Bibliographical References

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation (LREC) - Workshop on linguistics coreference*, volume 1, pages 563–566. ACL.

Bamman, D., Lewke, O., and Mansoor, A. (2020). An annotated dataset of coreference in english literature. In *Proc. of LREC*. European Language Resources

Association (ELRA), Association for Computational Linguistics (ACL).

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

Cohen, K. B., Lanfranchi, A., Choi, M. J.-y., Bada, M., Baumgartner Jr., W. A., Panteleyeva, N., Verspoor, K., Palmer, M., and Hunter, L. E. (2017). Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(372).

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

Eschenbach, C., Habel, C., Herweg, M., and Rehkämper, K. (1989). Remarks on plural anaphora. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, pages 161–167. Association for Computational Linguistics.

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

Hou, Y., Markert, K., and Strube, M. (2018). Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Hou, Y. (2020). Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online, July. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

Kantor, B. and Globerson, A. (2019). Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy, July. Association for Computational Linguistics.

Khosla, S., Yu, J., Manuvinakurike, R., Ng, V., Poesio, M., Strube, M., and Rosé, C. (2021). The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proc. of the CODI/CRAC Shared Task Workshop*.

Kobayashi, H. and Ng, V. (2021). Bridging resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1652–1659, Online, June. Association for Computational Linguistics.

Kolhatkar, V. and Hirst, G. (2014). Resolving shell nouns. In *Proc. of EMNLP*, pages 499–510, Doha, Qatar.

Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Liu, Q., Jiang, H., Evdokimov, A., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2017). Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2344–2350.

Lu, J. and Ng, V. (2018). Event coreference resolution: A survey of two decades of research. In *Proc. of IJCAI*, pages 5479–5486.

Luo, X. and Pradhan, S. (2016). Evaluation metrics. In Massimo Poesio, et al., editors, *Anaphora Resolution: Algorithms, Resources and Applications*, pages 147–170. Springer.

Luo, X., Pradhan, S., Recasens, M., and Hovy, E. (2014). An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland, June. Association for Computational Linguistics.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Marasović, A., Born, L., Opitz, J., and Frank, A. (2017). A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark, September. Association for Computational Linguistics.

Moosavi, N. S. and Strube, M. (2016). A proposal for

a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin.

Müller, M.-C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. In S. Braun, et al., editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, volume 3 of *English Corpus Linguistics*, pages 197–214. Peter Lang.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). Ancor_centre, a large free spoken french coreference corpus. In *Proc. of LREC*.

Nedoluzhko, A. (2013). Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In *Proc. of LAW*, pages 103–111.

O'Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K., and Palmer, M. (2018). Amr beyond the sentence: the multi-sentence amr corpus. In *Proc. of COLING*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paun, S., Yu, J., Moosavi, N., and Poesio, M. (2021). Scoring coreference chains with split-antecedent anaphors and other entities constructed from a discourse model. Submitted.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In A. Meyers, editor, *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, June.

Poesio, M. and Rieser, H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1(1):1–89.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to solve bridging references. In *Proc. of ACL*, pages 143–150, Barcelona, July.

Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., and Zinsmeister, H. (2018). Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana, June. Association for Computational Linguistics.

Poesio, M., Chamberlain, J., Kruschwitz, U., Paun, S., Uma, A., and Yu, J. (2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proc. of NAACL*, page 1778–1789, Minneapolis. Association for Computational Linguistics (ACL).

Poesio, M. (2016). Linguistic and cognitive evidence about anaphora. In M. Poesio, et al., editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 2. Springer.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Rahman, A. and Ng, V. (2012). Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July. Association for Computational Linguistics.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

Recasens, M. and Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. (2020). Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8732–8740, Apr.

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., and Poesio, M. (2020). Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.

Vala, H., Piper, A., and Ruths, D. (2016). The more antecedents, the merrier: Resolving multi-antecedent anaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2296, Berlin, Germany, August. Association for Computational Linguistics.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593, December.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and

Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Wiseman, S., Rush, A. M., Shieber, S., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, July. Association for Computational Linguistics.

Yu, J. and Poesio, M. (2020). Multitask learning based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yu, J., Moosavi, N. S., Paun, S., and Poesio, M. (2020). Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Yu, J., Moosavi, N. S., Paun, S., and Poesio, M. (2021). Stay together: A system for single and split-antecedent anaphora resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zhou, E. and Choi, J. D. (2018). They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.