The Universal Anaphora Scorer 2.0

Juntao Yu[1], **Michal Novák**[2], **Abdulrahman Aloraini**[3], **Nafise Sadat Moosavi**[4], **Silviu Paun**[5], **Sameer Pradhan**[6,7] and **Massimo Poesio**[5]
[1]Univ. of Essex; [2]Charles Univ.; [3]Qassim University; [4]Univ. of Sheffield;
[5]Queen Mary Univ.; [6]LDC, Univ. of Pennsylvania; [7]cemantix.org
j.yu@essex.ac.uk; mnovak@ufal.mff.cuni.cz;
m.poesio@qmul.ac.uk

Abstract

The aim of the Universal Anaphora initiative is to push forward the state of the art both in anaphora (coreference) annotation and in the evaluation of models for anaphora resolution. The first release of the Universal Anaphora Scorer (Yu et al., 2022b) supported the scoring not only of identity anaphora as in the Reference Coreference Scorer (Pradhan et al., 2014) but also of split antecedent anaphoric reference, bridging references, and discourse deixis. That scorer was used in the CODI-CRAC 2021/2022 Shared Tasks on Anaphora Resolution in Dialogues (Khosla et al., 2021; Yu et al., 2022a). A modified version of the scorer supporting discontinuous markables and the COREFUD markup format was also used in the CRAC 2022 Shared Task on Multilingual Coreference Resolution (Žabokrtský et al., 2022). In this paper, we introduce the second release of the scorer, merging the two previous versions, which can score reference with discontinuous markables and zero anaphora resolution.

1 Introduction

The objective of the **Universal Anaphora** initiative, or UA,[1] is to coordinate efforts to push forward the state of the art in anaphora and anaphora resolution beyond identity anaphora,[2] and also covering genres such as dialogue, exemplified by datasets such as ARRAU (Poesio et al., 2018; Uryupina et al., 2020), the CODI-CRAC 2021/2022 corpora (Khosla et al., 2021; Yu et al., 2022a) and GUM (Zeldes, 2017) for English, the Prague Dependency Treebank (its latest version in Hajič et al., 2020) for Czech, and ANCORA for Catalan and Spanish (Recasens and Martí, 2010). The initiative, modelled on Universal Dependencies (UD),[3] 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. The Universal Anaphora effort has proceeded in close collaboration with the COREFUD initiative (Nedoluzhko et al., 2021, 2022), whose objective is to facilitate research on coreference and anaphora (possibly along with morphology and dependency syntax) by converting corpora in various languages to a unified markup format, fully compatible with UD standards.

An essential prerequisite to make Universal Anaphora-compatible corpora usable in NLP is the availability of scorers that can evaluate the interpretation produced by a system for, 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, 2021). A first step in this direction was the introduction of the Universal Anaphora scorer for anaphoric interpretation (Yu et al., 2022b), 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 2021/2022 Shared Tasks in Anaphora Resolution in Dialogue (Khosla et al., 2021; Yu et al., 2022a) and a revised version supporting COREFUD was used in

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

[2] We use the term **identity anaphora** to refer to the subclass of anaphora in which the anaphor refers to the same discourse entity as the antecedent, also known in NLP as 'coreference'. E.g., in *[Geraint Thomas]$_i$'s Giro d'Italia challenge evaporated on the steep slopes of Monte Lussari in north-east Italy. [The Welsh rider]$_i$ was overtaken by his closest challenger, Primoz Roglic.* , the anaphor *The Welsh rider* refers to the same entity as its antecedent, *Geraint Thomas*.

[3] https://universaldependencies.org/

the CRAC 2022 Shared Task on Multilingual Coreference (Žabokrtský et al., 2022).

In this paper, we introduce the second version of the Universal Anaphora scorer. This release addresses two key limitations of the first release. The first limitation is the restriction to contiguous mentions, not allowing **discontinuous markables** such as *[a tanker] .. [of orange juice]* in (1.1), consisting of two chunks of text separated from S's uttering *yeah*. Discontinuous markables are common in spoken conversations, but are also used in the CRAFT-CR 2019 biomedical corpus (Cohen et al., 2017) and in corpora such as ARRAU to encode the conjuncts in noun phrases with coordinated heads such as *the students and lecturers from Queen Mary University*, which result in the discontinuous markables *[the students] [from Queen Mary University]* and *[the][lecturers from Queen Mary University]*.

Example 1.1

M : *... [a tanker]*
S : *yeah*
M : *[of orange juice]*

A second limitation of the UA scorer 1.0 is the inability to score the resolution of **zero anaphora** (unrealized arguments) as in (1.2), except in the 'gold' case in which the zero is explicitly marked in the test set.

Example 1.2 (IT) *[Giovanni]$_i$ è in ritardo, così [∅]$_i$ mi ha chiesto se posso incontrar[lo]$_i$ al cinema.*
[EN] [John]$_i$ is late so [he]$_i$ asked me if I can meet [him]$_i$ at the movies.

Zero anaphora is annotated in Arabic and Chinese ONTONOTES, and in several of the datasets in the COREFUD collection (Nedoluzhko et al., 2022). In Arabic and Chinese ONTONOTES, zeros are marked using an asterisk * to indicate the position of the empty category in the training data and in the test data in 'gold' mode, but not in the test data in 'predicted' mode, meaning that to evaluate this second mode the scorer must be able to handle 'insertion' of tokens, resulting in evaluation problems (Aloraini et al., 2022).

The new version of the scorer presented in this paper (i) incorporates the treatment of discontinuous markables developed for the COREFUD scorer, testing it also on the CRAFT-CR 2019 corpus; (ii) introduces a novel treatment for the basic form of zero anaphora; and (iii) supports both the COREFUD and UA markup formats.

2 Universal Anaphora And CorefUD

Achievements of the Universal Anaphora initiative so far include a first proposal concerning the range of phenomena to be covered, as well as a survey of the range of existing anaphoric annotations and two proposals for markup formats extending the CONLL-U format developed by **Universal Dependencies** with mechanisms for marking up the range of anaphoric information covered by UA.

2.1 Beyond Identity Anaphora

Most modern anaphoric annotation projects cover basic identity anaphora. However, many other types of identity anaphora exist, as well as other types of anaphoric relations that are annotated in a number of corpora (Novák et al., 2023).

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). These are also cases of plural identity coreference, but to sets composed of two or more entities introduced by separate noun phrases, as in *[John]$_1$ met [Mary]$_2$. [He]$_1$ greeted [her]$_2$. Then [they]$_{1,2}$ went to the movies.*

Discourse deixis (Webber, 1991; Kolhatkar et al., 2018) is the term used to cover both event anaphora, as in *John met Mary. [It]$_1$ happened at 3pm.*, as well as more general types of anaphoric reference to abstract objects not introduced by nominals, as in *John told Mary he was at the office. She didn't believe [that]$_1$ ..* Event anaphora is annotated in ONTONOTES and in corpora such as the multi-sentence AMR corpus (O'Gorman et al., 2018). The full range of discourse deixis is annotated in, e.g., ANCORA and ARRAU.

Possibly the most studied of non-identity anaphora is **bridging reference** or **associative anaphora** (Clark, 1977; Hawkins, 1978; Prince, 1981) as in *John looked at the house. [The roof] was thatched.*, where bridging reference / associative anaphora *the roof* refers to an object which is related to / associated with, but not identical to, the *the house*.

2.2 CONLL-UA

The markup format proposed in UA, called CONLL-UA,[4] is based on the CONLL-U-Plus tabular format

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

proposed in Universal Dependencies for corpora containing additional linguistic annotations.[5] 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.

The CONLL-UA format was designed to provide a way to specify anaphoric information independent from other layers, but compatible with the UD format. However, at present the UD parser used to validate documents included in UD datasets cannot process the CONLL-U-Plus format. Thus, UA collaborated with COREFUD to design a more 'compact' format that could be used to pack the anaphoric information representable in CONLL-UA in the '`Misc`' column of the CONLL-U format, and is fully compatible with the Universal Dependencies. We discuss COREFUD next.

2.3 The CorefUD format

The COREFUD initiative (Nedoluzhko et al., 2022) was launched in parallel with UA to create a collection of corpora annotated with coreferential and other anaphoric relations using a harmonized schema and format. Its current version COREFUD 1.1 (Novák et al., 2023) consists of 17 datasets for 12 languages in its publicly available edition.[6]

Whereas UA is primarily focused on anaphora, COREFUD has another objective besides harmonization of the coreference datasets, namely, to intersect the world of coreference with the world of syntax. This is achieved by augmenting the coreference data with morpho-syntax annotation compliant with the UD standards, which has been obtained automatically for the datasets that do not contain such manual annotation. This is motivated not only pragmatically (popularity of UD and standards for numerous technical issues), but it is also grounded theoretically. For instance, entity mentions often correspond to syntactically relevant notions (e.g. noun phrase, subject), some coreference relations are manifested mainly by syntactic means (e.g. reflexive and relative constructions), and zero expressions (e.g. pro-drops) are vital for coreference in many languages.

After developing a first format in COREFUD 0.1 (Nedoluzhko et al., 2021) independently from the UA initiative[7], a new format was jointly developed and introduced with COREFUD 1.0 (Nedoluzhko et al., 2022). This format can encode essentially the same information as CONLL-UA, but this information is encoded in the `Misc` column, which makes it possible to pass the official UD validation at level 2 (passing the higher levels is not possible with automatically predicted POS tags and dependency relations).[8] One remaining difference is that COREFUD has been from its very beginning designed to represent existing data in datasets including dependency graphs. Thus, it can capture zero expressions by stipulating 'empty tokens' and referencing them using enhanced dependency graphs, whereas in CONLL-UA, which does not require dependency layers, empty tokens are bound to the surface tokens by their relative position.

The COREFUD collection is accompanied with API implemented within the Udapi framework[9] that facilitates manipulation with the data in COREFUD format as well as its visualization.

3 The Universal Anaphora Scorer 1.0

The Universal Anaphora (UA) 1.0 scorer (Yu et al., 2022b) is a Python scorer for the varieties of anaphoric reference covered by the Universal Anaphora guidelines: identity anaphora, split antecedent plurals, identification of non-referring expressions, bridging reference, and discourse deixis.

For identify reference, the scorer builds on the original Reference Coreference scorer [10] (Pradhan

[5] https://universaldependencies.org/ext-format.html

[6] In total, 21 datasets for 13 languages, including datasets with non-public licences, e.g. ONTONOTES and ARRAU.

[7] This format, which substantially differs from the current format, is described in: https://ufal.mff.cuni.cz/~popel/corefud-1.0/corefud-1.0-format.pdf.

[8] https://universaldependencies.org/validation-rules.html#levels-of-validity

[9] https://github.com/udapi/udapi-python

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

et al., 2014) and its reimplementation in Python by Moosavi,[11] developed for the CRAC 2018 shared task (Poesio et al., 2018). The Reference Coreference scorer, developed for use in the CONLL 2011 and 2012 shared tasks on the ONTONOTES corpus (Pradhan et al., 2012), implemented the best known metrics for identity anaphora (coreference): MUC (Vilain et al., 1995), B^3(Bagga and Baldwin, 1998), CEAF (Luo, 2005), and BLANC (Recasens and Hovy, 2011). The Reference Coreference scorer popularized scoring by using the average F1 value of MUC, B^3 and CEAF, as originally proposed by (Denis and Baldridge, 2009)–so much so that this average, originally known as MELA, has since become known as the CONLL metric. Moosavi's CRAC 2018 scorer, apart from being written in Python, also implemented the LEA metric (Moosavi and Strube, 2016) and provided a separate score for the interpretation of non-referring expressions.

3.1 Identity Reference

In the CONLL-UA format, identity reference is specified in the `Identity` column, which specifies the cluster id (`EntityID`), markable id (`MarkableID`), the minimum span (`Min`) and the semantic type (`SemType`) (non-referring types, discourse new (`dn`) and discourse old (`do`)) 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. This is illustrated in the following example:

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

The UA 1.0 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.

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. The second option is the one used in the identity anaphora sub-task of the CRAC 2018 shared task (Poesio et al., 2018). This evaluation includes singletons, but not split-antecedents. Finally, the scorer can include both singletons and split-antecedent anaphors; this is the format used in CODI-CRAC 2021/2022 (Khosla et al., 2021; Yu et al., 2022a). Clusters include both split-antecedents and singletons. For split antecedents, a generalization of the existing coreference metrics was developed (Paun et al., 2023).

3.2 Split Antecedent Anaphora

The UA scorer implements a new method proposed by Paun et al. (2023), for scoring split-antecedent anaphora based on 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.

3.3 Non-referring expressions

A key aspect of anaphoric interpretation is correctly determining whether nominal phrases like markable *it* in Example 3.1 are referring or not, and to distinguish such noun phrases from singletons.

Example 3.1 *[It] was late at night.*

The semantic type (`SemType`) attribute is used to specify the non-referring type in detail for corpora such as ARRAU or CODI-CRAC 2021/2022 in which such distinctions are made (e.g. predicate, idiom). 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.

3.4 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.

In the UA markup, discourse deixis is specified in the `Discourse_deixis` column of the 'exploded' format, and the same attributes are used as

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

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.

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.

3.5 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). 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. Finally, anaphora recognition F1 is used to assess the system's ability to identify bridging anaphors.

4 The CorefUD Scorer 1.0

CorefUD scorer 1.0 was used in the CRAC 2022 Shared Task on Multilingual Coreference Resolution (Žabokrtský et al., 2022). It is based on the Universal Anaphora Scorer 1.0, reusing the implementations of all generally used coreferential measures without any modification. This guarantees that the measures are computed in exactly the same way. Nevertheless, CorefUD scorer is capable of processing the coreference annotation files in the CorefUD 1.0 format.

Among other things, it allows evaluation of coreference for zeros. Nonetheless, its version 1.0 is not able to handle a response document whose tokens are not completely identical to the tokens in the key document. This holds also for empty tokens, which virtually prevents the scorer to evaluate response documents where the zero expressions are automatically predicted.

Moreover, the CorefUD scorer re-defines matching of key and response mentions in the way to be able to process potentially discontinuous mentions, which are present in some CorefUD datasets. Instead of comparing mention boundaries, matching is based on set/subset relations between the tokens of the mentions in question.

Last but not least, the CorefUD scorer introduced two new scores. The MOR score measures to what extent key and response mentions match, no matter to which coreference entity they belong. The CorefUD scorer also implements the anaphor-decomposable scoring schema introduced by Tuggener (2014) and applies it to zeros. This allows for measuring the quality of predicting any of the antecedents of zero anaphors.

5 The UA Scorer 2.0

The UA scorer 2.0 merges the functions of the UA scorer 1.0 and CorefUD scorer 1.0 to make them a unified scorer. It also optimises/extends the scorer's ability on handling discontinuous markables and zeros, e.g. the new scorer can handle zeros in the predicted setting and can reproduce the CRAFT-CR 2019 shared task results. We introduce the details of the implementations in the next subsections.

5.1 Discontinuous Markables

In CONLL-UA, discontinuous markables can be used in both the Identity and Discourse_Deixis columns by sharing the MarkableID between the different sub-spans of a discontinuous markable. The scorer can then recognise the discontinuous markables from the text. For example, if a discontinuous markable consists of two continuous spans, the two spans will have the same Identity column, e.g. same EntityID, MarkableID, Min and SemType.

COREFUD format does not assign IDs to markables. Each continuous part of a discontinuous markable is thus labeled by its ordinal number

and the total number of parts in square brackets just after the cluster ID: `Entity=(10[1/2] ... Entity=10[1/2]) ... Entity=(10[2/2] ... Entity=10[2/2])`.

Since coreference evaluation metrics are developed based on the assumption that mentions in the key and response are aligned implicitly, the scorer provides two mention alignment strategies during the evaluation: 'strict' and 'partial'. In a 'strict' setting mentions are aligned only if all parts of the discontinuous markables are recognised correctly by the system. In the 'partial' setting, mentions can be aligned using a specified fuzzy matching algorithm. To use the 'partial' matching, the `Min/head` span for each mention needs to be specified in the key files. The `Min/head` span is specified as the minimum string that a coreference resolver must identify for the corresponding markable (either discontinuous or continuous). Allowing 'partial' mention alignment is especially useful for evaluating discontinuous mentions, given that it is more complex to predict, and most of the current coreference systems cannot predict the discontinuous markables.

To be more specific, the scorer provides two algorithms to align the mentions in 'partial' settings. By default, a mention in the response is considered a candidate for a gold mention if it contains the MIN/head string and does not go beyond the annotated maximum boundary. To align the mentions in the key and response, we first align the mentions based on the exact matching to exclude them from the partial matching step. Secondly, to align the remaining mentions, we compute the recall (the precision will always be 100% according to our definition of partial matching) between all remaining mention pairs between key and their corresponding candidates in the response to create a recall matrix. Finally, the recall matrix is used with the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) to find the best alignment between those mentions. After the alignment between the mentions is found, the coreference evaluation metrics can be used as normal.

To facilitate the research in the biomedical domain we also provide an option to align the mentions using the same algorithm as in CRAFT-CR 2019 shared task (Baumgartner et al., 2019) The CRAFT-CR 2019 corpus consists of biomedical files with coreference relations (including discontinuous markables) annotated. The algorithm used to align mentions in CRAFT-CR 2019 shared task considers a predicted mention correct if any continuous span of the predicted mention overlaps with and does not go beyond the first span of the key mention. Their algorithm does not impose a one-to-one alignment between mentions hence one key mention might be aligned with multiple predicted mentions and vice versa.

By default, if a corpus consists of discontinuous markables the system will use the 'strict' setting to evaluate them. The `-p|--partial-match` option can be used to enable the default partial matching algorithm. To use the CRAFT-CR 2019 algorithm, the `--partial-match-method` option needs to be set to `craft`.

5.2 Zeros

In both 'exploded' and 'compact' format, zeros are represented using the UD standard of empty nodes, in which the first column (`ID`, word index) is indicated using the decimal numbers. For instance, if we have a zero anaphora right after a token whose `ID` is 5, we index the zero with 5.1 instead of 6 used for a normal token. The scorer identifies the zeros by the decimal indexing and has the option to include zeros in the evaluation.

When zeros are included in the evaluation, again we need to align them between the key and response. Currently, the scorer performs the alignment based on the position of the zeros, i.e. zeros are aligned if they are located in the same position in the sentences. This is based on the assumption that the position of the zeros is not random, and the corpus which have zeros annotated has a consistent guideline on where should the zeros be positioned. We are also considering another approach that uses dependency relations to align the zeros, in which the position of zero does not need to follow a certain rule. However, due to the complication of this approach, we are not able to include it in this release and are planning to make it available in the next version of the scorer.

By default zeros are excluded in the evaluation, to include them the `-z|--keep-zeros` options can be specified.

5.3 Formats

The scorer supports three formats: CONLL 2012, CONLL-UA (UA 'exploded') and COREFUD (UA 'compact'). The CONLL-UA format is the default format for the scorer that support all anaphora relations assessed by the scorer e.g. singletons, non-referring expressions, split-antecedents, bridg-

ing reference and discourse deixis. The parser of the CorefUD format supports identity relations including discontinuous markables and zeros but does not support split-antecedents and non-referring expressions. The CONLL 2012 format only support continuous markables in the identity relation.

5.4 Shared Tasks Support

As the number of shared tasks supported by the scorer grows, the options also increase. To simplify the usage of the scorer we provide shortcuts for all coreference shared tasks supported by the scorer. The `-t|--shared-task` option can be used to specify the evaluation settings for the shared task in question. In total, the scorer supports 7 different settings used in 5 shared tasks:

- `conll12`: This evaluation mode is compatible with the coreference evaluation of the CONLL 2012 shared task in which only coreferring markables are evaluated.
- `crac18`: The evaluation method used in CRAC 2018 shared task. In this evaluation setting, coreference relations, singletons and non-referring mentions are taken into account for evaluation.
- `craft19`: This evaluation mode is used by the CRAFT 2019 shared task, it includes coreference relations, singletons and discontinuous markables.
- `crac22`: The evaluation method used as the primary metric by the CRAC 2022 shared task on multilingual coreference resolution. The evaluation applies partial matching and includes coreference relations, discontinuous markables, and zeros but excludes singletons and split-antecedents
- `codicrac22ar`: The evaluation method used by the anaphora resolution track of the CODI-CRAC 2021/2022 shared tasks. In this mode, both coreferring markables, split-antecedents and singletons are evaluated by the specified evaluation metrics.
- `codicrac22br`: The evaluation method used by the bridging resolution track of the CODI-CRAC 2021/2022 shared tasks. In this evaluation setting only bridging references will be evaluated.
- `codicrac22dd`: The evaluation method used by the discourse deixis track of the CODI-CRAC 2021/2022 shared tasks. The discourse deixis column is evaluated using the same method as `codicrac22ar`.

6 Results

In this section we demonstrate the scorer in practice by using it to score the submissions to two shared tasks that involved discontinous markables and zeros, CRAC 2022 and CRAFT-CR 2019.

6.1 CRAC 2022 Shared Task

We tested the new UA scorer on the submissions to the CRAC 2022 Shared Task on Multilingual Coreference Resolution (Žabokrtský et al., 2022), namely on the predictions of the winning setup of the CorPipe system (Straka and Straková, 2022).

Table 1 shows the performance of the winning submission evaluated on the shared task testset in terms of F-scores of multiple standard coreferential metrics macro-averaged over all datasets in the testset. We compare the measured performance to the scores calculated by the CorefUD scorer 1.0, the official scorer of the shared task, using 'strict' and 'partial' setting (denoted as exact and partial matching, respectively, in the CRAC 2022 shared task). Apart from the standard scores, it also compares the values of the anaphor-decomposable score for zeros and the MOR score, calculating the average overlap of key and response markables.

Firstly, note that all scores obtained with the 'strict' setting are significantly lower than those calculated with the 'partial' setting. It results from artificial reduction of system mentions to their heads done by the CorPipe system. They pursued this strategy in order to perform better in terms of the official metric, computed using partial matching.

Secondly, the comparison of pairs of corresponding scores measured by the two scorers confirms that the UA scorer implements processing of the CorefUD format including discontinuous markables correctly, exemplified by the identical scores with respect to the 'strict' setting. On the other hand, it also shows that partial matching is treated in a slightly different way, leading to consistently lower scores measured by UA scorer. The reason is that, unlike CorefUD scorer 1.0, the new UA scorer imposes one-to-one alignment when matching potentially overlapping markables.

Finally, the only mismatch for the 'strict' setting occurs in the MOR score. The two scorers in

Metrics	Exact		Partial	
	CorefUD	UA	CorefUD	UA
MUC	34.20	34.20	74.18	73.98
B^3	29.40	29.40	68.34	68.08
$CEAF_e$	35.93	35.93	69.64	69.40
$CEAF_m$	40.86	40.86	71.24	71.04
BLANC	28.39	28.39	64.86	64.35
LEA	22.68	22.68	65.02	64.78
CoNLL F1	33.18	33.18	70.72	70.49
Zero	60.42	60.42	83.65	83.15
MOR	45.37	26.76	45.37	44.75

Table 1: Comparison between the UA scorer and the CorefUD scorer.

fact use different mapping between key and system mentions. Whereas UA scorer uses the same mapping as for the other scores, which is based either on exact or partial matching, CorefUD scorer employs one-to-one mapping that maximizes the number of overlapping tokens regardless of the chosen matching. Two mentions that do not match even partially may still overlap. Consequently, the MOR scores outputted by CorefUD scorer are the same for each of the matching type as well as higher than those produced by the UA scorer.

6.2 CRAFT-CR 2019 Shared Task

Since the system outputs of the CRAFT-CR 2019 shared task are not publicly available, we have to find the system outputs elsewhere. We obtained the system output of the best-performing system from Lu and Poesio (2021) to compare the evaluation results between our scorer and the CRAFT-CR 2019 scorer[12] in both 'strict' and 'partial' mention matching settings.

Table 2 shows the comparison, as we can see from the 'strict' evaluation setting our scorer has the same results as their scorer. For the 'partial' setting we find their original scorer produces slightly different results if we run the scorer multiple times, whereas our scorer always produces the same results. The difference between the two scorers is within the range of the difference between two different runs of the original scorer. Hence we are convinced that the new scorer follows the same algorithm as the original scorer and can be used as a replacement for the original scorer.

[12] https://github.com/bill-baumgartner/reference-coreference-scorers

Metrics	Strict		Partial	
	CRAFT	UA	CRAFT	UA
MUC	57.69	57.69	59.74	59.78
B^3	45.43	45.43	48.03	48.02
$CEAF_e$	39.89	39.89	42.89	42.89
$CEAF_m$	51.26	51.26	53.19	53.20
BLANC	46.29	46.29	49.68	49.76
LEA	42.34	42.34	44.15	44.14
CoNLL F1	47.67	47.67	50.22	50.23

Table 2: The comparison between the UA scorer and the CRAFT-CR 2019 scorer.

7 Conclusion and Future Work

The new version of the Universal Anaphora scorer presented in this paper makes further progress towards the goal of providing the community with methods for evaluating systems carrying the full range of anaphoric interpretation. This version builds on the results of three separate shared tasks and additional research that enabled the Universal Anaphora community to test the scorer not only for a variety of types of anaphoric interpretation, but also for a range of genres covering dialogue (Khosla et al., 2021; Yu et al., 2022a) and biomedical text (Lu and Poesio, 2021), and for a variety of languages including Arabic (Aloraini et al., 2022) and the 13 languages covered in CorefUD (Žabokrtský et al., 2022). It revealed a number of limitations with the previous version of the scorer that needed addressing. We hope the community will take advantage of the new scorer to broaden the range of research on multilingual, multi-genre anaphoric interpretation.

Acknowledgements

Juntao Yu, Silviu Paun and Massimo Poesio were funded in part by the DALI project, ERC Grant 695662, and in part by the ARCIDUCA project, EPSRC grant EP/W001632/1. Michal Novák was funded by Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR) and by FW03010656 of the Technology Agency of the Czech Republic. The work described herein has also been using data provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Abdulrahman Aloraini, Sameer Pradhan, and Massimo Poesio. 2022. Joint coreference resolu-

tion for zeros and non-zeros in Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 11–21, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 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.

William Baumgartner, Michael Bada, Sampo Pyysalo, Manuel R. Ciosici, Negacy Hailu, Harrison Pielke-Lombardo, Michael Regan, and Lawrence Hunter. 2019. CRAFT shared tasks 2019 overview — integrated structure, semantics, and coreference. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 174–184, Hong Kong, China. Association for Computational Linguistics.

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.

Kevin Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner Jr., Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter. 2017. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(372).

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

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. 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.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpá nková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France. European Language Resources Association.

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

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

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proc. of the CODI/CRAC Shared Task Workshop*.

Hideo Kobayashi and Vincent Ng. 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. Association for Computational Linguistics.

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

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

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. In *Proc. of the CRAC Workshop*.

Xiaoqiang Luo. 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. Association for Computational Linguistics.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 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. Association for Computational Linguistics.

Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. 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. Association for Computational Linguistics.

Nafise S. Moosavi and Michael Strube. 2016. A proposal for a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin.

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

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. Coreference meets universal dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. ÚFAL Technical Report TR-2021-66, Charles University, Prague.

Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2023. Coreference in universal dependencies 1.1 (CorefUD 1.1). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 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.

Silviu Paun, Juntao Yu, Nafise Moosavi, and Massimo Poesio. 2023. Scoring coreference chains with split-antecedent anaphors and other entities constructed from a discourse model. *Dialogue and Discourse*.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 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. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 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. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 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.

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

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 231–235, Gothenburg, Sweden. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.

Hardik Vala, Andrew Piper, and Derek Ruths. 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. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 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*.

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

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Massimo Poesio. 2022a. The CODI/CRAC 2022 shared task on anaphora resolution, bridging and discourse deixis in dialogue. In *Proc. of CODI/CRAC Shared Task*.

Juntao Yu, Sopan Khosla, Nafise Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022b. The universal anaphora scorer 1.0. In *Proc. of LREC*.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 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). International Committee on Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 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.

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

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Ethan Zhou and Jinho D. Choi. 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. Association for Computational Linguistics.