Dialogue & Discourse 14(2) (2023) 1–48 doi: 10.5210/dad.2023.201

Scoring Coreference Chains with Split-Antecedent Anaphors

Silviu Paun* SPAUN3691@GMAIL.COM
School of Electronic Engineering and Computer Science,
Queen Mary University of London

Juntao Yu* JUNTAO.YU@QMUL.AC.UK
School of Electronic Engineering and Computer Science,
Queen Mary University of London

Nafise Sadat Moosavi N.S.MOOSAVI@SHEFFIELD.AC.UK
Department of Computer Science
University of Sheffield

Massimo Poesio M.POESIO@QMUL.AC.UK
School of Electronic Engineering and Computer Science,
Queen Mary University of London
Department of Information and Computing Science,
University of Utrecht

Editor: Barbara Di Eugenio

Submitted 12/2022; Accepted 04/2023; Published online 08/2023

Abstract

Anaphoric reference is an aspect of language interpretation covering a variety of types of interpretation beyond the simple case of identity reference to entities introduced via nominal expressions covered by the traditional coreference task in its most recent incarnation in ONTONOTES and similar datasets. One of these cases that go beyond simple coreference is anaphoric reference to entities that must be added to the discourse model via accommodation, and in particular split-antecedent references to entities constructed out of multiple discourse entities, as in split-antecedent plurals and in some cases of discourse deixis. Although this type of anaphoric reference is now annotated in many datasets, systems interpreting such references cannot be evaluated using the Reference coreference scorer (Pradhan et al., 2014). As part of the work towards a new scorer for anaphoric reference able to evaluate all aspects of anaphoric interpretation in the coverage of the Universal Anaphora initiative, we propose in this paper a solution to the technical problem of generalizing existing metrics for identity anaphora so that they can also be used to score cases of split-antecedents. This is the first such proposal in the literature on anaphora or coreference, and has been successfully used to score both split-antecedent plural references and discourse deixis in the recent CODI/CRAC anaphora resolution in dialogue shared tasks.

Keywords: Coreference, Evaluation, Split-Antecedent Anaphors

1. Introduction

Anaphoric reference is the use of language expressions to refer to entities already introduced in a discourse. The simplest case of anaphoric reference is **identity reference**, as in (1), where the

*. Equal contribution. Listed by alphabetical order

anaphoric expression *her* is a mention of the same **(discourse) entity** 1 earlier introduced with the nominal expression *Mary* (the **antecedent**),[1] whereas anaphoric expression *it* refers to the same discourse entity 2 first introduced with the nominal expression *a new dress*. Notice also that in both cases, the discourse entity is explicitly introduced with a nominal phrase, and that the anaphoric expressions refer to a single entity. We call this type of anaphoric reference **single-antecedent identity anaphora**. In Computational Linguistics (CL) / Natural Language Processing (NLP), identity anaphora is better known as **coreference** (see next Section), and the task is equivalently defined as that of clustering the sets of mentions referring to the same entity, or **coreference chains**.

(1) [Mary]$_1$ bought [a new dress]$_2$ but [it]$_2$ didn't fit [her]$_1$.

The dataset that currently serves as universally accepted reference testbed for some aspects of anaphoric interpretation including single-antecedent anaphora resolution is ONTONOTES (Pradhan et al., 2012). The performance of models for single-antecedent anaphora resolution on ONTONOTES has greatly improved in recent years (Wiseman et al., 2016; Clark and Manning, 2016; Lee et al., 2017, 2018; Kantor and Globerson, 2019; Joshi et al., 2020). As a consequence, the attention of the community has started to turn to cases of anaphora not annotated or not thoroughly covered in ONTONOTES. Examples of this trend include, for instance, research 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 on pronominal anaphors that cannot be resolved purely using gender, for which benchmarks such as GAP have been developed (Webster et al., 2018). In addition, more research has been carried out on aspects of anaphoric interpretation that go beyond identity anaphora but are covered by a great many existing datasets, such as ANCORA for Catalan and Spanish (Recasens and Martí, 2010), ARRAU (Poesio et al., 2018; Uryupina et al., 2020) and GUM for English (Zeldes, 2017), or the Prague Dependency Treebank for Czech (Nedoluzhko, 2013).[2] These aspects include, e.g., bridging reference (Clark, 1975; Hou et al., 2018; Hou, 2020; Yu and Poesio, 2020), discourse deixis (Webber, 1991; Marasović et al., 2017; Kolhatkar et al., 2018) and, finally, **split-antecedent anaphoric reference**, the type of anaphoric interpretation on which we are focusing in this paper. The best known example of split-antecedent are cases of **plural anaphoric references** such as pronoun *They* in (2) (Eschenbach et al., 1989; Kamp and Reyle, 1993), a plural anaphoric reference to a set composed of two or more discourse entities (John and Mary) introduced by separate noun phrases.

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

Split-antecedent plural references are found in all corpora and in all languages, so that more and more annotation schemes cover them, including ANCORA (Recasens and Martí, 2010), ARRAU (Poesio et al., 2018; Uryupina et al., 2020), FRIENDS (Zhou and Choi, 2018), GUM (Zeldes, 2017), *Phrase Detectives* (Poesio et al., 2019), the Prague Dependency Treebank (Nedoluzhko, 2013),

1. The term 'antecedent' is often used to refer to a *nominal expression*–one of the nominal expressions referring to the same discourse entity as the anaphoric expression (Lyons, 1977; Kamp and Reyle, 1993). In this paper we will however follow the less widely adopted use of Cornish (Cornish, 2006) and use the term to refer to the discourse entity itself. This is in part because the more traditional use of antecedent suggests that anaphora is a relation between linguistic expressions, instead of between linguistic expressions and a discourse model. In part because the antecedents of the anaphoric expressions studied here–split antecedent plurals and split-antecedent discourse deixis–are not antecedents in this sense.
2. See (Poesio et al., 2016a) for a more detailed survey and (Nedoluzhko et al., 2021) for a more recent, extensive update.

and the recently created CODI/CRAC 2021 Shared Task corpus of anaphora resolution in dialogue (Khosla et al., 2021; Yu et al., 2022a).[3] Split-antecedent plural references are not always common but e.g., in the FRIENDS corpus 9% of the mentions have more than one antecedent (Zhou and Choi, 2018). A number of computational models for the interpretation of split-antecedent plural anaphoric references have been proposed, but so far, every model has been tested using different evaluation scores (Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2020a, 2021).

But other types of anaphoric reference besides plural references can have multiple antecedents introduced ('evoked') by separate text segments in exactly the same way. This can happen, for instance, with **discourse deixis**, illustrated in (3), where the antecedent for discourse-deictic demonstrative *that* in 4.2 is the plan formed by the actions in 1.5/1.6 and 3.1, evoked by utterances from speaker M separated by utterances from speaker S.[4]

(3)

1.1	M	all right system
1.2		we've got a more complicated problem
1.3		uh
1.4		first thing I'd like you to do
1.5		is [send engine E2 off with a boxcar to Corning to pick up oranges]$_1$
1.6		uh [as soon as possible]$_1$
2.1	S	okay
3.1	M	and [while it's there it should pick up the tanker]$_2$
4.1	S	okay
4.2		and [that]$_{1,2}$ can get
4.3		we can get that done by three

Split-antecedent plural references and discourse deictic references are important from the point of view of our understanding of anaphora because they illustrate the fact that reference possibilities in discourse models are not limited to entities introduced in the discourse model via nominals–indeed, such cases were one of the reasons for, e.g., Webber's development of the idea of discourse model (Webber, 1979). Unlike simple cases of identity coreference, such cases of anaphora refer to entities that need to be **added to the discourse model at the point in which the anaphoric reference is encountered** via some form of inference. This operation on the discourse model is generally called **accommodation** (Lewis, 1979; Beaver and Zeevat, 2007). Assigning an antecedent to split-antecedent anaphors requires a particularly simple form of inference–creating a new plural object out of two atomic objects–but more complex cases are known requiring additional inferences (as in, e.g., **context change accommodation** (Webber and Baldwin, 1992; Fang et al., 2022)). These types of anaphoric reference thus test the ability of an anaphora resolution system to create antecedents *ex novo* instead of choosing them from the already introduced mentions, which is what differentiates proper discourse models from simple history lists of referents.

Assessing this ability, however, requires a scorer that can evaluate the interpretation produced by a system in cases that require accommodation. But even though split-antecedent plurals are cases of identity anaphora, they are not covered by the Reference Coreference Scorer (Pradhan et al., 2014). Simplified forms of evaluation for this type of references have been developed for discourse deixis and used, e.g., in the 2018 CRAC Shared Task (Poesio et al., 2018), but have not become part of

3. See the COREFUD report prepared for the Universal Anaphora initiative for an extensive discussion of the coverage of split-antecedent and other anaphors in the Universal Anaphora corpora (Nedoluzhko et al., 2021).

4. This example is from the TRAINS subset of the ARRAU corpus (Uryupina et al., 2020).

a standardized scorer for anaphora, and the existing metrics for discourse deixis do not work for split-antecedent discourse deixis cases such as (3). The previously proposed methods for scoring split-antecedent plural anaphors (Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2021), either work only for split-antecedent plurals in isolation, or require a substantial redefinition of the notion of coreference chain, or only generalize one of the existing metrics for coreference, as discussed in detail in Section 6.

The objective of this paper is to fill this gap in the literature, as part of the effort to develop the new Universal Anaphora scorer (Yu et al., 2022b),[5] an extension of the Reference Coreference scorer (Pradhan et al., 2014; Moosavi and Strube, 2016; Poesio et al., 2018) that can evaluate all aspects of anaphoric interpretation currently covered by the Universal Anaphora (UA) initiative,[6] including bridging reference, discourse deixis, and some cases of accommodation, and was used as the official scorer for the 2021 and 2022 CODI/CRAC Shared Tasks on Anaphoric Interpretation in Dialogue (Khosla et al., 2021; Yu et al., 2022a).[7] We start in Section 2 with a summary of the types of anaphoric interpretation the new scorer is meant to cover and a more detailed discussion of accommodation and split-antecedent plural reference. In Section 3 we review the current situation of evaluation for coreference and provide a brief description of the metrics we set to extend. Crucially, these include *all* the most widely used metrics for coreference evaluation (Luo and Pradhan, 2016): mention and entity-based metrics such as B^3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005) respectively, as well as link-based metrics such as MUC (Vilain et al., 1995), LEA (Moosavi and Strube, 2016), and BLANC (Luo et al., 2014; Recasens and Hovy, 2011). Our key contribution in this paper is a method for scoring references to accommodated objects created out of antecedents separately introduced that generalizes *all* existing metrics and thus can be used to score both single and split antecedent anaphoric reference in exactly the same way. Our proposed extensions of these metrics that can score both single and split-antecedent references are presented in Section 5 and illustrated in detail with an example in Section 4.4. Our solution is compared with all alternative proposals regarding the scoring of split antecedent anaphors in Section 6. For a more thorough demonstration of how the metrics can be used, we use the new UA scorer incorporating these extended metrics to score both split antecedent plural reference and discourse deixis with multiple antecedents. In this paper, we use this scorer to show how the proposed generalization, unlike the proposal by Zhou and Choi (2018) and our own proposal in (Yu et al., 2021), can be used to score both single and split antecedent anaphora in exactly the same way, as well as to further illustrate and analyze the behavior of our generalized metrics on the data used in previous work on split-antecedent plurals (Section 7). Finally, in Section 8 we discuss how the approach proposed here could be extended to cover other types of anaphoric interpretation requiring accommodation.

2. Anaphoric Phenomena Currently In the Scope of the Universal Anaphora Initiative

The ultimate objective of the Universal Anaphora initiative is to develop guidelines for annotating a broad range of anaphoric phenomena–not just identity anaphora–and evaluation methods to score systems carrying out the interpretation of all these types of reference. We briefly summarize here the

5. https://github.com/juntaoy/universal-anaphora-scorer
6. http://www.universalanaphora.org
7. https://competitions.codalab.org/competitions/30312. https://codalab.lisn.upsaclay.fr/competitions/614

types of anaphoric reference covered by the current version of the Universal Anaphora scorer (Yu et al., 2022b), and discuss in more detail the type of reference that is the focus of the present paper because of the lack of widely accepted evaluation methods: reference to accommodated entities, as exemplified by split-antecedent anaphora.

2.1 (Identity) Anaphora and Coreference

In much CL / NLP literature following the first Message Understanding Conference (MUC) shared tasks (Chinchor and Sundheim, 1995) a distinction is made between anaphora resolution and **coreference resolution**, and the term 'anaphora' is used to indicate pronominal anaphora only, whereas the term coreference is supposed to cover anaphoric reference with all types of nominals as well as, originally at least, predication. In this article, however, the terms 'anaphora' and 'anaphoric reference' is used in the more general sense of reference to entities in the discourse model adopted in linguistics (see, e.g., Lyons (1977); Kamp and Reyle (1993)) psycholinguistics (see, e.g., Garnham (2001)) and the pre-MUC computational work such as, e.g., Webber (1979); Carter (1987); Luperfoy (1991) (see (Mitkov, 2002; Poesio et al., 2016b; Gundel and Abbott, 2019; Poesio et al., 2023) for discussion). According to these theories of anaphoric reference, language is interpreted with respect to a **discourse model** (or **mental model**) which, in its simplest form, consists of **discourse entities** and their properties. Interpreting language involves, among other things, **updating** the discourse model with new entities as they are **mentioned**, or referring back anaphorically to the entities already introduced in the discourse model. For instance, in possibly the most widely known linguistic theory of interpretation in discourse, Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), the first mention of Mary in example (1), proper name *Mary*, introduces a new discourse entity 1 in the discourse model, and all subsequent mentions of Mary, whether using pronouns or proper names, are considered **(identity) anaphoric references** to that entity 1.

The term **coreference resolution** was introduced for the Message Understanding Conference (Chinchor and Sundheim, 1995) to specify a rather different task covering several interrelated aspects of language interpretation of interest for information extraction including not only identity anaphora resolution as in the example above but also, e.g., the association of properties with entities in cases such as (4), where the NP *an NLP researcher*, normally considered predicative from a linguistic perspective, would be considered 'co-referent' with *Mary*:

(4) [Mary]$_1$ is [an NLP researcher]$_1$.

This term 'coreference' was maintained in NLP even when criticism of the original definition of the coreference task (see, e.g., van Deemter and Kibble (2000)) led the field to adopt a revised definition focusing exclusively on (a subset of) identity anaphora in the (psycho-) linguistic sense (Passonneau, 1997; Poesio et al., 1999) adopted most notably in ONTONOTES (Pradhan et al., 2012). We will use here 'anaphora' in the traditional sense from (psycho-)linguistics, but we will adopt the characterization of identity anaphora interpretation and notation that have become standard in the computational literature, in particular in the literature defining metrics for the evaluation of this task.

Most modern anaphoric annotation projects cover the basic case of identity anaphoric reference to entities introduced via nominals in (1), although substantial differences exist regarding which types of identity anaphora are covered. Equally, the several proposed metrics for evaluating 'coreference' (reviewed in Section 3), and the Reference Coreference Scorer developed for the CONLL shared task and incorporating many of these metrics (Pradhan et al., 2014) all focus exclusively on

evaluating identity reference / identity anaphora in this sense. These metrics are defined by reference to **mentions**– the nominal phrases referring to a discourse entity–and a discourse entity i is identified with the set (cluster) K_i of mentions referring to that entity: the **coreference chain**. [8]

2.2 Beyond Identity Anaphora: The Universal Anaphora /CorefUD Specification of Anaphoric Phenomena

As already mentioned, many other types of anaphoric reference exist beyond basic identity anaphora, including, e.g., **bridging references** or **associative anaphora** (Clark, 1975; Hou et al., 2018; Hou, 2020; Yu and Poesio, 2020), and cases of anaphoric reference to entities not introduced using nominals, such as the cases of split-antecedent anaphoric reference to accommodated entities which are the focus of this paper. These other types of anaphoric reference are not covered, or are covered only partially, in ONTONOTES because of complexity and cost reasons (see (Pradhan et al., 2012) as well as the discussion in (Zeldes, 2022)), but as discussed above, they are annotated in the majority of the most recent anaphoric corpora, including ANCORA (Recasens and Martí, 2010), ARRAU (Poesio and Artstein, 2008; Poesio et al., 2018; Uryupina et al., 2020), GUM (Zeldes, 2017), *Phrase Detectives* (Poesio et al., 2019), the Prague Dependency Treebank (Nedoluzhko, 2013), the TÜBA-DZ corpus (Versley, 2008) the recently created CODI/CRAC corpus of anaphoric reference in dialogue (Khosla et al., 2021; Yu et al., 2022a), and many others (for a more complete list, see (Poesio et al., 2016a, 2023)). The objectives of the Universal Anaphora initiative include, first of all, identifying the types of anaphoric reference most frequently annotated in current corpora; second, to define shared guidelines for the annotation of these aspects of anaphoric reference; and third, to develop methods that can be used to evaluate anaphoric resolvers carrying out these types of interpretation. The list of phenomena currently considered includes the distinction between referring and non-referring expressions; identity anaphora also via zeros and split antecedent plurals; bridging reference; and discourse deixis. (Details can be found in the draft proposal on the Universal Anaphora website,[9] as well as in the CorefUD report (Nedoluzhko et al., 2021).) The Universal Anaphora scorer (Yu et al., 2022b) can evaluate the interpretation of all of the types of anaphoric reference covered by the current proposal.

2.3 Anaphoric References Requiring Accommodation

One of the types of anaphoric reference covered by the current Universal Anaphora proposal is split-antecedent anaphora: anaphoric reference to a set of entities previously mentioned separately from each other. In fact, split-antecedent anaphora is only one example of a more general class of anaphoric references that require so-called **accommodation** of a new antecedent (Lewis, 1979; van der Sandt, 1992; Beaver and Zeevat, 2007).[10] In this paper we propose a general method for scoring split-antecedent anaphora resolution that can be used both for split-antecedent plurals and for split-antecedent discourse deixis. In this Section we briefly discuss accommodation in general and the types of split-antecedent anaphoric references covered by the current proposal.

8. One of our reviewers pointed out that this use of sets of mentions as the 'referent' of anaphoric expressions is only really appropriate for anaphoric reference–a proper notion of entity is required for other types of reference, e.g., to objects in the scene in situated dialogue.
9. https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/Universal_Anaphora_1_0___Proposal_for_Discussion.pdf
10. Note that in fact bridging references discussed earlier require a form of accommodation as well: the entity related to the entity already introduced has to be added to the discourse model.

Accommodation One of the most powerful arguments for the discourse model view of anaphora, as opposed to older history-list approaches[11] is the fact that many cases of anaphoric reference cannot be interpreted with respect to the entities already introduced in the discourse model with a nominal, but require new entities to be added to the discourse model, or accommodated (Webber, 1979; Kamp and Reyle, 1993; Garnham, 2001). Accommodation as conceived by Lewis (1979) is a general operation on the discourse model which involves adding new content that is required to process a statement (Beaver and Zeevat, 2007). For instance, (5) presupposes that it was raining; when processing the statement, that fact has to be added to the discourse model.

(5) Mary realized it was raining.

Webber (1979); Kamp and Reyle (1993) and Garnham (2001) discuss a number of cases of anaphoric interpretation requiring new entities to be added to the discourse model to interpret anaphoric reference. We focus on the following two types of split-antecedent reference.

Split-antecedent plural reference In ONTONOTES, plural reference is only marked when the antecedent is mentioned by a single noun phrase. However, **split-antecedent plural reference** is also possible (Eschenbach et al., 1989; Kamp and Reyle, 1993), as in example (2). These are cases of plural identity reference, but where the antecedents are sets whose elements are two or more entities introduced by separate noun phrases, which have to be accommodated in (i.e., added to) the discourse model (Kamp and Reyle, 1993; Beaver and Zeevat, 2007). Such references are annotated in many modern datasets, including, e.g., ARRAU (Uryupina et al., 2020), GUM (Zeldes, 2017) and *Phrase Detectives* (Poesio et al., 2019; Yu et al., 2023b).

The complexities involved in evaluating systems interpreting split-antecedent plural references are illustrated in our reference (artificial) example (6). In this and other examples in the paper, mentions are delimited by square brackets; the discourse entity realized by a mention is indicated using a subscript; and the mention ID is indicated with a superscript. So, for instance, $[\text{him}^5]_1$ is a mention via noun phrase *him*; it is the 5th mention in the text; and realizes discourse entity 1. Split-antecedent references are indicated by listing all discourse entities that are part of the set, as in $[\text{their}^6]_{1,2}$. Example (6) was designed to illustrate in a concise way some of the key properties that a scorer for split-antecedent must have. First of all, that it would not be correct to interpret a split-antecedent plural such as $[\text{their}^6]_{1,2}$ in the third sentence in terms of links between mentions –say, stipulate that this plural pronoun should be interpreted by every system in terms of links between this mention and mentions $[\text{She}^4]_2$ and $[\text{him}^5]_1$. Clearly, a scorer must be able to interpret such a reference in terms of the constituent *entities* 1 and 2, so as to consider as equally correct system interpretations of the pronoun linking it to any of the mentions of these entities. Second, the example shows that plurals in a text may refer to sets of more than two entities. Third, different split antecedent plurals may refer to different sets. We use example (6) in later sections to provide a detailed illustration of how the extended metrics work.

(6) $[\text{John}^1]_1$ met $[\text{Mary}^2]_2$ after work and proposed to $[\text{her}^3]_2$ to go see a play.
$[\text{She}^4]_2$ liked the idea but suggested to $[\text{him}^5]_1$ to have dinner first.
On $[\text{their}^6]_{1,2}$ way to the restaurant, $[\text{they}^7]_{1,2}$ met $[\text{Bill}^8]_3$ and $[\text{Jane}^9]_4$.
$[\text{The two}^{10}]_{1,2}$ were very happy to see $[\text{Bill}^{11}]_3$, as $[\text{they}^{12}]_{1,2,3}$ go way back.
$[\text{He}^{13}]_3$ introduced $[\text{Jane}^{14}]_4$ and $[\text{all four}^{15}]_{1,2,3,4}$ agreed to have dinner together to catch up.

11. See e.g., Poesio et al. (2016b) for a review of early approaches to anaphoric interpretation.

Our readers might wonder whether references such as [their^6]$_{1,2}$ should be treated as bridging references–specifically, *element-inverse* associative references (Uryupina et al., 2020) to the entities referred to by the mentions [She^4]$_2$ and [him^5]$_1$. This interpretation is not inaccurate,[12] but it is incomplete, as it fails to capture the fact that John and Mary are the *entire* set of entities referred to by the plural pronoun [their^6]$_{1,2}$.

Split-antecedent plural references are not evaluated either by the standard Reference Coreference Scorer (Pradhan et al., 2014) or by our own CODI/CRAC 2018 scorer (Poesio et al., 2018), and therefore are not generally attempted by anaphoric resolvers. A few dedicated evaluation methods concerned with evaluating this type of reference were however proposed, as discussed in Section 6; but these proposals either work only for split-antecedent plurals in isolation, or require a substantial redefinition of the notion of coreference chain, or only generalize one of the existing metrics. As mentioned earlier, **the key contribution of this paper** is a method for scoring references to accommodated objects created out of antecedents separately introduced that generalizes *all* existing metrics and thus can be used to score both single and split antecedent plural anaphoric reference in exactly the same way. This extension also applies to the (much more frequent) phenomenon of split-antecedent discourse deixis, as discussed next.

Split-Antecedent Discourse deixis A second case of anaphoric reference that often also involves split antecedents, but that has been the focus of more NLP research than split-antecedent plural reference is **discourse deixis**, or anaphora with non-nominal antecedents (Webber, 1991; Byron, 2002; Gundel et al., 2003; Artstein and Poesio, 2006; Kolhatkar et al., 2018). Discourse deixis, exemplified by *this issue* in (7), which refers to an abstract object (Asher, 1993) whose type would not be easy to specify, is a type of abstract anaphora in which the antecedent is some type of abstract entity 'evoked' by the propositional content of a previous sentence. The evidence on discourse deixis interpretation suggests that these antecedents are not introduced in the discourse model immediately, but are accommodated upon encountering the anaphoric reference (Kolhatkar et al., 2018).

(7) The municipal council had to decide [whether to balance the budget by raising revenue or cutting spending]$_i$. The council had to come to a resolution by the end of the month. [This issue]$_i$ was dividing communities across the country. (Kolhatkar et al., 2018)

Discourse deixis was not annotated in many early coreference corpora (Artstein and Poesio, 2006; Kolhatkar et al., 2018), and when it was, the problem was simplified in a number of ways. In ONTONOTES, for instance, only **event anaphora**, a subtype of discourse deixis, is marked, as exemplified by *that* in (8), which refers to the event of a white rabbit with pink ears running past Alice.[13]

(8) ... 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!' ....

However, many of the more recent corpora–such as, e.g., ANCORA, GUM, ARRAU, and the CODI/CRAC corpus–do annotate the whole range of discourse deixis, covering, in addition to event anaphora, ref-

12. At least according to the annotation scheme for bridging references in the ARRAU corpus; [their^6]$_{1,2}$ would not be considered a bridging reference according to other annotation schemes such as, e.g., that for the ISNOTES corpus (Markert et al., 2012).

13. This example is from the annotated version of *Alice in Wonderland* in the *Phrase Detectives* corpus (Poesio et al., 2019; Yu et al., 2023b).

erences such as *this issue* in (7). But even in these corpora the task of annotating discourse deixis is simplified in a number of ways. One such simplification is that the annotators are not asked to mark as antecedent the accommodated abstract entity, but the list of sentences or clausal units that evoke the antecedent (e.g., the clause [whether to balance the budget by raising revenue or cutting spending] in (7)). The result of this simplification is that many if not most discourse deictic references annotated in such corpora are in fact cases of split antecedent discourse deictic reference, like the example in (3). (In fact, whereas split antecedent plurals are relatively rare–see Section 7–split antecedent discourse deictic references are very common in, e.g., ARRAU, from which example (3) was taken, and in the CODI/CRAC corpus.)

No standard metric for scoring discourse deixis resolution exists, but typically systems are scored by their ability to identify the sentence evoking the antecedent. The **success @ N** metric, introduced by Kolhatkar et al. (2013), considers a response as correct if the key sentence is among the top N candidate response sentences identified as 'antecedent' by the system. This metric was used by Marasović et al. (2017) and in the CRAC 2018 shared task (Poesio et al., 2018). The problem with this metric is that only one gold sentence per discourse deictic reference can be specified, meaning that this metric cannot be used to evaluate discourse deictic references with split antecedents, such as the example in (3).

In Universal Anaphora, discourse deixis is treated following the approach that has become standard for event anaphora (Lu and Ng, 2018). Discourse deictic references are marked in a separate discourse deixis layer, and differs from identity reference only in that the antecedents of discourse deixis are utterances rather than nominal phrases (Yu et al., 2023b). The interpretation of discourse deictic references can then be scored using the same metrics developed for the interpretation of identity reference (Yu et al., 2022b). Hence, the proposal in this paper for scoring split antecedent plural references can also be used without any change to handle split antecedent cases of discourse deixis such as the one in (3). (And was in fact used to score such cases in both editions of the CODI/CRAC shared task on anaphora resolution in dialogue.)

Context Change Accommodation As a final example of split antecedent anaphoric reference requiring accommodation we will mention the cases of **context change accommodation** discussed by Webber and Baldwin (1992) such as (9), where a new entity, the dough, is obtained by mixing together flour and water.

(9) Add [the water]$_i$ to [the flour]$_j$ little by little.
Then work [the dough]$_{i,j}$.

Context-change accommodation has recently become again the object of interest in the community and has been annotated in corpora including CHEMU-REF (Fang et al., 2021) and RECIPEREF (Fang et al., 2022). Although the type of accommodation required by this type of reference is not the focus of this paper as it involves more than combining separate entities in the discourse model in sets, we will argue in Section 8 that a version of the proposal in this paper could be used to score this type of anaphoric reference, as well, and could be argued to be more suited than the metrics used e.g., in (Fang et al., 2021, 2022).

3. Metrics for Scoring Identity Anaphora (Coreference)

One of the fundamental issues with anaphora resolution / coreference, and a reason of great dissatisfaction among many practitioners, in the fact that although the field has at various time points

converged on an 'official' metric developed for particular shared tasks, and one such metric–the CONLL score (Denis and Baldridge, 2007; Pradhan et al., 2014)–has become dominant in the last ten years (see below), it is far from clear that this metric captures our intuitions about how anaphoric interpretation should be evaluated–or indeed, what these intuitions are. This is not to say that the field is completely divided. For instance, it is universally accepted that coreference evaluation should be **entity-based**, in the sense that a system's interpretation of, e.g., (10) should be evaluated on the extent that system recognizes that 1, 2 and 3 are all mentions of the same entity (alternatively, elements of the same coreference chains), as opposed to merely its ability to identify, say, mention 1 as the previous mention of entity i that is mentioned by mention 2 ('link 2 to 1') or mention 3 as the previous mention for mention 2 ('link 3 to 2').

(10) [Mary^1]$_i$ woke up late that morning, so [she^2]$_i$ rushed out of bed– [she^3]$_i$ had an important meeting.

However, agreement that evaluation should be entity-centered still leaves many degrees of freedom, with the result that different ways have been proposed of computing precision, recall and F1 by comparing the entities (i.e., the coreference chains) in the gold annotation–in anaphora resolution these gold coreference chains are generally known as the **keys**–with those produced by a system, or **responses** (Vilain et al., 1995; Bagga and Baldwin, 1998; Luo, 2005) and no consensus has been reached on which metric is most appropriate. This impasse was broken by Denis and Baldridge (2007), who introduced a measure which was adopted in the CONLL 2011 and 2012 Shared Tasks and has since become known as the CONLL score and adopted as semi-standard (Pradhan et al., 2012, 2014), but this hasn't stopped the development of new metrics (Recasens and Hovy, 2011; Moosavi and Strube, 2016). Before introducing the proposed extensions to additionally evaluate split-antecedent references, we briefly discuss in this Section the metrics standardly used to evaluate single antecedent reference and their existing definitions. Please consult, e.g., (Luo and Pradhan, 2016) for more in-depth discussion.

3.1 Notation

The standard coreference evaluation metrics are based on the simplification discussed in Section 2.1 that discourse entities can be identified with coreference chains of mentions introduced in a text, and the assumption that each mention refers to a single entity. We use the following notation to indicate that an entity K_i is identified with a coreference chain of M_i mentions $m_{i,j}$:

$$K_i = \{m_{i,1}, m_{i,2}, ..., m_{i,M_i}\}$$

(As we will see in Section 5, one of the key proposals in this paper is a more complex representation of entities for entities *not* introduced via mentions.) Following standard convention, we use K and R to refer to entities from the key and from the response sets, respectively. The key entities represent the gold standard, whereas the response entities are the entities proposed by a system to be evaluated. The metrics to be described below adopt different approaches to comparing key and response entities.

3.2 The Standard Metrics for Anaphora / Coreference Resolution Evaluation

3.2.1 STANDARD MUC

MUC (Vilain et al., 1995) is a **link-based metric** that evaluates response entities on the basis of the number of links they have in common with the entities in the key. It is standard practice to compute this information indirectly, by counting the number of missing links, and discarding them from the maximum number of possible links, as we are about to see.[14] For the case of recall, this amounts to:

$$\text{Recall}_{MUC} = \frac{\sum_i |K_i| - |\mathcal{P}(K_i; R)|}{\sum_i |K_i| - 1}$$

In the equation above $\mathcal{P}(K_i; R)$ is a function called the **partition function**, that returns all the partitions of key entity K_i with respect to the response R of a system:

$$\mathcal{P}(K_i; R) = \{K_i \cap R_j \mid 1 \leq j \leq |R|\} \bigcup_{k_{i,u} \in K_i \setminus R} \{\{k_{i,u}\}\}$$

Notice how the number of partitions indicates the number of links found in the key entities but not in the response entities. To compute precision, we simply swap the key and response sets:

$$\text{Precision}_{MUC} = \frac{\sum_i |R_i| - |\mathcal{P}(R_i; K)|}{\sum_i |R_i| - 1}$$

The MUC metric reports as a final value an F1-measure, which is the harmonic mean between the precision and recall presented above:

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.2.2 STANDARD B^3

One problem with the MUC metric is that, by definition, it only scores a system's ability to identify links between mentions; its ability to recognize that a mention does *not* belong to any coreference chain–i.e., its ability to classify a mention as a *singleton*–does not get any reward (Bagga and Baldwin, 1998; Luo and Pradhan, 2016). The B^3 metric (Bagga and Baldwin, 1998) was proposed to correct this problem.

B^3 is a **mention-based metric**: the evaluation measures the number of mentions common between the entities in the key and in the response. Recall is computed by calculating recall for every mention, which is:

$$r(m) = \frac{|K_i \cap R_j|}{|K_i|}$$

And then summing all these mention recalls up. This can be done by finding all $|K_i \cap R_j|$ mentions in the intersection of key entity K_i and response entity R_j, summing up recall for these:

$$r(i,j) = \sum_{m \in K_i \cap R_j} r(m) = |K_i \cap R_j| * \frac{|K_i \cap R_j|}{|K_i|} = \frac{|K_i \cap R_j|^2}{|K_i|}$$

14. In MUC the maximum number of links in an entity is the minimum number of links needed to connect its mentions.

and then summing up across all i, j pairs and averaging by the total number of mentions. The result is:

$$\text{Recall}_{B^3} = \frac{\sum_{i,j} \frac{\left(|K_i \cap R_j|\right)^2}{|K_i|}}{\sum_i |K_i|}$$

Precision is computed in a similar way, again by swapping the key entities and the response entities. An F1 measure can then be computed from precision and recall in the usual way.

3.2.3 STANDARD CEAF

B^3 also suffers from a problem–namely, that a single chain in the key or response can be credited several times. This leads to anomalies–e.g., if all coreference chains in the key are merged into one in the response, B^3 recall is one (Luo, 2005; Luo and Pradhan, 2016). The solution proposed by Luo (2005), CEAF, is an **entity-level metric**: it aligns one to one the entities in the key with those in the response, and then it computes their similarity following the same function used for the alignment step. The metric comes in two flavors, depending on the similarity function used to align and compare the entities. Given a key entity K_i and a response entity R_j, **mention-based** CEAF is calculated using a similarity function measuring their mention overlap:

$$\phi_M(K_i, R_j) = |K_i \cap R_j|$$

Whereas in **entity-based** CEAF, the similarity between the two entities, is computed using the DICE coefficient:

$$\phi_E(K_i, R_j) = \frac{2\Big(|K_i \cap R_j|\Big)}{|K_i| + |R_j|}$$

At the alignment step, the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) is used to find the optimal one-to-one mapping between the entities in the key and the response such that their cumulative similarity is maximal. Let $K^* \subset K$ and $R^* \subset R$ be the key and the response entities for which an alignment was established, respectively, and let $g : K^* \rightarrow R^*$ be an alignment function storing the one-to-one mappings. Then recall is defined as the sum of the similarities over the maximal possible similarity:

$$\text{Recall}_{CEAF} = \frac{\sum_i \phi(K_i^*, g(K_i^*))}{\sum_i \phi(K_i, K_i)}$$

And again, precision is computed by swapping the entities from the key with those from the response set, and $F1_{CEAF}$ is computed from Recall_{CEAF} and Precision_{CEAF} as usual.

3.2.4 STANDARD LEA

LEA (Moosavi and Strube, 2016) is a link-based, but **entity-aware** metric which measures the resolution score of entities while taking into consideration their importance. So for instance for recall we have:

$$\text{Recall}_{LEA} = \frac{\sum_i \text{importance}(K_i) \times \text{resolution-score}(K_i)}{\sum_i \text{importance}(K_i)}$$

The importance of an entity could be defined in a different ways depending on the task, but for instance, it could be defined as being proportional to the entity's size so that more weight is given

to more frequently mentioned entities:

$$\text{importance}(K_i) = |K_i|$$

The resolution score, for recall, measures the proportion of links in the key that are recovered in the response:

$$\text{resolution-score}(K_i) = \sum_j \frac{links(K_i \cap R_j)}{links(K_i)}$$

In LEA the number of links in an entity is counted as the total number of links that can be formed between their mentions. For example, for an entity K_i, we have:

$$links(K_i) = \binom{|K_i|}{2}$$

As with the other metrics seen so far, Precision is evaluated by swapping the key with the response entities, and F1 is computed as usual.

3.2.5 STANDARD BLANC

Finally, BLANC (Recasens and Hovy, 2011; Luo et al., 2014) is an adaptation for coreference resolution of the Rand Index used in clustering (Rand, 1971). The key feature of the Rand Index is that it is computed assessing not only a clustering algorithm's decisions to put two entities in the same cluster, but also its decisions to put them in *different* clusters. BLANC adapts this approach to the case of coreference also taking into account the imbalance between the number of coreference (same cluster, aka coreference chain) vs. non-coreference (different cluster / coreference chain) links. To compute BLANC we need to determine the number of common coreference and non-coreference links found in the key and in the response entities. For a key K_i the coreference links are computed as follows:

$$C_K(i) = \{(m_{i,u}, m_{i,v}) \mid m_{i,u} \in K_i, m_{i,v} \in K_i, m_{i,u} \neq m_{i,v}\}$$

Whereas the non-coreference links between two key entities K_i and K_j are computed as follows:

$$N_K(i,j) = \{(m_{i,u}, m_{j,v}) \mid m_{i,u} \in K_i, m_{j,v} \in K_j\}$$

It follows that the set of all coreference and non-coreference links from the keys are:

$$C_K = \cup_i C_K(i), \quad N_K = \cup_{i \neq j} N_K(i,j)$$

The coreference links C_R and the non-coreference links N_R for the response entities are computed in a similar way. Precision and recall values are then computed for both coreference and non-coreference links. For example, for recall, we have:

$$\text{Recall}_C = \frac{|C_K \cap C_R|}{|C_K|}, \quad \text{Recall}_N = \frac{|N_K \cap N_R|}{|N_K|}$$

3.3 The Case for Diversity in Anaphora / Coreference Resolution Evaluation

Some subfields of NLP appear to have standardized on a single evaluation metric. The stereotypical example of this situation is perhaps Machine Translation, where the BLEU metric (Papineni et al., 2002) would appear to have been accepted as the *de facto* standard for measuring progress in the field. Another example is Summarization, where ROUGE (Lin, 2004) and its variants are equally dominant. The situation for coreference evaluation is apparently very different, and one might ask whether this should be a worry– shouldn't we just choose one of these metrics and generalize that?

We are not going to argue for choosing a particular metric in this paper. First of all, that would require providing experimental evidence that one metric is 'best' in some sense, which would be well beyond the scope of this paper. But we should also keep in mind, secondly, that the existence of multiple evaluation metrics in an NLP subfield is not unusual, and not necessarily problematic; other areas of NLP are also characterized by the existence of multiple metrics. The situation of coreference is very similar to the situation of parsing (Kakkonen, 2007), where in addition to the widely used PARSEVAL metric, developed in connection with a shared task (Black et al., 1991), a great number of other metrics exist, including, e.g., leaf-evaluation, cross-bracketing, minimum tree edit distance, and their variants. In fact, we would argue that the field has converged on a single metrics only for tasks which involve assigning a single label to a linguistic expression (e.g., part of speech tagging, named entity recognition), or for sub-fields in which the progress is driven by shared tasks organized by governments or industry, such as machine translation. Multiple metrics are the norm for tasks where the 'labels' to be compared are more complex, and it's far from obvious how these metrics should be compared. And even in fields which have converged on a single metric, such as machine translation, the debate is still raging–other metrics exist, and are often argued to be superior (see, e.g., the case of METEOR for MT (Banerjee and Lavie, 2005)).

Thirdly, we should point out that a *de facto* convergence *has* been reached in anaphora / coreference, on the CONLL metric proposed by Denis and Baldridge (2007) as the average among the F1 values obtained using MUC, B^3 and CEAF. This score was used in the CONLL Shared Tasks in 2011 and 2012 (Pradhan et al., 2012) and since then has become the standard for the field. But this metric is an average of three of the metrics introduced in this Section; so computing it requires generalizing all component metrics. We believe, therefore, that until the field reaches a new convergence, the best approach to promote the development of coreference resolvers able to resolve split antecedents is to extend all the current metrics in a way that is completely transparent to the developers of those systems in particular by extending all three metrics on which the computation of the current reference score, the CONLL score, is based, as done in this paper.

4. Generalizing Standard Metrics to Allow for Split-Antecedent References: Key Ideas and Terminology

The standard metrics for coreference resolution discussed in the previous Section all expect mentions to refer to a single entity. In this Section we describe an extension of these metrics that also allows for references to multiple entities, as in the cases of split-antecedent anaphors illustrated in (2). The generalization we propose follows the spirit of the existing metrics. The existing metrics assess the proposed single-antecedent references in the response on the basis of whether they refer to the same entity as the key; we propose that split-antecedent references should be evaluated in the same way, i.e., on whether the entities they refer to are the same. This ensures that two systems which propose as antecedents of a split-antecedent anaphor different mentions, but that refer to the

same entities, will be considered equivalent. Conversely, two systems will be considered different when a split antecedent anaphor is taken to refer to different entities by the two systems.

The key idea on which our generalisation is based on is to compare the entities referred to by a split-antecedent anaphor *using the very same metrics we set to extend*. So, for example, when scoring a system using MUC, we propose to score split-antecedent anaphora resolution according to how well the component entities of an accommodated set in the response match the key (gold) entities *according to the same* MUC *metric*. Similarly, for B^3 evaluation we use the B^3 metric, and so on for the other metrics. In this way, we can handle split-antecedent anaphora evaluation within coreference evaluation without altering the existing evaluation paradigm. We assess both single and split-antecedent references using the very same metrics, preserving their individual strengths and weaknesses, as evolved over years of research. A second important characteristic of our proposed extension is that when no split-antecedents are present, the scores produced by the extended metrics are identical to the scores obtained using their standard formulation.

In this Section, we introduce the technical ideas underlying our proposal–accommodated sets, their alignment, and the δ term responsible for comparing them–which will then used in Section 5 in which we introduce the proposed generalizations one by one, and demonstrate their computation with reference to our main example (6).

4.1 Accommodated Sets

In order to handle split-antecedent references, we generalize the notion of entity introduced in Section 3.1 to also allow entities consisting of the merge of an **accommodated object** K_i^o constructed from the discourse model (e.g., a set constructed from the existing entities in the case of split antecedent anaphors) and a traditional coreference chain K_i^m of mentions of that object. We indicate this using the following notation:

$$K_i = K_i^o \oplus K_i^m$$

Different types of accommodated objects are involved in anaphoric reference, as discussed earlier in Section 2.3 and then later in Section 8. In the case of split-antecedents anaphors, we use the notation K_i^s to refer to the accommodated object–the set that serves as antecedent for the split-antecedent reference. By definition, K_i^s is a set composed of two or more *entities*:

$$K_i^s = \{K_{i,1}, K_{i,2}, ..., K_{i,S_i}\}$$

We use the term **accommodated set** to refer to K_i^s, i.e., the set of two or more entities in the discourse model which was accommodated in the context to serve as the antecedent of a split antecedent *anaphor*.

The antecedent entities of a split-antecedent pronoun, in our representation above, are *atomic* entities, –i.e., entities all of whose mentions refer to a single antecedent. It is however possible for a split-antecedent anaphor to refer to an entity which in turn contains a split antecedent anaphor among its mentions. For instance, in the following example, the split antecedent anaphor *they all* has as split antecedents the entity Bill and the set consisting of John and Mary, which in turn was accommodated in the discourse as a consequence of the split antecedent anaphor *they* in the second utterance. (We omit mention indices in this example.)

(11) [John]$_1$ met [Mary]$_2$.
[They]$_{1,2}$ went to the movies, and met [Bill]$_3$.
Afterwards, [they all]$_{1,2,3}$ went to dinner.

In this case, we recursively replace the antecedents which are themselves accommodated sets (i.e., {1,2} = {John,Mary}) with their element entities, so that the larger accommodated set (the antecedent of *they all*) has only 'atomic' entities as its elements, i.e., only entities referred to using single-antecedent anaphors: ({1,2,3} = {John,Mary,Bill}).[15] Note that this does not lead to any loss of generality; we resolve the split-antecedent references to sets of atomic entities as those are the actual antecedents if you unpack the recursive references. (See Section 4.4 for a more detailed discussion of such cases.)

We use the following notation to indicate the set of all accommodated sets, and the set of all regular mentions (i.e., the single-antecedent references), respectively:

$$K^s = \bigcup_i K_i^s, \quad K^m = \bigcup_i K_i^m$$

The notation above was introduced for the entities in the key, but the corresponding notions will also be used for the entities in the response. Also, the formulation of the coreference metrics involves computing the cardinality of an entity. We generalize the notion of cardinality to complex entities $K_i^s \oplus K_i^m$ in the obvious way as follows:

$$|K_i^s \oplus K_i^m| = 1 + |K_i^m|$$

Finally, notice how an entity without a split-antecedent has the same representation as seen before in Section 3.1, where only single-antecedent references were allowed:

$$K_i = K_i^m = \{m_{i,1}, m_{i,2}, ..., m_{i,M_i}\}$$

4.2 Aligning Accommodated Sets

All the standard coreference evaluation metrics assume an implicit 'alignment' between the mentions in the key and in the response (the single-antecedent references). We say a mention in the key and one in the response are **aligned** if they share the same boundaries. We need to know which mentions align to compute the metrics: depending on the metric, the aligned mentions or the links between these mentions are used to compare the key and response entities following that metric's strategy, as discussed in Section 3. If only single-antecedent anaphors are present, and if response entities are well-formed, i.e., if mentions are not repeated across entities, each mention from the key is aligned with at most one mention in the response.

However, only aligning mentions is not sufficient if we also have split-antecedent anaphors. As discussed in Section 4.1, split-antecedent anaphors result in the introduction of entities containing accommodated sets. Thus, evaluating split-antecedent anaphora interpretation requires additionally aligning accommodated sets. We align the accommodated sets by aligning their element entities. If we did not specify a one-to-one alignment, the metrics would be ill-defined, i.e., contributions from multiple partially-overlapping accommodated sets may accumulate and inflate the scores.

15. This assumption amounts to adopting what has become known as the 'union' theory of plurals (Link, 1983; Schwarzschild, 1996). See (Link, 1984; Landman, 1989) for alternative views, and (Schwarzschild, 1996; Winter and Scha, 2015) for discussion.

As with the evaluation strategy briefly mentioned in the beginning of this section, we propose to align the accommodated sets using the very same metric that a system is to be evaluated with, for both single and split-antecedent anaphora. I.e., we propose to align the accommodated sets in the key entities and in the response entities for the purpose of computing metric μ using the $F1_\mu$ scores that the same metric μ returns for the element entities of those accommodated sets. For example, when computing a MUC score, we compute the alignment score between a key accommodated set K_i^s included in an entity i and a response accommodated set R_j^s from an entity j as follows:

$$\phi(K_i^s, R_j^s) = \text{F1}_{\text{MUC}}(K_i^s, R_j^s)$$

The alignment process involves finding the pairs of accommodated sets from the key and the response that lead to the largest cumulative $F1$ score. Since a brute-force approach to this problem can be computationally unfeasible, we use the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) adopted also in CEAF, which solves the alignment problem in polynomial time. Let $K^{s'} \subset K^s$ and $R^{s'} \subset R^s$ be the subsets of the aligned key and response accommodated sets; then we will use the following function (and its inverse for the reverse mappings) to access the aligned split-antecedent pairs:

$$\tau : K^{s'} \rightarrow R^{s'}$$

4.3 The δ Term

The generalization of the existing evaluation metrics that can be used to score both single and split-antecedent references proposed in this paper is uniform across all metrics and only requires an additional δ term responsible for the comparison between accommodated sets or between links involving accommodated sets, depending on the metric. As noted earlier in this Section, to compute the score between accommodated sets for metric μ we compute the score according to μ between the component entities of these accommodated sets. The accommodated sets will receive scores between 0 and 1 for how well they are resolved by a system.

As we will see in Section 5, the δ score has different interpretations for different metrics. However, two principles apply to all generalizations. The first is that if entity i contains no accommodated sets, the value of the generalized metric is equivalent to that of the original formulation (see Section 3). Second, when accommodated sets are perfectly resolved, the contribution from the accommodated sets equals that from regular mentions in the entity; they are after all just another element of the coreference chains.

4.4 The (Artificial) Illustrative Example, Revisited

We will illustrate the terminology just introduced using example (6), repeated here as example (12).

(12) [John^{1}]$_{1}$ met [Mary^{2}]$_{2}$ after work and proposed to [her^{3}]$_{2}$ to go see a play.
[She^{4}]$_{2}$ liked the idea but suggested to [him^{5}]$_{1}$ to have dinner first.
On [their^{6}]$_{1,2}$ way to the restaurant, [they^{7}]$_{1,2}$ met [Bill^{8}]$_{3}$ and [Jane^{9}]$_{4}$.
[The two^{10}]$_{1,2}$ were very happy to see [Bill^{11}]$_{3}$, as [they^{12}]$_{1,2,3}$ go way back.
[He^{13}]$_{3}$ introduced [Jane^{14}]$_{4}$ and [all four^{15}]$_{1,2,3,4}$ agreed to have dinner together to catch up.

Only the mentions of entities referred to by split-antecedent anaphors are marked in (12); other mentions that would be interpreted by coreference resolvers in a normal way, such as a *a play* or

the restaurant, are left out from our discussion, to keep things simpler, although their interpretation would also be scored by our (generalized) scorer of course.

The *key* entities mentioned in the example above are as follows (again, we use subscripts to indicate entities, and indicate the mention number using superscripts):

$$
\begin{aligned}
K_1 &= \{[\text{John}^1], [\text{him}^5]\} \\
K_2 &= \{[\text{Mary}^2], [\text{her}^3], [\text{She}^4]\} \\
K_3 &= \{K_1, K_2\} \oplus \{[\text{their}^6], [\text{they}^7], [\text{The two}^{10}]\} \\
K_4 &= \{[\text{Bill}^8], [\text{Bill}^{11}], [\text{He}^{13}]\} \\
K_5 &= \{[\text{Jane}^9], [\text{Jane}^{14}]\} \\
K_6 &= \{K_1, K_2, K_4\} \oplus \{[\text{they}^{12}]\} \\
K_7 &= \{K_1, K_2, K_4, K_5\} \oplus \{[\text{all four}^{15}]\}
\end{aligned}
$$

The first entities are atomic entities K_1 (John) and K_2 (Mary). The first accommodated set component appears with K_3: this is the set with elements entities K_1 and K_2. Notice next the accommodated set component of K_6. The anaphor [they^{12}] refers to entity K_4 (Bill) and to entity K_3 referred to by [The two^{10}] but entity K_3 in turn contains an accommodated set with elements the entities K_1 (John) and K_2 (Mary). As described in Section 4.1, we normalize the representation of accommodated sets so that they only contain atomic entities as constituents:

$$
\begin{aligned}
K_6 &= \{K_3, K_4\} \oplus \{[\text{they}^{12}]\} \\
&= \{K_1, K_2, K_4\} \oplus \{[\text{they}^{12}]\}
\end{aligned}
$$

In our illustrations of the metrics we will compare the responses produced by two hypothetical coference resolvers on example (12). Let us call the first coreference resolver 'system A'. System A outputted the following response:

$$
\begin{aligned}
R_{A,1} &= \{[\text{John}^1], [\text{him}^5]\} \\
R_{A,2} &= \{[\text{Mary}^2], [\text{She}^4]\} \\
R_{A,3} &= \{R_{A,1}, R_{A,2}\} \oplus \{[\text{their}^6], [\text{they}^7], [\text{The two}^{10}], [\text{they}^{12}]\} \\
R_{A,4} &= \{[\text{Bill}^8], [\text{Bill}^{11}], \} \\
R_{A,5} &= \{[\text{Jane}^9], [\text{Jane}^{14}]\} \\
R_{A,6} &= \{R_{A,1}, R_{A,2}, R_{A,5}\} \oplus \{[\text{all four}^{15}]\}
\end{aligned}
$$

System A made several mistakes. It did not include mention [her^3] in $R_{A,2}$ and [He^{13}] in $R_{A,4}$, respectively, and mistakenly interpreted [they^{12}] as a reference to John and Mary instead of to John, Mary and Bill. Also, System A only produced a partially-correct interpretation of the split antecedent anaphor [all four^{15}] which in the key refers to John, Mary, Bill, and Jane, whereas A only recovered 3 of the 4 constituent entities. Using the method discussed in Section 4.2, the accommodated sets in the response from system A align with those in the key as follows:[16]

16. This alignment is optimal irrespective of the similarity metric used.

$$\tau(K_3^s) = R_{A,3}^s, \quad \tau(K_6^s) = \emptyset, \quad \tau(K_7^s) = R_{A,6}^s, \quad \tau(R_{A,3}^s) = K_3^s, \quad \tau(R_{A,6}^s) = K_7^s$$

We will compare the score assigned to system A with those assigned to systems B, C and D whose output is a variation on how the accommodated set in K_7 may be resolved that raise interesting questions about the way our proposed generalizations operate:

$$R_{B,6} = \{R_{A,1}, R_{A,2}, R_{A,4}, R_{A,5}\} \oplus \{[\text{all four}^{15}]\}$$
$$R_{C,6} = \{R_{A,1}, R_{A,2}, R_{A,4}\} \oplus \{[\text{all four}^{15}]\}$$
$$R_{D,6} = \{R_{A,2}, R_{A,4}, R_{A,5}\} \oplus \{[\text{all four}^{15}]\}$$

Compared with $R_{A,6}$, the accommodated set in $R_{B,6}$ proposed by system B for mention [all four^{15}] correctly includes all 4 entities: John, Mary, Bill and Jane. The entity $R_{C,6}$ in C's response includes an accommodated set consisting of the entities John, Mary and Bill, that aligns better with the accommodated set in the key interpretation for mention [they^{12}] K_6, than with the key interpretation for mention [all four^{15}] K_7. System D's interpretation for [all four^{15}], $R_{D,6}$, also proposes 3 entities as antecedents for the anaphor, like $R_{A,6}$, but the interpretation is slightly worse than that proposed in A ($R_{A,1} = K_1$, but $R_{A,4}$ = $K_4 \setminus \{[\text{He}^{13}]\}$).

5. Generalized Definitions for the Coreference Metrics

We are now in the position to provide generalizations of the existing evaluation metrics that can be used to score both single and split-antecedent references. In this Section, we go through the standard coreference metrics one by one, showing how they are generalized using the δ term, and then illustrating their computation using example (12).

Note that it is not possible to evaluate the new generalized metrics in the traditional way, i.e., by showing that they produce more intuitive outputs than existing metrics on some examples, as done in papers introducing new coreference metrics such as (Bagga and Baldwin, 1998; Luo, 2005; Recasens and Hovy, 2011; Moosavi and Strube, 2016), for the simple reason that no generalization of all existing metrics to cover split antecedent references has been proposed before. (We discuss in Section 6 the existing proposals regarding scoring such cases, none of which involves generalizing all existing metrics, and all of which have other limitations, as discussed there.) Instead, we illustrate our generalizations and make a case for them as follows.

In this Section, we describe in detail how the scores for each metric are computed under the proposed extension with reference to example (12), with the dual objective of illustrating how our generalization works in practice and showing that the results obtained are sensible in the sense that systems producing intuitively 'better' responses get better scores. However, showing a step-by-step computation of all of the metric scores, for both single and split-antecedent references, would be tedious. Since the proposed generalization does not modify the computation of the metrics on single-antecedent anaphors, but just adds an additional δ term for the evaluation of split-antecedent references, we only illustrate here the computation of this term. For a step-by-step guide to the computation of the standard version of the metrics, i.e., defined only for single-antecedent references, see (Luo and Pradhan, 2016). Then, in Section 6, we compare our generalization metrics in

detail to the few existing and very partial previous proposals. Finally, in Section 7, we show that the extended metrics work in practice, in that they are effective at differentiating between systems which are intuitively better at the task and systems which are intuitively worse, and compare to the existing metrics, by using them to evaluate a system carrying out split-antecedent reference resolution on the datasets used in the papers in which the two most recent proposals for scoring split antecedent plural reference were made, Yu et al. (2021) and Zhou and Choi (2018).

5.1 Generalized B^3

5.1.1 DEFINITION

The simplest illustration of how the δ term is used to generalize an existing metric is our generalization of B^3. We illustrate this with $\text{Recall}^*_{B^3}$. Again, the aim is to give a system full credit when the component entities in an accommodated set in the response exactly match those in the key. (E.g., in example (12), a system would get full credit for their interpretation of [their^6] if this interpretation refers to an accommodated set R with two constituent entities R_1 and R_2 which perfectly match the two constituent entities K_1 and K_2 in the gold interpretation of [their^6].) Such perfectly resolved split-antecedent references should make the same contribution to $\text{Recall}^*_{B^3}$ as correctly identified single-antecedent references. When these conditions are not met, i.e., when the system either does not produce an accommodated set as the interpretation of a split-antecedent anaphor, or this accommodated set is not aligned with that in the key, no credit should be given. (See Section 4.2 for why alignment is necessary.) In intermediate cases, when the accommodated set in the response partially matches the accommodated set in the key, a recall score between 0 and 1 should be obtained for that mention. When the document does not contain any accommodated sets, generalized B^3 should be equivalent to standard B^3.

These goals are achieved by adding to the standard formula for Recall_{B^3} (discussed in Section 3.2.2) a δ term responsible for evaluating the resolution of split-antecedent references, if any.

$$r^*(m) = \frac{|K_i^m \cap R_j^m| + \delta_{i,j}}{|K_i|}$$

Notice that most of the B^3 formula stays unchanged (see Section 3.2.2 for a direct comparison). What has changed is that the cardinality of an entity, if it contains a split-antecedent reference, is one greater than the cardinality of the entities with single-antecedent references only.

δ for B^3 is defined as follows. When a key K_i and a response R_j contain an aligned accommodated set, i.e., $\tau(K_i^s) = R_j^s$, $\delta_{i,j}$ should specify how well the system resolved the component entities of the accommodated set. We do this using the very same Recall metric used for the evaluation of the single antecedents:

$$\delta_{i,j} = \text{Recall}_{B^3}\left(K_i^s, R_j^s\right)$$

In this way, the system is given full credit ($\delta_{i,j} = 1$) if the component entities in the accommodated set in the response exactly match (recall-wise) those in the key. I.e., perfectly resolved split-antecedent references make a contribution to recall identical to that of correctly identified single-antecedent references. By contrast, when the system does not produce an accommodated set as the interpretation of a split-antecedent anaphor or this accommodated set is not aligned with that in the key, $\delta_{i,j} = 0$. When the document does not contain any accommodated sets, $\delta_{i,j} = 0$ as well, so

that $\text{Recall}^*_{\text{B}^3}$ is equivalent to $\text{Recall}_{\text{B}^3}$. In the intermediate cases–the system response for a split-antecedent anaphor is aligned with an accommodated set in the key, but the match is not perfect– δ gets a intermediate value between 0 and 1 reflecting how well the response matches the key, as desired. (See 5.1.2.)

To get overall Recall, all the $r(m)$ are summed up as before, giving:

$$r^*(i,j) = \sum_{m \in K_i \cap R_j} r^*(m) = (|K_i^m \cap R_j^m| + \delta_{i,j}) * \frac{|K_i^m \cap R_j^m| + \delta_{i,j}}{|K_i|} = \frac{(|K_i^m \cap R_j^m| + \delta_{i,j})^2}{|K_i|}$$

and the following formula for overall recall:

$$\text{Recall}^*_{\text{B}^3} = \frac{\sum_{i,j} \frac{\left(|K_i^m \cap R_j^m| + \delta_{i,j}\right)^2}{|K_i|}}{\sum_i |K_i|}$$

To compute precision, we proceed in the usual way and replace the keys with the responses and vice-versa. $\text{F1}^*_{\text{B}^3}$ is also computed in the usual way.

5.1.2 COMPUTING GENERALIZED B^3 ON THE EXAMPLE

Let us now see in more detail how $\text{Recall}^*_{\text{B}^3}$ works with reference to example (12). Again, we focus on computing recall, as precision proceeds in the same way. The key entities K_1, K_2, K_4 and K_5 are only referred to using single-antecedent mentions. This means that there is no link to an accommodated set that the response should recover; thus, w.r.t. these entities, all the δ terms are 0, for all the systems we are considering, A, B, C and D. In these particular cases the value of the metric is not affected by our generalization.

Moving forward, let us focus first on system A. When a key and a response entity contain aligned accommodated sets, we need to credit the system for how well it resolved the accommodated set in the key. Looking at the mappings specified by the alignment function, this only happens in two cases. The first case involves the accommodated sets from the K_3 and $R_{A,3}$ entities:

$$\begin{aligned}
\delta_{A,3,3} &= \text{Recall}_{\text{B}^3}\left(K_3^s, R_{A,3}^s\right) \\
&= \text{Recall}_{\text{B}^3}\left(\{K_1, K_2\}, \{R_{A,1}, R_{A,2}\}\right) \\
&= \frac{\frac{|K_1 \cap R_{A,1}|^2}{K_1} + \frac{|K_2 \cap R_{A,2}|^2}{K_2}}{|K_1| + |K_2|} \\
&= \frac{2}{3}
\end{aligned}$$

The second is between the accommodated sets in K_7 and $R_{A,6}$:

$$\begin{aligned}
\delta_{A,7,6} &= \text{Recall}_{\text{B}^3}\left(K_7^s, R_{A,6}^s\right) \\
&= \text{Recall}_{\text{B}^3}\left(\{K_1, K_2, K_4, K_5\}, \{R_{A,1}, R_{A,2}, R_{A,5}\}\right) \\
&= \frac{8}{15}
\end{aligned}$$

In all other cases, i.e., $\forall(i,j) \in \{1,2,...,7\} \times \{1,2,...,6\} \setminus \{(3,3),(7,6)\}$, $\delta_{A,i,j} = 0$.

Let us now compare how system A resolved the split-antecedent references with the results of systems B, C, and D, similarly computed. The relevant δ terms are the following:

$$\delta_{A,7,6} = \frac{8}{15}, \quad \delta_{B,7,6} = \frac{2}{3}, \quad \delta_{C,7,?} = 0, \quad \delta_{C,6,6} = \frac{7}{12}, \quad \delta_{D,7,6} = \frac{7}{15}$$

Among the four systems, system B (correctly) gets the highest credit for identifying all 4 entity elements of the accommodated set in K_7. System C gets no credit for K_7 as that accommodated set does not align with any coreference chain in C's response (we indicate this using the notation $\delta_{C,7,?}$)–$R_{C,6}$ aligns best with entity K_6. Finally, system D gets a slightly lower score compared with system A which makes intuitive sense since entity $R_{A,1}$ is better resolved compared with $R_{A,4}$.

5.2 Generalized MUC

5.2.1 DEFINITION

The MUC metric, as well, can be generalized to score both single and split-antecedent references by including into the original formula an additional δ term responsible for scoring split-antecedent references using the MUC metric being generalized, but there is an additional complication.

As with B^3, we indicate generalized $\text{Recall}_{\text{MUC}}$ as $\text{Recall}^*_{\text{MUC}}$. When computing $\text{Recall}^*_{\text{MUC}}$, the additional δ term again measures how well the system resolves the links in the key involving accommodated sets, just as in the case of Generalized B^3. However, δ is used in a different way for MUC–instead of being added to recall, it is used to compute a **penalty term** by subtracting it from 1, defined as $1 - \delta$. If an entity K_i does not contain an accommodated set, then there is no such link for the response to recover, and therefore we want the penalty to be 0: this is obtained by having $\delta_{i,j} = 1$. When this is the case for all entities (i.e., when no split-antecedent anaphor is present in a document) then $\text{Recall}^*_{\text{MUC}}$ is equivalent to standard $\text{Recall}_{\text{MUC}}$. When a key entity does contain an accommodated set, however, we need to score the response for how well it recovers the link to this accommodated set. A response link is credited if one of its nodes matches a split-antecedent anaphor in the key and the other consists of the aligned accommodated set. When the match is perfect, we again want the penalty to be zero, but we want it to grow as the match becomes less perfect. This is done by again computing $\delta_{i,j}$ using $\text{Recall}_{\text{MUC}}$, as follows (where i is the index of a gold entity and j the index of a response entity). If $\exists(R^m_j, R^s_j)$ s.t. $R^m_j \in K^m_i$ and $\tau(K^s_i) = R^s_j$ then:

$$\delta_{i,j} = \text{Recall}_{\text{MUC}}\left(K^s_i, R^s_j\right)$$

Again, notice that we are comparing the accommodated sets in the key and in the response using the very same metric ($\text{Recall}_{\text{MUC}}$) used to evaluate the system for both single and split-antecedent references. If the key accommodated set K^s_i and the aligned response accommodated set R^s_j ($\tau(K^s_i)$) match perfectly, i.e., if their component entities are the same, we have $\delta_{i,j} = 1$, and so the system is fully rewarded for correctly producing the link to a split-antecedent contained by the key. A partial penalty is applied for an accommodated set in the response with $\text{Recall}_{\text{MUC}} < 1$, i.e., when the component entities of the accommodated set are not perfectly resolved by the system. If there is no link in the response that satisfies the aforementioned conditions, i.e., either its nodes do not contain a regular mention matching with the key, or the aligned accommodated set, then a missing link penalty is applied, and we set $\delta_{i,j} = 0$.

$\text{Recall}^*_{\text{MUC}}$ is computed by substracting the penalty term from the standard definition, as follows:

$$\text{Recall}^*_{\text{MUC}} = \frac{\sum_i |K_i| - |\mathcal{P}(K_i^m; R^m)| - (1 - \delta_{i,j})}{\sum_i |K_i| - 1}$$

The partition function $\mathcal{P}()$ takes as arguments the regular mentions portion of the entities, i.e., the single-antecedent references (cfr. Section 4.1), so that part of the MUC formula stays unchanged. As in the case of B^3, what changes is that the cardinality of an entity, if it contains a split-antecedent reference, is one greater than the cardinality of the entities with single-antecedent references only (Section 3.2.1).

To generalize MUC precision, we simply swap the entities in the key and the response, as per standard practice:

$$\text{Precision}^*_{\text{MUC}} = \frac{\sum_i |R_i| - |\mathcal{P}(R_i^m; K^m)| - (1 - \delta_i)}{\sum_i |R_i| - 1}, \quad \delta_i = \text{Precision}_{\text{MUC}}\left(K_i^s, R_j^s\right)$$

Note also that this time the δ term is computed using the precision of the metric, in accordance with the principle discussed earlier of using the very same metric for both single and split-antecedent references. $F1_{\text{MUC}}$ is computed as usual.

5.2.2 COMPUTING GENERALIZED MUC ON THE EXAMPLE

We will again focus on Recall. And again, we focus on the entities in the key that do contain split-antecedent references, whose interpretation must be found in the responses and assessed.

We start with the response provided by system A. The accommodated set in entity K_6 cannot be optimally aligned with any accommodated set in R_A; therefore, a missing link penalty is applied in this case, i.e., $\delta_{A,6,?} = 0$. In the case of the other two entities in the key containing an accommodated set, K_3 and K_7, links to an accommodated set are recovered in the response (as the conditions to have a matching regular mention and aligned accommodated sets are satisfied) and need to be evaluated. The accommodated set in $R_{A,3}$, $\{R_{A,1}, R_{A,2}\}$, is an example of a system identifying the correct number of antecedents for a split antecedent anaphor, but resolving imperfectly the antecedent entities. Its $\delta_{A,3,3}$ value is as follows:

$$\begin{aligned}
\delta_{A,3,3} &= \text{Recall}_{\text{MUC}}\left(K_3^s, R_{A,3}^s\right) \\
&= \text{Recall}_{\text{MUC}}\left(\{K_1, K_2\}, \{R_{A,1}, R_{A,2}\}\right) \\
&= \frac{\sum_{K_i \in \{K_1, K_2\}} |K_i| - |\mathcal{P}(K_i; \{R_{A,1}, R_{A,2}\})|}{\sum_{K_i \in \{K_1, K_2\}} |K_i| - 1} \\
&= \frac{2}{3}
\end{aligned}$$

The logic of this is that the entities included in the response accommodated set above have a recall of 2/3, so 1/3 is deducted from System A's Recall for the link to the accommodated set. (Note that the 'mention' component of $R_{A,3}$, $R_{A,3}^m$, is also incorrect as it includes an extra mention in comparison with K_3, but this aspect of the interpretation is not discussed here as it is not affected by our generalization.)

In the case of the accommodated set in K_7, system A recovers only 2 of the 3 antecedents; the missing link penalty is computed as follows:

$$\begin{aligned}\delta_{A,7,6} &= \text{Recall}_{\text{MUC}}\left(K_7^s, R_6^s\right)\\ &= \text{Recall}_{\text{MUC}}(\{K_1, K_2, K_4, K_5\}, \{R_{A,1}, R_{A,2}, R_{A,5}\})\\ &= \frac{1}{2}\end{aligned}$$

Let us now compare the penalty applied to system A for the interpretation of [all four[15]] with those applied to B,C and D. The relevant δ terms are:

$$\delta_{A,7,6} = \frac{1}{2}, \quad \delta_{B,7,6} = \frac{2}{3}, \quad \delta_{C,7,?} = 0, \quad \delta_{C,6,6} = 0, \quad \delta_{D,7,6} = \frac{1}{2}$$

Resulting in the following penalty (1 - δ) being deducted from the overall recall:

$$(1 - \delta_{A,7,6}) = \frac{1}{2}, \quad (1 - \delta_{B,7,6}) = \frac{1}{3}, \quad (1 - \delta_{C,7,?}) = 1, \quad (1 - \delta_{C,6,6}) = 1, \quad (1 - \delta_{D,7,6}) = \frac{1}{2}$$

We can see that system B receives a smaller penalty compared to system A, which makes sense considering B recovers all 4 entity references. System C is the one most heavily penalised by the scorer, because, first of all, the optimal alignment for the accommodated set produced by C is not the one in K_7, but that in K_6, so that C ends up completely missing the accommodated set in K_7 that the other systems get some credit for. Still, when computing recall with respect to K_6, system C continues to get a full penalty even though the accommodated sets align this time around, but a link cannot be determined because no regular mentions match. Finally, system D gets the same penalty as system A. This is because $R_{A,1}$ and $R_{A,4}$, the entities that are different between the accommodated sets from A and D, both contribute with one link when computing MUC recall. (We saw when discussing the B^3 metric earlier that system A does a boost in score over system D with that metric–this might perhaps be considered a problem with MUC).

5.3 Generalized CEAF

5.3.1 DEFINITION

As discussed in Section 3.2.3, central to computing the CEAF metric are the similarity functions used to both align and compare the key and response entities. To extend the metric to evaluate split-antecedent references as well we use, as in the other cases, an additional δ term which, as in the case of B^3, is added to both the similarity function used in the computation of mention-based CEAF, and to that used in entity-based CEAF:

$$\phi_M(K_i, R_j) = |K_i^m \cap R_j^m| + \delta_{i,j}^M, \quad \phi_E(K_i, R_j) = \frac{2\left(|K_i^m \cap R_j^m| + \delta_{i,j}^E\right)}{|K_i| + |R_j|}$$

It is only when a key K_i and a response R_j contain aligned accommodated sets –i.e., $\tau(K_i^s) = R_j^s$– that we need to evaluate how well their component entities were resolved. δ for recall is computed as follows:

$$\delta_{i,j}^M = \text{Recall}_{\text{CEAF}^M}\left(K_i^s, R_j^s\right), \quad \delta_{i,j}^E = \text{Recall}_{\text{CEAF}^E}\left(K_i^s, R_j^s\right)$$

When there are no accommodated sets or they are not aligned, $\delta_{i,j} = 0$. Otherwise, $\text{Recall}_{\text{CEAF}}$ is computed exactly as discussed in Section 3.2.3, and so for $\text{Precision}_{\text{CEAF}}$ and F1_{CEAF}.

5.3.2 COMPUTING GENERALIZED CEAF ON THE EXAMPLE

As with B^3, it is only when a key and a response entity contain aligned accommodated sets that we need to evaluate how well the two match. Again, we start with the computations for system A. We have seen there is an alignment in two cases, one between the accommodated sets in entities K_3 and R_3:

$$\delta^{M/E}_{A,3,3} = \text{Recall}_{\text{CEAF}^{M/E}}\left(K^s_3, R^s_{A,3}\right) = \begin{cases} \frac{4}{5} & \text{for mention-based CEAF} \\ \frac{9}{10} & \text{for entity-based CEAF} \end{cases}$$

The calculation is done for recall, precision is analogous. The other accommodated set alignment is between those in the K_7 and $R_{A,6}$ entities:

$$\delta^{M/E}_{A,7,6} = \text{Recall}_{\text{CEAF}^{M/E}}\left(K^s_7, R^s_{A,6}\right) = \begin{cases} \frac{3}{5} & \text{for mention-based CEAF} \\ \frac{7}{10} & \text{for entity-based CEAF} \end{cases}$$

All other cases either involve entities without accommodated sets, or whose accommodated sets do not align, so we have $\delta^{M/E}_{i,j} = 0$. We now introduce, for comparison, the relevant δ terms for the other systems:

$$\delta^{M}_{A,7,6} = \frac{3}{5}, \quad \delta^{M}_{B,7,6} = \frac{4}{5}, \quad \delta^{M}_{C,7,?} = 0, \quad \delta^{M}_{C,6,6} = \frac{3}{4}, \quad \delta^{M}_{D,7,6} = \frac{3}{5}$$
$$\delta^{E}_{A,7,6} = \frac{7}{10}, \quad \delta^{E}_{B,7,6} = \frac{9}{10}, \quad \delta^{E}_{C,7,?} = 0, \quad \delta^{E}_{C,6,6} = \frac{13}{15}, \quad \delta^{E}_{D,7,6} = \frac{13}{20}$$

As with the other metrics presented so far system B gets the highest score for identifying all four entities from the split-antecedent reference in K_7. System C is not allocated any credit for the accommodated set in K_7, only for the one from K_6, because of how the split-antecedents get aligned. And system D is found on par with system A when using mention-based CEAF and slightly worse (as it intuitively should) when the evaluation is conducted using entity-based CEAF. This is another illustration of the strengths and weaknesses of the existing metrics for coreference evaluation which our scorer inherit. For the computation of the rest of the metrics, the observations for systems B, C, and D will be similar to those expressed so far, and will be omitted. From now on, we will only calculate the scores for system A to illustrate the methods, as the calculations for the other metrics are the same.

5.4 Generalized LEA

5.4.1 DEFINITION

To extend LEA to evaluate both single and split-antecedent references we modify both the *importance* and the *resolution-score* functions. Starting with the former, we define the importance function to additionally include a β term to further reward entities which contain an accommodated set:

$$\text{importance}(K_i) = \beta_i |K_i|$$

The resolution score function is defined as in the standard version of the metric, but computed differently to also consider split-antecedent references:

$$\text{resolution-score}(K_i) = \sum_j \frac{links(K_i \cap R_j)}{links(K_i)}$$

Counting the number of links in K_i is trivial: $links(K_i) = \binom{|K_i|}{2}$. But special attention needs to be paid when counting the number of links between the set of mentions in common between a key K_i and a response R_j:

$$links(K_i \cap R_j) = \binom{|K_i^m \cap R_j^m|}{2} + \delta_{i,j} \times \Big(|K_i^m \cap R_j^m| - 1\Big)$$

We distinguish two types of links between the mentions common to both a key and a response entity. First, we have links between regular mentions present in both entities, i.e., links between single-antecedent references. The number of these links is expressed in the first term of the equation above. The second type are links involving an accommodated set. When a key K_i and a response R_j contain aligned accommodated sets, i.e., $\tau(K_i^s) = R_j^s$, we need to assess how well their element entities compare. Again, this can be done just using LEA's notion of Recall:

$$\delta_{i,j} = \text{Recall}_{\text{LEA}}\Big(K_i^s, R_j^s\Big)$$

After evaluating how well the key and the response accommodated sets compare, we use this information to weigh the number of links that can have an accommodated set as a node; this is expressed in the second term from the link counting formula presented earlier. When there are no accommodated sets in the entities, or when they are not aligned, we set $\delta_{i,j} = 0$.

The generalized version of the metric is equivalent to the standard version presented in Section 3.2.4 when the entities in the key and response do not contain any accommodated sets (when $\delta_{i,j} = 0$, we have $link(K_i \cap R_j) = \binom{|K_i^m \cap R_j^m|}{2}$). When accommodated sets do exist, however, and they were perfectly resolved by a system, the scorer allocates full credit to each of these links involving accommodated sets, just as it does for the links between correctly identified single-antecedent references: when $\delta_{i,j} = 1$, $links(K_i \cap R_j) = \binom{|K_i^m \cap R_j^m|+1}{2}$. For imperfectly resolved accommodated sets the credit allocated to a system lies in-between the two extremes.

5.4.2 COMPUTING GENERALIZED LEA ON THE EXAMPLE

When computing LEA, for those key and response sets that contain aligned accommodated sets, we need to evaluate how well these accommodated sets compare. For recall, we have:

$$\delta_{i,j} = \begin{cases} \text{Recall}_{\text{LEA}}\Big(K_3^s, R_{A,3}^s\Big) & \text{for } i = j = 3 \\ \text{Recall}_{\text{LEA}}\Big(K_7^s, R_{A,6}^s\Big) & \text{for } i = 7, j = 6 \\ 0 & \text{otherwise} \end{cases}$$

When computing precision, LEA precision is used instead to evaluate accommodated sets.

5.5 Generalized BLANC

5.5.1 DEFINITION

We saw back in Section 3.2.5 that to compute BLANC we need to establish the *coreference* and the *non-coreference* links found in the key and in the response entities. In the standard version of the metric these links are only between regular mentions, i.e., between single-antecedent references. In the generalized version, in which entities may also include accommodated sets, we additionally distinguish two types of links: (i) links where both nodes are accommodated sets,[17] and (ii) links where one node is an accommodated set, and the other is a single-antecedent reference.

BLANC evaluates a response by comparing the coreference and non-coreference links in the response set with those in the key. In standard BLANC this is done by simply computing the intersection of these sets. In the generalized version of the metric, however, we cannot do this anymore, due to the introduction of accommodated sets and the additional types of links that get created, as discussed above. To help us compare different types of links we introduce a new function $\delta()$ that takes as argument two links–one from the key (m_1^k, m_2^k), the other from the response (m_1^r, m_2^r)–and specifies how to allocate them credit. In short, this function will allocate full credit (i.e., a value of 1) to links between regular mentions that match, and partial credit (a value between 0 and 1) to those key and response links whose nodes involve aligned accommodated sets. The partial credit in this case will depend on how well the element entities in the accommodated sets compare. The function will not allocate any credit (a value of 0) to all other pairs of links. We can assess how the coreference and the non-coreference links in the response compared to those in the key by evaluating the credit allocated by the function above to all pairs of links found between these sets. If no accommodated sets are present in the entities in the key and the response, using the function as described has the same effect as the set intersection operation mentioned before used in standard BLANC.

We illustrate below how the $\delta()$ function is used to compute recall for the non-coreference links:

$$R_N = \frac{1}{|N_K|} \sum_{\substack{(m_1^k, m_2^k) \in N_K \\ (m_1^r, m_2^r) \in N_R}} \delta\Big((m_1^k, m_2^k), (m_1^r, m_2^r)\Big)$$

A link from the key and one from the response whose nodes are regular mentions receive a credit of 1 if their mentions match. Formally, if $m_1^k, m_2^k \in K^m, m_1^r, m_2^r \in R^m$, and $m_1^k = m_1^r, m_2^k = m_2^r$ (or $m_1^k = m_2^r, m_2^k = m_1^r$), then:

$$\delta\Big((m_1^k, m_2^k), (m_1^r, m_2^r)\Big) = 1$$

Two links one of whose nodes is a regular mention while the other is an accommodated set are scored based on how well the accommodated sets compare, assuming they are aligned, and that the regular mentions match. In line with the rest of the metric extensions, we compare accommodated sets by comparing their element entities using the very same metric we are evaluating the system with overall, for both single and split-antecedent references. The current example is for the recall of the non-coreference links, so this is the metric that is used here as well. Formally, if $\exists m_s^k, m_m^k \in \{m_1^k, m_2^k\}$ s.t. $m_s^k \in K^s, m_m^k \in K^m$, and $\exists m_s^r, m_m^r \in \{m_1^r, m_2^r\}$ s.t. $m_s^r \in R^s, m_m^r \in R^m$, and

17. These can occur among the non-coreference links.

$m_m^k = m_m^r, \tau(m_s^k) = m_s^r$ then:

$$\delta\Big((m_1^k, m_2^k), (m_1^r, m_2^r)\Big) = \text{Recall}_N(m_s^k, m_s^r)$$

Two links whose nodes are aligned accommodated sets receive credit on the basis of how well their element entities compare. Formally, if $m_1^k, m_2^k \in K^s, m_1^r, m_2^r \in R^s$, and $\tau(m_1^k) = m_1^r, \tau(m_2^k) = m_2^r$ $\big($or $\tau(m_1^k) = m_2^r, \tau(m_2^k) = m_1^r\ \big)$ then:

$$\delta\Big((m_1^k, m_2^k), (m_1^r, m_2^r)\Big) = \text{Recall}_N\Big(m_1^k, \tau(m_1^k)\Big) \times \text{Recall}_N\Big(m_2^k, \tau(m_2^k)\Big)$$

All other links in the key and response are unrelated and receive no credit. For these, we have:

$$\delta\Big((m_1^k, m_2^k), (m_1^r, m_2^r)\Big) = 0$$

All the other computations required by BLANC, i.e., the precision of the non-coreference links, together with both the precision and the recall of the coreference links, are computed in a similar fashion. BLANC then reports as its final value the arithmetic mean of the F1 values for the coreference and the non-coreference links.

5.5.2 COMPUTING GENERALIZED BLANC ON THE EXAMPLE

For this metric we need to determine how the coreference and the non-coreference links in the key and response entities compare. The link space is large, but let us look, for example, at $N_K(3, 7)$ and $N_R(3, 6)$, the sets of non-coreference links between keys K_3 and K_7, and response entities $R_{A,3}$ and $R_{A,6}$, respectively. Starting with the former, we have:[18]

$$N_K(3,7) = \Big\{(K_3^s, K_7^s), (K_3^s, 15), (6, K_7^s), (6, 15), (7, K_7^s), (7, 15), (10, K_7^s), (10, 15)\Big\}$$

The non-coreference links between the response entities $R_{A,3}$ and $R_{A,6}$ are:

$$\begin{aligned} N_R(3,6) = \Big\{&(R_{A,3}^s, R_{A,6}^s), (R_{A,3}^s, 15), (6, R_{A,6}^s), \\ &(6, 15), (7, R_{A,6}^s), (7, 15), (10, R_{A,6}^s), (10, 15), (12, R_{A,6}^s), (12, 15)\Big\} \end{aligned}$$

Let us now consider how matching links are determined in a recall-based evaluation. Two links, one from the key, and the other from the response, whose nodes are regular mentions, get full credit if their mentions match. In our example that happens in 3 cases:

$$\begin{aligned} \delta\Big((6, 15), (6, 15)\Big) &= 1 \\ \delta\Big((7, 15), (7, 15)\Big) &= 1 \\ \delta\Big((10, 15), (10, 15)\Big) &= 1 \end{aligned}$$

18. We shall use the id of the mentions, single or split-antecedent, for a more concise representation.

Two links, where one of the nodes is a regular mention, and the other an accommodated set, are given credit if the regular mentions match and the accommodated sets are aligned. The allocated credit depends on how well the accommodated sets evaluate:

$$\delta\Big((K^s_3, 15), (R^s_{A,3}, 15)\Big) = \text{BLANC}_{R_N}(K^s_3, R^s_{A,3})$$
$$\delta\Big((6, K^s_7), (6, R^s_{A,6})\Big) = \text{BLANC}_{R_N}(K^s_7, R^s_{A,6})$$
$$\delta\Big((7, K^s_7), (7, R^s_{A,6})\Big) = \text{BLANC}_{R_N}(K^s_7, R^s_{A,6})$$
$$\delta\Big((10, K^s_7), (10, R^s_{A,6})\Big) = \text{BLANC}_{R_N}(K^s_7, R^s_{A,6})$$

Two links both of whose nodes are accommodated sets receive credit if the accommodated sets are aligned, and the score depends on how well the response evaluates against the key:

$$\delta\Big((K^s_3, K^s_7), (R^s_{A,3}, R^s_{A,6})\Big) = \text{BLANC}_{R_N}\Big(K^s_3, R^s_{A,3}\Big) \times \text{BLANC}_{R_N}\Big(K^s_7, R^s_{A,6}\Big)$$

There is no alignment for all other key and response links, so no credit can be allocated in these cases, and $\delta = 0$. Finally, notice we used R_N to compare the entities included in the accommodated sets, as the computations above were used to determine the credit allocated to non-coreference links in a recall-based evaluation. Computing the credit for the coreference links involves the same steps, but using R_C instead. And when turning to precision, P_N and P_C, the precision related metrics from BLANC are used.

6. Alternative proposals to score split-antecedent anaphora

There is limited previous work on split-antecedent anaphora resolution and its evaluation. We are aware of four proposals, two of which we put forward ourselves in previous work.

Metrics for gold evaluation only Vala et al. (2016) and Yu et al. (2020a) only evaluate their models on split-antecedent selection task that assumes gold anaphors and mentions were provided. They compute precision, recall, and F1 measures based on the links between split-antecedent anaphors and their antecedent. Recall is defined as the percentage of gold split-antecedent anaphoric links that are correctly identified by the system (where each atomic antecedent of a split-antecedent is counted as a separate link). Precision is the percentage of split-antecedent anaphoric links proposed by a system that is correct. Because they only evaluate on the split-antecedent selection task, the evaluation is much simpler, i.e. it does not require any form of alignment. However, such an evaluation has a number of limitations: it is not entity-based, and is not realistic, as it is based on the assumption that the gold mentions and gold anaphors will be provided.

Distributing the plural mentions among the coreference chains for singular entities Zhou and Choi (2018) propose a method to evaluate split-antecedent plural references resolution using the standard CONLL scorer. This is done by adding the plural mention to each of the clusters for its atomic elements: for example, they would represent the coreference chain

$$\{\{\text{John}_i, \text{Mary}_j\}, \text{They}^t\}$$

encoded in this paper as the discourse entity K_t including an accommodated set :

$$K_i = \{\text{John}_i^{i_1}, ...\text{John}_i^{i_m}\}$$
$$K_j = \{\text{Mary}_j^{j_1}, ...\text{Mary}_j^{j_n}\}$$
$$K_t = \{K_i, K_j\} \oplus \{\text{They}_t^{t_1}\}$$

by adding [They^{t_1}]$_t$ to the coreference chains for John and Mary, i.e., as the following two gold coreference chains each consisting of all mentions in one individual entity plus the split-antecedent reference:

$$\{[\text{John}^{i_1}]_i, \ldots [\text{John}^{i_m}]_i, [\text{They}^{t_1}]_t\}$$
$$\{[\text{Mary}^{j_1}]_j, \ldots [\text{Mary}^{j_n}]_j, [\text{They}^{t_1}]_t\}.$$

First of all, this representation clearly violates the fundamental assumption behind the notion of coreference chain, i.e., that all mentions in the chain refer to the same entity. In the two coreference chains, mention [They^{t_1}]$_t$ does not refer to the same entity as mention [John^{i_k}]$_i$ of *John* or mention [Mary^{j_l}]$_j$ of *Mary*: it refers to the set consisting of John and Mary. In addition, this approach also causes problems with the evaluation and may produce counterintuitive results.

One example of a problem is that the proposed representation cannot be used with the MUC metric. MUC relies on the partition function to compute the number of links in common between the key and response entities, and the function requires all mentions to participate in a single cluster. As a result, the authors do not use the MUC metric in their evaluation. Regarding counterintuitive results, the authors themselves point out that the approach may have unpredictable effects on the alignment of the CEAF^E score. The following example illustrates this. Suppose we have a text which first introduces the entity John, mentioned four times; the entity Mary, mentioned once; the entity Bill, also mentioned once; followed by three mentions of the set {John,Mary,Bill}. This situation is schematically encoded by the following key entities:

$$K_1 = \{\text{John}^1, \text{John}^2, \text{him}^3, \text{he}^4\}$$
$$K_2 = \{\text{Mary}^5\}$$
$$K_3 = \{\text{Bill}^6\}$$
$$K_4 = \{K_1, K_2, K_3\} \oplus \{\text{they}^7, \text{them}^8, \text{their}^9\}$$

And suppose that the response entities predicted by a system are as follows:

$$R_1 = \{\text{John}^1\}$$
$$R_2 = \{\text{Mary}^5\}$$
$$R_3 = \{R_1, R_2\} \oplus \{\text{they}^7, \text{them}^8, \text{their}^9\}$$

In the approach proposed by Zhou and Choi (2018), the plural mentions [they^7], [them^8] and [their^9] would be represented by distributing them into the singular clusters (e.g. K_1, R_1) –i.e., the key and response entities would be represented as follows:

$$K_1 = \{\text{John}^1, \text{John}^2, \text{him}^3, \text{he}^4, \text{they}^7, \text{them}^8, \text{their}^9\}$$
$$K_2 = \{\text{Mary}^5, \text{they}^7, \text{them}^8, \text{their}^9\}$$
$$K_3 = \{\text{Bill}^6, \text{they}^7, \text{them}^8, \text{their}^9\}$$

$$R_1 = \{\text{John}^1, \text{they}^7, \text{them}^8, \text{their}^9\}$$
$$R_2 = \{\text{Mary}^5, \text{they}^7, \text{them}^8, \text{their}^9\}$$

Intuitively, the system's coreference chain for John, R_1, should be aligned with K_1; distributing the plurals should not change this alignment. But in reality, R_1 will be aligned with K_3, because after adding the plurals the similarity score between R_1 and K_3 is 3/8, which is higher than the score of 4/11 between R_1 and K_1. By contrast, our approach will correctly align the singular and plural clusters (i.e. K_1 and R_1; K_2 and R_2, and K_4 with R_3).

Generalizing LEA Finally, in our own previous work (Yu et al., 2021), we proposed an extension of the LEA metric to score split-antecedent plural references. A first obvious difference between the proposal in this paper and that earlier proposal is that back then we only proposed a generalization for LEA, whereas here we generalize every one of the most used coreference metrics discussed in Section 5. But another reason for the current proposal is that after publishing that paper we found issues with the approach to alignment we had used.

The generalization proposed in that paper scores split-antecedent references in two steps, each of which requires an alignment step. In the first step, the singular coreference chains in the response are aligned with the gold coreference chains. The atomic antecedents of the split-antecedent references found in the response are then replaced by the aligned gold coreference chains. The alignment between the coreference chains in the key and response in this step is established using CEAF^E. After that, a second alignment step is used between the converted split-antecedents when computing the link scores in the LEA. Using the notation introduced in Section 4.1, our early approach can be formulated as follows. A gold accommodated set K_i^s is aligned with an accomodated set R_j^s in the response, i.e., $\tau(K_i^s) = R_j^s$, if R_j^s is the accommodated set in the system entities that has the largest number of coreference chains in common with K_i^s, i.e., $R_j^s = \text{argmax}_{R_{j*}^s} |K_i^s \cap R_{j*}^s|$. Following our presentation of LEA in Section 5.4, the aligned accommodated sets, when evaluating recall, are assessed based on the recall of their mentions:

$$\delta_{i,j} = \frac{|K_i^s \cap R_j^s|}{|K_i^s|}$$

However, a one-to-one alignment between the accommodated sets was not imposed, hence there might be multiple accommodated sets in the key being aligned to the same accommodated sets in the response, and vice-versa. This is potentially problematic as the coreference metrics are based on the assumption that a mention can only participate in one coreference chain. The following example shows how the alignment used in our early approach might cause some instability in the metrics.

Suppose we have the following key and response entities:

$$K_1 = \{\text{John}^1, \text{John}^2, \text{him}^3, \text{he}^4\}$$
$$K_2 = \{\text{Mary}^5\}$$
$$K_3 = \{\text{Bill}^6\}$$
$$K_4 = \{K_1, K_2\} \oplus \{\text{they}^7\}$$
$$K_5 = \{K_1, K_2, K_3\} \oplus \{\text{all three}^8\}$$

$$R_1 = \{\text{John}^1, \text{John}^2, \text{him}^3, \text{he}^4\}$$
$$R_2 = \{\text{Mary}^5\}$$
$$R_3 = \{\text{Bill}^6\}$$
$$R_4 = \{R_1, R_2\} \oplus \{\text{they}^7\}$$
$$R_5 = \{R_1, R_3\} \oplus \{\text{all three}^8\}$$

(I.e., the system correctly resolved atomic entities 1, 2 and 3 and plural entity 4, but then only identified two of the antecedents for split antecedent reference [all three^8].) In our approach in Yu et al. (2021), we will first align the singular coreference chains (e.g. K_1, R_2) without accommodated sets using CEAF^E, and then replace the 'atomic' coreference chains that are part of accommodated sets in the response with the aligned gold atomic coreference chains. As a result, R_4 and R_5 will become:

$$R_4 = \{K_1, K_2\} \oplus \{\text{they}^7\}$$
$$R_5 = \{K_1, K_3\} \oplus \{\text{all three}^8\}$$

After that, when computing the actual LEA score, the accommodated set part of K_4, K_4^s ($\{K_1, K_2\}$), will be aligned to R_4^s, which is expected as they are identical. K_5^s ($\{K_1, K_2, K_3\}$) intuitively should be aligned to R_5^s, since R_4^s has already been aligned. But because there is no one-to-one alignment restriction, K_5^s is equally likely to be aligned to either R_4^s or R_5^s, as both have the same $\delta_{i,j}$ of 2/3 (two out of three clusters are overlapping). The actual outcome solely depends on the order in which the response split-antecedents are presented. By contrast, with the new method proposed in this paper, K_5^s will be aligned correctly with R_5^s due to the one-to-one alignment constraint.

A second drawback of Yu et al. was caused by the pre-alignment step on the atomic entities using the CEAF^E score. The pre-alignment step is deterministic, and imposes a hard alignment between key and response coreference chains; and the later scoring step does *not* take into account how good the alignments are. This is problematic as the hard alignment in this stage is not robust enough to align multiple equivalent response clusters, as illustrated by the following example. Suppose we have the following key entities:

$$K_1 = \{\text{John}^1, \text{John}^2, \text{him}^3, \text{he}^4\}$$
$$K_2 = \{\text{Mary}^5\}$$
$$K_3 = \{K_1, K_2\} \oplus \{\text{they}^6\}$$

And consider the following two responses, which differ on the interpretation of [they^6]:

$$R_1 = \{\text{John}^1, \text{him}^3\}$$
$$R_2 = \{\text{John}^2, \text{he}^4\}$$
$$R_3 = \{\text{Mary}^5\}$$
$$R_4 = \{R_1, R_3\} \oplus \{\text{they}^6\}$$
$$R'_4 = \{R_2, R_3\} \oplus \{\text{they}^6\}$$

The two system predictions for [they^6], R_4 and R'_4, should be equivalent, since they both get half of K_1 and all of K_2. However, because the pre-alignment step requires a hard alignment, either R_1 or R_2 will be aligned with K_1. Suppose R_1 is aligned with K_1; the Yu et al.'s scorer will then give a score of 100% to $\{R_1, R_3\}$ but 50% to $\{R_2, R_3\}$. By contrast, the new scorer proposed in this paper will correctly assign the same scores to both interpretations.

In order to provide a more empirical comparison between our new approach to evaluation and the one we proposed in Yu et al. (2021), in the next Section we use our new generalizations to score the same system tested by Yu et al. (2021) on the same corpus. We also test the same systems on the FRIENDS corpus used by Zhou and Choi (2018) to provide a baseline for future work.

7. Using the Metrics for Scoring Generalized Anaphoric Reference

Because no generalizations of the existing coreference metrics to cover split-antecedent anaphoric reference was previously proposed, it is not possible to compare the proposal in this paper to previous ones except on individual examples, as done in the previous Section. However, it is possible to show that, unlike the existing and partial solutions discussed in the previous Section, the generalized metrics proposed here can be used to compare anaphoric resolvers in exactly the same way as done using the existing coreference metrics.

Our generalized metrics were incorporated in the new Universal Anaphora scorer, a new scorer for anaphora that can also score split antecedent anaphora resolution (as well as non-referring mentions identification, bridging reference resolution and discourse deixis resolution) (Yu et al., 2022b, 2023a), and this scorer was then used to score the performance of a state of the art system able to carry out split-antecedent anaphora resolution on real data (Yu et al., 2021) and to compare it with simpler baselines, both on the corpus used by Yu et al. (2021) and on the corpus used by Zhou and Choi (2018). In this Section we present the results of this evaluation.

Split-antecedent anaphoric references are rarer compared to single entity anaphoric references, at least for the case of plural split antecedent references (split discourse deictic references are much more common). As a result, they typically do not make a significant contribution to the overall evaluation score. Thus, to offer a clear picture of the performance of a system on split-antecedents, our scorer also reports separate scores for the split-antecedent references only. The scores for the separate evaluation of split-antecedents are the micro-average F_1 of all the aligned gold and system pairs. Those accommodated sets for which an alignment could not be found (e.g., missing or spurious accommodated sets) were paired with empty sets when computing the scores.

7.1 The models being compared

In order to evaluate the generalized coreference metrics on the output of a real system, we obtained the best-performing output and all the baselines from the system by Yu et al. (2021), the only mod-

	MUC			B^3			$CEAF^E$			CoNLL
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**F1**
Recent 2	76.6	77.3	76.9	78.8	76.1	77.5	78.0	74.3	76.1	76.8
Recent 3	76.7	77.3	77.0	78.9	76.2	77.5	78.0	74.4	76.1	76.9
Recent 4	76.8	77.3	77.0	79.0	76.2	77.5	78.0	74.4	76.1	76.9
Recent 5	76.7	77.3	77.0	79.0	76.2	77.5	78.0	74.3	76.1	76.9
Random	76.5	77.1	76.8	78.8	76.0	77.4	77.9	74.2	76.0	76.7
Single Ant	76.3	78.4	77.4	78.7	76.9	77.8	78.0	74.4	76.2	77.1
Yu et al.	77.1	77.9	77.5	79.1	76.5	77.8	78.1	74.5	76.3	77.2
Oracle	77.6	78.4	78.0	79.5	76.8	78.1	78.3	74.6	76.4	77.5

(a) MUC, B^3, $CEAF^E$ and CONLL F1.

	$CEAF^M$			BLANC			LEA (β=1)			LEA(β=10)		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Recent 2	77.2	75.4	76.3	75.3	71.5	73.4	70.3	66.9	68.6	59.7	61.2	60.5
Recent 3	77.3	75.4	76.3	75.3	71.5	73.4	70.4	66.9	68.6	60.3	61.3	60.8
Recent 4	77.3	75.4	76.3	75.3	71.6	73.4	70.4	66.9	68.6	60.7	61.5	61.1
Recent 5	77.3	75.4	76.3	75.3	71.6	73.4	70.4	66.9	68.6	60.7	61.1	60.9
Random	77.2	75.3	76.2	75.3	71.5	73.3	70.3	66.7	68.4	59.3	59.9	59.6
Single Ant	77.2	75.8	76.5	75.2	72.3	73.7	70.3	67.4	68.8	59.0	67.4	62.9
Yu et al.	77.4	75.7	76.6	75.5	71.9	73.6	70.7	67.2	68.9	62.8	64.7	63.7
Oracle	77.8	75.9	76.8	75.9	72.1	74.0	71.2	67.6	69.4	65.9	68.4	67.1

(b) $CEAF^M$, BLANC and LEA with different split-antecedent importance (β).

Table 1: Evaluation of Yu et al. (2021)'s model and of the baselines on ARRAU using the new Universal Anaphora scorer.

	MUC	**B^3**	**$CEAF^E$**	**$CEAF^M$**	**BLANC**	**LEA**	**CoNLL**
Recent 2	25.6	22.2	15.9	23.6	15.8	21.9	21.2
Recent 3	27.1	24.2	20.8	26.3	18.2	23.5	24.0
Recent 4	28.0	25.2	21.0	27.5	17.5	24.4	24.7
Recent 5	26.6	23.6	19.0	26.0	15.9	22.9	23.1
Random	19.6	15.3	7.9	16.8	11.4	14.8	14.3
Yu et al.	35.8	31.9	37.0	32.8	18.2	30.9	34.9
Oracle	70.1	63.7	62.9	68.1	68.6	61.4	65.6

Table 2: Split-antecedent F1 scores only for Yu et al. (2021)'s system on ARRAU evaluated using our extension of the coreference scorers.

ern system that can process both single- and split-antecedent anaphors, and ran our scorer on all six predictions. The results are illustrated in Tables 1 -3. Yu et al. (2021)'s model is an extension of the system proposed by Yu et al. (2020b) which further interprets split-antecedents. The model shares most of the network architecture proposed by Yu et al. (2020b), but in addition, it includes a dedicated feed-forward network for split-antecedents. The baselines are based on heuristic rules or random selection. The same candidate split-antecedent anaphors/singular clusters are used in

	Yu et al.			Our approach		
	Lenient$_{split}$	**LEA (β=1)**	**LEA (β=10)**	**LEA$_{split}$**	**LEA (β=1)**	**LEA (β=10)**
Recent 2	12.4	68.7	61.4	21.9	68.6	60.5
Recent 3	17.0	68.7	61.4	23.5	68.6	60.8
Recent 4	16.0	68.7	61.5	24.4	68.6	61.1
Recent 5	13.1	68.7	61.3	22.9	68.6	60.9
Random	4.5	68.5	60.4	14.8	68.4	59.6
Yu et al.	36.4	69.0	64.1	30.9	68.9	63.7

Table 3: Comparison between our generalized scores and the score proposed in Yu et al. (2021) on the Yu et al.'s system output evaluated on the ARRAU corpus.

the baselines as in the Yu et al. (2021) model. Then these baselines attempt to interpret the mentions belonging to a small list of plural pronouns which could be interpreted as split antecedent anaphor (e.g., *they*, *their*, *them*, *we*) but were classified as not having an atomic antecedent by the single-antecedent anaphoric resolver, and attempt to resolve them as split-antecedent anaphors. The **random** baseline randomly assigns two to five antecedents to these candidate split antecedent anaphors. After that, the **recent-x** baseline uses the x closest singular clusters as antecedents to each chosen anaphor. In addition, we also include scores for the system only resolving single antecedent anaphors (Single Ant) and system augmented with the gold split-antecedent anaphors (Oracle)[19].

7.2 Evaluation on ARRAU

Table 1 shows the overall scores (single and split-antecedent anaphors) on ARRAU evaluated using the new scorer.[20] The general direction of the results doesn't change from those reported in Yu et al. (2021) (see below). What changes is that whereas in Yu et al. (2021) the system's performance on split antecedents could only be evaluated using a single, *ad-hoc* metric (see, e.g., Tables 3 and 5 in that paper), thanks to the extension proposed in this paper it is now possible to score both single-antecedent and split-antecedent anaphors using the same metrics.

The results with all the metrics confirm the results obtained by Yu et al. with their specialized metric. First of all, as already observed by Yu et al., the difference between the baselines and the best model is small when single- and split-antecedent anaphors are evaluated together, because the number of split-antecedents is low (only 0.8% of the clusters containing split-antecedents). However, the difference becomes very clear when considering the performance on split antecedents only (see Table 2): up to 20 percentage points in CONLL score. The Oracle setting has again much better split-antecedent F1 scores, and this results in considerable improvements on all the scores when evaluated with singular clusters. Confirming our hypothesis on the need for partial credit on split-antecedents, we find that only 16% of the split-antecedents were fully resolved by the Yu et al. model. The vast majority of the split-antecedents that were partially resolved rely on the partial credit to get a fair assessment.

19. For oracle setting we allow the system to use the gold split-antecedent anaphors annotations when possible. Please note the system is still constrained by the quality of singular clusters. This simulates a better system on resolving the split-antecedent anaphors.
20. Following Yu et al. (2021), we only report scores for documents in which at least one split-antecedent anaphor is annotated.

We further compare our new and previous approaches (Yu et al., 2021) in Table 3. Since Yu et al. only generalised the LEA score, the comparison is primarily based on LEA. For each approach, LEA scores were computed with two different split-antecedent importance ($\beta \in \{1, 10\}$). A large β means more weight is given to the split-antecedents. We additionally include the most relevant split-antecedent-only score (Lenient$_{split}$ and LEA$_{split}$ for Yu et al. and our approach respectively) to assess the correlation between the split-antecedent-only scores and the LEA scores. A s we can see from the Table, when $\beta = 1$ both approaches successfully register a difference between the Yu et al. and the baselines, but do not show a visible difference between the baselines apart from the 'Random' setting that has a lower score overall. This is not surprising given that the number of split-antecedent anaphors is small. When we increase the split-antecedent importance (i.e. $\beta = 10$) the score differences become more visible; however, we noticed that the LEA score from our previous approach does not correlate well with the split-antecedent only score. Even though the 'Recent 3' baseline has a higher Lenient$_{split}$ than 'Recent 2' and 'Recent 4', it has the same LEA score (61.4%) as 'Recent 2' and lower than 'Recent 4' 's 61.5%. On the other hand, the new approach reports LEA scores that follow the same trend as its split-antecedent scores (i.e. LEA$_{split}$), which makes the evaluation more consistent between the split-antecedent only and overall scores.

7.3 Evaluation on FRIENDS

In order to compare our metrics in practice with the evaluation approach proposed by Zhou and Choi (2018), we further tested our extended metrics on the FRIENDS corpus used in their paper, which contains a larger percentage of split-antecedent plural references. We follow Zhou and Choi (2018) in using episodes 1 - 19 for training, 20, 21 for development and 22, 23 for testing. The original corpus is annotated for entity linking, so the coreference clusters are created by grouping the mentions that refer to the same entity (a character in the show FRIENDS). 14.5% of those clusters contain split-antecedents. In contrast, in the original annotation, split-antecedent anaphors represent 9% of all mentions; this is because in the original version of the FRIENDS corpus all subsequent mentions of a set accommodated using a split-antecedent anaphor are also marked as split-antecedents instead of being coreferent with the first mention. (E.g. in our illustrative example (12), [they^{7}] and [The two^{10}] are treated as single-antecedent by putting them in the same cluster as [their^{6}], whereas in the FRIENDS corpus would be annotated as split-antecedent references as well.) We transformed all of these cases into single-antecedent anaphors; after this transformation, 4.1% of the mentions remain split-antecedent anaphors.

To obtain the system predictions, we trained the Yu et al. (2021) system on the FRIENDS corpus and computed all the baselines in the same way as for the ARRAU corpus.[21] As shown in Table 4, the Yu et al. model outperforms the baselines by a large margin according to all the metrics even though the performance improvements on split-antecedent anaphors (see Table 5) are smaller than those we observed with the ARRAU evaluation. This was expected, as the FRIENDS corpus contains many more split-antecedent anaphors. When comparing the Yu et al. model with the single-antecedent only system (Single Ant), the model has a better recall but a lower precision, overall having similar F1 scores for most of the matrices. This is because system performance on the split-antecedent part is not good enough to make a clear difference. With the oracle setting, however , the better performance on split-antecedent anaphors contributed to a robust improvement

21. We contacted the authors of Zhou and Choi (2018) to obtain their system's outputs but did not get a reply. It should also be noted that they evaluated their system in a gold mention setting which is not realistic.

	MUC			B^3			$CEAF^E$			CoNLL
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**F1**
Recent 2	80.0	80.4	80.2	71.0	72.0	71.5	59.1	61.7	60.4	70.7
Recent 3	80.0	80.6	80.3	71.1	72.1	71.6	59.3	61.6	60.4	70.8
Recent 4	80.1	81.0	80.6	71.1	72.5	71.8	59.9	61.9	60.9	71.1
Recent 5	79.6	81.3	80.4	70.6	72.9	71.7	60.2	61.8	61.0	71.1
Random	79.8	79.9	79.9	71.0	71.4	71.2	60.6	61.3	61.0	70.7
Single Ant	77.9	85.0	81.3	69.1	77.0	72.8	64.0	63.0	63.5	72.6
Yu et al.	80.7	82.2	81.4	71.7	73.5	72.6	63.6	64.8	64.2	72.7
Oracle	81.5	83.9	82.7	72.5	75.3	73.9	66.5	65.3	65.9	74.2

(a) MUC, B^3, $CEAF^E$ and CONLL F1.

	$CEAF^M$			BLANC			LEA (β=1)			LEA(β=10)		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Recent 2	70.8	71.8	71.3	73.1	75.9	74.5	60.9	62.0	61.4	43.1	42.8	42.9
Recent 3	70.8	71.9	71.3	73.3	76.2	74.7	61.1	62.0	61.5	43.4	42.6	43.0
Recent 4	70.8	72.2	71.5	73.3	76.5	74.9	61.3	62.3	61.8	43.5	43.8	43.7
Recent 5	70.5	72.3	71.4	73.2	76.9	75.0	60.9	62.5	61.7	40.9	43.9	42.3
Random	71.1	71.5	71.3	73.5	75.7	74.6	60.8	61.1	61.0	42.5	39.8	41.1
Single Ant	70.4	74.7	72.5	72.4	80.1	76.0	60.5	65.5	62.9	34.8	65.5	45.5
Yu et al.	72.0	73.3	72.6	73.6	76.8	75.2	63.0	64.2	63.6	47.6	50.5	49.0
Oracle	73.0	74.1	73.6	74.6	78.1	76.3	64.6	66.2	65.4	53.4	61.2	57.1

(b) $CEAF^M$, BLANC and LEA with different split-antecedent importance (β).

Table 4: Evaluation of Yu et al. (2021)'s model and of the baselines on the FRIENDS Corpus using the new Universal Anaphora scorer (Yu et al., 2022b, 2023a).

	MUC	**B^3**	**$CEAF^E$**	**$CEAF^M$**	**BLANC**	**LEA**	**CoNLL**
Recent 2	43.4	37.5	30.9	42.7	28.3	36.1	37.3
Recent 3	45.1	38.9	32.0	43.8	30.7	37.3	38.6
Recent 4	44.0	37.7	31.2	42.7	30.2	36.0	37.6
Recent 5	42.8	36.3	28.5	41.4	30.6	34.5	35.9
Random	43.5	36.9	34.2	42.2	31.1	35.3	38.2
Yu et al.	52.3	44.9	43.3	50.8	37.8	43.3	46.8
Oracle	65.5	57.1	65.6	65.1	50.9	54.8	62.7

Table 5: Split-antecedent F1 scores only for Yu et al.'s systems evaluated on the FRIENDS corpus using our generalization of the coreference scorers.

on overall performance on both single- and split-antecedent anaphors. This indicates a better split-antecedent anaphora resolver is needed to achieve a significant improvement when compared with systems that only resolve single-antecedent anaphors. If split-antecedent anaphors are the main focus of the evaluation, one can use the split-antecedent F1 scores to look at the split-antecedent only scores (see Table 5) or the LEA metric with appropriate split-antecedent importance (e.g. $\beta = 10$) to prioritise split-antecedent anaphors.

	Zhou&Choi			Our approach		
	B^3	**CEAF^E**	**BLANC**	**B^3**	**CEAF^E**	**BLANC**
Recent 2	71.2	53.1	76.5	71.5	60.4	74.5
Recent 3	71.2	53.1	76.4	71.6	60.4	74.7
Recent 4	71.2	52.9	76.3	71.8	60.9	74.9
Recent 5	70.8	52.4	76.0	71.7	61.0	75.0
Random	70.7	51.8	75.8	71.2	61.0	74.6
Sing Ant	70.2	54.1	75.0	72.8	63.5	76.0
Yu et al.	74.2	59.1	79.0	72.6	64.2	75.2
Oracle	75.9	62.6	80.3	73.9	65.9	76.3
Gold Sing Ant	89.3	82.2	89.1	96.1	95.0	97.0

Table 6: Comparison between our generalization of the coreference scorers and the Zhou and Choi (2018) method on the Yu et al.'s system output evaluated on the FRIENDS corpus.

In Table 6 we compare in detail our scorer with that of Zhou and Choi (2018). We use the scorer developed by Zhou and Choi to score the baselines and system output and compare all three metrics proposed by them with our equivalents. As we can see from the Table, the main difference is that the Zhou and Choi approach heavily penalises systems that do not predict split-antecedents: the 'Sing Ant' system that only predicts references to coreference chains without accommodated antecedents scores lower than the baselines in most of the cases. We suggest this is a result of the decision to put the split-antecedent anaphors into the relevant singular clusters. By doing so, the plural mentions are credited multiple times in the evaluation, hence the split-antecedent anaphors gain more weight than normal mentions. To give a clearer picture of the scorer's behaviour on penalising the missing split-antecedents, we also provide in Table 6 the scores on evaluating the gold singular clusters (Gold Sing Ant). You can see this as a system that predicts all singular clusters correctly and only misses the links between the split-antecedent anaphors with their antecedents. Given that about 4% of mentions in the corpus are split-antecedent references, we would expect the scores to be close to 96%. The Zhou and Choi scorer, however, penalises the system by 10% - 18%, way beyond the credit that should be assigned to split-antecedent anaphors. In contrast, our scorer gives scores between 95% - 97% ,which is close to the expectation. Although in some circumstances one might want to give more credit to split-antecedent anaphors if this is the main focus, it might not be a good idea to assume this is the mainstream. Instead of permanently boosting the importance of split-antecedent anaphors as done in Zhou and Choi, we posit that it would be better to allow users (e.g. the shared task organizers) to decide how important should the split-antecedent anaphors be, as we have done in LEA, using the split-antecedent importance parameter (β) to configure the importance of split-antecedent plurals flexibly.

8. Scoring other types of split-antecedent anaphora and of anaphora involving accommodation

As discussed in the Introduction and in Section 2, split-antecedent plurals are just one example of anaphoric reference referring to an entity which wasn't previously mentioned, thus requiring accommodation of a new antecedent (Beaver and Zeevat, 2007; van der Sandt, 1992). In all of these

cases, the new entity is composed of a part constructed out of the pre-existing discourse model, together with a 'coreference chain' part–i.e., the structure proposed here for split antecedent plurals:

$$K_i = K_i^o \oplus K_i^m$$

The difference is the relation linking the new entity to existing entities in the context. For split antecedents, this relation is set membership: the new set is the set of the entities mentioned by the split antecedents. This type of accommodation is required not just for plurals, but for discourse deixis as well. In the case of bridging references, the relation is associative, not coreference. In the case of context change accommodation, the new entity is typically the result of an action carried out over the entities in the context. So, the proposed notation could potentially also serve as the basis for extensions covering these cases.

Split-antecedent discourse deixis As discussed in Section 2, discourse deixis (Webber, 1991; Kolhatkar et al., 2018), may also involve split-antecedents, as shown by example (3). We also mentioned there that in Universal Anaphora, discourse deixis is treated following the approach that has become standard for event anaphora (Lu and Ng, 2018)–i.e., the interpretation of discourse deictic references is scored using the same metrics developed for the interpretation of identity reference (Yu et al., 2022b). (Discourse deictic references are marked in a separate discourse deixis layer, but this layer differs from the identity reference only in that the antecedents of discourse deixis are utterances rather than nominal phrases.) The extension of these metrics proposed in this paper can therefore be immediately used for cases of split-antecedent discourse deixis such as (3). And indeed, this approach was used to score discourse deixis, including split-antecedent discourse deixis, in the CODI/CRAC shared tasks (Khosla et al., 2021; Yu et al., 2022a).

Context change accomodation A more complex case of split-antecedent anaphoric reference are the cases of **context change accommodation** discussed by Webber and Baldwin (1992), where a new entity, the dough, is obtained by mixing together flour and water.

(13) Add [the water]$_i$ to [the flour]$_j$ little by little.
Then work [the dough]$_{i+j}$

Context change accommodation was not annotated in any of the best-known datasets for anaphoric reference discussed in Section 2, so no evaluation method was proposed to our knowledge. However, as mentioned earlier, the community has started to study context-change accommodation again recenly, and it has been annotated in CHEMU-REF (Fang et al., 2021) and RECIPEREF (Fang et al., 2022). In these corpora, context-change anaphoric references are treated as cases of bridging references, and systems interpreting them have been evaluated in this way. This approach suffers however from the same problem as treating split-antecedent plural references as cases of bridging discussed in Section 2.3–namely, that it fails to capture the fact that the water and the flour are *all* of the antecedents of 'the dough,' and therefore there is no way of rewarding a system for identifying *all* antecedents, or penalizing a system for recovering only some of them. Using the approach proposed in this paper would better reflect the semantics of these cases, although some work is needed to work out the implications of the fact that the water and the flour are not simply elements of a set, but the two components of a piece of matter.

9. Conclusions

In order to push forward the state of the art in anaphora resolution beyond the simplest form of identity anaphora it is not sufficient to create suitable datasets annotated with the more general cases of anaphora, although that is an important effort. It is also necessary to develop methods for evaluating the performance of anaphoric resolvers on these cases. In this paper we proposed a method for evaluating one of these more general cases–the case of anaphoric reference to entities that need introducing in a discourse model via accomodation, exemplified by split-antecedent anaphors but including other cases as well, such as discourse deixis–that is a straightforward extension of existing proposals for coreference evaluation and thus does not require introducing additional metrics, an issue in a field already over-rich with proposals in this direction.

Acknowledgments

This research was supported in part by the DALI project, ERC Grant 695662, and in part by the ARCIDUCA project, EPSRC grant number EP/W001632/1.

References

Ron Artstein and Massimo Poesio. Identifying reference to abstract objects in dialogue. In *BRANDIAL 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)*, Potsdam, 2006.

Nicholas Asher. *Reference to Abstract Objects in English*. D. Reidel, Dordrecht, 1993.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Workshop on Linguistics Coreference at the First International Conference on Language Resources and Evaluation (LREC)*, volume 1, pages 563–566. Association for Computational Linguistics, 1998.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

David Beaver and Henk Zeevat. Accommodation. In G. Ramchand and C. Reiss, editors, *The Handbook of Linguistic Interfaces*, pages 503–536. Oxford, 2007.

Ezra Black, Steve Abney, Dan Flickinger, C. Gdaniec, Ralph Grishman, Paul Harrison, Don Hindle, Rob Ingria, Fred Jelinek, Judith Klavans, Mark Liberman, Mitch Marcus, S. Roukos, Beatrice Santorini, and Tomek Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*. Association for Computational Linguistics, 1991. URL https://aclanthology.org/H91-1060.

Donna K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Philadelphia,

Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073099. URL https://aclanthology.org/P02-1011.

David M. Carter. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK, 1987.

Nancy A. Chinchor and Beth Sundheim. Message Understanding Conference (MUC) tests of discourse processing. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford, 1995.

Herbert H. Clark. Bridging. In *Theoretical Issues in Natural Language Processing*, 1975.

Kevin Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1061. URL https://www.aclweb.org/anthology/P16-1061.

Francis Cornish. Discourse anaphora. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 631–638. Oxford University Press, 2nd edition, 2006.

Pascal Denis and Jason Baldridge. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester, New York, 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1030.

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. Remarks on plural anaphora. In *Proceedings of the 4th Conference of the European chapter of the Association for Computational Linguistics*, pages 161–167. Association for Computational Linguistics, 1989.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. ChEMU-Ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 1362–1375. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.116. URL https://aclanthology.org/2021.eacl-main.116.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 3481–3495. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.275. URL https://aclanthology.org/2022.findings-acl.275.

Alan Garnham. *Mental models and the interpretation of anaphora*. Psychology Press, 2001.

Jeanette K. Gundel and Barbara Abbott, editors. *The Oxford Handbook of Reference*. Oxford University Press, 2019.

Jeanette K. Gundel, Michael Hegarty, and Kaja Borthen. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic Language and Information*, 12(3):281–299, 2003.

Yufang Hou. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.132. URL https://www.aclweb.org/anthology/2020.acl-main.132.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl_a_00300. URL https://aclanthology.org/2020.tacl-1.5.

Tuomo Kakkonen. *Framework and Resources for Natural Language Parsing Evaluation*. PhD thesis, University of Joensuu, 2007.

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

Ben Kantor and Amir Globerson. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1066. URL https://www.aclweb.org/anthology/P19-1066.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-sharedtask.1. URL https://aclanthology.org/2021.codi-sharedtask.1.

Varada Kolhatkar, Heike Zinsmeister, and Graeme Hirst. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 300–310, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1030.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612, 2018. doi: 10.1162/coli_a_00327. URL https://www.aclweb.org/anthology/J18-3007.

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

Fred Landman. Groups I. *Linguistics and Philosophy*, pages 559–605, 1989.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://www.aclweb.org/anthology/D17-1018.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 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, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2108. URL https://www.aclweb.org/anthology/N18-2108.

David K. Lewis. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359, 1979.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Godehard Link. The logical analysis of plurals and mass terms: A lattice- theoretical approach. In R. Bäuerle, C. Schwarze, and A. von Stechow, editors, *Meaning, Use and Interpretation of Language*, pages 302–323. Walter de Gruyter, 1983.

Godehard Link. Hydras: On the logic of relative clause constructions with multiple heads. In F. Landman and F. Veltman, editors, *Varieties of Formal Semantics*. Foris, Dordrecht, 1984.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2344–2350, 2017. doi: 10.24963/ijcai.2017/326. URL https://doi.org/10.24963/ijcai.2017/326.

Jing Lu and Vincent Ng. Event coreference resolution: A survey of two decades of research. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization, 2018. doi: 10.24963/ijcai.2018/773. URL https://doi.org/10.24963/ijcai.2018/773.

Xiaoqiang Luo. 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, 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1004.

Xiaoqiang Luo and Sameer Pradhan. Evaluation metrics. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, pages 147–170. Springer, 2016.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 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, 2014. As-

sociation for Computational Linguistics. doi: 10.3115/v1/P14-2005. URL https://www.aclweb.org/anthology/P14-2005.

Susan Luperfoy. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*. PhD thesis, The University of Texas at Austin, Dept. of Linguistics, Austin, TX, 1991.

John Lyons. *Semantics*. Cambridge University Press, 1977.

Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. 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, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1021. URL https://www.aclweb.org/anthology/D17-1021.

Katja Markert, Yufang Hou, and Michael Strube. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Juju island, Korea, 2012. URL http://www.aclweb.org/anthology/P12-1084.

Ruslan Mitkov. *Anaphora Resolution*. Longman, 2002.

Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1060. URL https://www.aclweb.org/anthology/P16-1060.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Anna Nedoluzhko. Generic noun phrases and annotation of coreference and bridging relations in the Prague dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2313.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. Coreference in universal dependencies 0.1 (CorefUD 0.1), 2021. URL http://hdl.handle.net/11234/1-3510. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Rebecca J. Passonneau. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript., December 1997.

Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/297_paper.pdf.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Workshop Towards Standards and Tools for Discourse Tagging*, 1999. URL https://aclanthology.org/W99-0309.

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

Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors. *Anaphora Resolution*. Springer Berlin Heidelberg, 2016b. doi: 10.1007/978-3-662-47909-4. URL https://doi.org/10.1007%2F978-3-662-47909-4.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0702. URL https://www.aclweb.org/anthology/W18-0702.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1176. URL https://aclanthology.org/N19-1176.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. Computational models of anaphora. *Annual Review of Linguistics*, 2023.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-4501.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 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, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2006. URL https://www.aclweb.org/anthology/P14-2006.

Altaf Rahman and Vincent Ng. 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, 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1071.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846—-850, 1971. doi: 10.2307/2284239.

Marta Recasens and Eduard Hovy. BLANC: Implementing the Rand Index for coreference evaluation. *Natural language engineering*, 17(4):485–510, 2011.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Roger S. Schwarzschild. *Pluralities*. Kluwer, 1996.

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

Hardik Vala, Andrew Piper, and Derek Ruths. 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, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1216. URL https://www.aclweb.org/anthology/P16-1216.

Kees van Deemter and Rodger Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637, 2000. URL https://aclanthology.org/J00-4005.

Rob A. van der Sandt. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9 (4):333–377, 1992.

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

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 1995. URL https://www.aclweb.org/anthology/M95-1005.

Bonnie L. Webber. *A Formal Approach to Discourse Anaphora*. Garland, New York, 1979.

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

Bonnie Lynn Webber and Breck Baldwin. Accommodating context change. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Newark, Delaware, USA, 1992. Association for Computational Linguistics. doi: 10.3115/981967.981980. URL https://aclanthology.org/P92-1013.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. doi: 10.1162/tacl_a_00240. URL https://www.aclweb.org/anthology/Q18-1042.

Yoad Winter and Remko Scha. Plurals. In Shalom Lappin and Chris Fox, editors, *The Handbook of Semantic Theory*, chapter 3, pages 77–113. Wiley Blackwell, 2nd edition, 2015.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1114. URL https://www.aclweb.org/anthology/N16-1114.

Juntao Yu and Massimo Poesio. Multitask learning based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.315. URL https://www.aclweb.org/anthology/2020.coling-main.315.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online), 2020a. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.538. URL https://www.aclweb.org/anthology/2020.coling-main.538.

Juntao Yu, Alexandra Uma, and Massimo Poesio. A cluster ranking model for full anaphora resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 11–20, Marseille, France, 2020b. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.2.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 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, 2021. URL https://arxiv.org/abs/2104.05320.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–14, Gyeongju, Republic of Korea, 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.codi-crac.1.

Juntao Yu, Sopan Khosla, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. The universal anaphora scorer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4873–4883, Marseille, France, 2022b. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.521.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. The universal anaphora scorer 2.0. In *Proceedings of the International Workshop on Computational Semantics (IWCS)*, 2023a.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Carretero Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and wikipedia texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 767–781, Dubrovnik, Croatia, 2023b. Association for Computational Linguistics (ACL). URL https://aclanthology.org/2023.eacl-main.54.

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

Amir Zeldes. Can we fix the scope for coreference? *Dialogue and Discourse*, 13(1):41–62, 2022. doi: https://doi.org/10.5210/dad.2022.102. URL https://journals.uic.edu/ojs/index.php/dad/article/view/11706.

Ethan Zhou and Jinho D. Choi. 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, 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1003.