Natural Language Engineering (2023), 1–30
doi:10.1017/S1351324923000293

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Linguistically aware evaluation of coreference resolution from the perspective of higher-level applications

Voldemaras Žitkus, Rita Butkienė and Rimantas Butleris

Department of Information Systems, Kaunas University of Technologies, Kaunas, Lithuania
Corresponding author: Voldemaras Žitkus; E-mail: voldemaras.zitkus@ktu.lt

(Received 11 February 2021; revised 29 April 2023; accepted 17 May 2023)

Abstract

Coreference resolution is an important part of natural language processing used in machine translation, semantic search, and various other information retrieval and understanding systems. One of the challenges in this field is an evaluation of resolution approaches. There are many different metrics proposed, but most of them rely on certain assumptions, like equivalence between different mentions of the same discourse-world entity, and do not account for overrepresentation of certain types of coreferences present in the evaluation data. In this paper, a new coreference evaluation strategy that focuses on linguistic and semantic information is presented that can address some of these shortcomings. Evaluation model was developed in the broader context of developing coreference resolution capabilities for Lithuanian language; therefore, the experiment was also carried out using Lithuanian language resources, but the proposed evaluation strategy is not language-dependent.

Keywords: Coreference resolution; Linguistic information; Dominant mentions; Coreference evaluation

1. Introduction

Coreference resolution has been a topic of research since the late 1970s (Hobbs 1978) and remains an active field of investigation. An advance in coreference resolution gives more opportunities to increase the quality of such applications like semantic search, summarization, question answering, and others. The quality of these applications highly depends not only on information retrieval methods applied but also on information extraction methods used as well. In general, information extraction is known as an activity of automatically extracting structured information from unstructured information source. A classical information extraction model requires standard text preprocessing. It includes lexical analysis, morphological analysis, and named entity recognition, while some information extraction solutions are getting complemented by more advanced ones such as coreference resolution.

In natural language texts, coreference occurs when two or more different linguistic structures refer to the same discourse-world entity. Resolving a relationship between these structures is an important part of the natural language processing (NLP) task and can greatly improve the performance of information retrieval systems.

The following sentence can be used to illustrate this:

- Tom skipped school today. He was sick and called the ambulance. When the ambulance arrived, the boy was already outside.

Here the words "Tom," "He," and "boy" refer to the same discourse-world entity. Without resolving the relationship between these structures, we would not be able to determine, why Tom skipped school, who was sick, and who was waiting outside when the ambulance arrived. In such cases, we would lose semantic information and some end user queries would be left not answered. Therefore, the quality of the information search service implemented by the higher-level application depends on the quality of coreference resolution.

Usually, coreference resolution tools return only chains of linked pairs of linguistic structures, which serves as the input to other applications. The quality of resolution (usually measured by precision, recall, and F-measure) is the main indicator for the application developer and also for the resolution tool developer. The higher-level application developer relies on the tool's evaluation scores when deciding whether to use a particular coreference resolution tool, while a tool developer relies on it when identifying the weak points of the annotator. Therefore, it is very important to have a reliable evaluation approach of the coreference resolution. However, this field remains unsolved fully yet. Since the 1990s, many different metrics have been suggested: MUC (Vilain *et al.* 1995), B^3 (Bagga and Baldwin 1998), CEAF (Luo 2005), LEA (Moosavi and Strube 2016), BLANC (Recasens 2010), ARCS (Tuggener 2014), and PARENT (Kaczmarek and Marcińczuk 2015). But none of them has been accepted as the standard for the field. Different approaches often solve coreferences of different types (nominal and pronominal) and use different methods (deterministic, statistical, and machine learning), due to that they cannot be formally compared to each other in order to determine the superior one. It is also not entirely clear how their outputs should be compared. Moreover, the higher-level application can make specific requirements for the implementation of coreference resolution. For example, the semantic search application needs coreferences resolved to the same entity. But the same entity can be referenced to by multiple different noun phrases, and it is not efficient to show all of them for end user in query results. It would be useful to know which noun phrase best represents the entity. Therefore, a tool satisfying such requirement should include additional data to coreference annotations, like representative mentions in Stanford CoreNLP (Lee *et al.* 2011). Naturally, evaluation metrics created before such additions were made are not able to evaluate them. Due to that, there is a continuing need to improve the evaluation process.

In this paper, a new linguistically aware coreference evaluation strategy is proposed. It expands on already existing linguistically aware metrics by combining preference for the most relevant mention of the discourse-world entity with the addition of coreference type identification to the evaluation process. Type identification (in accordance with the selected annotation scheme) can improve the error analysis process and help in identifying the weak points of the annotator. It also allows to better gauge how rich coreference annotations are with linguistic information (information that can be extracted from the text based on its linguistic properties such as part of speech of the words or relations between different words or phrases present in the text). Richer annotation gives more control to the higher-level applications on how to use their provided data.

The rest of the paper is structured as follows. In Section 2, main definitions of coreference resolution concepts used in this paper are explained. Related works in the coreference evaluation field are covered in Section 3. The proposed evaluation strategy and relevant concepts to it are presented in Section 4. Section 5 concludes and gives insights for future works.

2. Definitions

Coreference occurs when two or more expressions refer to the same discourse-world entity (Mitkov 2014). For example, Tom skipped school today. He was sick.

Coreference resolution is the process of linking together expressions that refer to the same discourse-world entity (Pradhan *et al.* 2012). It is a complex task that is relevant to multiple different domains, including NLP tasks and linguistic research. Due to these reasons, often it is not

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

entirely clear what exactly is meant by coreference and other related terms (Krahmer and Piwek 2000). In this section, commonly used definitions are explained, while definitions specific to our evaluation strategy are covered in Section 4.

- Anaphora is the use of an expression, the interpretation of which depends on another word or phrase presented earlier in the text (Elango 2005; Mitkov 2014). In the aforementioned example, the word "Tom" would be considered an antecedent in such a case. Usually, anaphoric objects are expressed with pronouns and cannot be independently interpreted without going back to their antecedents.
- Cataphora is identical to anaphora with the only difference being that reference is made to a phrase that will be present later in the text (Elango 2005; Mitkov 2014).
- In certain cases, anaphoric (or cataphoric) relations might not be coreferential: "*Every man* has *his* own destiny" (Mitkov 2014). But since the proposed evaluation strategy also covers, such expressions distinction is not made.
- Associative (or bridging) and evolutive expressions (when connection is inferred and not stated directly (Mitkov 2014) – although *the store* had only just opened, *the food hall* was busy and there were long queues at *the tills*.) are not covered in this paper. But certain constructions (like hypernym/hyponym) that are often used in such expressions are covered if they form a coreferential relationship: "My *cat* was not feeling well. You could see that the *animal* was hurting."
- Appositional expressions where two noun phrases are placed side by side and one of them supplements the other are not covered as well. In our opinion, such expressions should be instead covered by fact extraction solutions and (or) semantic annotators.

Coreferences can further be divided into two groups:

- Endophoric references – referring to something present in the same text (Gardelle 2012).
- Exophoric references – referring to something outside of the text and usually requiring some additional information (like the context in which the text was written or the author's other works) to make a correct interpretation. Deixis, or deictic expression, is an example of such reference. To interpret the phrase we need to know, for example, who is speaking or writing the text (Gardelle 2012). In this paper, evaluation of exophoric coreferences is not addressed.

In the context of this work, referring expression (in this case "he") will be called a **referent**. Expression to which it is pointing (in this case "Tom") is usually called the **antecedent**. In this work, **antecedent** is further specialized into **dominant mention** and **non-dominant mention** depending if the specific **antecedent** best describes the discourse-world entity. Process of selecting such **antecedent** is detailed in Section 4.2.

Coreference resolution creates a data structure where all mentions of the same discourse-world entity are stored. Usually, such a structure is called a chain. During the evaluation of the annotations created by the automated coreference resolution approach, they are compared against manually annotated (ideally by more than one human annotator) corpus, often called **gold corpus**. Manual annotations are sometimes referred to as key chains, key sets, and gold coreference chains, while automated annotations are referred to as response sets and system chains. In this paper, manual annotations will be referred to as a **gold set** and automatic annotations as a **response set**. Each coreference evaluation metric, or strategy, proposes how to compare these two sets against each other to best evaluate the coreference resolution approach.

Instructions explaining what expressions are counted as coreferences and how they should be marked are usually called **annotation scheme**.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

3. Related works

Over the years there have been multiple evaluation strategies suggested, but none of them has been adopted as the standard of this field. In this section, the majority of the evaluation strategies found in related literature are covered. Existing evaluation approaches can be broadly divided into two groups: linguistically agnostic, covered in Section 3.1, and linguistically aware metrics, covered in Section 3.2. In Section 3.3, error analysis of the annotations created by the coreference resolution approaches is covered.

3.1 Linguistically agnostic evaluation metrics

Linguistically agnostic metrics can be further classified into mention, link, optimal mapping-based or link-based and entity-aware metrics (Sukthanker *et al.* 2018). But a common feature of all of these metrics is that they treat coreference evaluation task as a clustering task and do not take in mind linguistic information of coreference chain elements into consideration.

3.1.1 Basic precision, recall, and F-measure metric

Basic metrics of precision (P), recall (R), and F-measure (F) are often used to evaluate various NLP tasks. Precision, (1) formula, shows the percentage of correctly resolved (C) coreference expressions against the actual number of provided coreferences (A) by the annotator:

$$P = \frac{C}{A} \tag{1}$$

Recall, (2) formula, shows the percentage of correctly resolved (C) coreference expressions against the total amount (T) of expressions pre-annotated in the text:

$$R = \frac{C}{T} \tag{2}$$

While the precision metric is rather straightforward, the recall has certain limitations. For example, not all coreferences might be in the scope of the research. As a result, the total amount of pre-annotated expressions might vary even if the evaluation is run against the same text corpus.

There have been suggestions made for result reporting guidelines proposing additional metric called resolution rate (RR), (3) formula, that would replace recall and next to the total amount of pre-annotated expressions would add coreferences that were excluded (E) by the annotator that is being evaluated (Byron 2001):

$$\mathrm{RR} = \frac{C}{T + E} \tag{3}$$

The obvious question here is, how to define what is excluded and what should not be covered by coreference resolution in general? In related literature, there are many different types of expressions (presuppositions, appositions, etc.) that sometimes are treated as coreferences by some researches and not treated as such by others (Ceberio *et al.* 2018; Hou, Markert, and Strube 2013; Rösiger and Teufel 2014; Souza *et al.* 2017; Van Deemter and Kibble 1999).

Lastly, F-measure, (4) formula, is a harmonic mean of precision and recall. There is also a possibility, (5) formula, to assign an additional weight to either precision or recall. Newer metrics usually provide different formulas for precision and recall calculations, while F-measure calculations remain largely unchanged:

$$F = \frac{2PR}{P + R} \tag{4}$$

$$F_\beta = \frac{(\beta^2 + 1)PR}{(\beta^2)P + R} \tag{5}$$

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

3.1.2 Trivial, nontrivial, and critical success rates

Mitkov proposed to calculate three separate scores depending on the complexity of the coreference expression (Mitkov 2014). He defines N as a set of all coreferences present in the evaluation and S as a set of correctly resolved coreferences by the annotator. Following that K is defined as a set of resolved trivial coreferences where each referent has only one candidate antecedent. And M is defined as a set of resolved coreferences where each referent has more than one candidate, but the correct resolution is made by applying gender and number agreement. With these variables *success rate*, (6) formula, *nontrivial success rate*, (7) formula, and *critical success rate*, (8) formula, are calculated:

$$\text{success rate} = \frac{S}{N} \tag{6}$$

$$\text{non-trivial success rate} = \frac{S-N}{N-K} \tag{7}$$

$$\text{critical success rate} = \frac{S-N-M}{N-K-M} \tag{8}$$

Since the *critical success rate* evaluates more complex coreference expressions, it is considered to be the most important metric of the proposed three.

3.1.3 MUC link-based metric

An alternative strategy was proposed for MUC-6 evaluation (Vilain *et al.* 1995). The proposed model is link-based and assumes that each reference links two mentions, and it attempts to solve the transitivity problem. For example:

- In the text, there are three coreferences marked: A → B, B → C, and D → E.
- Annotator that is being evaluated marks: A → C and D → E.

Going with the basic precision metric, it would get 1/2 score and basic recall would get a 1/3 score. But assuming that A, B, and C have a transitive relationship, it is reasonable to claim that A → C marking is correct as well. The proposed model constructs two groups of equivalence classes. Gold set:

- {A B C} and {D E}

Response set:

- {A C} and {D E}

For recall calculation, the size of the gold set class is taken as S, and it is subtracted by the number of partitions required from response set class $p(S)$ to match it. Then each equivalence class is added up for total recall. The author proposes to calculate recall, (9) formula, by using the size of S and $p(S)$ sets:

$$R_T = \frac{\sum |S| - |p(S)|}{\sum |S| - 1} \tag{9}$$

For the precision process is inverse, (10) formula, the size of response set class is taken as S' and it is subtracted by the number of partitions required from gold set class $p(S')$ to match it. Then each equivalence class is added up for total precision:

$$P_T = \frac{\sum |S'| - |p(S')|}{\sum |S'| - 1} \tag{10}$$

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

As the author notes, such an approach works only with equivalence classes and when the transitivity is enforced. It also favors results that over-merge entities. For example, if all mentions of different discourse-world entities in the text were merged into one coreference chain, then it would result in a 100% recall and very high precision score.

3.1.4 B-cubed (B^3) metric

This mention-based metric was proposed as a response to MUC-6's evaluation model. It raises the important question of not all precision errors being equal when looking at the task from the information extraction viewpoint (Bagga and Baldwin 1998). For example, we have three different equivalence classes: {A B}, {C E}, and{D F G}. If the annotator would make a mistake by marking B → C relationship and as such merging the first two classes, then it should be considered as a smaller error than if it marked E → D relationship and merged the last two classes. The weight of the error is based on the size of the newly created class. This is relevant to information extraction tasks, since the size of the error can greatly impact the result. On the other hand, from a linguistic standpoint, both errors could be treated equally.

Precision (P_i) for each entity is calculated, (11) formula, by taking the number of correctly annotated entities in the equivalence class (C) and the total number of entities in the equivalence class (A):

$$P_i = \frac{C}{A} \tag{11}$$

Recall (R_i) for each entity is calculated, (12) formula, by taking the number of correctly annotated entities in the equivalence class (C) and the total number of entities in the pre-annotated equivalence class (T):

$$R_i = \frac{C}{T} \tag{12}$$

Then final precision, (13) formula, and recall, (14) formula, are calculated. All (N) previously calculated precision (P_i) and recall (R_i) values are added up and multiplied by their assigned weight (w_i). By default, the author suggests dividing 1 by the number of total entities present in the text. But if required weights can be altered for each specific entity:

$$\text{Precision} = \sum_{i=1}^{N} (w_i * P_i) \tag{13}$$

$$\text{Recall} = \sum_{i=1}^{N} (w_i * R_i) \tag{14}$$

One of the arguments in favor of B-cubed against the MUC link-based approach was that MUC evaluation did not deal with singleton mentions. Yet, if the annotator marked all mentions as singletons B-cubed evaluation would result in 100% precision.

3.1.5 ACE value

This metric was created for the ACE task (Doddington *et al.* 2004). It scores entities, relations, and events separately. The value is computed by counting the number of following errors: missed mentions, false positives, and misclassifications. Calculation of entity, composed of all (N) its mentions, value can be seen in (15) formula:

$$\text{Value}_{\text{response\_entity}} = \text{Entity\_value(response\_entity)} \cdot \sum_{i=1}^{N} (\text{Mention\_value(response\_mention}_i)) \tag{15}$$

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Entity_value assigns value depending on whether the same entity is present in the gold set. *Mention_value* assigns value to each mention of the entity based on its presence in the gold set. Each error is assigned a weight based on the type of entity (person, location, etc.) and mention type (nominal, pronominal, etc.); therefore, it does include some linguistic information in the evaluation process.

ACE value is considered to be task-specific and not very useful for general purpose coreference evaluation (Zitouni 2014). Furthermore, it has been criticized for being hard to interpret (Luo 2005).

3.1.6 CEAF metric

CEAF is an entity-based metric (Luo 2005) and attempts to evaluate similarities between entities. Entities, in this case, are similar to coreference chains – all mentions of one object in the text form one entity. It provides two ways of scoring the coreference resolution approach, mention-based and entity-based. Both use R (gold set) and S (response set) for calculations.

Entity-based approach (CEAFE) measures, (16) formula, how many same mentions two entities (R and S) share. It can also function as F-measure, (17) formula:

$$\phi = |R \cap S| \tag{16}$$

$$F(R, S) = \frac{2|R \cap S|}{|R| + |S|} \tag{17}$$

The mention-based approach (CEAFM) calculates recall, (18) formula, and precision, (19) formula, separately:

$$\text{Recall} = \frac{\phi(g^*)}{\sum_{i=1}^{N} \phi(R_i, R_i)} \tag{18}$$

$$\text{Precision} = \frac{\phi(g^*)}{\sum_{i=1}^{N} \phi(S_i, S_i)} \tag{19}$$

Here, g^* represents Kuhn–Munkers algorithm, (20) formula, that is used to find the best mapping of the two entities. R^*_m refers to gold set subset where g^* is attained:

$$\phi(g^*) = \sum_{R \epsilon R^*_m} \phi(R, g^*(R)) \tag{20}$$

One of the flaws of this approach is that it does not take into consideration unaligned entities in the response set. The annotator might make a mistake and create two entities instead of one. CEAF would ignore the second entity even if it had multiple right mentions linked.

3.1.7 CoNLL score

CoNLL score is not a separate evaluation metric, but it is often used in the coreference evaluation. During CoNLL-2012 shared task on coreference resolution, it was decided that MUC, B^3, and CEAF metrics have their benefits and drawbacks (Pradhan *et al.* 2012). Therefore, instead of introducing a completely new metric, it was decided to take an average of their F-measures, (21) formula, as the evaluation score:

$$\text{CoNLL} = \frac{F_{\text{MUC}} + F_{B^3} + F_{\text{CEAF}}}{3} \tag{21}$$

This approach was originally proposed earlier (Denis and Baldridge 2009), but it is not clear why the average of three flawed numbers would not result in a fourth flawed number (Moosavi and Strube 2016).

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

3.1.8 BLANC

BLANC is a link-based approach that adapts the Rand index (Luo *et al.* 2014; Pradhan *et al.* 2014; Recasens 2010). It constructs four sets. There are two gold sets, one representing all coreference links (C_k) in the text and another representing all non-coreference links (N_k). Same for response set, C_r and N_r.

Recall, (22) and (23) formulas, and precision, (24) and (25) formulas, are calculated for both coreference and non-coreference links:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|} \tag{22}$$

$$R_n = \frac{|N_k \cap N_r|}{|N_k|} \tag{23}$$

$$P_c = \frac{|C_k \cap C_r|}{|C_k|} \tag{24}$$

$$P_n = \frac{|N_k \cap N_r|}{|N_k|} \tag{25}$$

After this F-measure is calculated for both corefering and non-corefering links. Their average is used as BLANC's final score. The problem with this approach lies in the fact that if the text has a high number of corefering links, then naturally it also will have a very high number of non-corefering links due to different coreference chains not referring to each other. This might result in higher precision and recall values than if the same annotator marked the text with fewer coreference expressions in it.

3.1.9 LEA metric

LEA attempts to combine link and entity-based approaches for coreference resolution (Moosavi and Strube, 2016). It is one of the newest evaluation methods and attempts to tackle various issues with previously covered metrics. It has a weighting mechanism called importance, but it functions similarly to B^3 weights. It is based on the size of the entity ($|e|$) but can be adjusted according to domain needs. Additionally, the number of links, (26) formula, for each entity (e) with the number of mentions (n) is calculated:

$$\text{links}(e) = n * \frac{n-1}{2} \tag{26}$$

As in other approaches, gold set (k) and response set (r) sets are used for recall, (27) formula, and precision, (28) formula, calculations. For precision, the role of gold and response sets are reversed:

$$R = \frac{\sum_{k_i \epsilon K} \left(|k_i| * \sum_{r_j \epsilon R} \frac{\text{links}|k_i \cap r_j|}{\text{links}|k_i|} \right)}{\sum_{k_z \epsilon K} (|k_z|)} \tag{27}$$

$$P = \frac{\sum_{r_i \epsilon R} \left(|k_i| * \sum_{k_j \epsilon K} \frac{\text{links}|r_i \cap k_j|}{\text{links}|r_i|} \right)}{\sum_{r_z \epsilon R} (|r_z|)} \tag{28}$$

A common issue with linguistically agnostic metrics is that they treat coreference evaluation as a generic clustering problem. But in practice, a hierarchy of importance can be established, where a definitive noun phrase stating the full name of the person has a higher value than a generic pronoun that might refer to that person (Chen and Ng 2013). This results in the possible loss of semantic information being not addressed in the linguistically agnostic evaluation metrics (Holen 2013).

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

3.2 Linguistically aware evaluation metrics

One of the noted issues with linguistically agnostic metrics is that they treat coreference evaluation as a generic clustering problem. This results in the possible loss of semantic information being not addressed in the linguistically agnostic evaluation metrics. While not as important in linguistic researches, this is a vital problem for information extraction tasks.

3.2.1 LMetrics

LMetrics (Chen and Ng 2013) proposes to solve this problem by extending some of the already existing metrics. Authors identified the common weight functions in MUC, B^3, and CEAF metrics:

- Weight of common subset in response and gold sets (w_c).
- Weight of gold set (w_k).
- Weight of response set (w_s).

Then these weight functions are redefined into linguistically aware versions: w_c^L, w_k^L, and w_s^L. For example, w_c is redefined, in (29) formula:

$$w_c^L(C_j^i) = \begin{cases} \sum_{e_l \epsilon E} w_l(e_l) & \text{if}(C_j^i) > 1 \\ w_{\text{sing}} & \text{if}|(C_j^i)|, |K_i|, |S_j| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

Here, e_l is coreference link, and C_j^i is a common subset in response (S_j) and gold (K_i) sets. W_l is additional weight based on the coreference link: name, nominal, or pronominal. If the coreference set is singleton, then a different weight is used (w_{sing}).

3.2.2 ARCS

Instead of redefining existing metrics, Tuggener proposes new ARCS evaluation framework (Tuggener 2014, 2016). To determine if the suggested link is correct, separate strategies are suggested for different types of higher-level applications that would use coreference annotations:

- An application that investigates distributions and patterns of entity occurrences in discourse. In such a case, immediate antecedent should be selected for referent.
- Summarization and machine translation applications. In such case, the closest nominal antecedent should be selected for referent.
- Query-driven applications. In such a case, anchor mention should be selected for referent. Anchor mention is the first nominal mention in the coreference chain. It is assumed that the first nominal mention in the text best describes the underlying discourse-world entity.

Four scores are aggregated over gold and response sets:

- *TP*, true positive, where the referent is in the gold and response sets and the suggested link is correct.
- *WL*, wrong linkage, where the referent is in the gold and response sets, but the suggested link is incorrect.
- *FP*, false positive, where the referent is in the response set, but not in the gold one.
- *FN*, false negative, where the referent is in the gold set, but not in the suggested one.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

F-measure is the standard harmonic mean of precision, (30) formula, and recall, (31) formula:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{WL}} \tag{30}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{WL}} \tag{31}$$

3.2.3 Prague Anaphora Score

Similar to Tuggener's approach is Prague Anaphora Score (Novák 2018). Next to four ARCS scores, it adds spurious zero positive (SZP) variable that deals with ellipses that should not be resolved. It is language-dependent and is used in precision, (32) formula, calculation. Recall and F-measure calculations are the same as in ARCS:

$$P = \frac{\text{TP} + \text{SZP}}{\text{TP} + \text{FN} + \text{WL} + \text{SZP}} \tag{32}$$

However, Prague Anaphora Score does not use any of the three strategies outlined in ARCS to determine the correctness of the response set. Instead, a link to any antecedent in the chain, which does not refer to another antecedent, is considered valid. Ideally, the coreference chain should contain only one such element.

3.2.4 PARENT

PARENT metric (Kaczmarek and Marcińczuk 2015) also attempts to better evaluate correct information returned by the coreference resolution approaches. It divides all mentions present in the text into two disjoint subsets: *defining* and *non-defining*. Defining mentions are those that carry enough semantic information that allows them to identify as discourse-world entities. A *non-defining* subset can be further divided into *referring* and *ignored*, not relevant for the evaluation process, subsets. This provides certain flexibility for the evaluation process. For example, if we want to evaluate pronoun linkage to definitive nouns, then all other types of mentions would be contained in an *ignored* subset. It focuses on finding relations between referring and defining mentions, since they are more valuable than relations between two different referring mentions.

All mentions of one entity constitute one gold set cluster (C_i^{key}) and a response set cluster (C_i^{sys}). Relations for gold (G) set are defined in (33) formula:

$$G = \{(m_{rl}^i, C_i^{\text{key}}) | \forall C_i^{\text{key}} \epsilon C^{\text{key}} \forall m_{rl}^i, \epsilon C_i^{\text{key}}\} \tag{33}$$

Here, m_{rl}^i is referring to mention that belongs to the gold set cluster. Relations for response (S) set are defined in (34) formula:

$$S = \{(m_{rl}^i, [[m_{dk}^i]]\text{key}) | \forall C_i^{\text{sys}} \epsilon C^{\text{sys}} \forall m_{rl}^i, \epsilon C_i^{\text{sys}} \forall m_{dk}^i, \epsilon C_i^{\text{sys}}\} \tag{34}$$

Here, m_{rl}^i is referring mention that belongs to response set cluster and m_{dk}^i is defining mention that it links to. Precision, (35) formula, and recall, (36) formula, are calculated using G and S. F-measure is calculated in a standard way:

$$P = \frac{|G \cap S|}{|S|} \tag{35}$$

$$R = \frac{|G \cap S|}{|G|} \tag{36}$$

While the problem of all mentions being treated equally has been addressed by these linguistically aware metrics, they still lack in extendibility. Extendibility in this context determines how

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

easy it would be to add additional linguistic information into the evaluation process. The need to introduce new coreference data might come from the requirements of the higher-level applications that are being developed. New data would lead to potential new types of errors that would not be covered by these metrics. Due to that linguistically aware evaluation metrics should be easily expandable and be able to address potential new types of errors.

3.3 Coreference resolution error analysis

During the development of coreference resolution approach phase, many different types of coreferences are identified and analyzed (Delmonte 2002; Fischer 2015; Hou *et al.* 2013; King and Lewis 2018; Saeboe 1996; Van Deemter and Kibble 1999). Different types often require different techniques to resolve. But this information is relevant not only for the development of new approaches but also for interpretations of the results. Currently, coreference type identification is not addressed by any coreference evaluation metric.

This problem is partially addressed by coreference resolution error categorization tools (Kummerfeld and Klein 2013). They take the response sets from the annotators that participated in ConLL 2011 task and attempt to categorize underlying error types. But since coreference resolution evaluation metrics do not require providing information on what kind of coreference was resolved, such tools are limited to categorizing along the part of speech and span errors. Often this does not provide the required detail to properly evaluate the performance, since very different types of coreferences can be constructed with the same type of speech. For example, a name repetition, certain feature, synonym, metonym, or hypernym/hyponym can be used to refer to the same discourse-world entity. In all cases, noun phrase would be used. Therefore, stating that a specific coreference resolution approach has incorrectly resolved a certain number of noun phrases would not be informative enough to determine the exact weak points of the coreference resolution approach.

Another option would be adding such information to coreference resolution response sets (and gold sets) after the fact with automated tools, but it would introduce additional errors since the determination of what kind of coreference relationship two phrases have is not a trivial task. Additionally, after adding such information to response sets, it becomes questionable if it is the coreference resolution system being evaluated or these were the tools that added this information.

In many coreference resolution systems, especially rule-based ones, such information is already available internally since usually different strategies are used for solving different coreference constructions. Therefore, the main issue here is that coreference resolution evaluation metrics do not require providing this information in response sets.

4. Proposed coreference resolution evaluation strategy

In this section, the coreference resolution evaluation strategy is presented. We start by defining the overall life cycle of the coreference resolution approach development in Figure 1. It shows that the development of the coreference resolution approach depends on the annotation scheme, gold corpus selected or created. The evaluation helps developers to make a decision, whether the developed solution is sufficient or it should be improved. An evaluation itself is based on evaluation strategy and gold corpus, which are dependent on the annotation scheme. Thus, when proposing a new evaluation strategy, it is important to understand the whole process of coreference resolution approach development, its course, the artifacts used and the dependencies between them.

The rest of the section is divided into four subsections. In Section 4.1, we introduce the coreference annotation scheme that was used for evaluation and which forms a context of our research. In Section 4.2, a special attention is paid to the concept of dominant mention and its implementation in our coreference annotations. In Section 4.3, compatibility with other annotation schemes

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

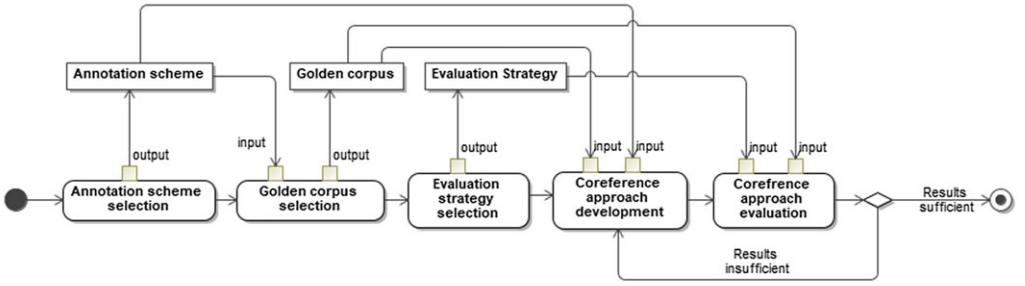

Figure 1. Coreference resolution approach development life cycle

is addressed. In Section 4.4, we present our coreference resolution evaluation model and strategy. Finally, in Section 4.5, application of the proposed evaluation approach is demonstrated.

4.1 Coreference annotation scheme

To provide a context for our proposed evaluation strategy at first, we introduce the coreference annotation scheme that was used during the development. Usually, the coreference annotation scheme does not influence coreference resolution evaluation when applying existing evaluation approaches. Our proposed evaluation strategy is a bit different. It requires dominant mentions (or other, similar concept, covered in Section 4.2) and that annotation scheme would classify different coreference expressions (covered in this section). Different annotation schemes might use different classification, important point is that classification is done.

Our coreference annotation scheme and the first version of coreference resolution algorithm were implemented as part of NLP tools for the Lithuanian language in 2014. It allowed us to create a Semantic Search Framework for Lithuanian Language (SSFLL) Internet corpus extracted from public news portals (Vileiniškis, Šukys, and Butkienė 2015). Initial resolution approach was limited to pronoun resolution only but eventually was expanded to cover nouns as well. For this reason, coreference annotation scheme for the Lithuanian language was updated as well and the last version can be seen in Table 1 (Žitkus and Butkienė 2018).

A distinguishable feature of our annotation schema is that it requires to specify to which class each identified coreference belongs. This information can be valuable to a higher-level application. Furthermore, enforcing such classification allows us to evaluate implemented coreference resolution more precisely, especially in cases when for different classes of coreferences a separate resolution algorithm (or model) is implemented.

Proposed coreference annotation scheme covers only endophoric expressions and is of four levels. In the first level, coreference dependency to one of the five broad classes is specified: pronominal (p), nominal (further divided into generic (g) and definitive nouns (d)), ellipsis (e), and adverbs (a). Further in the second level, more specific classes, specifying their parent classes, are indicated. At the moment, in the second level, only nominal, pronominal, and ellipsis coreferences are specified. Third and fourth levels are global and are used by all types of coreferences. The third level determines if the referring object is pointing backward (a), forward (p), or is the direction irrelevant (i). The fourth level defines if the referring object is referring to one antecedent (s) or group of antecedents (g). Third option is for those cases where, due to authors intention or mistake, it is ambiguous (a) to which antecedent referring object is referring to.

Technically, all four level features of the specific coreference are encoded using certain letters specified in brackets (see Table 1). These letters are combined to define a specific code for each type of coreference. For example, if in cases where the coreference is pronominal (p), personal

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Table 1. Structure of coreference annotation scheme

First level	Second level	Third level	Fourth level
Pronominal (p)	Personal (p)		
	Reflexive (r)		
	Possessive (o)		
	Relative (e)		
Nominal (g/d)	Repetition (t)		
	Partial repetition (a)		
	Abbreviations (b)	Position (a/p/i)	Group (g/a/s)
	Feature (f)		
	Hyponymy/hypernymy (h)		
	Metonym (m)		
	Synonym (s)		
Adverbial (a)	–		
Ellipsis (e)	Same object (i)		
	Same type of object (y)		
	Verb phrase (v)		

(p), pointing backward (a) and refers to multiple antecedents (g), then the final combination can be reduced to "ppag" as the code of that specific coreference:

- *Tom* and *Jim* are very good friends. *They* know each other since second grade.
- Word "they" refers to "Tom" and "Jim."

In the case of adverbs, since they do not have the second level specified, they would form codes like this: "*a-is*." Adverbs do not have three-letter code because we want to preserve a common format in case of future research that would make second-level classification of adverbs meaningful.

Same entity can be mentioned multiple times in the text with different phrases and if we linked them in different order, then different annotations might have different coreference types specified for the same text. Due to that we recommend linking each referent to the closest, more dominant antecedent. Determination of the more dominant antecedent is covered in Section 4.2.

Since Lithuanian language has free word order, one of the goals of this annotation scheme was to have a scheme that would produce comparable results for different coreference expressions regardless of the used word order. This fits the overall trends in annotation schemes. For example, new time expressions and named entity annotation scheme TOMN/UGTO (Zhong *et al.* 2020) attempts to solve similar problem—same date written in different formats.

It is important to mention that in our evaluation model when we refer to coreference class we mean first-level values. When we refer to coreference type, we mean full four-letter code.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

As can be seen, our annotation scheme does not cover cases where the discourse-world entity is mentioned only once, such cases are usually called singletons. We assume that the scope of the coreference resolution task requires at least two mentions (one antecedent and one referent) of the same entity in the text that is being analyzed. Singletons can be successfully identified by other NLP components such as mention identification, which is very important to coreference resolution approaches, but at the same time should not necessarily be considered as part of coreference resolution itself.

It could be argued that such information could be added at a later date by semantic annotator (or another higher-level application), but it would be just duplicating work already done by coreference resolution approach and likely have worse results since it is not designed for identifying and resolving coreferences.

4.2 Dominant mentions

The same discourse-world entity can be referenced in the text by different mentions: proper noun, generic noun, pronoun, or a gap. In linguistically agnostic evaluation strategies, all of these different mentions are treated as equal. Looking from the information extraction viewpoint and the needs of higher-level applications, it is obvious that it should be possible to determine which of those mentions best describes a discourse-world entity that it refers to. Naturally, such mentions tend to be semantically richest mentions. Due to that, other linguistically aware evaluation metrics introduced concepts like anchor mentions and defining mentions that were overviewed in Section 3.2. They help in evaluating whether the coreference resolution approach correctly identified the most relevant mention(s) of the entity. For the same purpose, we use the concept of dominant mention.

A dominant mention is an expression that carries the richest semantics or describes most precisely the discourse-world entity (Ogrodniczuk *et al.* 2013). The authors of this paper propose ordering expressions by their dominance in the following order: full proper noun, abbreviated proper noun, partial proper noun, NP, and ellipsis referring back to the same object. Certain expressions like pronouns, adverbs, and other types of ellipsis cannot be dominant, since they do not carry much semantic information on their own. If two or more expressions are of the same dominance level, then preference should be given to expression that appeared earlier in the text.

Let's take this abridged text fragment for an example:

- Early in the morning *president* declared that. . . *He* hopes. . .. This was unusual for *B. Obama*. Earlier in the year, *Obama* was criticized. . . But *his* position. . . It looks like *Barack Obama* is not doubting himself. But critics said that *Barack Obama* is. . .

From this text fragment, we can create the following collection of coreferences referring to the same discourse-world entity: President ← He, He ← B. Obama, B. Obama ← Obama, Obama ← His, His ← Barack Obama, Barack Obama ← Himself, Himself ← Barack Obama.

To determine the dominant mention, all mentions of the same entity should be listed first. This can be done either at the same time as individual coreference relationships are resolved or after it. It depends entirely on the specific implementation. Listed mentions have to be ordered by their appearance in the text starting from the earliest to the latest. Going with the previous example, we would have this list of mentions created:

- { [President]$_1$ [He]$_2$ [B. Obama]$_3$ [Obama]$_4$ [His]$_5$ [Barack Obama]$_6$ [Himself]$_7$ [Barack Obama]$_8$}

Next pronouns are filtered out since they do not carry much semantic information on their own:

- { [President]$_1$ [B. Obama]$_3$ [Obama]$_4$ [Barack Obama]$_6$ [Barack Obama]$_8$}

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Then elements are ordered by their dominance:

- { [Barack Obama]$_6$ [Barack Obama]$_8$ [B. Obama]$_3$ [Obama]$_4$ [President]$_1$ }

After these steps, the first element in the list is selected as the dominant mention. The sixth overall mention gets preference over the eighth one due to it being present earlier in the text. Same mention is selected as dominant mention for the entire chain. All remaining mentions are treated as referents of the selected dominant mention. But their types of coreference that were initially assigned to them are preserved, this is done so that type validity can be checked when performing the evaluation. Asides from coreference type preservation, dominant mentions could be integrated into any other annotation scheme without any issues.

If anchor mentions were used instead of proposed dominant mentions, then the first mention present in the text, in this case, "President," would be selected. Certainly, it is not the mention present in the text that best describes the discourse-world entity. Defining mentions would take all named entities in this case and treat them as equally important, yet it is natural to assume that "Barack Obama" better describes the discourse-world entity than "B. Obama" and due to that is more important. Asides from evaluation strategies, Stanford CoreNLP solution has incorporated representative mentions into their coreference annotation scheme (Lee *et al.* 2011).

Representative mentions are similar to dominant mentions but do not take in mind the order in which mentions were presented in the text. Additionally, instead of making a distinction between full, abbreviated and partial named entities preference is given to the longest mention. While most of the time it provides the same results, it can also cause inaccuracies. For example, the last name of the person usually better describes him than his first name, yet the first name can be longer and, in such case, a less descriptive proper noun would be selected. Due to these reasons, we believe that dominant mentions provide an advantage over the anchor, representative and defining mentions.

Outside of evaluation purposes, dominant mentions can also be useful for the resolution of exophoric expressions. If we can determine semantically richest mentions of the same entity from two or more different data sources, then it is easier to determine if they refer to the same discourse-world entity and are coreferent or not.

4.3 Proposed evaluation strategy

In this section, a new, linguistically aware evaluation strategy is presented. Compared to other linguistically aware evaluation strategies, it adds additional linguistic information with coreference type identification and dominant mention usage. Furthermore, classification of errors (with added coefficients) allows adjusting the evaluation process if certain errors are deemed to be more, or less, severe.

Concepts and their relationships which we use to explain the evaluation strategy are depicted in Figure 2 using UML class diagram notation. Concepts relevant to the coreference annotations themselves (the *Annotation* package) are defined in Appendix A, Table A1. The *Evaluation* package covers concepts relevant to the process of the evaluation, and it is defined in Appendix A, Table A2.

As shown in Figure 2, annotations are classified into six classes: *Correct annotation*, *Correct annotation with the wrong type*, *Correct annotation with the wrong dominant mention*, *Correct annotation with wrong dominant mention and type*, *Missed annotation*, and *False positive annotation*. While other linguistically aware evaluation metrics have a similar classification as dominant mentions (Tuggener 2016; Kaczmarek and Marcińczuk 2015), to our knowledge they are not being combined with coreference type classification like in our evaluation strategy.

Asides from dominant mentions, we do not make any distinctions between types of speech used in coreference expression. In the information extraction context, it is important to determine not only what phrases are used to construct coreference relationships but also what kind of semantic information can be extracted near them. Therefore, a generic noun, or even a pronoun, can be

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

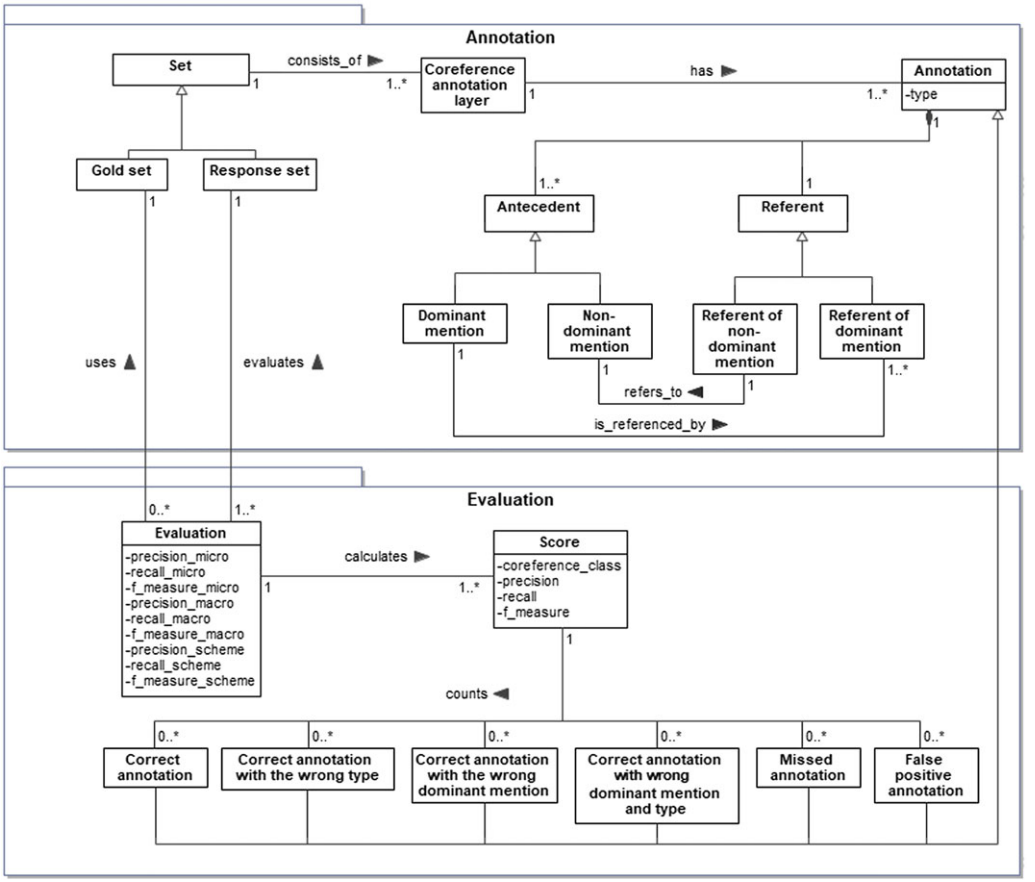

Figure 2. Conceptual data model of coreference evaluation

linked to more valuable semantic information than a definitive noun. Different types of systems might also value the same semantic information differently. For example:

- *Tom* was feeling sick. The *boy* wandered the house for a while until *he* decided to call the ambulance.

In this example, "The boy" mention itself is semantically richer than "he" mention. The pronoun "he" implies gender and number of the discourse-world entity, while the noun phrase "The boy" has those too and additionally implies the age of the discourse-world entity. Yet semantic information near "he" mention is more valuable: fact that he called the ambulance would be more important in most cases than the fact that he wandered the house for a while. It depends entirely on the sentence and the context, therefore, assigning a higher value to, for example, noun than to pronoun is not useful and can be counterproductive as in the given example.

As explained in Section 3.3, including coreference type into annotation would increase the overall quality of the error analysis. But such information can be valuable for higher-level application as well.

For our example, we used English text and Stanford CoreNLP coreference annotator (text and annotation available in Appendix B). One of the created coreference chains has the following structure: We ← We ← us ← We ← We ← We ← We ← We.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

It is unlikely that someone is going to search for information with "we" or "us" keywords. And even in such a case, it would be difficult to find useful information since pronouns are very common in various documents. For this reason, such coreference annotation is not very useful from the higher-level application (like semantic search) perspective. If coreference type would be included in the annotation, then the application could be optimized by filtering out coreferences that are not relevant to its tasks.

In our opinion, having coreference type in coreference annotations can provide many fine-tuning opportunities to higher-level applications. For example, coreference annotator has high precision with synonyms, but noticeably lower precision with metonyms. With coreference type provided semantic search application can rank query results by giving priority to synonyms over metonyms since they are more likely to be correct. Therefore, it is worthwhile to include such information in coreference annotations and evaluate them during the evaluation process.

However, naturally, some coreference types like repetition, partial repetition, or abbreviation might be deemed to be not informative enough and coreference resolution approach should not be penalized for incorrectly identifying them. For example:

- *Barack Obama* was elected the 44th president of the United States on Tuesday. *B. Obama* is the first black US president. The newly elected *president* will be in charge of the armed forces. *He* is scheduled to appear at the press conference tomorrow.

Here we have "B. Obama," being an abbreviation of "Barack Obama" and "president" a feature (referring to his specific occupation). Coreference resolution approach might label "B. Obama" as a partial repetition of "Barack Obama," which would be technically incorrect, but it still allows us to identify that this is an alternate name for the same discourse-world entity. On the other hand, if "president" was mislabeled as a partial repetition then it could cause problems since it is not an alternate name, but a specific feature of the discourse-world entity by which it was referred to. Which type identification errors can be ignored should be determined either by a specific annotation scheme being used or higher-level application needs.

Going with previous example, coreference annotator might make a mistake and create two separate coreference clusters instead of one:

- Cluster A: Barack Obama, B. Obama
- Cluster B: president, He

Items in the second cluster will be assigned *Correct annotation with the wrong dominant mention* class since they are not linked to "Barack Obama." Additionally, there will be one missing annotation connecting "president" to "B. Obama." Additional penalties for creating misleading clusters are not applied. Calculation of precision and recall for this example are provided in Appendix C.

One of the issues with the evaluation process that is usually not addressed in other evaluation metrics is the overrepresentation of certain expressions in the corpus that evaluation was performed on. Usage of coreference expressions can vary depending on the type and style of the text. For example, technical manuals tend not to have many such expressions and avoid complex constructions in general, while literary works often employ them for stylistic and other purposes. Let's assume that we have two coreference resolution approaches. The first one focuses on pronouns and solves them very well yet struggles with other expressions. The second one does well with all types of expressions. If the selected corpus is dominated by pronominal coreference expressions, then the first coreference resolution approach can score higher despite the second coreference resolution approach being more useful in general.

One solution to this problem is to run an evaluation with multiple different corpora and compare their results. This approach is not straightforward and unfortunately not applicable to less researched languages. For example, recent study on text classification lists 24 datasets for English

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

language (Minaee *et al.* 2021). Less researched languages often have only few, or even one, such datasets. Another possibility is the usage of micro and macro F-measure averages.

Micro average pools the performance over the smallest possible unit, in the context of coreference resolution it would be all coreference annotations. High micro F score indicates that the coreference resolution approach has a good overall performance. On the other hand, macro average pools the performance from large groups, in the context of coreference resolution that would be different coreference classes (Manning *et al.* 2010). High macro F score indicates that the CR approach has good performance for each coreference class. Unfortunately, in related literature, it is rarely specified if micro or macro evaluation should be performed and it is usually assumed that micro is used by default. Some papers present micro and macro evaluation numbers for the developed coreference resolution approaches, but usually, it is also not specified and can cause confusion when interpreting provided results.

One of the main advantages of macro average is that it adjusts for imbalanced coreference class distribution. It could be argued that this is the case with natural language texts. For example, we are bound to find noticeably more nouns and pronouns in texts than an ellipsis. It is not realistic to expect that corpus would be constructed in such a way that each possible coreference class would be equally represented. Moreover, if such a corpus would be constructed, then it would have to be altered each time the annotation scheme was changed to adjust for possible imbalances created by that change. On the other hand, it could be argued that such imbalance actually represents discourse-world data and as such micro average is preferable. Due to that, we propose to use both micro and macro averages when evaluating coreference resolution approaches.

The separate precision, recall, and F-measure calculations for each coreference class are useful in case we want to find a specialized coreference resolution approach that is suitable for a specific task. While papers presenting new coreference resolution approaches often tend to detail how well certain types of coreferences are solved, coreference evaluation metrics themselves usually do not provide recommendations on how it should be done, or should it be done at all. We propose to always follow the classification of the annotation scheme that is being used and provide a separate evaluation for each coreference class when reporting results.

For the calculation of precision and recall, additional weighting coefficients are assigned to each coreference depending on which of the six annotation classes (see Figure 2 and Appendix A, Table A2) it was assigned to:

- A number of annotations assigned to *Correct annotation (True Positive, TP)* concept get k_1 coefficient.
- A number of annotations assigned to *Correct annotation with the wrong type (Wrong Type, WT)* concept get k_2 coefficient.
- A number of annotations assigned to *Correct annotation with the wrong dominant mention (Wrong Linkage, WL)* concept get k_3 coefficient.
- A number of annotations assigned to *Correct annotation with wrong dominant mention and type (Wrong Type and Linkage, WTL)* concept get k_4 coefficient.
- A number of annotations assigned to *Missed annotation (False Negative, FN)* and *False positive annotation (FP)* concepts do not get any coefficients.

These coefficients allow differentiating among different types of errors by assigning different values to them. The values of the coefficients range from 0 to 1. In our proposal, the values of the coefficients are proportionally lowered depending on how severe the error of a certain type is. These values are not universal as they could be a starting point for further refinements to the evaluation model. Current coefficients' values are as follows:

- $k_1 - 1$;

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

- k_2 – 0.75;
- k_3 – 0.5;
- k_4 – 0.25.

The evaluation process is not tied to six annotation classes that are shown in the conceptual data model (Figure 2). Depending on the annotation scheme, it can have different numbers of classes. For example, if we wanted to evaluate how well coreference annotator resolves exophoric coreferences, then we would add new classes related to their evaluation. Or if we decided that type identification is not relevant to our evaluation, then we can remove second and fourth classes. Naturally, after such additions or subtractions number of coefficients, and their values, would also have to be adjusted accordingly.

The annotation scheme that was used in this work has five coreference classes, but the evaluation model is not tied to that number. There can be from 1 to n different coreference classes defined. Calculations of precision (P_i), (37) formula, recall (R_i), (38) formula, and F-measure (F_i), (39) formula, for each coreferences class, are identical:

$$P_i = \frac{k_1\mathrm{TP} + k_2\mathrm{WT} + k_3\mathrm{WL} + k_4\mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FP}} = \frac{\mathrm{TP} + 0.75 * \mathrm{WT} + 0.5 * \mathrm{WL} + 0.25 * \mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FP}} \tag{37}$$

$$R_i = \frac{k_1\mathrm{TP} + k_2\mathrm{WT} + k_3\mathrm{WL} + k_4\mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FN}} = \frac{\mathrm{TP} + 0.75 * \mathrm{WT} + 0.5 * \mathrm{WL} + 0.25 * \mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FN}} \tag{38}$$

$$F_i = \frac{2P_iR_i}{P_i + R_i} \tag{39}$$

To diminish the impact of overrepresented classes of coreferences for final evaluation scoring, macro precision (P_{macro}), (40) formula, recall (R_{macro}), (41) formula, and F-measure (F_{macro}), (42) formula, are used. Here, n_a is a number of coreference classes that coreference resolution approach attempted to resolve:

$$P_{\mathrm{macro}} = \frac{\sum_i^{n_a} P_i}{n_a} \tag{40}$$

$$R_{\mathrm{macro}} = \frac{\sum_i^{n_a} R_i}{n_a} \tag{41}$$

$$F_{\mathrm{macro}} = \frac{2P_{\mathrm{macro}}R_{\mathrm{macro}}}{P_{\mathrm{macro}} + R_{\mathrm{macro}}} \tag{42}$$

Next, we also calculate micro precision (P_{micro}), (43) formula, recall (R_{micro}), (44) formula, and F-measure (F_{micro}), (45) formula:

$$P_{\mathrm{micro}} = \frac{\sum_i^{n_a} k_1\mathrm{TP}_i + k_2\mathrm{WT}_i + k_3\mathrm{WL}_i + k_4\mathrm{WTL}_i}{\mathrm{TP}_i + \mathrm{WT}_i + \mathrm{WL}_i + \mathrm{WTL}_i + \mathrm{FP}_i} \tag{43}$$

$$R_{\mathrm{micro}} = \frac{\sum_i^{n_a} k_1\mathrm{TP}_i + k_2\mathrm{WT}_i + k_3\mathrm{WL}_i + k_4\mathrm{WTL}_i}{\mathrm{TP}_i + \mathrm{WT}_i + \mathrm{WL}_i + \mathrm{WTL}_i + \mathrm{FN}_i} \tag{44}$$

$$F_{\mathrm{micro}} = \frac{2P_{\mathrm{micro}}R_{\mathrm{miro}}}{P_{\mathrm{micro}} + R_{\mathrm{micro}}} \tag{45}$$

The purpose of these scores is to evaluate how well the coreference resolution approach resolves coreferences that it attempts to resolve. Naturally, the annotation scheme might have more coreference classes than the specific coreference resolution approach attempted to resolve.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

To determine how well the proposed coreference resolution approach covers the used annotation scheme, separate calculations should be made. For that purpose, we introduced precision (P_{scheme}), (46) formula, recall (R_{scheme}), (47) formula, and F-measure (F_{scheme}), (48) formula, values for annotation scheme coverage:

$$P_{\text{scheme}} = \frac{\sum_i^{n_a} P_i}{n} \tag{46}$$

$$R_{\text{scheme}} = \frac{\sum_i^{n_a} R_i}{n} \tag{47}$$

$$F_{\text{scheme}} = \frac{2P_{\text{scheme}}R_{\text{scheme}}}{P_{\text{scheme}} + R_{\text{scheme}}} \tag{48}$$

These look similar to macro formulas, but the difference is that division is performed not only by n_a but also by n —the number of coreference classes present in the annotation scheme. Scheme coverage score heavily penalizes coreference resolution approaches that do not attempt to solve certain coreference classes.

Overall, our presented evaluation strategy provides the following advantages:

1. The use of both macro and micro averages allows diminishing the impact of imbalanced classes to the final score and at the same time provides a score that is more representative of the discourse-world data.
2. Performing separate calculation for scheme coverage allows distinguishing between how well coreferences resolution approach is doing what it attempts to do and how well does it cover the annotation scheme.
3. Addition of coreference type identification in the evaluation process allows to better identify the weak points of the evaluated coreference resolution approach.
4. Combination of coreference type and dominant mentions to the evaluation process allows to better evaluate to what extent additional semantic information is added by the coreference resolution approach.

4.4 Compatibility with other annotation schemes

The evaluation process will be demonstrated with our annotation scheme, but it can easily work with other annotation schemes as well. The only requirements are that the annotation scheme would classify different coreference expressions and that mention that best describes the entity would be selected.

Our annotation scheme is divided into four levels and together they form a four-letter code that is used in the evaluation process. But for the process to work code can be of any length. This means that the annotation scheme can have two or three levels, and the evaluation process will function without problems. For example, Basque language EPEC-KORREF coreference corpus (Ceberio *et al.* 2018) uses annotation scheme with three levels: *type*, *subtype*, and *semantic relation*. Not every type of coreference has a subtype or semantic relation value. In such cases, hyphen would be used for the second or third letter to indicate the lack of value. As a result, three-letter codes could be formed and used for the proposed evaluation strategy. For example, if coreference in this corpus was marked as nominal, no repetition, hyponym then, based on the first letters of the types, it could be transformed into "nnh" coreference type code. This code would then be used in exact same way as our presented codes in earlier sections.

Evaluation process would work as well with one level, for example, making distinction only between nominal and pronominal coreferences. However, authors of the paper believe that such

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Table 2. Evaluation with different metrics

Metric	P_{micro}	R_{micro}	F_{micro}	P_{macro}	R_{macro}	F_{macro}	P_{scheme}	R_{scheme}	F_{scheme}
Proposed metric	91.5	67.1	77.4	85.8	59.1	70	51.5	34.44	41.99
ARCS	89.6	65.7	75.8	82.4	57.4	67.7	–	–	–
MUC	90.6	74.9	82	84.2	69.9	76.4	–	–	–
B^3	93.1	75.4	83.6	–	–	–	–	–	–
CEAFE	66.3	58.4	62.1	–	–	–	–	–	–

Table 3. Evaluation of different coreference classes

Coreference class	TP	WT	WL	WTL	FN	FP	S*	P	R	F
Pronominal	289	30	27	25	182	29	533	82.8%	59.9%	69.5%
Generic nominal	123	4	21	9	380	23	537	77.1%	25.8%	38.7%
Definitive nominal	973	0	14	0	84	17	1071	97.6%	91.5%	94.5%

classification would not be informative enough. More classification levels annotation scheme has more detailed evaluation process, and more valuable annotations can be to higher-level application.

4.5 Evaluation process

The aim of the experiment was to demonstrate the proposed evaluation strategy by evaluating existing coreference resolution component against Lithuanian Coreference Corpus (LCC) (Žitkus 2018). These tools were chosen because they implement the annotation scheme that was presented in Section 4.1 and due to that make the evaluation process straightforward. For the experiment, the SSFLL NLP pipeline (Vileiniškis *et al.* 2015) was used, which includes these tools. The evaluation was made by analyzing 100 Lithuanian Internet news sites articles of politics and economy domains present in LCC. At the time, additional experiments were not carried out using other datasets due to them not providing detailed coreference type classification and (or) equivalent for dominant mentions. One of the goals for future work related to this research could be adapting existing English language datasets for this evaluation strategy.

The used coreference resolution approach (Žitkus *et al.* 2019) attempts to solve certain pronominal and nominal coreferences. Additionally, evaluation of the same coreference resolution approach was done using ARCS, MUC, B^3, and CEAFE. Full results of this evaluation are displayed in Table 2. As can be seen, results of our evaluation metric are not out of line with results achieved with other evaluation metrics. But in this section, we will highlight what advantages and opportunities our approach provides.

More detailed results, using proposed metric, are provided in Table 3. First column lists coreference classes: pronominal, generic nominal, and definitive nominal. The next six columns correspond to six classes of annotations that were detailed in Section 4.3. S^* is a sum of *TP*, *WT*, *WL*, *WTL*, and *FN*. Last three columns show precision, recall, and F-measure for each coreference class. Results for adverbial and ellipsis coreference classes are not provided, since the proposed approach does not attempt to solve them. Raw, not aggregated, data is available via GitHub (Žitkus 2020).

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Evaluation is made using the data from six annotation classes. As an example, calculations of P, R, and F for pronominal coreferences are presented in (49)–(51) formulas:

$$P_1 = \frac{k_1\text{TP} + k_2\text{WT} + k_3\text{WL} + k_4\text{WTL}}{\text{TP} + \text{WT} + \text{WL} + \text{WTL} + \text{FP}} = \frac{289 + 0.75 * 30 + 0.5 * 27 + 0.25 * 25}{289 + 30 + 27 + 25 + 29} = 82.8\% \quad (49)$$

$$R_1 = \frac{k_1\text{TP} + k_2\text{WT} + k_3\text{WL} + k_4\text{WTL}}{\text{TP} + \text{WT} + \text{WL} + \text{WTL} + \text{FN}} = \frac{289 + 0.75 * 30 + 0.5 * 27 + 0.25 * 25}{289 + 30 + 27 + 25 + 182} = 59.9\% \quad (50)$$

$$F_1 = \frac{2P_1R_1}{P_1 + R_1} = \frac{2 * 82.8 * 59.9}{82.8 + 59.9} = \frac{9919.44}{142.7} = 69.5\% \quad (51)$$

Main advantage of the proposed metric is that we can drill deeper and see more detailed results based not on coreference class but on coreference type (4-letter code). These data are provided in Appendix D, Tables D1–D3.

By analyzing this data for pronominal coreference resolution (Appendix D, Table D1), one can see that the coreference resolution approach resolves relative pronouns (*peas*) very well, but that it misses many (*M* column) personal pronouns (*ppas*). Additionally, reflexive pronouns that function as cataphora (*prps*), group, and ambiguous coreferences are not resolved at all. Such detailed information helps with error analysis and allows us to see which parts of the coreference resolution approach are sufficient enough (relative pronouns) and which need additional work (personal pronouns) or a new approach entirely (cataphoric reflexive pronouns, group, and ambiguous references).

For macro average calculations, we use coreference class value—pronominal, generic nominal, definitive nominal, adverbial, or ellipsis. But as we can see (Appendix D, Table D1), we could use full 4-letter code instead and get more specific results. This is potential direction for future researches.

Generic nominal coreference resolution gets a rather low score (38.7%) and at first glance evaluated coreference resolution approach is not very useful for resolving such coreferences. But looking closer at the data (Appendix D, Table D2), we can see that it struggles with most of these coreferences but performs much better when the discourse-world entity is referred to by its feature (*gfas*). For example, if a politician is being referenced by using his occupation (minister, president, and parliamentarian) or another feature (veteran, firebrand politician, man, and women), then this coreference resolution approach might be suitable for annotating texts that heavily use such type of coreferences. Therefore, coreference type identification can help in selecting the right coreference resolution approach for narrow and specific tasks. With other evaluation metrics, such information might be either lost entirely or difficult to determine.

Evaluated coreference resolution approach solves definitive nominal coreferences rather well, but as can be seen in the detailed data (Appendix D, Table D3) it missed all metonyms (*dmis*). This is a rather important detail if the coreference resolution approach would be used in the annotation of foreign policy texts where metonym relationship is rather common. For example, Russia, Moscow, and Kremlin are often used in such texts to refer to the same discourse-world entity (Russian government), and our coreference resolution approach is likely to miss all of them. Due to that, this coreference resolution approach might not be suitable for texts focusing on international politics. Knowing this is very valuable when selecting tools for NLP pipeline, but without coreference type identification we would not have such information.

As we can see from these results, doing a separate evaluation for each coreference class gives more clarity to the overall performance of the coreference resolution approach and helps in identifying with what specific types of coreferences it struggles with. Next, we calculate macro precision, (52) formula, recall, (53) formula, and their F-measure, (54) formula. And we do the same for micro averages in (55)–(57) formulas:

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

$$P_{\text{macro}} = \frac{\sum_i^{n_a} P_i}{n_a} = \frac{82.8 + 77.1 + 97.6}{3} = 85.8\% \tag{52}$$

$$R_{\text{macro}} = \frac{\sum_i^{n_a} R_i}{n_a} = \frac{59.9 + 25.8 + 91.5}{3} = 59.1\% \tag{53}$$

$$F_{\text{macro}} = \frac{2P_{\text{macro}}R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} = \frac{2 * 85.8 * 59.1}{85.8 + 59.1} = 70\% \tag{54}$$

$$P_{\text{micro}} = \frac{\begin{matrix}(289+0.75*30+0.5*27+0.25*25)+\\ (123 + 0.75 * 4 + 0.5 * 21 + 0.25 * 9) + (973 + 0.5 * 14)\end{matrix}}{\begin{matrix}(289+30+27+25+29)+\\ (123 + 4 + 21 + 9 + 23) + (973 + 14 + 17)\end{matrix}} = 91.5\% \tag{55}$$

$$R_{\text{micro}} = \frac{\begin{matrix}(289+0.75*30+0.5*27+0.25*25)+\\ (123 + 0.75 * 4 + 0.5 * 21 + 0.25 * 9) + (973 + 0.5 * 14)\end{matrix}}{\begin{matrix}(289+30+27+25+182)+\\ (123 + 4 + 21 + 9 + 380) + (973 + 14 + 84)\end{matrix}} = 67.1\% \tag{56}$$

$$F_{\text{micro}} = \frac{2P_{\text{micro}}R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} = \frac{2 * 91.5 * 67.1}{91.5 + 67.1} = 77.4\% \tag{57}$$

It can be seen that the micro F-measure has a 7.4% higher score than macro F-measure. The main reason for this disparity is the fact that definitive nominal coreferences were present in 49.5% of the analyzed cases, and our evaluated coreference resolution approach solves them very well (94.5% F-measure). Therefore, definitive nominal coreference class skews micro average results in its favor.

Since these scores show how well coreference resolution approach is solving coreferences that it attempts to resolve, the following scores, (58)–(60) formulas, have been calculated to show how well it covers the used annotation scheme:

$$P_{\text{scheme}} = \frac{\sum_i^{n_a} P_i}{n} = \frac{82.8 + 77.1 + 97.6}{5} = 51.5\% \tag{58}$$

$$R_{\text{scheme}} = \frac{\sum_i^{n_a} R_i}{n} = \frac{59.9 + 25.8 + 91.5}{5} = 35.44\% \tag{59}$$

$$F_{\text{scheme}} = \frac{2P_{\text{scheme}}R_{\text{scheme}}}{P_{\text{scheme}} + R_{\text{scheme}}} = \frac{2 * 51.5 * 35.44}{51.5 + 35.44} = 41.99\% \tag{60}$$

Overall results of our proposed metric are not out of the line with what other evaluation metrics scored our coreference resolution approach. But the advantage of our proposed metric is the additional dimensions added to the inclusion of dominant mentions and coreference type tracking. This improves the quality of error analysis and provides a clearer picture for coreference resolution approach integration into higher-level applications. Coefficients for each error class also provide a fine-tuning option if certain errors are deemed to be less, or more, important than others.

5. Conclusions and future works

The currently popular coreference evaluation metrics are linguistically agnostic and do not take into account that different types of mentions tend to carry a different amount of semantic information. Some newly developed evaluation metrics started incorporating linguistic information into the valuation process, but they are still lacking. To improve the situation, we have presented a new linguistically aware evaluation strategy for coreference resolution evaluation.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

The proposed evaluation strategy combines dominant mentions with coreference type identification and can measure performance among different coreference classes. This provides more valuable information for error analysis. Next to the commonly used micro averages for scoring, the coreference resolution approaches authors suggest adding macro averages and scheme coverage values. Macro averages diminish the impact of the overrepresented classes of coreference in the given corpus. Combined with type identification macro averages also help in selecting the coreference resolution approach for a narrow or specific task. While, scheme coverage values provide a bigger picture view and show how well the coreference resolution approach covers annotations scheme in general.

The proposed evaluation strategy is not language-dependent. For the presentation, we have used our developed annotation scheme for the Lithuanian language, but the evaluation strategy itself is not tied to it. Evaluation strategy can easily be adapted to another annotation scheme as long as it provides the classification of coreference expressions.

For future works, dominant mentions needs to be further fleshed out to cover exophoric mentions, and a streamlined process is required to determine which dominant mention from two, or more, different sources is the most dominant. After further testing coefficients, assigned to different coreference annotations based on their correctness, could be further adjusted.

References

Bagga A. and Baldwin B. (1998). Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, vol. **1**. Granada, pp. 563–566.

Byron D. K. (2001). The uncommon denominator: a proposal for consistent reporting of pronoun resolution results. *Computational Linguistics* **27**(4), 569–577.

Ceberio K., Aduriz I., de Ilarraza A. D. and Garcia-Azkoaga I. (2018). Coreferential relations in Basque: the annotation process. *Journal of Psycholinguistic Research* **47**(2), 325–342.

Chen C. and Ng V. (2013). *Linguistically aware coreference evaluation metrics*. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1366–1374.

Delmonte R. (2002). *Relative clause attachment and anaphora: a case for short binding*. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 6)*, pp. 84–89.

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

Doddington G. R., Mitchell A., Przybocki M. A., Ramshaw L. A., Strassel S. M. and Weischedel R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Lrec*, vol. **2**. Lisbon, pp. 1–4.

Elango P. (2005). *Coreference Resolution: A Survey*. Madison, WI: University of Wisconsin.

Fischer S. (2015). Pronominal anaphora. In *Syntax-Theory and Analysis: An International Handbook*, vol. **1**. Berlin: Mouton de Gruyter, pp. 446–477.

Gardelle L. (2012). Anaphora" "anaphor" and "antecedent" in nominal anaphora: definitions and theoretical implications. *Cercles* **22**, 25–40.

Hobbs J. R. (1978). Resolving pronoun references. *Lingua* **44**(4), 311–338.

Holen G. I. (2013). *Critical reflections on evaluation practices in coreference resolution*. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pp. 1–7.

Hou Y., Markert K. and Strube M. (2013). *Global inference for bridging anaphora resolution*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 907–917.

Kaczmarek A. and Marcińczuk M. (2015). Evaluation of coreference resolution tools for Polish from the information extraction perspective. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pp. 24–33.

King J. C. and Lewis K. S. (2018). *Anaphora*. Stanford, CA: Metaphysics Research Lab, Stanford University.

Krahmer E. and Piwek P. (2000). Varieties of anaphora. Introduction. In *Reader ESSLLI*, pp. 1–15.

Kummerfeld J. K. and Klein D. (2013). *Error-driven analysis of challenges in coreference resolution*. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 265–277.

Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M. and Jurafsky D. (2011). *Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task*. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 28–34.

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

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Luo X., **Pradhan S.**, **Recasens M. and Hovy E.** (2014). An extension of BLANC to system mentions. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. **2014**. NIH Public Access, pp. 24–29.

Manning C., **Raghavan P. and Schütze H.** (2010). Introduction to information retrieval. *Natural Language Engineering* **16**(1), 100–103.

Minaee S., **Kalchbrenner N.**, **Cambria E.**, **Nikzad N.**, **Chenaghlu M. and Gao J.** (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)* **54**(3), 1–40.

Mitkov R. (2014). *Anaphora Resolution*. London: Routledge.

Moosavi N. S. and Strube M. (2016). *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)*, pp. 632–642.

Novák M. (2018). *Coreference from the Cross-lingual Perspective*. PhD Thesis, Faculty of Mathematics and Physics, Charles University.

Ogrodniczuk M., **Głowińska K.**, **Kopeć M.**, **Savary A. and Zawisławska M.** (2013). *Polish coreference corpus*. In *Language and Technology Conference*. Cham: Springer, pp. 215–226.

Pradhan S., **Luo X.**, **Recasens M.**, **Hovy E.**, **Ng V. and Strube M.** (2014). Scoring coreference partitions of predicted mentions: a reference implementation. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. **2014**. NIH Public Access, pp. 30–35.

Pradhan S., **Moschitti A.**, **Xue N.**, **Uryupina O. and Zhang Y.** (2012). *CoNLL-2012 shared task: modeling multilingual unrestricted coreference in OntoNotes*. In *Joint Conference on EMNLP and CoNLL-Shared Task*. Association for Computational Linguistics, pp. 1–40.

Recasens M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation*. Unpublished PhD Dissertation. University of Barcelona.

Rösiger I. and Teufel S. (2014). *Resolving coreferent and associative noun phrases in scientific text*. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 45–55.

Saeboe K. J. (1996). Anaphoric presuppositions and zero anaphora. *Linguistics and Philosophy* **19**(2), 187–209.

Souza M., **Glauber R.**, **de Oliveira L. S.**, **Sena C. F. L. and Claro D. B.** (2017). Nominal coreference annotation in IberEval2017: the case of FORMAS Group. In *IberEval@ SEPLN*, pp. 92–101.

Sukthanker R., **Poria S.**, **Cambria E. and Thirunavukarasu R.** (2018). Anaphora and coreference resolution: a review. *Information Fusion* **59**, 139–162.

Tuggener D. (2014). *Coreference resolution evaluation for higher level applications*. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*, pp. 231–235.

Tuggener D. (2016). *Incremental Coreference Resolution for German*. PhD Thesis. University of Zurich.

Van Deemter K. and Kibble R. (1999). *What is coreference, and what should coreference annotation be?* In *Proceedings of the Workshop on Coreference and Its Applications*. Association for Computational Linguistics, pp. 90–96.

Vilain M., **Burger J.**, **Aberdeen J.**, **Connolly D. and Hirschman L.** (1995). *A model-theoretic coreference scoring scheme*. In *Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics, pp. 45–52.

Vileiniškis T., **Šukys A. and Butkienė R.** (2015). *Searching the web by meaning: a case study of Lithuanian news websites*. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Cham: Springer, pp. 47–64.

Zhong X., **Cambria E. and Hussain A.** (2020). Extracting time expressions and named entities with constituent-based tagging schemes. *Cognitive Computation* **12**(4), 844–862.

Žitkus V. (2018). Lithuanian coreference corpus. In *CLARIN-LT Digital Library in the Republic of Lithuania*. Available at http://hdl.handle.net/20.500.11821/19

Žitkus V. (2020). Coreference resolution annotator raw data. Available at https://github.com/volzitk/CoreferenceAnnotatorData

Žitkus V. and Butkienė R. (2018). *Coreference annotation scheme and corpus for Lithuanian Language*. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, pp. 243–250.

Žitkus V., **Butkiene R.**, **Butleris R.**, **Maskeliunas R.**, **Damaševičius R. and Woźniak M.** (2019). Minimalistic approach to coreference resolution in Lithuanian medical records. *Computational and Mathematical Methods in Medicine* **2019**, 1–14.

Zitouni I. (2014). *Natural Language Processing of Semitic Languages*. Berlin/Heidelberg: Springer.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Appendix A

Table A1. Definition of the Annotation package concepts

Concept	Definition
Set	Collection of coreference annotation layers that are being used for the evaluation. *Set* consists of one or more *Coreference annotation layers*. *Set* is specialized by two types: *Gold set* and *Response set*.
Gold set	Set of manually specified coreference annotations that CR approaches are evaluated against.
Response set	Set of coreference annotations created by automated annotator that has to be evaluated.
Coreference annotation layer	Complete coreference annotation for one given text document. Each *Coreference annotation layer* is made out of multiple *Annotations*.
Annotation	Notation that describes one specific case of coreference found in the text. *Annotation* has a property type that specifies four letters coreference type code combining letters from four different coreference classification levels (Table 1). Each *Annotation* is composed of one *referent* and at least one *antecedent*.
Antecedent	Expression to which another expression is pointing to in the context of one text document. *Antecedent* is specialized by two separate classes: *Dominant mention* and *Non-dominant mention*.
Dominant mention	Antecedent that can be dominant according to guidelines presented in Section 4.2. Each *Dominant mention* is referenced by one or more *Referent of dominant mention*.
Non-dominant mention	Antecedent that cannot be dominant according to guidelines presented in Section 4.2. Each *Non-dominant mention* is referenced by one *Referent of non-dominant mention*.
Referent	Expression pointing to another expression in the context of one text document. *Referent* is specialized by *Referent of non-dominant mention* and *Referent of dominant mention* concepts.
Referent of dominant mention	Referent that points to dominant mention. Each *Referent of dominant mention* refers to one *Dominant mention*.
Referent of non-dominant mention	Referent that points to antecedent that is not dominant. Each *Referent of non-dominant mention* refers to one *Non-dominant mention*.

Table A2. Definition of the Evaluation package concepts

Concept	Definition
Evaluation	Process of the evaluation. *Evaluation* uses annotations from *Gold set* to evaluate annotations that are present in *Response set*. *Evaluation* has *precision_micro*, *recall_micro*, *f_measure_micro*, *precision_macro*, *recall_macro*, *f_measure_macro*, *precision_scheme*, *recall_scheme,* and *f_measure_scheme* properties that store final evaluation values. Each *Evaluation* calculates evaluation values from one or more *Scores*. Each Score has *coreference_class* property declaring for which coreference class *precision*, *recall,* and *f_measure* properties were calculated.
Score	Calculated evaluation value for specific coreference class. Each *Score* counts annotations from six different concepts: *Correct annotation*, *Correct annotation with the wrong type*, *Correct annotation with the wrong dominant mention*, *Correct annotation with wrong dominant mention and type*, *Missed annotation*, and *False positive annotation*.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Table A2. (continued)

Concept	Definition
Correct annotation	Annotation that has correct coreference type specified and is linked to the correct dominant mention. Specializes *Annotation* concept.
Correct annotation with the wrong type	Annotation that is linked to the correct dominant mention but has wrong coreference type specified. Specializes *Annotation* concept.
Correct annotation with the wrong dominant mention	Annotation that has correct coreference type specified but is linked to the wrong dominant mention. Specializes *Annotation* concept.
Correct annotation with the wrong dominant mention and type	Annotation that is linked to the wrong dominant mention and has wrong coreference type specified. Specializes *Annotation* concept.
Missed annotation	Annotation that is present in a *Gold set* but is not found in a *Response set*. Specializes *Annotation* concept.
False positive annotation	Annotation that is present in a *Response set* but is not found in a *Gold set*. Specializes *Annotation* concept.

Appendix B

B.1 Text used for coreference resolution

Prime Minister of Greece as the country is taking the helm of the EU: "Cassandra's prophecy did not come true." Greece presents its program and priorities for the Presidency of the Council of the European Union (EU) at a plenary session of the European Parliament (EP) in Strasbourg. According to Prime Minister Antonis Samaras, Greece and the whole EU were on the brink of disaster, but "Cassandra's prophesies did not come true." "Greece has not gone bankrupt. A year and a half ago, my country was on the brink of disaster. There were talks that Greece would leave the eurozone and that could lead to the collapse of the EU as a whole, but that has not materialized. Greece still stays in the EU, the Union still exists and remains reliable. We have taken over the Presidency of the Council of the EU and I hope that my country will become a symbol that Europe is making progress, working hard and is capable of delivering results," Samaras said in Strasbourg.

Jose Manuel Barroso, the head of the European Commission (EC), welcomed the fact that the apocalyptic scenarios for Greece had not come true. "What do we see? The Greek Prime Minister has arrived, he is committed to Europe. All this proves that the doomsayers and scaremongers were wrong. Greece's experience since the beginning of the crisis makes us work even harder and strive for a successful Presidency," the head of the EC noted. The Greek prime minister said he would also strive for dialog between the EU institutions. The role of the EP should also be strengthened, he said. "We are going through difficult times, the EU has suffered greatly. Some mistakes have been made in the past, but we are gradually overcoming the crisis and Europe is learning from its mistakes. I think it was during the crisis that we showed that the EU works, that it can work together. We should finish the work that we started a couple of years ago," the Prime Minister said.

According to Samaras, failure is not when you fall, but rather being unable to get up. "If you fall and get up again, you are resilient and able to overcome the problem maintaining your dignity. I think my people have sacrificed a lot, but the Greeks have got up and they preserved their dignity. It overcame the biggest problems, and implemented the financial corrections," said Samaras.

The priorities of the Greek Presidency of the Council of the EU are job creation, economic and social cohesion and structural reforms, further EU integration, the establishment of economic and monetary union, maritime policy and EU enlargement. Greece takes over the EU Presidency from

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Lithuania, which held the Presidency for 6 months from July last year. Presenting the results of the presidency in Strasbourg on Tuesday, President Dalia Grybauskaitė stressed that Lithuania had successfully overcome the challenges it had faced during the 6-month term.

B.2 Relevant coreference resolution results provided by Stanford CoreNLP 4.2.2

```
<coreference<
  <mention representative="true">
  <sentence> 8< /sentence> < start> 1< /start> < end> 2< /end>
  <head> 1< /head> < text> We< /text>
  </mention>
  <mention>
  <sentence> 10< /sentence> < start> 4< /start> <end> 5< /end>
  <head> 4< /head> <text> we< /text>
  </mention>
  <mention>
  <sentence> 13< /sentence> < start> 11< /start> <end> 12< /end>
  <head> 11< /head> <text> us< /text>
  </mention>
  <mention>
  <sentence> 16< /sentence> < start> 2< /start> < end> 3< /end>
  <head> 2< /head> < text> We< /text>
  </mention>
  <mention>
  <sentence> 17< /sentence> < start> 11< /start> < end> 12< /end>
  <head> 11< /head> <text> we< /text>
  </mention>
  <mention>
  <sentence> 18< /sentence> < start> 9< /start> < end> 10< /end>
  <head> 9< /head> < text> we< /text>
  </mention>
  <mention>
  <sentence> 19</sentence> < start> 1< /start> < end> 2< /end>
  <head> 1</head> <text> We< /text>
  </mention>
  <mention>
  <sentence> 19</sentence> < start> 7< /start> < end> 8< /end>
  <head> 7< /head> <text> we< /text>
  </mention>
  </coreference>
```

Appendix C

Provided example:

Barack Obama was elected the 44th president of the United States on Tuesday. *B. Obama* is the first black US president. The newly elected *president* will be in charge of the armed forces. *He* is scheduled to appear at the press conference tomorrow.

Coreference cluster in gold set: Barack Obama, B. Obama, president, He. “Barack Obama” is marked as the dominant mention.

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Clusters in response set:

Cluster A: Barack Obama, B. Obama. "Barack Obama" is marked as the dominant mention.
Cluster B: president, He. "president" is marked as the dominant mention.

Assumption is made that response set has correctly identified coreference types. Due to small example size, using macro calculations is not efficient and only micro calculations for precision (61) and recall (62) are provided:

$$P = \frac{k_1\mathrm{TP} + k_2\mathrm{WT} + k_3\mathrm{WL} + k_4\mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FP}} = \frac{1 * 1 + 0.75 * 0 + 0.5 * 1 + 0.25 * 0}{1 + 0 + 1 + 0 + 0} = \frac{1.5}{2} = 0.75 \tag{61}$$

$$R = \frac{k_1\mathrm{TP} + k_2\mathrm{WT} + k_3\mathrm{WL} + k_4\mathrm{WTL}}{\mathrm{TP} + \mathrm{WT} + \mathrm{WL} + \mathrm{WTL} + \mathrm{FN}} = \frac{1 * 1 + 0.75 * 0 + 0.5 * 1 + 0.25 * 0}{1 + 0 + 1 + 0 + 1} = \frac{1.5}{3} = 0.5 \tag{62}$$

Appendix D

Coreference type column has 4-letter codes indicating what type of the coreference (according to the presented annotation scheme in Section 4.1) was attempted to solve. Coreference types that resolution approach did not attempt to solve would not provide additional insight; therefore, they are aggregated under "Other" label. "All" lab sums up totals for each column.

Table D1. Experiment results for pronominal coreference resolution

Coreference type	TP	WT	WL	WTL	FN	FP	S*
ppas	103	19	12	14	83	20	231
ppps	9	4	15	6	7	2	41
pras	4	3	0	0	13	0	20
prps	0	0	0	0	2	0	2
poas	18	4	0	5	48	5	75
pops	1	0	0	0	3	2	4
peas	141	0	0	0	3	0	144
peag	13	0	0	0	6	0	19
Other	0	0	0	0	17	0	17
All	289	30	27	25	182	29	553

https://doi.org/10.1017/S1351324923000293 Published online by Cambridge University Press

Table D2. Experiment results for generic nominal coreference resolution

Coreference type	TP	WT	WL	WTL	FN	FP	S∗
gais	0	0	0	0	54	0	54
gtis	0	0	0	0	109	0	109
gfas	107	0	15	0	62	13	184
gfps	3	0	4	0	7	3	14
ghas	4	2	2	6	47	1	61
gmis	0	0	0	0	34	0	34
gsis	9	2	0	3	34	6	48
Other	0	0	0	0	33	0	33
All	123	4	21	9	380	23	537

Table D3. Experiment results for definitive nominal coreference resolution

Coreference type	TP	WT	WL	WTL	FN	FP	S*
dtis	728	0	0	0	40	8	768
dais	22	0	5	0	14	4	41
dbis	223	0	9	0	13	5	245
dmis	0	0	0	0	14	0	14
Other	0	0	0	0	3	0	3
All	973	0	14	0	84	17	1071

Cite this article: Žitkus V, Butkienė R and Butleris R. Linguistically aware evaluation of coreference resolution from the perspective of higher-level applications. *Natural Language Engineering* https://doi.org/10.1017/S1351324923000293