Natural Language Processing Journal 13 (2025) 100184

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

# Position-aware end-to-end cross-document event coreference resolution for Dutch

Loic De Langhe *, Orphée De Clercq, Veronique Hoste

*LT3, Language and Translation Technology Team, Ghent University, Ghent, 9000, Belgium*

ARTICLE INFO

*Keywords:*
Coreference Resolution
Discourse Processing
Events

ABSTRACT

Natural language understanding entails the ability to comprehend the relations between various people, objects or events throughout one, or multiple, text(s). Event coreference resolution (ECR) is a discourse-based natural language processing (NLP) task which aims to link those textual events, be they real or fictional, that refer to the same conceptual event. In this paper, we introduce a novel end-to-end approach for cross-document ECR which combines expert-level positional knowledge and graph-based representations in order to create a memory-efficient and accurate system meant for the detection and resolution of events in large document collections. We make three fundamental architectural changes to a current state-of-the-art cross-document ECR system and show that our approach outperforms this earlier model (+ 4% CONLL F1) on a large Dutch ECR dataset. Moreover, we show through in-depth qualitative and quantitative analysis that our proposed approach consistently detects more relevant events and suffers notably less from the typical issues models exhibit when predicting coreference chains.

## 1. Introduction

Being able to connect various mentions of people, objects and events in written and oral discourse is a fundamental aspect of human language understanding. Moreover, the knowledge that certain parts of a text or a conversation can be intrinsically linked – not only to other texts or conversations, but also to real-world concepts – is at the very centre of what enables us humans to communicate with one another. Yet, this critical aspect of human communication plays only a marginal role in most modern-day studies on Computational Linguistics (CL) and Natural Language Processing (NLP).

The advent and success of Large Language Models (LLMs) and transformer-based architectures has profoundly shifted the field's overall research focus to word-level lexical semantics, prioritizing local context understanding over document or multi-document level structural processing (Jiang et al., 2023). While a large variety of tasks which require thorough syntactic and semantic knowledge, such as sentiment analysis, machine translation and question-answering, have immensely benefited from the transformer-based revolution (Tenney et al., 2019; Koroteev, 2021), most tasks involving discourse processing remain a tough nut to crack, even with access to powerful LLMs (Nair et al., 2023; Maekawa et al., 2024).

Cross-Document Event Coreference Resolution (CD-ECR) is one of the foremost tasks with which modern-day LLMs tend to struggle (Lu and Ng, 2021). In this discourse-based task, the goal is to connect all references of events or happenings in a written document collection if, and only if, they refer to the same real-world conceptual event. Consider the two sentences below in which the textual span of the event, often designated as the event *mention*, is indicated in between square brackets:

1. Hochul [elected] as 57th Governor of New York
2. She [received] her ticket to Albany street in 2021.

For a human reader with adequate background knowledge (i.e. The New York Governor's official residence is on Albany street), it is clear that the same real-world event is being referenced. However, for most present-day LLMs, which often lack up-to-date real-world knowledge and situational context, connecting these two events is more problematic. A specific problem with generative LLMs is their tendency to produce false negatives, primarily due to their reliance on general pre-training data, which lacks the implicit knowledge necessary for tasks like coreference resolution (Liu et al., 2023). Interestingly, fine-tuned transformer architectures suffer less from this issue, making them more commonly used for specialized tasks where accuracy is critical, though they are still far from perfect, and deployment at large scale is not yet possible (Lu and Ng, 2021).

Nevertheless, the potential of integrating event coreference resolution systems into downstream applications can be of great benefit:

* Corresponding author.
*E-mail address:* loic.delanghe@ugent.be (L. De Langhe).

https://doi.org/10.1016/j.nlp.2025.100184
Received 25 February 2025; Received in revised form 27 June 2025; Accepted 11 September 2025

more specifically, CD-ECR algorithms have been used in the past in tasks such as template filling (Humphreys et al., 1997a), automated population of knowledge bases (Ji and Grishman, 2011), question answering (Narayanan and Harabagiu, 2004) and contradiction detection (De Marneffe et al., 2008). Incorporating event-level information and resolving the coreference between these events has also enhanced the performance of automatic summarization systems (Alsaedi et al., 2016) and content-based news recommenders (Kompan and Bieliková, 2010).

The integration of CD-ECR systems in most applications is complicated as for most downstream tasks an algorithm would ideally need to be able to not only find coreferential links between events, but also first detect the events themselves. This joint detection and resolution of events is typically designated as end-to-end (e2e) event coreference resolution and remains one of the primary challenges within the field of discourse processing (Lu and Ng, 2018a). Such a streamlined joint system would remove the need for cumbersome intermediary processing steps and complex model pipelines, which are often used in cross-document applications today (Elov et al., 2023).

In this paper, we propose such an e2e system for cross-document event coreference resolution. It learns to both detect and resolve event mentions in the Dutch ENCORE corpus (De Langhe et al., 2022a). This corpus consists of a large collection of online news articles with high-quality cross-document event annotations and is currently the largest human-annotated cross-document event corpus available. The choice to work exclusively with news data in this corpus is because most proposed downstream applications of CD-ECR mentioned above can be situated within the news domain. Additionally, having a collection of news articles reporting on a wide variety of present-day issues allows us to research CD-ECR in an unrestricted setting with a broad topical scope.

For our overall model architecture, we drew inspiration from earlier work on e2ecoreference resolution models in cross-document settings (Cattan et al., 2021a) and integrate both structural document-level information and a method to explicitly model graph-based representations of coreference chains. We perform an extensive quantitative and qualitative analysis of the model's output and show that our proposed model architecture consistently detects and classifies harder cases of coreference compared to a series of solid LLM-based baseline algorithms.

The remainder of this paper is structured as follows: first, Section 2 provides an extensive overview of widely-used resources and methods in ECR studies. In this section, we discuss the overall motivation, methodology and flaws of the current generation of event coreference resolution models. Then, we propose three concrete changes in model architecture that we believe will combat some of the issues plaguing earlier developed CD-ECR models and discuss our overall model training and testing strategy in Section 4. After reporting the results (Section 5), we then perform a series of extensive qualitative analyses of our proposed model in Section 5.2. In contrast to most earlier studies, we not only analyse the predicted coreference links, but also the individual events which have been detected by the models. Finally, Section 7 contains a brief summary of our most important findings and results, as well as an outlook for future cross-document event coreference resolution research.

## 2. Related work

### 2.1. Resources

Available resources for CD-ECR are few and far between, especially for languages other than English. Moreover, most of the available corpora adopt different definitions of what events constitute, as well as diverging annotation strategies, making model comparison across various datasets a complicated endeavour. The section below attempts to provide a complete overview of publicly existing resources for (cross-document) event coreference resolution, while briefly discussing their key characteristics and current place in the field. A schematic overview of all corpora can be found at the end of this section in Table 1.

Some of the first corpora specifically developed for event coreference resolution were the ACE corpora, and more specifically ACE 2005 (Anon, 2008). In this collection of documents, events are seen as a set of 100 predefined type actions i.e. *Life - BeBorn* or *Employment - Start*, which results in a fairly restricted event taxonomy as a whole. Coreference between events is annotated at the within-document level, with coreffering events needing to be compatible in (sub)type. In addition to the annotations at the event-level, entity mentions are also included in this dataset. Furthermore, the ACE corpora distinguished themselves from other early event-based resources through the inclusion of a (relatively small) set of annotated Chinese documents, making this corpus the first known multilingual corpus for ECR.

As mentioned, the overall number of event types in ACE is somewhat restricted. With this in mind, there have been efforts to create ACE-like corpora geared towards domain-general applications. For instance, the TAC-KBP corpora (Mitamura et al., 2015) build upon the ACE methodology by not only expanding the event typology, but also by including a more complex and information-rich annotation style for the events contained in the documents. This multilingual dataset includes documents for English, Arabic, Chinese and Spanish and considers coreferential links at the within-document level.

A third resource that should be discussed, is the OntoNotes corpus (Pradhan et al., 2007). For a significant period in time, this corpus was considered the standard dataset for both entity and event coreference resolution systems in the English language domain. This large-scale multi-domain text collection includes Treebank (Taylor et al., 2003), Propbank (Kingsbury and Palmer, 2002, 2003) and within-document coreference annotations both at the entity and event level. Additionally, both entities and event types are largely unrestricted, meaning that this corpus covers documents and topics in a large variety of domains. A potential hazard of the OntoNotes corpus is that no distinction has been made between coreferential links at the entity and event level. While the tasks of entity and event coreference are certainly related, it should be noted that, over time, methodologies for entity and event coreference resolution have started to diverge (Lu and Ng, 2018a; Liu et al., 2023). For instance, event-based resolution approaches have started to heavily rely on discourse and meta-textual knowledge rather than on sentence and document level information which is typically used for the detection of coreferential entities (Chen et al., 2023; De Langhe et al., 2023b). In addition, the OntoNotes corpus handles most verbal (event-level) mentions in an entirely different manner than the earlier discussed ACE and TAC-KBP corpora. Concretely, a verbal mention is only annotated if there exists a noun phrase equivalent of the same real-world or fictional event in the dataset. It has been suggested that this annotation strategy has an inherent impact on data consistency and is therefore less optimal for the development of practical real-word applications of event coreference resolution (De Langhe et al., 2023d).

In the three aforementioned corpora, the annotation of coreferential links has been restricted to the within-document level. Where resources on event coreference are already scarce, resources for cross-document event coreference are even harder to come by. In the English language domain, two datasets of notable size currently exist. The first is the monolingual English ECB+ corpus (Cybulska and Vossen, 2014b), which was based on the small-scale ECB corpus (Bejan and Harabagiu, 2010). Since its release, the ECB+ corpus has been used for numerous cross-document coreference studies (Lu and Ng, 2018b, 2021) and has become the standard benchmark for newly developed (English-language) CD-ECR models. The corpus includes a significant number of news articles from different domains in which multi-word event spans are annotated using a schema based on the Rich ERE guidelines (Song et al., 2015). A large part of the ECB+ corpus' popularity is owed to the fact that its creators explicitly included events of the same type which were not coreferent in an effort to diversify the data. Consider Lists 3 and 4 below:

**Table 1**
Overview of widely-used event coreference corpora , both at the within-document (WD) level and cross-document (CD) level.

| Corpus | #Documents | Genre | Coref | Languages |
|---|---|---|---|---|
| *ACE 2005* | 600, 500 | news articles, conversations | WD | EN, CH |
| *OntoNotes* | 600 | financial news articles | CD | EN |
| *TAC KBP* | 1000, 800, 400 | news articles, forum discussions | WD | EN, SP, CH |
| *ECB* | 480 | news articles | CD | EN |
| *ECB+* | 982 | news articles | CD | EN |
| *Newsreader Meantime* | 120 | news articles | CD | EN, DU, IT, SP |

3. Tara Reid committed to rehab
4. Singer Britney Spears entered an LA-based rehab facility last month

For a human reader it is immediately clear that even though the two events are similar, they do not refer to the same event because the participant(s), time and place are different. The emphasis on including similar, but not coreferent, events results in a dataset which is arguably highly representative of the proposed ECR applications for real-word data.

A second and final cross-document corpus for English is the automatically extracted WEC-Eng dataset (Eirew et al., 2021). In this dataset, Wikipedia data is leveraged where both event mentions and coreference links between events are not restricted to pre-defined topics.

While all resources previously discussed are situated either in the English language domain or comparatively higher-resourced domains such as Chinese and Arabic, there do exist two noteworthy CD-ECR datasets for Dutch. First, the MeanTime Newsreader (Minard et al., 2016) corpus. This richly annotated event dataset contains cross-document level coreference annotations for English, Dutch, Spanish and Italian and deals with largely unrestricted events. However, the documents collected in this dataset are limited to four specific topics, and the Dutch, Spanish and Italian sections of the dataset have been machine-translated from the original English texts. Moreover, the event and coreference annotations of the Italian and Spanish sections have not been carried out manually but through cross-lingual projection. The second, and primary, dataset for Dutch CD-ECR is the large-scale ENCORE dataset, which serves as the backbone for this paper and will therefore be discussed in more detail in Section 3.

### 2.2. Methodology

When it comes to operationalizing coreference resolution, most models fall within the mention-pair paradigm, in which each possible anaphor-antecedent pair in a text is evaluated in isolation and a binary (coreferent/non-coreferent) decision in made for each pair (Soon et al., 2001). This paradigm has been by far the most popular and effective approach for coreference resolution of entities and events alike in the past 20 years (Liu et al., 2023). Even within transformer-based studies on anaphora resolution, the mention-pair approach remains the primary methodology of choice.

Nonetheless, there exists a non-insignificant body of work aiming to break away from binary representation learning and which attempts to represent coreference chains through other structures (i.e. computational graphs or trees). In the sections below, we aim to provide the reader with a comprehensive overview of widely-used (end-to-end) event coreference methodologies, both within and beyond the mention-pair paradigm.

#### 2.2.1. Mention-pair models

The mention-pair paradigm is a relatively fixed structure. First, a number of candidate mention pairs is generated, either at the document or cross-document-level. Then, representations of each mention in the pair are jointly passed through a classifier which predicts whether or not the two mentions are coreferent. Then, from the binary output of the coreference model, coreference chains are formed. In the past, this was often done through either cluster-ranking (Rahman and Ng, 2009) or iterative learning of the optimal coreference chains (Clark and Manning, 2015), whereas more recent approaches tend to prefer agglomerative clustering methods (Barhom et al., 2019).

Early research on mention-pair coreference resolution was primarily inspired by classical feature-based methods, such as support vector machines (Chen and Ng, 2014), decision trees (Cybulska and Vossen, 2015) and deep neural networks (Nguyen et al., 2016). Features used for these approaches centred around string-based similarities and role-filling. Additionally, explicitly modelling document and discourse aspects, such as the textual distance between events, were also thought to be important for the resolution of coreferential events (Lu and Ng, 2018b; De Langhe et al., 2022b).

As is the case for most domains within the NLP research sphere, feature-based methods in ECR have nearly entirely been replaced by transformer architectures. In most cases, these take the form of BERT-based encoders, which create contextualized embedding representations of the individual events which are subsequently passed through a simple feedforward classification layer (Lu et al., 2022; Cattan et al., 2021b). Among others, transformer-based ECR models have taken the form of span or long-format encoders such as SpanBERT (Joshi et al., 2020) and Longformer (Beltagy et al., 2020).

A second, less frequent, approach to transformer-based ECR is by directly prompting generative models. Some preliminary studies have attempted to leverage the real-world implicit knowledge contained within GPT-like decoders (Zhang et al., 2023), usually focusing on chain-of-thought explanations (Nath et al., 2024).

For end-to-end (e2e) settings, the mention-pair approach is changed in one significant aspect: instead of candidate pairs being created from gold-standard event mentions, they are being predicted (Lu et al., 2022). The most straightforward approach to e2e event coreference resolution is by employing a pipeline of different systems. In a pipeline setup, the detection and resolution of event mentions are considered as two entirely different tasks. The output of any event detection system can then be fed directly into any of the resolution methods described above. While some studies have extensively focused on the development of pipeline approaches (Singh et al., 2019; Zhu et al., 2024), it should be noted here that these systems come with a notable drawback. For one, pipeline models notoriously suffer from error propagation, meaning that errors made in one of the components get passed as the input for the next component with no method to rectify them. This is even more apparent in systems which add extra steps in the event detection step, such as the detection of task-specific event properties (Singh et al., 2019).

This issue has inspired many researches to shift their attention to joint coreference models, in which detection and resolution are considered a joint optimization task. Early joint methods to coreference typically used segment-based decoding (Araki and Mitamura, 2015), Markov logic networks (Lu et al., 2016) or integer linear programming (Chen and Ng, 2016) to perform joint inference. However, since the arrival of transformer-based architectures, most e2e coreference resolution systems, for entities and events alike, have been based on Lee et al. (2017)'s seminal *e2e-coref* algorithm. In this approach, strong contextual span representations are passed through an event scorer module, before being combined, pruned and passed through a pairwise scoring layer. Since its inception, many variants and optimizations of this model have been proposed, ranging from removing span representations (Kirstain et al., 2021) to specialized pairwise encoders (Kobayashi et al., 2023). One such variant for Cross-Document event coreference resolution (Cattan et al., 2021a) is seen as the backbone for the current state of the art on that particular task (Lu and Ng, 2021). This approach will be discussed in closer detail in Section 4.2.

#### 2.2.2. Other approaches

While mention-pair approaches remain the preferred method for (e2e) coreference resolution, the models have been shown to suffer from an over-reliance on superficial lexical similarity (Cattan et al., 2021a; Ahmed et al., 2023; De Langhe et al., 2023d). This leads to similar mentions consistently being classified as part of a coreference relation, even when they are not, resulting in a form of topic classification rather than actual coreference resolution.

In an effort to reduce this over-reliance on lexical similarity between event pairs, some have sought to step away from pairwise classification altogether. Graph-based models have been increasingly popular is this regard. These methods' primary goal is to create a representation of coreference based on the actual structure of the chains by integrating the overall document (Fan et al., 2022; Tran et al., 2021) or discourse (Huang et al., 2022) structure. Other graph-based methods have focused on commonsense reasoning (Wu et al., 2022). In addition to this, efforts have been made to create topology-aware frameworks for inductive relation prediction, where mentions are represented not by their identities but through the structural patterns of their neighbouring relations in the knowledge graph (Chen et al., 2022). This type of model constructs relation-aware subgraphs and encodes them using a graph neural network that captures both the types and configurations of adjacent relations, allowing the system to generalize effectively to previously unseen mentions by relying on shared topological features rather than explicit embeddings. Finally, it has also been suggested that small graph-based models can be integrated into existing mention-pair methods in order to learn chain structure as a whole and scale down on the pairwise dependence (De Langhe et al., 2023a).

Another approach to model coreference is to model coreferential relations using trees instead. Tree-based coreference representations have ranged from the exploration of universal dependency trees (Popel et al., 2021) to latent tree-based classifiers (Lassalle, 2015). In addition to this, modelling global discourse structure as a tree in coreference resolution has also been shown an interesting direction to pursue (Gordon and Hendrick, 1998).

Finally, work on coreference resolution has explored the use of generative LLMs for both entity and event coreference, often through prompt-based approaches. Some methods such as CorefPrompt, reframe event coreference as a cloze-style task using templates to model event and argument compatibility (Xu et al., 2023). In addition, some frameworks focus aim to better handle long documents by resolving references within sub-contexts, mitigating the context-length limitations of LLMs (Liu et al., 2024). Hybrid methods combining LLM-generated summaries with smaller fine-tuned models have also been proposed (Min et al., 2024). These approaches leverage the flexibility of prompting to encode coreference rules without retraining, but generative models still struggle with false negatives and lack the robustness of traditional fine-tuned models, especially in large-scale or precision-critical settings (Min et al., 2024).

## 3. Data

The ENCORE corpus (De Langhe et al., 2022a), currently the largest cross-document unrestricted event coreference corpus for Dutch, contains 1,015 full newspaper documents, sourced from a variety of national (*De Morgen*, *Het Nieuwsblad*, *Het Laatste Nieuws*, *De Standaard*) and regional (*Het Belang van Limburg*) online newspapers, as well as articles published by the Flemish public broadcasting agency (*VRT News*).

In the ENCORE corpus, the annotation of events and their coreferential links was mostly inspired by the general design methodology of the previously mentioned benchmark (Cybulska and Vossen, 2014a) corpus. The data is structured in topical news clusters, where each cluster is a self-contained collection of news articles reporting on a single conceptual event. These clusters were obtained by grouping documents if they share a threshold number (> 5) individual named entities. Each cluster is assigned an id and a descriptor which covers the theme of the cluster, examples of which are presented in Table 2.

In each of these clusters, events were manually annotated relying on specific annotation guidelines (De Langhe et al., 2021) At the hearth of the event is the event trigger, which can either be a nominal or verbal expression denoting the central action of the event. In the ENCORE corpus, there is no predefined event typology under which triggers must fall to be considered an event, resulting in a highly diverse unrestricted event corpus. For each event, a series of event arguments are additionally annotated (if present): *location*, *time*, *human participant* and *non-human participant*. Consider the example below as a prototypical event annotation in the ENCORE dataset:

5. [De [finale]$_{Action}$ van het WK voetbal 2022$_{TIME}$.]$_{Event}$ *EN: The final of the 2022 World Cup.*

In addition, three key event properties are specified for each event trigger. First, the event sentiment, which is self-explanatory and can either be positive, negative or neutral. Second, the event realis, which denotes whether or not the event is happening, either now or in the future. Finally, the prominence of the event, which denotes the status of an event in a given document. The prominence property can either take the value "main", in which case the event is central to the news being reported upon in the given news article, or "background", in which case it is seen as background information.

Coreference was annotated at two levels. First, coreference chains were established between event arguments in the same document, which is actually an entity coreference annotation step. Second, and more importantly for this study, within-document and cross-document coreference links were drawn between all annotated events. Note here that cross-document annotations were performed only if two events were found in the same topical cluster. For two events to be considered coreferential, three conditions had to be fulfilled: the two events must happen at the same time (1), in the same place (2), and involve the same participants (3).

In total, the corpus comprises 1,018 within-document and 1,578 cross-document event coreference chains (i.e., a coreference chain of size >= 2), making it comparable in size to the large-scale English-language datasets that were discussed in Section 2. Furthermore, each chain contains, on average, seven events and the cross-document chains span an average of 5–6 documents.

## 4. Experiments

In the following sections we will first discuss certain issues plaguing popular modern-day coreference algorithms after which we provide an in-depth description of our proposed end-to-end model architecture and how our proposed changes from earlier coreference models can aid in mitigating some of these issues. Finally, we also provide implementation details related to the training process, data setup, evaluation and hardware used in the experiments.

### 4.1. Rationale

There are two substantial issues with the current generation of Cross-Document Event Coreference Models.

First, an increasing number of recent studies has highlighted the over-reliance of present-day coreference models on lexical similarity (Cattan et al., 2021a; Ahmed et al., 2023). Concretely, it is regularly

**Table 2**
Three randomly drawn clusters from the ENCORE corpus, with their overarching topic name and the number of documents included.

| Cluster id | Topic | # of documents |
|---|---|---|
| 47 | Tim Burton exposition in Genk | 11 |
| 75 | Royal Wedding Prince Harry | 24 |
| 87 | Election of Cuban president | 12 |

the case that mentions with a high degree of (superficial) string similarity are deemed coreferent, while mentions with a comparatively low degree of similarity are often not seen as such. While lexical similarity is indeed a logical predictor for coreference, it is far from the only characteristic of coreferential relations (Ahmed et al., 2023; De Langhe et al., 2023b). The inherent dependence on lexical similarity subsequently leads to most coreference algorithms making consistent and highly predictable mistakes: either similar – but not coreferent – events are seen as coreferent or non-similar – but coreferent – mentions are deemed to be entirely different events. For event coreference, which is a fine-grained task in which attention to nuance and detail is highly important, these consistent errors are problematic, especially since similar – but not coreferent – mentions, such as in examples 6 and 7, are highly prevalent.

6. Pope Francis visits Argentina later this month
7. Pope John Paul II visited Argentina in 1982

Several hypotheses exist with respect to the models' over-reliance on superficial similarity ranging from an inherent lack of lexical diversity in most coreference datasets (Eirew et al., 2021) to issues with the underlying transformer models (De Langhe et al., 2023). Most often, however, the use of the mention-pair paradigm is highlighted as the primary problem. While mention-pair algorithms are by far the most widespread approach for coreference resolution, they reduce the task from a complex interaction between several instances to a simplified binary problem. Several studies have therefore argued that using graphs or trees, rather than pairwise combinations, are a more accurate and less problematic manner of representing coreference chains (Liu et al., 2018).

A second issue is that most coreference models are typically only optimized for within-document coreference resolution as they assume a strict textual ordering of the proposed events. Concretely, a coreferential link can only be proposed and established between two mentions if one is a known antecedent of the other. However, in cross-document settings the order of two events in different documents is ambiguous, meaning that this approach cannot be reliably used. In order to circumvent this ambiguity, earlier studies on cross-document event coreference have opted to compare a candidate mention not only to its document-level antecedents, but to all candidate mentions across all documents (Cattan et al., 2021a,b). While this approach is generally accepted as the most viable and effective manner of performing cross-document event coreference resolution, there exists one significant caveat. The total number of proposed mention-pairs in a document collection for these cross-document models is defined as:

$$\frac{n!}{2!(n-2)!}$$

where $n$ is the number of total candidate events in the collection. This means that both the number of mentions to compute and inference time can increase exponentially over time for large document collections. Given that most CD-ECR datasets are comparatively small, considerations with respect to computational resources are typically limited. Nonetheless, for more practical applications such as the ones discussed in Section 1, coreference resolution would need to be performed over large document collections, potentially comprising thousands of (largely) unstructured texts, making efficiency and resources important factors in the algorithms' design.

### *4.2. Architecture*

In order to tackle these issues we start from a popular joint cross-document model (Cattan et al., 2021a). This approach identifies and links coreferring events across multiple documents by leveraging contextual embeddings obtained through transformer-based encoders. The model uses a two-stage approach: first, it selects all possible spans up to a given token length and predicts potential event mentions using a binary classifier. Then, some pruning measures are applied to the detected events. In this base model, pruning consists of retaining the highest $x$ scoring spans from the mention scorer, where $x$ is a parameter to be manually specified. Then, the model assigns (pairwise) coreferential relations between these mentions. As a postprocessing step, the mention-pairs are clustered using a general hierarchical clustering method, as described in Section 2. We make three significant changes to the overall architecture of this model. First, we supplement the model with information with respect to the proposed events' position in a given document as a manner to reduce the model's over-reliance on lexical matching. Second, we opt for a setting in which the model only performs inference on within-document pairs in order to remove strain on computational resources. Finally, we add a small Variational Graph Auto-Encoder network (VGAE) to the coreference model, which is both tasked with learning graph representations of the event chains (stepping away from the reliance on pairwise computation) and connecting the individual within-document chains in order to build cross-document event graphs. A full visualization of our proposed architecture can be found in Fig. 1. In the sections below, we discuss each of the model's individual components in more detail.

#### *4.2.1. Mention scorer*

Similarly to the standard e2e-coref coreference model (Lee et al., 2017) and the cross-document event coreference model (Cattan et al., 2021a), the initial stages of our proposed approach consist of a mention detection module which scores individual candidate spans. First, all documents in the dataset are encoded in their entirety through the Dutch transformer-based encoder BERTje (de Vries et al., 2019). In case a document exceeds the predefined limit of 512 wordpiece tokens, it is split into several non-overlapping segments instead. Then, all spans up to a length of $\lambda$ wordpiece tokens are sampled from the encoded documents. Spans' starts and ends that do not respect word boundaries are removed from the candidates pool. Additionally, spans exceeding sentence boundaries are also pruned as an event mention will never span multiple sentences.

Following earlier work on mention representation (Lee et al., 2017; Cattan et al., 2021a; Lu and Ng, 2021), span representations $R$ are obtained by concatenating the encoded start token $t_s$, end token $t_e$ and the mean of all individual tokens in the span $t_{mean}$ for each of the candidate spans. In addition to this, we add to the representation vector a domain-specific positional embedding which indicates the span's position in the structure of the document as a whole. While early feature-based work on coreference resolution (cfr. Section 2) often hinted at mention distance as a key predictor for both entity and event coreference, this information is often neglected in transformer-based setups. Nonetheless, the intuition that coreferential mentions are often found in close proximity of each other corresponds well with established general linguistic theories on discourse structure (Hoeken and Van Vliet, 2000; Glasbey, 1994). Given that our dataset, and indeed most data involving practical applications to Cross-Document Event

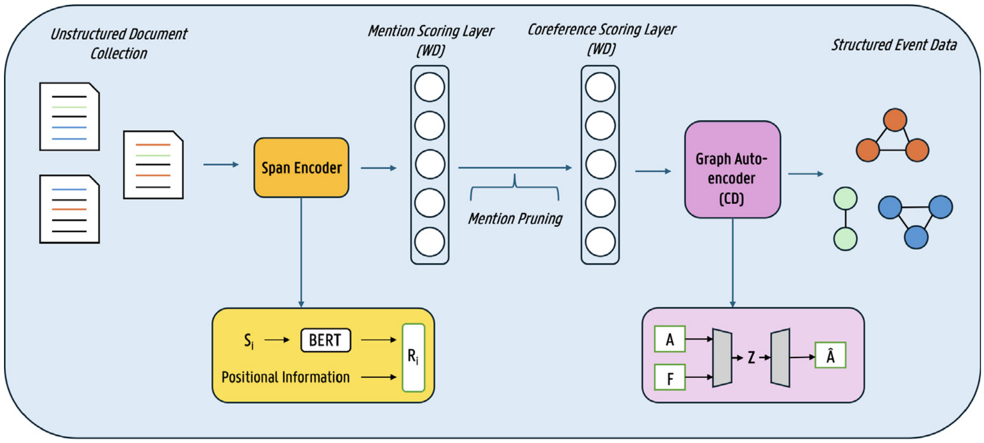



**Fig. 1.** Main architecture of our position-aware end-to-end event coreference model. Each candidate span is passed through an encoder, which combines the contextual embedding and document position of the event into one representation $R_i$. This vector is subsequently passed through a mention scoring layer and only the top $k$ scoring spans for each document are retained. Pairs are then generated and passed through a coreference scoring layer. Finally, the binary output of this scoring layer is passed to a graph reconstruction network in order to obtain the predicted (cross-document) coreference clusters.

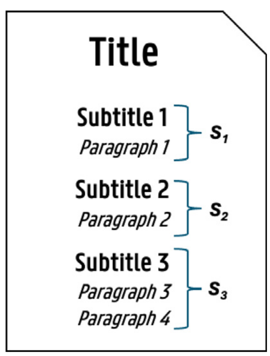



**Fig. 2.** Typical layout of a news document in the ENCORE corpus. For the positional information added to our end-to-end model, we exclusively use the subsection-level position of the event mentions, designated here as $S_1$, $S_2$, and $S_3$.

Coreference Resolution, is composed of newspaper articles, we opt to encode distance between mentions as a domain-specific feature rather than simply adding information with respect to mention or sentence distance between the candidate spans. We assume here that most (online) newspaper articles will have a fairly prototypical structure, as shown in Fig. 2.

Each article naturally has a title and is further composed of different subsections, with the start of a subsection often being signalled with a subtitle. These subsections are further composed of different paragraphs, separated by whitespace. Subsections will typically have a more narrow topical scope when compared to the article as a whole, meaning that it is safe to assume that events that are related, and indeed coreferential, will have a comparatively higher chance of occurring in the same subsection of the document they are in. Given that the subsection level of the candidate mentions may be valuable information, we add a specialized subsection embedding to the mention representation. Unlike previous work, which tied the positional information directly to the input representations of BERT-based models (De Langhe et al., 2023b), we now opt for a model-independent approach to include this kind of information. Our primary motivation for this choice is an effort to future-proof our method and not inherently rely on BERT-based architectures. Our positional embeddings are obtained by initializing a trainable embedding layer of dimension $exS$, where e is the length of the embedding and a trainable parameter and S is the number of subsections in the longest document (on the subsection level) in the data collection. The subsection information $\phi_{Subsection}$ is then concatenated with the aforementioned token representation of the span, resulting in a mention representation vector $m_i$ for each generated candidate span $i$:

$$m_i = [t_s; t_e; t_{mean}; \phi_{Subsection}]$$

Finally, the mention scoring algorithm itself is parameterized by a single-layer feedforward neural network with layer normalization (Ba et al., 2016) and a ReLu activation function, initialized using Glorot initialization (Glorot and Bengio, 2010):

$$MentionScore_i = LayerNorm(ReLU(Wm_i))$$

During training and inference, all candidate spans are passed through the mention scorer, after which a pruning heuristic is applied to obtain the highest scoring candidate mentions. For each document, we only retain the top-$k$ scoring spans, where $k$ is defined as the product of the pruning coefficient $\lambda$ with the number of tokens in a given document.

#### 4.2.2. Pairwise model

The next step in our architecture consists of a pairwise scoring algorithm which takes two candidate spans as input and computes a coreference score for each pair. In contrast to earlier cross-document studies (Cattan et al., 2021a), our pairwise model only scores mentions and their antecedents which occur within the same document in order to limit computational cost. Note here that events belonging to different segments of the same document are still scored together, as they can be logically ordered.

The pairwise scorer, as seen in Fig. 1, is parameterized by a single-layer feedforward neural network identical to the one described in the previous section. This scoring network takes as input a vector which is a concatenation of the two candidate spans' representations and the dot product of those representations:

$$m_{ij} = [m_i; m_j; m_i \cdot m_j]$$

$$CoreferenceScore_{ij} = LayerNorm(ReLU(Wm_{ij}))$$

Finally, for each pair the respective mention scores and coreference score are jointly optimized through a binary cross-entropy loss function:

$$PairScore_{ij} = MentionScore_i + MentionScore_j + CoreferenceScore_{ij}$$

$$L = -\frac{1}{|N|} \sum_{i,j \in N} y \cdot log(PairScore(i,j))$$

For this loss function, $N$ corresponds to all candidate mention-pairs and the class label $y$ reflects the coreference label 0 (no coreferential relation) or 1 (coreferential relation).

#### 4.2.3. Graph reconstruction model

The final step in our proposed approach to CD-ECR consists of learning to convert the binary output of the pairwise scorer network to proper coreference chains and taking this beyond the document level. Most modern-day pairwise coreference models construct coreference chains from binary output by means of an agglomerative clustering system (Barhom et al., 2019). In this setting, clusters are continuously merged based on the average of distances between all event pairs in those clusters. For two coreference clusters A and B, the distance between them is defined as:

$$d_{AB} = \frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l} d(X_i, Y_j)$$

Where $X_i$ and $Y_j$ are objects within clusters A and B, respectively, and $d(.)$ is the distance function. Note that this type of clustering is, at heart, fairly rudimentary. For one, there is no active learning of the structure of the data, despite coreference chains having some interesting learnable properties. Additionally, the end state of the clusters is determined by a, sometimes arbitrarily, chosen similarity threshold value.

We believe that the step of transforming pairwise predictions to coreference clusters is often overlooked in modern literature on coreference resolution and believe that exploring it further can be worthwhile. One interesting property of (identity) coreference chains is that they are transitive, i.e. if mention A and B are corefent and mention B and C are corefent, it automatically follows that mentions A and C are also corefent. With coreference chains having such fundamental and predictable properties, it would be interesting to make an effort towards trying to actively learn such characteristics as a standard part of any coreference model. Additionally, having coreference models automatically learn the optimal clustering without relying on (often) data-specific clustering threshold values may ultimately lead to more generally applicable models.

As a final step in our proposed approach, we thus forgo the use of the popular agglomerative clustering algorithm and instead rely on a small graph neural network to transform the pairwise output of our coreference model. In the past, computational graphs were often used as an intuitive manner of representing coreference chains (Nicolae and Nicolae, 2006). More recently, it has been shown that supplementing existing pairwise models with graph networks can immensely benefit the learning of coreference at the structural level (De Langhe et al., 2023a). Additionally, it has been shown that modelling coreference chains as either graphs or trees results in highly robust models as a whole (Moosavi, 2020).

For our architecture, we base ourselves on earlier work on graph-based coreference, where the output of a pairwise model is directly fed in a small graph-autoencoder (GAE) model (Kipf and Welling, 2016; De Langhe et al., 2024) in order to learn the optimal coreference chain structure from the binary predictions. We selected this approach for event graph modelling primarily due to its computational efficiency. Unlike alternative approaches such as Convolutional Graph Networks (CG) (Zhang et al., 2019) or Graph Attention Networks (CAT) (Veličković et al., 2017), which typically require deeper and more costly (self-attention) layers and more extensive training data to achieve strong performance, GAE relies on a lightweight structure – often just a single neural layer – while still effectively capturing the latent structure of the graph. This makes GAE especially well-suited for scenarios with limited data and processing resources, as in our case. Its efficiency and ability to model relational information with minimal overhead justified its use over more complex alternatives. In contrast to earlier applications of this model in the domain of coreference resolution, we do not consider this graph network as a standalone model or separate post-processing step. Instead, we fully integrate it in our architecture. This results in a streamlined model which optimizes all its constituent networks (event detection, pairwise coreference classification and clustering)as part of a single optimization step at the end. For the integration of the GAE model, we assume that a coreference chain can be modelled by an undirected, unweighted graph $\mathcal{G}(V, E)$ with V nodes, where each node represents an event and each edge $e \in E$ between two nodes denotes a coreferential link between those events. We construct a partially masked adjacency matrix $A$, of dimension $nxn$, and a node-feature matrix $F$, of dimension $fxn$ which are used to predict all original edges in the graph, where $n$ denotes the total number of events in the test set and $f$ denotes feature length. A high-level visualization of this process can be seen in Fig. 1. Each node $E$ (event) is furthermore represented by a feature vector which is a concatenation of its in-context encoded start token $t_s$, end token $t_e$ and the mean of all individual tokens in the span $t_{mean}$:

$$E_i = [t_s; t_e; t_{mean}]$$

The adjacency matrix $A$ is created from the output of the pairwise coreference scorer. In total, 90% of the predicted edges are used for training and 10% are sampled for validation. Then, an equal amount of non-edges is sampled to balance the validation data. Note here that the graph network's function is not to detect new events, but to assign and find additional coreferential links from the already predicted events, as well as learning some fundamental properties (i.e. transitivity) of event coreference chains. Note that adjacency matrix created here is based on the predicted chains from all documents processed earlier in the algorithm and is effectively a event-based cross-document representation of the entire document collection. Additionally, since the graph network is trained on predicted output, it is unavoidable that some faulty coreference links will be used in the training data. However, it has been demonstrated that including (some) faulty predictions in the training does not significantly impact the reconstruction model's performance (De Langhe et al., 2024). We obtain graph representations of the predicted coreference chains by passing the adjacency matrix $A$ and node-feature matrix $X$ through a Graph Convolutional Neural Network (GCN) encoder and then compute the reconstructed matrix $\hat{A}$ from the latent embeddings $Z$:

$$Z = GCN(X, A)$$

$$\hat{A} = \sigma(ZZ^{\tau})$$

The reconstructed adjacency matrix $\hat{A}$ and its resulting coreference graphs then serve as the final output of our end-to-end coreference model.

### 4.3. Experimental setup

#### 4.3.1. Data partition

As stated in Section 2, the ENCORE dataset contains a total of 1,115 documents and is split into 91 topical clusters. Note that for an optimal data partition, the number of events in each document as well as the number of within-document and cross-document event coreferential links in each topical cluster in the training set would ideally need to be in proportion to those in the development and test sets. However, enabling such proportions would lead to an experimental dataset that could be considered 'too artificial'. Therefore, we determined that, in order to create a set of experimental conditions which align best with most practical use cases of CD-ECR, it would be best to base the size of the training, development and test sets on the total number of documents in the ENCORE dataset, rather than on the amount of events/coreference chains in those documents.

In total, we reserved 70%, or around 780 documents for training, and around 15% – or around 167 and 168 documents – for development and testing set respectively. In Table 3 we illustrate the salience with relation to the relevant data points in each partition, we include the average number of events and average number of within-document chains for each document, as well as the average number of cross-document coreference chains per topical cluster.

**Table 3**
The number of topical clusters, average number of event mentions (i.e. gold-standard events) as well as average number of within-document and cross-document (i.e. coreference clusters of size >= 2) for the training, development and test set respectively.

| | Training | Development | Test |
|---|---|---|---|
| Number of Topics | 53 | 18 | 20 |
| Avg. Event Mentions (WD) | 11.2 | 11.8 | 11.5 |
| Avg. Coreference Chains ( WD) | 1.3 | 1.3 | 1.4 |
| Avg. Coreference Chains (CD) | 33.7 | 34.4 | 36.2 |

**Table 4**
Training configuration of both the feedforward networks used for scoring and classifying candidate mentions (left) and of the graph reconstruction neural network (right).

| Hyperparameter | Value |
|---|---|
| *Epochs* | 5 |
| *Batch Size* | 64 |
| *Optimizer* | Adam |
| *Dropout* | 0.3 |
| *Learning Rate* | 1e−4 |
| *Hidden Layer Size* | 1024 |

| Hyperparameter | Value |
|---|---|
| *Epochs* | 200 |
| *Optimizer* | Adam |
| *Learning Rate* | 1e−3 |
| *Hidden Layer Size* | 64 |
| *Latent Layer Size* | 32 |

### *4.3.2. Training process and parameter tuning*

Hyperparameter configurations for the mention span scorer, pairwise scorer and the graph reconstruction model were obtained through grid searching the parameter space and optimizing performance on the development set described in Section 4.3.1. The span scorer and pairwise scorer modules were jointly optimized and the GAE algorithm was trained separately on its own validation set, as described in Section 4.2.

Table 4 contains the optimized parameter values that were used for the final evaluation of the test set for the mention scorer/pairwise model (left) and GAE model (right), respectively. Note that in all experiments the mention scorer and pairwise model were always identical in architecture and training setup.

### *4.3.3. Meaningful comparison*

We compare our position-aware e2e system with five other approaches, which we believe will serve as a series of solid baselines for gauging the capabilities of e2e coreference models in the Dutch language domain. Baseline 1 is the feature-based pipeline approach which combines a CRF-based event detection model with an XGBoost head for coreference classification (De Langhe et al., 2023d). Baseline 2 is a transformer-based version of the same pipeline architecture, which combines syntactic parsing with encoder finetuning. Baseline 3 is the joint model (Cattan et al., 2021a), which was used as a basis for our own position-aware algorithm. Note that training of this model was done according to the specifics given in the original paper, i.e., candidate pairs were generated across documents, rather than within-document. Baselines 4 and 5, based on LLaMA-3.3-70B-instruct and GPT-4o, serve as strong benchmarks for evaluating generative modelling performance in coreference resolution. They follow a two-step prompting strategy: events are first extracted from the entire article, and then all extracted events are combined and resolved into clusters. Based on earlier work, we opt for a few-shot approach in both steps of the task (Min et al., 2024). Concretely, we provide the model with 3 examples in the prompt. The following type of prompt (with only 1 example as an illustration) was used to guide the model through event mention extraction and coreference resolution in thec two-step process:

```
<SYS> You are an expert in information extraction. Given a paragraph of text, you are able to extract and connect crucial events.
<USER> Step 1: Event Mention Extraction
Identify all spans of text that describe events---these are actions, occurrences, or changes of state. Extract each event as a short span of text (usually including the verb and relevant context). Assign each event a unique identifier like E1, E2, etc.
Step 2: Event Coreference Resolution
Look at the list of event mentions. Determine which events refer to the same underlying real-world event, even if they are expressed differently. Group co-referring events into clusters, labelling each cluster with a number.
```

**Output Format:**

```
Text: "The company launched a new product last month. This release was praised by critics. They said the launch was a major success."

Event Mentions:
E1: "launched a new product"
E2: "release"
E3: "launch"
E4: "was praised by critics"
E5: "said"

Coreference Clusters:
Cluster 1: [E1, E2, E3]
Cluster 2: [E4]
Cluster 3: [E5]
```

### *4.3.4. Evaluation*

Coreference resolution is typically evaluated through a series of cluster-based or link-based metrics, as opposed to weighed F1 or accuracy scores. The choice of which method to use is a topic of discussion among researchers investigating coreference resolution, and many different metrics have been proposed throughout the years. Usually, an average F1 is computed from three different metrics (MUC, B3, and CEAF), often referred to as the unified CONLL F1 score (Chang et al., 2012).

Another bone of contention in coreference studies is the inclusion of singleton clusters in the evaluation. Singletons consist of isolated mentions without coreferential links to other mentions in the document collection (i.e., coreference chains with a length of 1). Some research has suggested that the inclusion of these mentions tends to distort certain evaluation methods, leading do hugely inflated scores (Moosavi and Strube, 2016). Other studies have argued that including singleton clusters in the evaluation in of end-to-end event coreference tasks is in fact beneficial, as it provides an implicit manner of evaluating a model's event detection capabilities (Upadhyay et al., 2016).

For our own evaluation procedure, we follow earlier standards set for CD-ECR, which support the inclusion of singleton clusters (Cattan et al., 2021b). In the sections below, we further discuss the advantages and drawbacks for each of the mentioned coreference metrics. Additionally, we also make a case for the inclusion of the more recently proposed LEA evaluation metric within the CONLL framework.

*MUC.* The MUC evaluation metric (Vilain et al., 1995) is one of the earliest proposed methods for reliably evaluating coreference. It is inherently based on the number of missing coreferential links that need to be inserted in the response clusters in order to match the gold-standard coreference clusters. A major benefit of the MUC metric is its robustness with respect to singleton clusters. While singleton distortion is kept to a minimum, the metric does not discriminate between different types of errors and more severe errors are punished as hard as smaller ones (Moosavi and Strube, 2016). This, in turn, can result in scores which do not fully correspond to human judgment and are more difficult to interpret.

*B3.* In contrast to the MUC metric, B3 precision and recall are calculated based on correct entities/events in the response clusters. While the B3 metric addresses several inconsistencies introduced by MUC, it also suffers from singleton distortion due to the entity-based calculation. Additionally, it has been argued that repeated gold entities or events will boost the scores artificially and that this results in problematic cases when dealing with parse trees where an NP node has a single, pronominal, child (Pradhan et al., 2014). In that case, both nodes (i.e., child and NP node) are resolved and result in an inflated score.

*CEAF.* The CEAF coreference evaluation metric (Luo, 2005) is based on finding an optimal alignment between entity/event chains in the system output and corresponding gold-standard chains. Concretely, the Kuhn–Munkres algorithm is used to align chains of coreferring entities in key and response based on a given similarity metric. Being an entity-based evaluation, it suffers from singleton distortion, even more so than the aforementioned B3 metric because in the calculation of the CEAF metric, the total size of the aligned coreference chains is not taken into account. This implies that a correctly predicted singleton cluster contributes as much to the final score as a cluster with a large number of entities (Stoyanov et al., 2009).

*LEA.* More recently, the link-based entity-aware (LEA) metric was proposed (Moosavi and Strube, 2016). Learning from the shortcomings of the previously discussed metrics, LEA computes a relative importance for each of the coreference chains in the output based on the size of the respective chain. Furthermore, each of the response chains is checked for a minimal partial overlap ($> 1$) with the gold-standard chains, reducing the influence of singletons on the overall score. It has to be noted that recent studies suggest that singleton mentions can still have an effect in this metric, although the distortion is significantly smaller than for other metrics (Poot and van Cranenburgh, 2020). We believe that the improvements introduced in the LEA metric are significant and can help in a more thorough understanding and better interpretation of research in coreference resolution. We will therefore include LEA scores in our own evaluation.

#### 4.3.5. Hardware specifications

The end-to-end coreference algorithms were trained and evaluated on 1 T V100-SXM2-16 GB GPU. The graph encoder models were all trained and evaluated on a 16 Intel(R) Xeon(R) Gold 6142 2.60 GHz CPU. For the prompt-based approaches, we utilized the official OpenAI API to access GPT-4o. For LLaMA 3.3, we ran the models on the Vlaams Supercomputer Centre (VSC) Tier-1 cluster, utilizing 4 NVIDIA A100-SXM4 GPUs.

## 5. Results

### 5.1. Quantitative results

#### 5.1.1. Main experiments

In this section we evaluate our proposed approach and compare it to the baseline and SOTA models dicussed in 4.3.3. For completeness, we choose to report both within-document and cross-document results in our paper to provide a comprehensive evaluation of model performance across different contexts.

*Within-document coreference.* Note that each of the algorithms discussed are still trained in cross-document fashion here. In order to obtain within-document scores we applied an additional post-processing step which iterates over each predicted cluster and divides it further into smaller clusters in order to enforce document boundaries. Table 5 below details an evaluation using all metrics discussed in Section 4.3.4.

*Cross-document coreference.* Table 6 then details the results of the end-to-end event coreference experiments. We find that our model outperforms the pipeline-based approaches (Baseline 1 and Baseline 2), without having to rely on an extensive collection of handcrafted features (baseline 1) or an additional syntactic parsing step (baselines 1 and 2). Additionally, our approach shows a moderate improvement over the state-of-the-art cross-document e2e model (Baseline 3) on which its general architecture was based, indicating that relatively small domain-specific adjustments can lead to overall improvements in the model's performance. Finally, while both LLM baselines (4 and 5) outperform the feature-based models, the two end-to-end models (baseline 3 and our own) consistently perform better across all metrics. As a whole, we do note that the overall performance of cross-document event coreference models remains relatively low, especially when compared to state-of-the-art models for entity coreference tasks, which typically attain CONLL F1 scores of around 80% (Kirstain et al., 2021) on the cross-document Ontonotes corpus (Pradhan et al., 2007).

#### 5.1.2. Ablation studies

As detailed in Section 4.2, we made three significant changes to the Baseline 3 model (Cattan et al., 2021a): the inclusion of subsection-level information for the candidate events to make it position-aware (1), a change from cross-document to within-document candidate generation (2) and replacement of the standard agglomerative clustering step by a graph-based model (3). Table 7 displays the model performance and changes in CONLL F1 score ($\delta$) when compared to the performance of our proposed model. We train versions of the algorithm using both within-document and cross-document candidate generation when leaving out position-aware information or the GAE model respectively. For the sake of comparison we also trained and evaluated a model which generated within-document candidates(WD) using neither positional information nor the GAE model. It should be noted that all experiments in this section exclusively report on the evaluation in a cross-document setting.

As can be inferred from the table, including either a small reconstruction model (GAE) (+0.03) or positional event information (+0.03) in the model both yield an increase in performance over the Baseline 3 model, confirming our assertion that supplementing general coreference models with task- and domain-specific position information can be of great benefit. We also note that in the case where no positional information is given and pairs are only generated at the cross-document level, performance is still slightly higher than for the baseline e2e-coref model. This indicates that including a small graph model which learns coreference chain structure can be an attractive alternative to costly candidate pair generation in cross-document applications. Evidently, in the case where pairs are only generated at the within-document level and no positional information or GAE reconstruction algorithm is present, the performance is noticeably worse than that of the baseline algorithm.

**Table 5**
MUC, B3, CEAFe, CONLL and LEA F1 scores for each of the evaluated (within-document) models. Note that CONLL F1 score is the average of the first three scores.

| | MUC | B3 | CEAFe | CONLL | LEA |
|---|---|---|---|---|---|
| *Baseline 1* | 0.28 | 0.21 | 0.17 | 0.22 | 0.28 |
| *Baseline 2* | 0.33 | 0.35 | 0.32 | 0.33 | 0.24 |
| *Baseline 3* | 0.49 | 0.38 | 0.34 | 0.40 | 0.33 |
| *Baseline 4* | 0.31 | 0.32 | 0.29 | 0.30 | 0.21 |
| *Baseline 5* | 0.44 | 0.38 | 0.32 | 0.38 | 0.31 |
| *Our approach* | 0.53 | 0.39 | 0.39 | **0.43** | **0.36** |

**Table 6**
MUC, B3, CEAFe, CONLL and LEA F1 scores for each of the evaluated (cross-document) models. Note that CONLL F1 score is the average of the first three scores.

| | MUC | B3 | CEAFe | CONLL | LEA |
|---|---|---|---|---|---|
| *Baseline 1* | 0.21 | 0.13 | 0.11 | 0.15 | 0.05 |
| *Baseline 2* | 0.30 | 0.25 | 0.29 | 0.28 | 0.12 |
| *Baseline 3* | 0.42 | 0.27 | 0.30 | 0.33 | 0.24 |
| *Baseline 4* | 0.31 | 0.24 | 0.29 | 0.28 | 0.18 |
| *Baseline 5* | 0.39 | 0.25 | 0.27 | 0.30 | 0.20 |
| *Our approach* | 0.44 | 0.34 | 0.33 | **0.37** | **0.31** |

**Table 7**
CONLL F1 scores for the ablation experiments on the ENCORE test set. For the final two models, no position-aware information is passed along and potential candidate pairs are either generated across all documents (CD) or within-document (WD).

| | CONLL F1 | delta |
|---|---|---|
| Baseline 3 | 0.33 | – |
| Our approach | 0.37 | – |
| *- Position-aware information (WD)* | 0.36 | −0.01 |
| *- Position-aware information (CD)* | 0.35 | −0.02 |
| *- GAE model (WD)* | 0.35 | −0.02 |
| *- GAE model (CD)* | 0.34 | −0.03 |
| *- Position-aware information and GAE model (WD)* | 0.31 | −0.06 |

**Table 8**
CONLL F1 scores when including various types of positional information on the mention triggers in the model.

| | CONLL F1 | delta |
|---|---|---|
| Baseline 3 | 0.33 | – |
| Our approach | 0.37 | – |
| *Encoded Mention Position* | 0.31 | −0.06 |
| *Encoded Token Position* | 0.30 | −0.07 |
| *Encoded Sentence Position* | 0.35 | −0.02 |
| *All Positional Information* | 0.33 | −0.04 |
| *All Positional Information + Structural Features* | 0.35 | −0.02 |

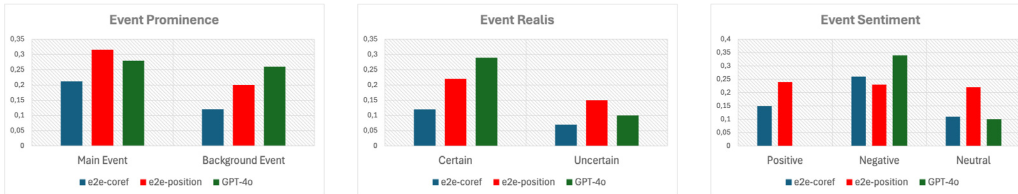
In addition to the ablation experiments mentioned above, we also evaluated a series baseline variants of our proposed model in which we replace the domain-specific positional information with more general distance features often used in earlier research on coreference. Concretely, we explicitly encode mention, token and sentence position for each event mention in a given document in a similar manner as was described in Section 4.2. For completeness, we also include an evaluation of two models in which we first stack the three types of positional information on their own (1) and then combine this stacked information with the domain-specific features of our proposed approach (2). The results of these experiments can be found in Table 8.

As can be deduced from these results, the sole type of positional encoding that demonstrates a positive impact on the model when applied in isolation is sentence-level information. This type of encoding lead to a slight but noticable improvement of 0.02 in the CONLL F1 score, specifically in comparison to Baseline 3. Moreover, it is worth highlighting that while the combination of all forms of positional encodings with our domain-specific structural features consistently results in an overall increase in model performance – also marked by a +0.02 improvement in the CONLL F1 score relative to baseline 3 – the structural information on its own remains the most effective. This indicates that even though additional types of information are integrated, the standalone use of structural information provides the highest gains in performance.

### 5.2. Qualitative results

The interpretation of results in coreference models is hindered by the fact that the most widespread coreference evaluation metric (CONLL) is in itself an average of three other distinct metrics with their individual flaws and advantages. Moreover, it has been argued that relying on a single-value score for evaluation may obscure crucial model flaws (Moosavi and Strube, 2016; Poot and van Cranenburgh, 2020). In addition to this, most end-to-end studies tend to focus solely on the coreference aspect of the task when evaluating model quality, while overlooking the mention detection component. We believe that a more thorough evaluation of the quality of the detected events, as well as the quality of the generated coreference chains can lead to a more complete understanding of the model as a whole.

In the following sections, we aim to provide such a comprehensive and quality-oriented evaluation of our position-aware approach to




(a) (b) (c)

**Fig. 3.** The total percentage of detected gold-standard events based on the Event Prominence (**a**), Event Realis (**b**) and Event Sentiment (**c**) for the two end-to-end joint coreference models.

coreference. First, we perform a set of standard ablation experiments in which we disassemble our position-aware approach. Next, we perform a series of qualitative analyses on the model's output, with a focus on the generated events as well as the quality and completeness of the coreference chains.

*5.2.1. Event mention detection*

The mention detection component of end-to-end models is rarely examined thoroughly. Nonetheless, having insight into which type of events are more likely to be detected (and which ones are not) can be of tremendous value in the development of future coreference algorithms.

As described in Section 3, each news event in ENCORE is labelled with three key properties: *prominence*, *realis* and *sentiment*. For the three best performing models in Section 5 (Baseline 3 (e2e-coref), baseline 5 (GPT-4o) and our own approach (e2e-position)) we display the percentage of detected gold-standard events adhering to each specific property label value in the test set in Fig. 3.

In general, we observe that our position-aware method (red bars) consistently detects more relevant events compared to the baseline e2e model (blue bars), while being slightly outperformed the generative LLM. Regarding Prominence, we observe that the event types which are detected, are largely consistent for the three systems. On average, main events are detected more frequently than background events, possible because they are featured more prominently and more often in a given text. Looking at Event realis, events which have/will happen(ed) with certainty are detected at higher rates than those containing explicit markers of uncertainty in all approaches. In this respect, it should be noted that uncertainty within an event is often expressed in a variety of ways, whereas 'regular' *Certain* events follow the more typical structure of *Actor-Action-Object*. This is illustrated in the events expressed in the sentences 8 (*Certain*) and 9 (*Uncertain*). Consequently, the models are better at learning the fixed structure found in the majority of events, whereas events deviating from that pattern seem to be less detectable.

8. Koning Filip [bezoekt] kerncentrale Doel *EN: King Filip visits Doel powerplant*
9. Als dat tien jaar voor de gevreesde crash [gebeurt] *EN: If that happens 10 years before the feared crash*

When examining the distribution of detected event sentiments, we find that the baseline e2e model detects a disproportionately high amount of negative events, whereas our position-aware approach returns a more balanced spread of events. Interestingly, the generative approach detects positive and negative events at a fairly even rate, while neutral events are detected at much lower rates.

In addition to examining these three key event properties, we also investigate the detection of events for each individual news cluster in the test set, as comparing the performance on some news topics with others may provide additional insight in a model's strengths and weaknesses. From Fig. 4 we learn that for two clusters in particular, namely C37 (*Brussels University guest professor sentenced to death in Iran*) and C87 (*Air Belgium starts with direct flights to Hong Kong*) a relatively high amount of gold-standard events are found with both the baseline and position-aware approach. These outliers can be accounted for by considering that the training set contained several clusters, namely C29 (*American Politics in the Middle East and the Iranian Hostage crisis* and C72 (*Strikes and reforms at Belgian Airports*, which dealt with somewhat similar domains. In our opinion, this highlights interesting future potential in which unrestricted cross-document models can be further enhanced with domain-specific data, depending on the application for which they will be used.

We further remark that our proposed model predicts a large amount of correct events for C22 (*2008 Banking Crisis accountability trials*), where the baseline model does not. One possible explanation for this difference is the length of the documents in C22, which are significantly longer compared to the others in the collection (an average token count of 2,087 in C22 compared to only 1,196 tokens in the rest of the collection). As a result, most documents contained in C22 are divided in and structured through a large amount of subsections. Because subsection information is explicitly passed to our model, this might indicate that our approach, because of its structural awareness, is particularly fit for long-document collections.

For a full legend and topic descriptions, for both the training and test set, we refer the reader to Appendix.

*5.2.2. Lexical diversity*

It is generally assumed that transformer-based encoders capture a significant amount of implicit textual information. However, coreference models employing these encoders tend to greatly suffer from the aforementioned over-reliance on lexical similarity.

Recent work has increasingly pointed out that the classification of coreference in most commonly-used approaches is more akin to string matching, rather than actually discovering the implicit web of textual links which form coreferential chains (Ahmed et al., 2023). A simple, but effective manner of gauging whether or not a model inherently suffers from this problem is to determine string similarity between each predicted candidate pair and to compare it across all (gold) coreferential and not-coreferential pairs.

For this analysis, we propose two methods to compare string similarity between predicted mentions: the normalized Levenshtein distance (Yujian and Bo, 2007) and the cosine distance between spans based on static word2vec embeddings (Mikolov et al., 2013). The Levenshtein distance corresponds to the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other, normalized for string length. Static word2vec embeddings map words to a single vector, which does not change based on the word's sense (as opposed to transformer-based contextual embeddings).

In order to ensure domain compatibility and limit Out-Of-Vocabulary words, we used a set of static embeddings generated from the Dutch SoNaR corpus (Oostdijk et al., 2013a), which contains a large amount of news texts (Tulkens et al., 2016). Span representations were obtained by averaging the word-level vector representations of each word in an event's span. Since both methods are inherently non-contextual, they form a good measuring unit for prediction quality. That is, it would be reasonable to assume that the model suffers from an intrinsic lack of

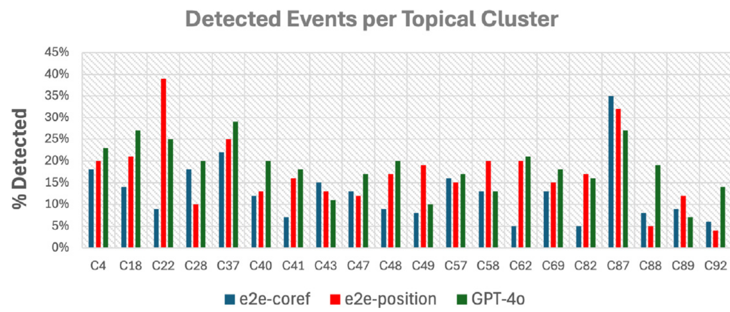
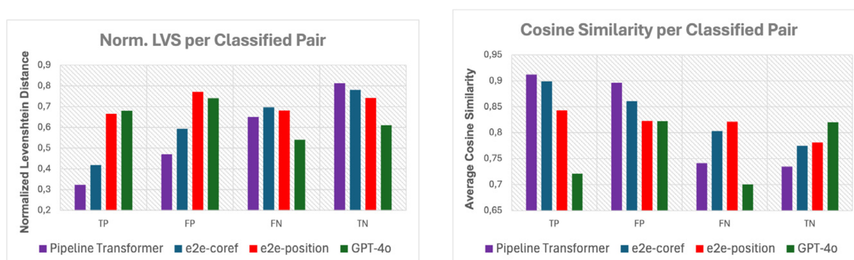



**Fig. 4.** Distribution of correctly predicted events per topical cluster in the ENCORE test set.




(a) (b)

**Fig. 5.** Lexical similarity analysis of the transformer models' output expressed in: (**a**) Normalized Levenshtein Distance and (**b**) Average cosine distances for the static word2vec embeddings.

knowledge w.r.t. coreference if all pairwise predictions are also highly correlated with the value of these superficial similarity metrics.

Fig. 5 plots the average normalized Levenshtein distance and word2vec distance of the models' predictions for each of the transformer-based models (Baseline 2, Baseline 3 and our approach) and the best performing generative model (baseline 5) discussed in Section 5. For each model, we display the True Positive or TP predictions (coreferential mentions classified as such), True Negative or TN predictions (non-coreferential mentions classified as such), False Positive of FP predictions (Non-coreferential mentions classified as coreferential) and False Negative or FN predictions (coreferential mentions classified as non-coreferential). Additionally, we show the average distance for gold-standard coreferential and non-coreferential pairs in the ENCORE test set.

From the Figure, we can infer that the Levenshtein distance and cosine similarity are respectively lower and higher for TP mentions compared to the other classifications in all transformer-based classification models. This supports earlier work (Ahmed et al., 2023) which often argues that less similar, but coreferent, mentions are rarely classified as such. For our position-aware approach, however, the normalized Levenshtein distance is clearly lower and the average cosine similarity markedly higher for TP mentions compared to the earlier transformer pipeline and *e2e-coref* methods. This indicates our approach is not only able to detect more events (see Section 5.2), but that it also manages to consistently classify harder – non-similar – mentions as coreferential.

This is further supported by examining the normalized Levenshtein distance and cosine similarity for erroneously classified coreferential pairs (FP). For these mentions, our approach displays higher distance and lower similarity scores, indicating that our model is better at generalizing that non-similar mentions can also be coreferential. Nonetheless, it is important to emphasize that, although our model demonstrates improved performance with respect to mitigating an overreliance on lexical similarity, the issue still persists and remains one of the primary reasons for the low performances of NLP models often observed in the domain of (cross-document) event coreference resolution.

Finally, consistent with earlier findings, the generative LLM baseline is less reliant on lexical similarity when identifying positive cases, as evidenced by the higher average cosine similarity and LVS scores among its false positives. However, we observe a notable drop in both metrics for false negative classifications, suggesting that the model consistently struggles with harder cases that require more implicit reasoning—an issue hinted at in Section 1.

## 6. Discussion

### 6.1. Generalizability

Ensuring that models perform well across diverse datasets and real-world scenarios is of critical importance in the modern-day NLP landscape. While many models achieve high accuracy on specific benchmark tasks, their utility is often limited by poor performance when applied to new domains, languages, or unseen data. This gap between in-domain success and out-of-domain applicability underscores the need for models that can generalize effectively. In the following sections we discuss potential strengths and pitfalls for both cross-domain and cross-lingual applications of our developed model.

#### 6.1.1. Generalizing to other domains

Given the specific nature of the data worked with in our research (news texts), one might ask if the proposed approach would be applicable to other textual genres as well. After all, the modifications included in the algorithm primarily leverage the distinct structure of newspaper articles to help resolve coreferring mentions. We foresee two potential issues with porting the developed architecture to other domains that we will elaborate on in the following paragraphs.

*Data sparseness.* Newspaper data is by-and-large the most thankful resource for studying event coreference resolution. As was also detailed in Section 2, all corpora dealing with ECR are effectively composed of news data. News articles serve as an excellent reservoir of coreferential relations, owing to their predictable structure of *Title - Lead - Main Text* in which the most important events reported upon are usually mentioned several times. Additionally, cross-document coreference links can easily be found by collecting articles from different sources, reporting on the same central event. Despite this, sparseness remains a persistent issue within the field. At the entity-level, coreference occurs constantly throughout language. This is reflected in the composition of many resources concerned with entity coreference, as they come from a large variety of different domains such as news texts (Oostdijk et al., 2013b), literature (Bamman et al., 2019) or biomedical reports (Pyysalo et al., 2007). For events, coreference occurs much less frequently. From a discourse perspective, this sparseness can be explained by the fact that most of our narratives tend to centre around people or objects going through a sequence of events, rather than repeated mentions of a single event (Choubey and Huang, 2021).

We believe that this issue might be even more prominent for texts outside of the (online) news domain, resulting in very few coreferential links from which a system can learn. While this issue might be (in part) resolved through general sparseness mitigation strategies such as up- or downsampling methods (Singh et al., 2021) or even generating additional artificial in-domain data (He et al., 2022), we believe that it should be taken into serious consideration when applying the methods discussed in this paper to other textual domains.

*Textual structure.* A second possible issue when applying our model to texts outside of the news domain is its inherent reliance on subsection and paragraph-level information.

In general, the use of paragraphs and subsections in long texts is widespread across various domains, as they are essential for structuring information clearly. Paragraphs help organize ideas by focusing on specific points, while subsections provide a hierarchical division that enhances readability and coherence. These structural elements are crucial in academic and technical writing to maintain logical flow and guide the reader through complex arguments (Biber et al., 2021). We therefore foresee relatively little issues when applying our own model on genres and domains that are typically made up of long(er) texts.

However, shorter texts, such as the ones often found in online spaces and on social media platforms, might pose more of an issue. In short-form text organizing content with paragraphs and subsections is more difficult due to the limited space available. Writers must convey key points concisely, leaving little room for the detailed separation of ideas that longer texts allow. Moreover, short texts often require the compression of complex arguments, making it harder to create clear divisions (Gopen, 2004; Williams and Bizup, 2014). Given that our system actively makes use of structural and discourse markers for finding and resolving coreference, the model might struggle with a multitude of short, unstructured documents.

#### 6.1.2. Generalizing to other languages

While our present study is entirely focused on CD-ECR in the Dutch language domain, our general architecture can be easily ported to other languages. In contrast to many earlier studies on coreference resolution, our method is fully language-agnostic both in its general setup and in the features it employs. Prior research, both from a journalistic and NLP perspective, has argued that the general structure of news texts (title-summary-subsections) is near-universal across languages (Thomson et al., 2008; Rupnik et al., 2016). While subtle differences between news reports in various languages can occur, these are primarily semantic and not at the general discourse level (Károly, 2012). We believe that our model can be easily retrained for CD-ECR in other languages by simply swapping the Dutch encoder model (BERTje) for a different (multilingual) encoder.

While there are no (architectural) problems with cross-lingual applications of our model, it is the lack of available resources in event coreference resolution that still poses significant challenges for advancing research in this direction. For now, as mentioned in Section 2 large-scale corpora only exist for English, Chinese, Spanish and Dutch, while other languages are regrettably left behind. The lack of available resources is arguably the primary bottleneck in event coreference resolution (ECR) today, significantly hindering advancements in the field. Unlike entity coreference, which has seen substantial progress due to the availability of large annotated corpora, ECR suffers from a scarcity of high-quality, standardized datasets.

### 6.2. Potential integration in downstream tasks

The development of event coreference systems has always been largely inspired by the idea of leveraging them for information extraction (IE) frameworks. One of the foundational studies in ECR research (Humphreys et al., 1997b) extended the LaSie (Large Scale Information Extraction) system (Gaizauskas et al., 1995) with event-based coreference knowledge in order to further enhance its capabilities with respect to the recognition and extraction of event scenarios (Gaizauskas and Humphreys, 2000). More recently, event coreferential links have been used in biomedical texts to identify implicit and sentence-crossing relations between arguments of said events (Yoshikawa et al., 2011). Additionally, temporal event relations and knowledge on event coreference links have also been used in efforts to extract clinical event timelines from electronic health records (Tourille, 2018).

Beyond general Information Extraction (IE), Event Coreference Resolution (ECR) has been effectively applied to more user-focused tasks, such as Question-Answering (QA). In earlier studies, answers to complex questions were retrieved by identifying entity and event coreference relations between the user's query and a set of documents (Morton, 1999). Further research leveraged temporal relationships between events (including coreference) to enhance QA performance, particularly for complex and nuanced questions in multi-document contexts (Harabagiu and Bejan, 2005).

Recent applications of Event Coreference Resolution (ECR) systems have primarily targeted multi-document scenarios aimed at organizing and extracting information from large volumes of online sources. For instance, coreference links between events have been used in news articles to identify the most prominent events within a text collection (Choubey et al., 2018). Similarly, it has been suggested that ECR links can aid in detecting contradictions across online news sources (Bejan and Harabagiu, 2008).

Many studies on event coreference resolution emphasize its potential applications as a key motivation for exploring these intricate event-event relationships. The prospect of embedding such fine-grained textual relations into tools is undeniably exciting, as successfully resolving these links could transform our comprehension of natural language. However, we must be cautious not to overreach. For ECR systems to be fully integrated into practical applications, the performance gap between ECR systems and those in other NLP fields must first be bridged. Currently, most proposed ECR applications – such as news recommendation, event tracking, and information extraction – are more effectively handled by approaches like coarse-grained document similarity or other dedicated methods, which are both more accurate and memory-efficient than ECR in its current form.

This of course does not mean that the study of event coreference is not worthwhile or less interesting. As mentioned multiple times throughout this paper, discourse relations, including (co)reference, are at the very heart of communication and gaining insight in them can reveal many things about natural language understanding in both humans and machines. However, when it comes to using ECR as a practical tool, there is a lot of catching up to do with respect to other NLP disciplines.

## 7. Conclusion and future work

In this paper, we proposed a novel model architecture for end-to-end cross-document event coreference resolution. We drew inspiration from earlier feature-based studies on coreference resolution and integrated subsection-level structural knowledge in a transformer-based coreference algorithm. In addition, we proposed to replace the standard agglomerative clustering algorithm typically found as the final component of a mention-pair coreference model by a small graph-based network which actively learns the general structure and properties of coreference chains beyond the document level.

Our experiments show that our position-aware approach scores higher when evaluated on a large cross-document coreference corpus when compared with three competitive baseline algorithms. Moreover, in a series of ablation experiments we revealed that each of the proposed architectural changes contributes equally to said performance increase. The last sections of this paper were dedicated to an extensive qualitative analysis of the model's output. With respect to event detection, we show that our model can consistently detect more correct events in the raw input data. Additionally, our results lead us to believe that the inclusion of subsection-level knowledge results in a more effective processing of longer documents. As a second part of our analysis, we showed that our proposed model suffers less from the typical issues mention-pair models exhibit when predicting coreference chains: our model was able to consistently classify harder (non-similar) event pairs and was able to generalize certain structural properties of coreference better than the other models under scrutiny.

Overall, we believe that directly encoding task-specific information (such as subsection-level information) in transformer-based models holds interesting potential, not only for coreference resolution, but for the entire field of cross-document processing. Even though earlier feature-based studies heavily emphasized the importance of this type of knowledge, the inclusion of it has largely been lost in more recent work. This can be explained by the intensive and therefore impractical aspect of extracting task-specific information in large document collections. However, we have shown that including even fairly rudimentary document-level information such as the subsection-level of a candidate mention, which can be easily obtained by simply detecting subsection titles or whitespaces, already has a noticeable effect on model performance. This comes with the added benefit that the graph reconstruction is fast and can be trained on CPU (as mentioned earlier), unlike larger models today, making the approach both effective and computationally accessible.

In addition, the success of combining the – almost universally adopted – mention-pair approach to coreference with a more structurally informed graph-based model, leads us to believe that this combination of the strengths of both methods is a viable strategy for moving forward.

Finally, we wish to emphasize the inherent value of qualitative model analysis. Performance improvements in coreference studies are often somewhat obscured by the use of different evaluation metrics, making it hard to pinpoint why one algorithm scores better than another. Through a direct analysis of the predicted output, we were able to show concrete areas in which our model improved upon the current state of the art.

In future work, we aim to further infuse end-to-end coreference models with easily obtainable document and discourse-level information in order to improve upon the proposed architecture. In addition, we will explore the use of other coreference chain representations (e.g. trees) and use them to extend current mention-pair methods. We also envisage to fully integrate CD-ECR in a series of downstream tasks such as summarization, news recommendation and contradiction detection and perform extrinsic evaluation through these tasks in an effort to verify whether or not the current generation of CD-ECR models can make a meaningful contribution to these practical and commercial applications.

## CRediT authorship contribution statement

**Loic De Langhe:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Orphée De Clercq:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Veronique Hoste:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Loic De Langhe reports financial support was provided by Research Foundation Flanders. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Research Foundation–Flanders under project grant number FWO.OPR.2020.0014.01.

## Appendix. Example appendix section

1. London Terror Attack 29/11/2019
2. Trial of Fouad Belckacem
3. Bribery of Belgian Football Referees
4. Health Inspection Closes Slaughterhouse in Geel
5. Soyuz Emergence Landing
6. Earthquake Lombok
7. Death of actor Frans Aendenboom
8. Herman De Conickprijs
9. Death of Refugee in High speed chase
10. Plane crash Indonesia
11. Controversy on Meghan Markle's father
12. Hairdresser discovers his lover is a priest
13. Nobel Prize in literature schandal
14. Disappearance MH370
15. Release of Belgian movie Kursk
16. Migrant crisis at Brussel Noord
17. Ensor awards
18. Studio 100 sues Gaia for images of smoking Maya de Bij
19. First politician with Down Syndrome in Belgium
20. Kevin Spacey allegations and aftermath
21. IVAGO strikes
22. FORTIS trials
23. Purchase of F-35 planes by government
24. Volcanic eruption in Guatemala
25. Mia Doornaert selected as VAF chairwoman
26. Italian elections
27. Meeting Trump and Kim Jong-Un
28. Migrant crisis Guatemala
29. American Middle East policies
30. Alex Jones Banned from Twitter
31. Guy Van Sande pedophilia allegations
32. Release of Belgian movie Girl
33. Future plans Hedwigepolder
34. Elections European Commission
35. Death of Taliban leader
36. Nomination of Cavanaugh for Supreme Court US
37. Trial VUB Professor in Iran
38. ABBA holograms
39. New production line Volvo

40. Violence in Gaza
41. Stormy Daniels Confession
42. Tim Burton exposition Genk
43. Berlin forest fires
44. Energy policy Belgian federal government
45. Life on Mars
46. Sexual harassment lawsuits at Google

47. Terrorist attack Luik
48. Death of Renate Dorenstein
49. Elephants in Planckedael zoo
50. Pussy Riot disrupts FIFA world cup final
51. RyanAir strikes
52. Brabant killers investigations
53. Carrefour Strikes
54. Sexual harrasment allegations Asia Argento
55. Trump Russsia collusion
56. Lauryn Hill concerts Brussels
57. Camping Kitsch 2018
58. Riots in Anderlecht
59. Zimbabwe elections
60. Dismissal of Roseanne Barr
61. Bpost financial reports
62. Publication Book Trump
63. Eurostadium building plans
64. Release of Belgian movie 'Niet Schieten'
65. death of Jean-Pierre Van Rossem
66. Royal Wedding
67. Star Wars exposition in Brussels
68. Greek debt crisis
69. New Lars Von Trier movie
70. Venezuala migrant crisis
71. David Goffin at Roland Garros
72. Dismissal Brussels Airlines CEO

73. Ghent municipal council issues
74. Debate on return of IS wives and children
75. Discovery of the Bennu astroid
76. Controversy at Belgian Art Prize
77. Brazilian presedential elections
78. Starvation crisis in Yemen
79. Migrant crisis Kruibeke
80. American midterm elections
81. Antisemetic murder in France
82. New Cuban president
83. Hungary political reforms
84. Trump visits Iraq
85. Fraud in Belgian meat industry
86. Wolves killing livestock in Flanders
87. Air Belgium long distance flights
88. Camebridge Analytica schandal
89. Second meeting Trump and Kim Jong-Un
90. Turkey elections
91. Trump aide expelled from restaurant
92. Publication book Michelle Obama

## References


Ahmed, S.R., Nath, A., Martin, J.H., Krishnaswamy, N., 2023. $2 * n$ is better than $n^2$: Decomposing event coreference resolution into two tractable problems. arXiv preprint arXiv:2305.05672.

Alsaedi, N., Burnap, P., Rana, O., 2016. Automatic summarization of real world events using twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 10, (1), pp. 511–514.

Anon, 2008. ACE English Annotation Guidelines for Events (v5.4.3). Linguistics Data Consortium.

Araki, J., Mitamura, T., 2015. Joint event trigger identification and event coreference resolution with structured perceptron. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp. 2074–2080. http://dx.doi.org/10.18653/v1/D15-1247, URL http://aclweb.org/anthology/D15-1247.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bamman, D., Lewke, O., Mansoor, A., 2019. An annotated dataset of coreference in english literature. arXiv preprint arXiv:1912.01140.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., Dagan, I., 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. arXiv preprint arXiv:1906.01753.

Bejan, C.A., Harabagiu, S.M., 2008. A linguistic resource for discovering event structures and resolving event coreference. In: LREC.

Bejan, C., Harabagiu, S., 2010. Unsupervised event coreference resolution with rich linguistic features. In: Proceedings of the 48th Annual Meeting ofthe Association for Computational Linguistics. (July), pp. 1412–1422. http://dx.doi.org/10.1162/COLI_a_00174, URL http://www.aclweb.org/anthology/P10-1143.

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Biber, D., Johansson, S., Leech, G.N., Conrad, S., Finegan, E., 2021. Grammar of Spoken and Written English. John Benjamins.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., Dagan, I., 2021a. Cross-document coreference resolution over predicted mentions. arXiv preprint arXiv:2106.01210.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., Dagan, I., 2021b. Realistic evaluation principles for cross-document coreference resolution. arXiv preprint arXiv:2106.04192.

Chang, K.-W., Samdani, R., Rozovskaya, A., Sammons, M., Roth, D., 2012. Illinois-coref: The UI system in the conll-2012 shared task. In: Joint Conference on EMNLP and CoNLL-Shared Task. pp. 113–117.

Chen, C., Ng, V., 2014. SinoCoreferencer : An end-to-end Chinese event coreference resolver. In: Lrec 2014. pp. 4532–4538.

Chen, C., Ng, V., 2016. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. February 12-17, 2016, Phoenix, Arizona, USA, pp. 2913–2920, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12413.

Chen, X., Xu, S., Li, P., Zhu, Q., 2023. Cross-document event coreference resolution on discourse structure. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4833–4843.

Chen, Z., Yu, H., Li, J., Luo, X., 2022. Entity representation by neighboring relations topology for inductive relation prediction. In: Pacific Rim International Conference on Artificial Intelligence. Springer, pp. 59–72.

Choubey, P.K., Huang, R., 2021. Automatic data acquisition for event coreference resolution. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1185–1196.

Choubey, P.K., Raju, K., Huang, R., 2018. Identifying the most dominant event in a news article by mining event coreference relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 340–345.

Clark, K., Manning, C.D., 2015. Entity-centric coreference resolution with model stacking. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1405–1415.

Cybulska, A., Vossen, P., 2014a. Guidelines for ECB+ Annotation of Events and their Coreference. Tech. Rep. NWR-2014-1, VU University Amsterdam.

Cybulska, A., Vossen, P., 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In: Proceedings of the Ninth InternationalConference on Language Resources and Evaluation. LREC'14, pp. 4545–4552.

Cybulska, A., Vossen, P., 2015. Translating granularity of event slots into features for event coreference resolution. In: Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Association for Computational Linguistics, Denver, Colorado, pp. 1–10. http://dx.doi.org/10.3115/v1/W15-0801, URL http://aclweb.org/anthology/W15-0801.

De Langhe, L., De Clercq, O., Hoste, V., 2021. Guidelines for Annotating Events and Event Coreference in Dutch News Articles. Tech. Rep..

De Langhe, L., De Clercq, O., Hoste, V., 2022a. Constructing a cross-document event coreference corpus for Dutch. Lang. Resour. Eval. 1–30.

De Langhe, L., De Clercq, O., Hoste, V., 2022b. Investigating cross-document event coreference for Dutch.

De Langhe, L., De Clercq, O., Hoste, V., 2023. What does BERT actually learn about event coreference? Probing structural information in a fine-tuned dutch language model. In: Proceedings of the Fourth Workshop on Insights from Negative Results in NLP. pp. 103–108.

De Langhe, L., De Clercq, O., Hoste, V., 2023a. Filling in the gaps: Efficient event coreference resolution using graph autoencoder networks. In: Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference. CRAC 2023, pp. 1–7.

De Langhe, L., De Clercq, O., Hoste, V., 2023b. Leveraging structural discourse information for event coreference resolution in Dutch. In: Proceedings of the 4th Workshop on Computational Approaches to Discourse. CODI 2023, pp. 48–53.
De Langhe, L., De Clercq, O., Hoste, V., 2024. Enhancing unrestricted cross-document event coreference with graph reconstruction networks. In: Proceedings International Joint Conference on Computational Linguistics, Language Resources and Evaluation.
De Langhe, L., Desot, T., De Clercq, O., Hoste, V., 2023d. A benchmark for dutch end-to-end cross-document event coreference resolution. Electronics 12 (4), 850.
De Marneffe, M.-C., Rafferty, A.N., Manning, C.D., 2008. Finding contradictions in text. In: Proceedings of ACL-08: HLT. pp. 1039–1047.
de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M., 2019. Bertje: A Dutch bert model. arXiv preprint arXiv:1912.09582.
Eirew, A., Cattan, A., Dagan, I., 2021. WEC: Deriving a large-scale cross-document event coreference dataset from wikipedia. arXiv preprint arXiv:2104.05022.
Elov, B., Khamroeva, S.M., Xusainova, Z., 2023. The pipeline processing of NLP. In: E3S Web of Conferences, vol. 413, EDP Sciences, p. 03011.
Fan, C., Li, J., Luo, X., Xu, R., 2022. Enhancing structure preservation in coreference resolution by constrained graph encoding. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 2557–2567.
Gaizauskas, R., Humphreys, K., 2000. Quantitative evaluation of coreference algorithms in an information extraction system. pp. 145–169, Corpus-based and computational approaches to discourse anaphora.
Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., Wilks, Y., 1995. University of Sheffield: Description of the LaSIE system as used for MUC-6. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.
Glasbey, S.R., 1994. Event Structure in Natural Language Discourse (Ph.D. thesis). University of Edinburgh.
Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.
Gopen, G.D., 2004. The sense of structure: writing from the reader's perspective.
Gordon, P.C., Hendrick, R., 1998. The representation and processing of coreference in discourse. Cogn. Sci. 22 (4), 389–424.
Harabagiu, S., Bejan, C.A., 2005. Question answering based on temporal inference. In: Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering. pp. 27–34.
He, X., Nassar, I., Kiros, J., Haffari, G., Norouzi, M., 2022. Generate, annotate, and learn: NLP with synthetic text. Trans. Assoc. Comput. Linguist. 10, 826–842.
Hoeken, H., Van Vliet, M., 2000. Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. Poetics 27 (4), 277–286.
Huang, C., Xu, S., He, L., Li, P., Zhu, Q., 2022. Incorporating generation method and discourse structure to event coreference resolution. In: International Conference on Neural Information Processing. Springer, pp. 73–84.
Humphreys, K., Gaizauskas, R., Azzam, S., 1997a. Event coreference for information extraction. In: Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robus Anaphora Resolution for Unrestricted Texts. pp. 75–81.
Humphreys, K., Gaizauskas, R., Azzam, S., 1997b. Event coreference for information extraction. In: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.
Ji, H., Grishman, R., 2011. Knowledge base population: Successful approaches and challenges. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1148–1158.
Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., Qiu, L., 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. arXiv preprint arXiv:2310.06839.
Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O., 2020. Spanbert: Improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. 8, 64–77.
Károly, K., 2012. News discourse in translation: Topical structure and news content in the analytical news article. Meta 57 (4), 884–908.
Kingsbury, P.R., Palmer, M., 2002. From treebank to propbank. In: LREC. pp. 1989–1993.
Kingsbury, P., Palmer, M., 2003. Propbank: the next level of treebank. In: Proceedings of Treebanks and Lexical Theories, vol. 3, Citeseer.
Kipf, T.N., Welling, M., 2016. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308.
Kirstain, Y., Ram, O., Levy, O., 2021. Coreference resolution without span representations. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 14–19.
Kobayashi, H., Hou, Y., Ng, V., 2023. Pairspanbert: An enhanced language model for bridging resolution. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6931–6946.
Kompan, M., Bieliková, M., 2010. Content-based news recommendation. Lect. Notes Bus. Inf. Process. 61 LNBIP, 61–72. http://dx.doi.org/10.1007/978-3-642-15208-5_6.
Koroteev, M., 2021. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.
Lassalle, E., 2015. Structured Learning with Latent Trees: A Joint Approach to Coreference Resolution (Ph.D. thesis). Univeristé Paris Diderot Paris 7.
Lee, K., He, L., Lewis, M., Zettlemoyer, L., 2017. End-to-end neural coreference resolution. arXiv:1707.07045[cs].
Liu, R., Mao, R., Luu, A.T., Cambria, E., 2023. A brief survey on recent advances in coreference resolution. Artif. Intell. Rev. 1–43.
Liu, Z., Mitamura, T., Hovy, E., 2018. Graph based decoding for event sequencing and coreference resolution. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3645–3657, URL http://www.aclweb.org/anthology/C18-1309.
Liu, Y., Peng, X., Cao, J., Bo, S., Shen, Y., Du, T., Cheng, S., Wang, X., Yin, J., Zhang, X., 2024. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. arXiv preprint arXiv:2410.01671.
Lu, Y., Lin, H., Tang, J., Han, X., Sun, L., 2022. End-to-end neural event coreference resolution. Artificial Intelligence 303, 103632.
Lu, J., Ng, V., 2018a. Event coreference resolution: A survey of two decades of research. In: IJCAI. pp. 5479–5486.
Lu, J., Ng, V., 2018b. Event coreference resolution: A survey of two decades of research. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, pp. 5479–5486. http://dx.doi.org/10.24963/ijcai.2018/773, URL https://www.ijcai.org/proceedings/2018/773.
Lu, J., Ng, V., 2021. Conundrums in event coreference resolution: Making sense of the state of the art. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 1368–1380.
Lu, J., Venugopal, D., Gogate, V., Ng, V., 2016. Joint inference for event coreference resolution. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3264–3275.
Luo, X., 2005. On coreference resolution performance metrics. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 25–32.
Maekawa, A., Hirao, T., Kamigaito, H., Okumura, M., 2024. Can we obtain significant success in RST discourse parsing by using large language models? In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2803–2815.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 26.
Min, Q., Guo, Q., Hu, X., Huang, S., Zhang, Z., Zhang, Y., 2024. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. arXiv preprint arXiv:2406.02148.
Minard, A.-L., Speranza, M., Urizar, R., van Erp, M., Schoen, A., van Son, C., 2016. Meantime, the NewsReader multilingual event and time corpus. In: Proceedings of the 10th Language Resources and Evaluation Conference. LREC 2016, European Language Resources Association (ELRA), Portorož, Slovenia, p. 6.
Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., Strassel, S., 2015. Event nugget annotation: Processes and issues. In: Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Association for Computational Linguistics, Denver, Colorado, pp. 66–76. http://dx.doi.org/10.3115/v1/W15-0809, URL http://aclweb.org/anthology/W15-0809.
Moosavi, N.S., 2020. Robustness in Coreference Resolution (Ph.D. thesis).
Moosavi, N.S., 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.
Morton, T.S., 1999. Using coreference for question answering. In: Coreference and Its Applications.
Nair, I., Somasundaram, S., Saxena, A., Goswami, K., 2023. Drilling down into the discourse structure with LLMs for long document question answering. In: Findings of the Association for Computational Linguistics. EMNLP 2023, pp. 14593–14606.
Narayanan, S., Harabagiu, S., 2004. Question Answering Based on Semantic Structures. Tech. Rep., International Computer Science Inst Berkeley CA.
Nath, A., Manafi, S., Chelle, A., Krishnaswamy, N., 2024. Okay, let's do this! Modeling event coreference with generated rationales and knowledge distillation. arXiv e-prints arXiv–2404.
Nguyen, T.H., Meyers, A., Grishman, R., 2016. New york university 2016 system for KBP event nugget: A deep learning approach. In: Text Analysis Conference. p. 7.
Nicolae, C., Nicolae, G., 2006. Bestcut: A graph algorithm for coreference resolution. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 275–283.
Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I., 2013a. The construction of a 500-million-word reference corpus of contemporary written Dutch. pp. 219–247. http://dx.doi.org/10.1007/978-3-642-30910-6_13.
Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I., 2013b. The construction of a 500-million-word reference corpus of contemporary written Dutch. Essent. Speech Lang. Technol. Dutch: Results By the STEVIN Program. 219–247.
Poot, C., van Cranenburgh, A., 2020. A benchmark of rule-based and neural coreference resolution in dutch novels and news. arXiv preprint arXiv:2011.01615.

Popel, M., Žabokrtskỳ, Z., Nedoluzhko, A., Novák, M., Zeman, D., 2021. Do UD trees match mention spans in coreference annotations? In: Findings of the Association for Computational Linguistics. EMNLP 2021, pp. 3570–3576.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., 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, p. 30.

Pradhan, S.S., Ramshaw, L., Weischedel, R., MacBride, J., Micciulla, L., 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In: ICSC 2007 International Conference on Semantic Computing. pp. 446–453. http://dx.doi.org/10.1109/ICSC.2007.93.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T., 2007. Bioinfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 8, 1–24.

Rahman, A., Ng, V., 2009. Supervised models for coreference resolution. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 968–977.

Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M., 2016. News across languages-cross-lingual document similarity and event tracking. J. Artificial Intelligence Res. 55, 283–316.

Singh, S., Bhattacharjee, K., Darbari, H., Verma, S., 2019. Analyzing coreference tools for NLP application. Int. J. Comput. Sci. Eng. 7 (5), 608–615.

Singh, B., Desai, R., Ashar, H., Tank, P., Katre, N., 2021. A trade-off between ML and DL techniques in natural language processing. In: Journal of Physics: Conference Series, vol. 1831, (1), IOP Publishing, 012025.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X., 2015. From light to rich ERE: Annotation of entities, relations, and events. In: Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015. ACL, Denver, Colorado, pp. 89–98.

Soon, W.M., Ng, H.T., Lim, D.C.Y., 2001. A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. 27 (4), 521–544.

Stoyanov, V., Gilbert, N., Cardie, C., Riloff, E., 2009. Conundrums in noun phrase coreference resolution : Making sense of the state-of-the-art. Differences (August), 656–664. http://dx.doi.org/10.3115/1690219.1690238, URL http://www.aclweb.org/anthology/P/P09/P09-1074.

Taylor, A., Marcus, M., Santorini, B., 2003. The penn treebank: an overview. Treebanks 5–22.

Tenney, I., Das, D., Pavlick, E., 2019. BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.

Thomson, E.A., White, P.R., Kitley, P., 2008. "Objectivity" and "hard news" reporting across cultures: Comparing the news report in English, French, Japanese and Indonesian journalism. Journal. Stud. 9 (2), 212–228.

Tourille, J., 2018. Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records (Ph.D. thesis). Université Paris Saclay (COmUE).

Tran, H.M., Phung, D., Nguyen, T.H., 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4840–4850.

Tulkens, S., Emmery, C., Daelemans, W., 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation. LREC 2016/Calzolari, Nicoletta [Edit.]; Et Al., pp. 1–7.

Upadhyay, S., Gupta, N., Christodoulopoulos, C., Roth, D., 2016. Revisiting the evaluation for cross document event coreference. In: COLING. URL http://cogcomp.org/papers/revisiting-evaluation-cross.pdf.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

Vilain, M., Burger, J.D., Aberdeen, J., Connolly, D., Hirschman, L., 1995. A model-theoretic coreference scoring scheme. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.

Williams, J.M., Bizup, J., 2014. Lessons in Clarity and Grace. Pearson.

Wu, X., Huang, K.-H., Fung, Y., Ji, H., 2022. Cross-document misinformation detection based on event graph reasoning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 543–558.

Xu, S., Li, P., Zhu, Q., 2023. Corefprompt: Prompt-based event coreference resolution by measuring event type and argument compatibilities. arXiv preprint arXiv:2310.14512.

Yoshikawa, K., Riedel, S., Hirao, T., Asahara, M., Matsumoto, Y., 2011. Coreference based event-argument relation extraction on biomedical text. J. Biomed. Semant. 2, 1–14.

Yujian, L., Bo, L., 2007. A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 1091–1095.

Zhang, H., Ke, W., Zhang, J., Luo, Z., Ma, H., Luan, Z., Wang, P., 2023. Prompt-based event relation identification with constrained prefix attention mechanism. Knowl.-Based Syst. 111072.

Zhang, S., Tong, H., Xu, J., Maciejewski, R., 2019. Graph convolutional networks: a comprehensive review. Comput. Soc. Networks 6 (1), 1–23.

Zhu, Y., Peng, S., Pradhan, S., Zeldes, A., 2024. SPLICE: A singleton-enhanced pipeline for coreference resolution. arXiv preprint arXiv:2403.17245.