Towards Fine(r)-grained Identification of Event Coreference Resolution Types

Loic De Langhe* LOIC.DELANGHE@UGENT.BE
Orphée De Clercq* ORPHÉE.DECLERCQ@UGENT.BE
Veronique Hoste* VERONIQUE.HOSTE@UGENT.BE

**LT3 - Language and Translation Technology Team, Groot-Brittaniëlaan 45 9000 Ghent, Belgium*

Abstract

In this paper we present initial efforts to study complex event-event relations or event coreference in the Dutch language. We are primarily interested in the event-subevent relations between event pairs, in which one event is part of another (larger) encompassing event. We detail how event coreference is defined and annotated in the Dutch ENCORE corpus, after which the corpus is used as training data. Two experiments are conducted in order to gauge the possibility of integrating event-event relationships in ongoing research on Dutch event coreference resolution. The first experiment consists in classifying the nature of the coreferential relations between two gold-standard events. This task is used as a stepping stone for the second experiment, in which we attempt to predict whether pairs of textual events corefer and, if so, what the nature of this coreferential relation is. Our baseline experiments consist of fine-tuning various transformer language models, after which model ensembles are created to gauge the combined performance. Initially, the best results were achieved with these ensembles. However, in a second step, we also applied self-ensembling and self-distillation techniques to improve the fine-tuning process of the existing monolingual language models. Here we demonstrated that adding a warmup parameter in the self-ensembling process and a temperature in the self-distillation algorithm can have a noticeable effect on model performance, leading to on par or better performance than the ensembles.

1. Introduction

The ability to analyze how written language is structured and how it refers or relates to entities, happenings and situations outside of the text is of paramount importance to the human capacity of natural language understanding. Knowing this, it is no surprise that research in the field of Natural Language Processing (NLP) has since long been trying to mimic this skill at the computational level. The inter- and intra-textual relationships, be it at the word, paragraph or discourse level seem to be key to a breakthrough in the true algorithmic understanding of natural language.

Currently, these efforts are still mostly guided by modelling superficial lexical (Reiter et al. 2014) and contextual similarity (Benedetti et al. 2019), rather than by investigating the internal discourse and narrative structure of text. Nonetheless, in recent years the direction of research has been steadily shifting towards the latter and a hoist of practical applications that would immensely benefit from more fine-grained structural text analysis have been proposed. These include, but are not limited to, the automatic construction of narrative arcs in literary texts (Bhyravajjula et al. 2022), automatic summarization (Huang 2021), monitoring developing crisis events through social media coverage (Girish et al. 2022) and timeline generation (Huang et al. 2013).

Earlier work on fine-grained textual analysis was conceptualized by breaking down larger texts and perform analyses on smaller, more contained, text parts (Mitra et al. 1997), whereas at present an event-centric approach is gaining traction. In this latter view, texts are seen as a sequence or collection of textual events rather than separated blocks of discourse. These events can either refer to real-world occurrences or fictional happenings and are uniquely defined by the time and place at which they occur (Pustejovsky et al. 2003). Events are furthermore defined by a set of event characteristics or arguments. These arguments can refer to the action, spatio-temporal aspects and participants of the happening. Note that event arguments are not always explicitly lexicalised in the text as sometimes these have to be inferred from context. The two examples below represent a real-world and fictional event, respectively, in which the arguments have also been annotated.

1. [[Het vliegtuig van vlucht MH17]^{Participant} werd [op 17 juli 2014]^{Time} boven [Oost-Oekraïne]^{Location} uit de lucht [geschoten]^{Action} door [een Buk-raket, een wapen van Russische makelij]^{Participant}]^{Event} *EN: The airplane of flight MH17 was shot down on July 17th 2014 above eastern Ukraine by a Russian-made BUK-missile.*

2. Ze vertelde de vogelverschrikker alles over Kansas en hoe][de cycloon]^{Participant} [haar]^{Participant} naar [het vreemde land van Oz]^{Location} had [gebracht]^{Action}]^{Event}. *EN: She told the scarecrow all about Kansas and how the cyclone had carried her to the strange land of Oz.*

From a discourse analysis perspective, events form an interesting area of study. More specifically, one can consider the relations between events in a given text and how these relations contribute to its narrative structure or form. In this respect the analysis of coreferential relations in particular can provide valuable insights. We distinguish two possible coreferential event relations. First, there is the scenario where two events can fully refer to the same real-world or fictional event. This is the case when the temporal and geographical information match and when the same participants partake in the event. We denote this relationship between two events as an *Identity* relation (Example 3). Additionally, we also consider events that refer only in part to one another, for example when one of the events is fully contained by the other (Example 4), this is known as a part-whole relation and in the framework of events this can also be referred to as an event-subevent relation.

3. [De wedstrijd]^{Event} eindigde uiteindelijk in een 0-0 gelijkspel. Na [de match]^{Event} gaf de coach zijn spelers een uitbrander van jewelste. *EN: The game eventually ended in a 0-0 draw. After the match the coach spoke firmly to his players.*

4. [De politieke crisis]^{Event} bereikte zijn hoogtepunt toen [de premier zijn ontslag indiende]^{Event}. *EN: The political crisis reached its summit when the prime minister resigned.*

In earlier studies, subevent detection has often been perceived as a separate task (Pohl et al. 2012), where the goal was to detect subevent(s) for one previously specified key event or as part of a general event extraction task (Araki 2018). We believe, however, that this task can be tackled more adequately through the broader scope of event coreference resolution (ECR), in which the aim is to correctly couple textual events that refer to the same real-world or fictional event (Lu and Ng 2018a)), for two main reasons. First, integrating subevent detection into ECR enables a more complex textual analysis; when multiple key events are present in a text, the ability to also detect relations between these adds an additional layer of complexity to the analysis. Second, by considering the possibility of coreference resolution in a cross-document setting – which means relations can be defined between events spanning multiple documents – this enables the extraction and categorisation of large-scale information networks. In such networks events can be seen as nodes whose edges denote the various relations between events. These can then be used to construct elaborate knowledge bases, which in turn can be used for the development of promising NLP applications such as timeline generation (Bedi et al. 2017) and detailed automatic summarization (Altmami and Menai 2020).

The main goal of this paper is two-fold. First, we argue that the inclusion of event-subevent relations in the broader task of event coreference resolution is both logical from a theoretical perspective and can provide extra insights into discourse analysis as a whole. Second, we also aim to establish, for the first time in the Dutch language domain, baseline scores for the task of cross-document subevent detection and integrate this task in the already existing domain of event coreference. We use the recently developed ENCORE dataset (De Langhe et al. 2022a), which is currently the only available Dutch large-scale cross-document event coreference corpus. This corpus distinguishes identity relations between events and so-called ”part-whole” (or subevent) relations. For our exact definitions of event-event relationships, we base ourselves on earlier work in the field of Dutch entity coreference (Oostdijk et al. 2013), which distinguishes several relations between coreferring textual entities. In the sections below we give a broad overview of the current corpora, trends and research methods for subevent detection and ECR. We then describe in more detail the corpus that was used for the experiments in this paper, including a thorough discussion of the design choices and annotation strategies. This is followed by two experiments to enable Dutch subevent detection. In a first experiment our aim is to correctly predict the relation between two textual events that are known to corefer. Subsequently, we perform a second experiment in which the goal is to automatically determine whether or not two events corefer and only then predict the relation, using both a pipeline and joint approach. We experiment with both transformer-based and established feature-based models, after which we present the results and perform a detailed error analysis. Additionally, following the success of specialized transformer architectures in the English language domain (Joshi et al. 2020), we add a number of lightweight modifications to extend widely-used Dutch models on this newly defined task.

Our results reveal that while transformer-based methods attain the best results, they do not necessarily learn the correct underlying principles of coreferential relations.

2. Related work

As stated above, many studies regard the detection of subevents as a task in itself (Chowdhury et al. 2022) where the primary objective is to find events that follow from, or are part of, a predefined larger event without taking coreference resolution into consideration. In recent years, however, there has been a growing interest in the integration of subevents into the broader task of event coreference resolution, at least for English. This because both tasks are naturally connected and because of the advances made in including non-identity relations in the domain of entity coreference resolution. The sections below will mostly focus on such studies when discussing recent advances in subevent detection, but first an overview of existing ECR resources will be provided.

2.1 Resources for Event Coreference Resolution

Compared to entity coreference, event coreference resolution is much less studied. Available resources are few and far between and corpora that distinguish between different coreferential relations at the level of events are even harder to find. This is further complicated by the fact that the creation of new ECR resources is often a time-consuming task, due to the many layers in the annotation process and the expert knowledge that is required. Nonetheless, five major English large-scale ECR corpora exist which are commonly used for benchmarking event coreference studies: ECB (Bejan and Harabagiu 2010), ECB+ (Cybulska and Vossen 2014), OntoNotes (Pradhan et al. 2007), TAC-KBP (Mitamura et al. 2015) and ACE (*ACE English Annotation Guidelines for Events (v5.4.3)* 2008). These large-scale corpora comprise coreference relations, but no event-subevent links. Moreover only the ECB+, TAC-KBP and Ontonotes corpora include cross-document coreference links.

Despite the absence of large-scale ECR corpora that include event-subevent links, there exist some smaller resources that do take into account more complex event-event relationships.The first one is the HiEve corpus by Glavaš et al. (2014). This dataset consists of a total of 100 randomly selected documents from the larger GraphEVE event corpus, which was annotated with coreference. The possible event-event relations of this corpus are defined as *COREF*, *SubSuper* and *SuperSub*. These last two denote event-subevent relations. On average, each document contains 23 *SubSuper/SuperSub* relationships.

A second resource is the Richer Event Description (RED) corpus (O'Gorman et al. 2016). This dataset was specifically developed to display all possible interactions between textual events and presents as such one of the only corpora which includes all commonly studied event-event relations.

The corpus contains a total 95 documents and 4969 event-event relationships, annotated at the within-document level.

Hong et al. (2016) compiled another resource for event-event relationships, similar to the RED corpus. However, this corpus comprises annotations in a cross-document setting and a more fine-grained event schema. This schema allows for five main types of event-event relations, each overarching several subrelations, including coreferential and subevent relations. The complete corpus comprises 125 documents and 863 events resulting in a total of 25610 event pairs for which a possible relations can be defined.

When it comes to Dutch not much work has been done on compiling ECR corpora comprising event-event relations. A final corpus that should be mentioned in this respect is the Dutch Meantime Newsreader corpus (Minard et al. 2016). While limited in size (120 news articles), we highlight this corpus as the first dataset in the Dutch language domain that includes both coreferential and spatio-temporal event-event relations. In addition to its Dutch component the corpus also includes event-event annotations in a cross-document setting for English, Italian and Spanish. It should be noted that the articles in Dutch, Spanish and Italian were translated from the original English source news articles and that the Italian and Spanish data was not annotated directly, but rather through cross-lingual projection.

2.2 Methodologies for ECR and subevent detection

For general event coreference resolution, i.e the detection of identity links between event mentions, early methodologies consisted mainly of rule-based approaches and methods based on traditional feature-based machine learning. Due to the emergence of large-scale corpora, the use of large-scale neural approaches also became commonplace (Lu and Ng 2018a). Typically, mention-pair models were used as a first step in the classification process. These models consider two event mentions in a binary decision on whether or not they corefer and often take the form of widely-used machine learning algorithms such as decision trees (Cybulska and Vossen 2015), support vector machines (Chen and Ng 2014) or maximum entropy models (Ahn 2006). Because of this binary output, which could possibly violate transitivity [1], a defining property of coreference, the output was subsequently fed to clustering algorithms, ranging from best-first to graph-based partitioning and spectral clustering in order to obtain coreference chains (Chen and Ji 2010). The availability of corpora such as ECB+ (Cybulska and Vossen 2014), TAC-KBP (Mitamura et al. 2015) and OntoNotes (Pradhan et al. 2007) has resulted in a shift from traditional machine learning algorithms towards larger neural models (Lu and Ng 2018b). While the mention-pair approach is still widely used for resolving event coreference, the current classification algorithm often include deep neural networks (Nguyen et al. 2016) or transformer-based architectures (Lu and Ng 2021).

A second paradigm within coreference studies are mention-ranking algorithms. Rather than classifying mentions two at a time, this approach constructs a ranking of all possible antecedents of a given event, based on the likelihood of coreference (Lu and Ng 2018a). Additionally, some mention-ranking models rank coreferential chain partitions of a given document as a whole, taking full advantage of a text's discourse structure (Lu and Ng 2017). A notable disadvantage of mention-ranking approaches however is their scalability as far as cross-document coreference is concerned.

Please note that the above-mentioned approaches only determine coreferential relations between events and are not concerned with the identification of the events themselves. Recently however, event mention identification is increasingly being integrated into the task of ECR. The first of such end-to-end coreference resolution models came in the form of pipeline structures, where events are extracted from raw text prior to applying a dedicated coreference resolution algorithm (Choubey and Huang 2017). More recently, English ECR research has primarily focused on the development of joint models which perform event extraction and resolution in one go, to mitigate the problem of error propagation which often plagues pipeline architectures. This newer generation of joint models

[1] if the relation holds between A and B and between the B and C, it holds between A and C

can take on many forms ranging from techniques involving Integer Linear Programming (ILP) (Chen and Ng 2016) and Markov Logic Networks (Lu and Ng 2016) to span-based transformer language models (Lu and Ng 2021).

3. From Entity to Event Coreference Relations

Before discussing the Dutch ENCORE corpus and our experimental setup in more detail, it is useful to take a look at subevents and other coreferential relations from a more theoretical perspective. In the following sections, we first briefly list the different coreferential entity relations, after which these are extrapolated to the event level in order to create a more nuanced scheme for annotating coreference between events.

3.1 Entity Coreference in Dutch

When considering Dutch corpora annotated with coreference, generally three relations have been distinguished between coreferring entities (Hoste 2005). First, there is the fairly straightforward identity relation, in which two entities refer to exactly the same real-world person or object (Example 5).

5. De laatste e-mails van **de leraar** aan de schooldirectie voor **hij** werd onthoofd. *EN: **The teacher's** final e-mails to the school board before **he** was decapitated.*

Second, there exist part-whole relations, when one of the entities is connected to another entity, but only to a part of it. This relation often occurs when one entity denotes a group and the other a person that is part of said group. This type of relation, however, is not exclusive to entities referring to people. Non-living (object) entities can also, as shown in Example (6), constitute a part-whole relation.

6. **De Amerikaanse regering** keurde de actie in ieder geval al af. **President Trump** schreef op Twitter zeer teleurgesteld te zijn in zijn Noord-Koreaanse collega. *EN: **The American government** disapproved of the action. **President Trump** tweeted that he was very disappointed in his North-Korean colleague*

Finally, two entities can also be linked through a type/token relation, in which they refer to the same type of object, but to a different (real-life) token. In other words, the entities do not refer to the same real world object, but to one of a similar description, as illustrated in Example (7).

7. Premier Michel koos op het fotomoment voor **de blauwe vlag**, terwijl Tom Van Grieken naar **de gele** greep. *EN:Prime minister Michel chose a **blue flag** for the photo op, while Tom Van Grieken went for **the yellow one***

3.2 Extrapolating Entity Relations to Event Relations

The different relations between coreferential events used in the ENCORE corpus are largely based on the entity coreference relations discussed in the section above. In this section, we provide a brief discussion on how the relations between various entities can be extrapolated to the event level. We also discuss the practical and conceptual difficulties encountered when trying to impose the entity relations on event structures.

3.2.1 Identity Relations

First, we distinguish the standard identity relation for two coreferring events. In this case, the two events in question refer to exactly the same real-world or fictional event. In order for this relation to be assigned to a pair of events, they must happen at the same time (1), same place (2) and must have the same participants involved (3). It is useful to mention that often, not all event arguments (time, location and participants) are explicitly mentioned in the text. However, most of the time, one can infer the (implied) participants from the context with relative certainty. If we consider the example below, it would be safe to assume that the two events *De spelen in London* and *het Britse Olympische feest* refer to the same event, as the wider context makes clear that it is the 2012 London Olympics that are being discussed and not the 1904 or 1948 editions. In this case a coreference link can be safely established, even though we can only verify a match in geographical location for both events.

8. **De spelen in London** werden afgesloten met een indrukwekkende ceremonie. Voor de Engelsen zullen de prestaties van Mo Farah op **het Britse Olympische feest** waarschijnlijk nog lang nazinderen. *EN: The games came to a close on sunday with a spectacular ceremony. The English will most likely remember the Olympic feast for the accomplishments of Mo Farah.*

3.2.2 Part-whole Relations

Part-whole relations between events are more difficult to accurately define compared to their identity counterparts. In general, one could say that for an event to be considered part of another event it should: have (1) at least partial temporal overlap with and (2) be a direct contributing factor to its overarching event. These two properties are needed to distinguish the part-whole coreferential relation from other event-event relations such as causality (O'Gorman et al. 2016). This can be illustrated by the two examples below, where Example (9) denotes a part-whole relation between the two events and Example (10) denotes causality.

9. (a) **Politieke aardbeving in Israël** *EN: Political earthquake in Israel.*

   (b) **De Israëlische premier Ariel Sharon heeft zijn lidmaatschap van de Likoedpartij opgezegd** en **het ontslag van zijn regering aangeboden** *EN: The Israeli premier Ariel Sharon has terminated his membership of the Likud party and proposed the resignation of his government.*

10. Door **de vulkaanuitbarsting** moesten **alle omwonenden geëvacueerd worden**. *EN: Those living in the area had to be evacuated due to the volcanic eruption.*

Note that the line between causality and part-whole can be muddled, especially in reference to the term 'subevent'. Social media subevent extraction studies (Chowdhury et al. 2022) often include both causal and part-whole relationships under the name 'subevent', while existing ECR literature uses the term 'subevent' to refer to part-whole structures exclusively (Hong et al. 2016). As we approach this problem from the point of view of coreference, we adhere to the latter terminology.

An interesting theoretical observation is that the principle of transitivity that holds between coreferring identity events does not necessarily hold true for all part-whole relations. As an illustration, we see three events in Example (11). If there is an identity relation between event A and B and there is an identity relation between event A and C, we can automatically assume that there is an identity relation between events B and C as well. However, looking at Example (11) we can see that the principle does not necessarily hold for part-whole links, as there is no part-whole relation between the events *de atletiekcompetitie* and *het zwemmen.*

11. (a) **De Olympische spelen in Rio** gaan door ondanks hevige tegenstand van klimaatgroepen. *EN: The Olympic games in Rio will go ahead as planned, despite heavy opposition by climate groups.*

    (b) Usain Bolt begon sterk aan zijn afscheidsnummer in **de atletiekcompetitie**. *EN: Usain Bolt had a strong start in his final race in the athletics discipline.*

(c) Bij **het zwemmen** kroonde Michael Phelpps zich opnieuw tot de onbetwiste kampioen. *EN: Michael Phelps became the undisputed champion again during the swiming competition*

3.2.3 Type/token Relations

A final relation between entities that was defined for Dutch is the type/token link, in which two entities refer to an object of similar description. Much like with the previous two relations, entity type/token links have an equivalent at the event level. Consider Example (12) below as an illustration:

12. (a) Bart Tommelein verkozen tot burgemeester van Oostende. *EN: Bart Tommelein reelected as mayor of Ostend.*

    (b) Matthias De Clercq wordt de nieuwe burgemeester van Gent. *EN: Matthias De CLercq becomes new mayor of Ghent.*

We decided not to include this coreferential relation in our experiments for the following reasons. First and foremost, it can be argued that type/token relations at the event level are not as useful as other event-event relations with regards to information extraction and structural text analysis. Unlike identity or part-whole links they cannot be integrated in textual knowledge bases or in the creation of narrative timelines. At most, they can be used to establish some parallelism between event chains that have different participants and arguments. In this case however, we would argue that type/token relations are much more related to the domain of semantic role labeling (SRL) than to coreference. Additionally, we feel that the approach to resolve type/token relationships between events would differ significantly from the methodologies that were outlined in Section 2.3. It would be hard to incorporate this relation in existing coreference algorithms such as knowledge-graphs or mention-pair and mention-ranking approaches due to conceptual differences between a type/token link on the one hand and identity or part-whole links on the other. In order to detect the type/token parallelism it would perhaps be better to use fairly rudimentary lexical matching of the verbal component, coupled with a resource such as Wordnet (Miller 1995) or, if indeed approaching the problem through the lens of SRL, using established frameworks such as PropBank (Kingsbury and Palmer 2003) or FrameNet (Fillmore et al. 2002).

A second and more conceptual problem is how to accurately define the boundaries of an event type/token relation. For the identity relation this was fairly simple: only if two events refer to exactly the same real-world or fictional event an identity link can be established. For part-whole relations a concrete definition was slightly harder to formulate mainly due to the high degree of similarity between part-whole and causal links. However, as was illustrated by Example (11) in the previous section, as well as from the existing body of work on subevent detection, we could set clear boundaries. The problem with doing this for type/token relations is nicely illustrated by Example (13).

13. (a) [Voor het eerst kwamen de Amerikaanse president en de Noord-Koreaanse leider samen in een historische top]. *EN: American president Trump and North Korean leader meet fur the first time in historic summit*

    (b) [President Macron en de Iraanse Ayatollah spreken elkaar morgen op een nieuw topoverleg]. *EN: President Macron and the Iranian Ayatollah will hold talks tomorow at the summit*

    (c) [De tweede ontmoeting tussen Trump en Kim-Jong-Un] wierp dus voorlopig nog geen vruchten af. *EN: The second meeting of Trump and Kim-Jong-Un has lead to no breakthroughs for now*

Going by the definition of entity type/token relations we could draw a type/token link between the events in Example (13a) and Example (13b), as well as between the events in (13a) and (13c). Nonetheless, it is apparent from the examples that there is a conceptual difference between the two relations. By including the type/token relation in our dataset we would also import a hoist of conceptual problems as to what exactly constitutes this relation.

We therefore conclude that it would be best to only focus on identity and part-whole coreference links, because these are clearly defined and can thus be reliably annotated, offering more perspectives with respect to the creation of large-scale event corpora. The availability of large datasets in itself opens up the possibility of introducing deep neural architectures to the problem of event coreference and subevent detection. Something that was, as mentioned before, not viable in the past.

4. The ENCORE Corpus

While a full in-depth description of the Dutch ENCORE corpus can be found in De Langhe et al. (2022a), we will now provide a brief overview of its contents, as well as the reasons why this dataset in particular can give a boost to coreference studies and subevent analysis in the Dutch language. Note that in De Langhe et al. (2022a) the event-subevent relations of ENCORE are only discussed marginally, as they are not the intended focus of the corpus. In this paper, however, we take the opportunity to highlight certain design choices with respect to the annotation of subevents, based on the theoretical observations made in Section 3. Additionally, we include a more in-depth Inter-Annotator Agreement (IAA) analysis of coreference annotations in the corpus, emphasizing cases that could be problematic to categorize based on the current annotation guidelines for Dutch event coreference (De Langhe et al. 2021).

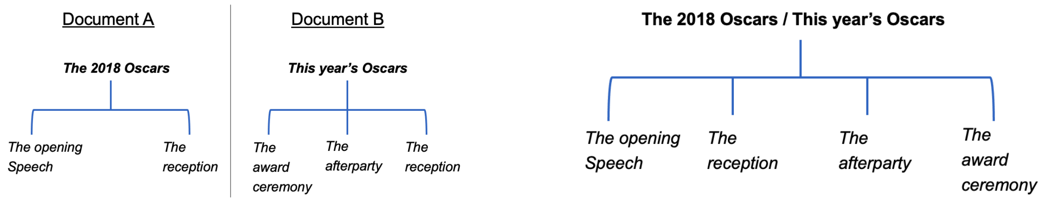
4.1 General corpus description

The Dutch ENCORE corpus contains 1,115 newspaper articles which were sourced from a larger database encompassing over 631,559 news documents. All articles were retrieved from the online versions of a number of Dutch (Flemish) national (*De Morgen*, *Het Nieuwsblad*, *Het Laatste Nieuws*, *De Standaard*) and regional (*Het Belang van Limburg*) newspapers. In addition to this, articles published on the news website of the Flemish public broadcasting agency (*VRT News*) were included. All news was collected throughout the 2018 calendar year, making this dataset especially suited for extracting cross-document coreference resolution data. Typically, newspapers will publish multiple articles describing the same overarching event and while the exact topic may be different it is often the case that the same events will be present in these articles as background information.

A second reason as to why this collection is interesting for the collection of coreference data, is that multiple newspapers are included. As the articles are all published within the same calendar year, the different newspapers will cover (largely) the same events, resulting in many cross-document coreferential links. Those newspapers will often report events from a given angle, sometimes putting emphasis on different background events and context. Furthermore, also the fact that different authors will describe the same event differently is interesting in itself from a lexical richness viewpoint. Unlike some of the corpora mentioned in Section 2.3 the goals of the ENCORE corpus specifically included the collection of diverse events, ranging from financial news, to local curiosities that made the headlines. In total, the corpus contains over 15,000 events (De Langhe et al. 2022a) between which coreference relations can be drawn. Note however that cross-documents coreferential links were only established between events contained by the same overarching topic. Each of the 91 topics contains on average 11-12 newspaper articles. This measure was devised in order to create a more overseeable annotation task. Finally, a total of 1,018 intra-document event coreference chains and 1,587 cross-document chains were annotated in the corpus, where each coreference chain comprises at least 2 events.

4.2 Subevents and subevent annotation in the ENCORE corpus

The inclusion of multiple sources discussing the same events is also particularly useful if we consider subevent coreferential relations. Consider a case where article A might refer to a number of consequences from a given event, while article B focuses on subevents that were not covered in A. By resolving coreference between these two documents we can thus merge and, by extent, improve knowledge about a given situation or event. Consider Figures 1a and 1b below for an illustration. In Figure 1a, the main events of two newspaper articles are presented. Both articles report on the 2018 Oscars, but discuss different aspects of said event. By resolving coreference between all the events in those articles, i.e knowing that for example *The 2018 Oscars* and *This year's Oscars* refer to one another and that *The opening speech* is a subevent of *The 2018 Oscars*, and by result of *This year's Oscars*, we can create a much more detailed representation of the real-life event (Figure 1b).



(a) Within-document event coreference

(b) Cross-document event coreference

Figure 1: Illustration of the potential of cross-document event coreference

As was discussed in detail in Section 3, we believe that we can reliably define two event coreference relations: identity and part-whole (or subevents). Six annotators (all graduate students in linguistics) were hired over a two-month period. Each annotator worked part-time and was presented around 200 news articles of varying length. Annotators were first tasked with identifying all events and their arguments and then annotated coreference between those events, both at a within-document and cross-document level. In a final step, the annotators had to specify the relation between the coreferring events, such as *identity* or *part-whole*. The annotation guidelines for determining coreference were based on the observations outlined in Section 3. For an *identity* link events had to take place at exactly the same time (1) in the same place (2) and should involve the same participants (3). For a part-whole link to be established one event should take place within the duration of a larger event (1) and that same event should be a contributing factor to or part of the overarching event (2). In total, 44,148 coreferential links between event pairs were drawn. 15,587 (or around 35 percent) of those were part-whole relations or subevents while the other 28,561 were identity relations.

In order to determine the quality of the annotations, we calculate IAA for two subtasks of the annotation process: the annotation of coreferential links between events and of the relations. Note that for the calculation of IAA for the coreference annotation tasks, annotators were given documents containing gold-standard annotations for the events in order to eliminate the potential influence of mistakes made in the event mention annotation task. The agreement between annotators was calculated using a pairwise Cohen's Kappa score (Cohen 1960) between each of the possible 15 annotator pairings. A final score was then obtained by taking the average of the pairwise scores for each of the two tasks. The choice for a pairwise agreement rather than a multi-annotator metric such as Fleisch's Kappa (Fleiss and Cohen 1973) was motivated by evidence that such metrics can become distorted and much less interpretable for larger amounts of (> 4) annotators (Gwet 2008). The calculated agreement was 0.80 for the general event coreference task and 0.83 for the task where the coreferential relation had to be determined, with both scores indicating a very strong agreement. The plotted pairwise scores for each of the tasks (Figure 2) show very little variance between annotators and no outliers.

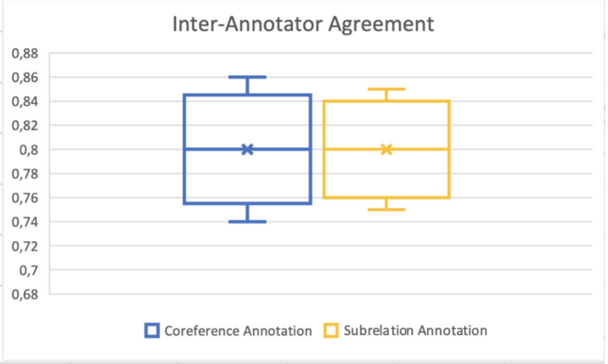


Figure 2: Inter-Annotator Agreement for event coreference annotation and relation annotation

5. Experiments

In this section we describe two experiments to perform Dutch subevent detection in a cross-document setting. Our main objective is to test the viability of integrating subevent detection as a subsidiary task to general event coreference resolution. Please note that we do not perform event mention detection itself, as for this paper our interest lies purely in the coreferential relations between (already detected) events. Contrary to earlier studies on resolving event-event relations (Hong et al. 2016) we have access to a large-scale dataset (see Section 4) allowing us to make full use of recent advancements in NLP, such as deep neural networks and transformer architectures (Devlin et al. 2018). The current state of the art in both entity and general event coreference studies revolves around span-based transformer architectures (Joshi et al. 2020). However, given that no Dutch or multilingual version of these models currently exist and that training one would be far beyond the scope of this paper, we make use of existing monolingual Dutch and multilingual transformer models, as well as try to integrate these into ensemble architectures for optimal performance. Moreover, since most state of the art (SOTA) algorithms for entity and event coreference are based on slightly modifying the pretraining and fine-tuning steps in regular transformer models, we also experiment with several methods for parameter averaging such as self-ensembling and self-distillation as recently proposed by Xu et al. (2020).

5.1 Experimental setup

We evaluate all models described below using a traditional train-test split in our data. In our case, 70 percent of the data is reserved for training, 15 percent for development and hyperparameter tuning and the final portion for evaluation. Due to the pairwise setup of our task we do not shuffle our data before splitting it in train/dev/test subsets. In addition to this, we ensure that topic cluster boundaries are respected when dividing the data, i.e no events from the same topic will be in both the training and test sets. We take this measure so that we can ensure that the model is applicable to unrestricted multi-domain contexts, without any additional domain-specific fine-tuning or post-training.

5.1.1 Coreference relation classification

First, we attempt to correctly classify the coreference relation (identity/part-whole) using gold standard coreferring event mentions. To this purpose, all annotated events and gold-standard event coreference chains are retrieved from the corpus, after which a pairwise classification approach is applied, similar to the mention-pair approaches described in Section 2. 3 provides an illustration of the model setup.

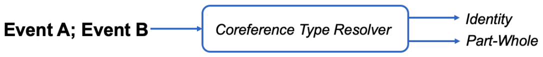


Figure 3: Visual representation of the mention-pair coreference relation task

5.1.2 Event coreference resolution

For the second experiment we start from all gold standard event mentions present in the dataset and perform event coreference resolution. In other words, we try to determine whether they refer to the same real-world event, whether one of the events is a subevent of the other or whether there is no relation at all. To this purpose a pipeline approach is used, based in the first place on an existing Dutch event coreference resolution algorithm (De Langhe et al. 2022b). The output of this step is then directly presented to the coreference relation classification models. It should be noted that the coreference resolver itself was retrained from scratch to ensure that none of the data that was used for testing the coreference relation classification was used for training. Training parameters were set to be identical to those used in De Langhe et al. (2022b). Figure 4 gives a schematic representation of the pipeline approach.

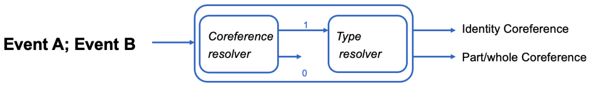


Figure 4: Schematic representation of the pipeline coreference model

While pipeline architectures are widely used in ECR studies, they are also notoriously susceptible to error propagation, which is why we also experimented with a joint model which considers the task as a pairwise multilabel classification task (see Figure 5 for a schematic overview of this joint approach).

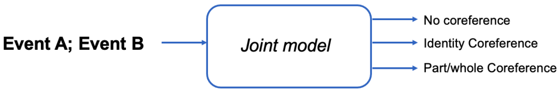


Figure 5: Schematic representation of the joint coreference model

5.2 Baseline Experiments

5.2.1 Models

5.2.1.1 Traditional machine learning We developed one feature-based baseline model for both the coreference relation classification and ECR tasks. We opt for a support vector machine (SVM) due its reliable track record when it comes to robustness and general performance in a variety of NLP tasks (Daumé III 2004). We use a collection of lexical similarity features, such as event span and action similarity based on cosine distance, dice coefficient and minimum edit distance (MED). Additionally, we use a number of discourse and logical constraining features that have been applied in both English (Lu and Ng 2018a) and Dutch (De Langhe et al. 2022b) ECR studies.

5.2.1.2 BERT-based Transformer Architectures For both tasks, several transformer-based mention-pair models are employed. Most notably, BERTje (de Vries et al. 2019) and RobBERT (Delobelle et al. 2020), which are monolingual Dutch versions of the BERT-base and RoBERTa-base models, respectively. BERTje was pre-trained on a total of around 2.4B tokens of high-quality Dutch texts which include the Dutch Sonar-500 (Oostdijk et al. 2013) and TwNC (Ordelman et al. 2007) corpora, Wikipedia data, historical fiction and a large collection of Dutch online newspaper articles collected over a 4 year period. As a significant portion of the BERTje pretraining data is made out of newspaper articles, we believe this model is particularly fit for event-related tasks. RobBERT on the other hand was pretrained on 6.6B tokens of Commoncrawl webdata (Suárez et al. 2019). Since the Commoncrawl data consists of individual lines and not every line contains more than one sentence, we anticipate that this model might be less effective.

We also finetune the monolingual RobBERTje models for this task. The RobBERTje models include a series of distilled language models (Sanh et al. 2019), employing both the aforementioned BERTje and RobBERT as teacher models. The distillation model has previously been shown to outperform the two other Dutch language models on coreference-based tasks (Allein et al. 2020) and pronoun prediction (Delobelle et al. 2022). In addition to these three monolingual models, we finetune the multilingual XLM-ROBERTa (Lample and Conneau 2019), as it contains a substantial amount of Dutch data and has been shown to be quite effective in a number of Dutch NLP tasks (Bouma 2021).

The fine-tuning strategy is identical across all transformer-based language models described above. Both experiments are set up as a text pair classification task. Event mention pairs are concatenated, tokenized and fed to the encoder. Note that for the fine-tuning process, we only make use of the annotated event mention spans, without their encompassing discourse context. For classification, the final state of the input token [CLS] is used as an aggregate representation of the text pair and used as input for a standard classification function.

5.2.1.3 Ensemble Methods It is often the case that model performance can be increased by combining multiple algorithms, either by conducting a hard vote between model predictions or by averaging their output classification logits (Zhou et al. 2002). As we have a large number of trained models at our disposal, we propose to see whether or not we can increase model performance by combining the transformer models with the traditional feature-based approach. A hard vote three-way and five-way ensemble has been created for the feature-based SVM, BERTje, RobBERT, RobBERTje and XLM-RoBERTa models. In addition, a weighed voting algorithm has also been created. In the latter case, classification predictions for each of the models are fed in a one-layer feedforward neural network, which is then trained on the development set in order to obtain optimal voting weights for each model in the ensemble.

5.2.2 Results

Before discussing the results of our experiments it is useful to mention that coreference is usually evaluated through cluster- and link-based metrics such as MUC (Vilain et al. 1995), B3 (Bagga and Baldwin 1998), CEAFe (Luo 2005), Lea (Moosavi and Strube 2016) or an aggregation of those. However, it is not possible to meaningfully assess model performance through these standard means for our event-event relation tasks. If we consider our output coreference chains as graphs, the aforementioned metrics would evaluate our systems based on the presence/absence of edges between event nodes (= coreference), rather than evaluating the edge labels (= subevents) themself, which is the goal of our classification task. For the sake of interpretability we will therefore evaluate the task using standard macro-averaged F1.

5.2.2.1 Coreference Relation Classification Table 1 details the results for the coreference relation classification task, where the goal was to predict the nature of the relation (identity/part-whole) between two events. Note that for the 3- and 5-way (weighted) ensembles only the best ensemble model is presented in the table. For a detailed overview of all ensembles that were used, we refer to Appendix A.

From the table we can infer that, unsurprisingly, the three monolingual transformer models outperform both the feature-based SVM and the multilingual XLM-RobBERTa model. Note specifically,

Model	Macro F1 score
SVM	0.55
BERTje	0.60
RobBERT	0.58
RobBERTje	0.61
XLM-RobBERTa	0.52
Ensemble (BERTje, RobBERTje, SVM)	0.61
Ensemble$_w$ (BERTje, RobBERTje, SVM)	**0.63**

Table 1: Results for the coreference relation classification task

however, that the XLM-RobBERTa model performs worst out of all models. This is in line with earlier work regarding Dutch event-event relationships, which has observed that multilingual models typically do not perform well with respect to coreference-based tasks (De Langhe et al. 2022b). The ensemble models, with the weighed ensemble model in particular, seem to improve on the performance of the monolingual fine-tuned BERT-based models. This is of course at the cost of a significant increase in training time, as three large-scale models are trained instead of one.

5.2.2.2 Event coreference resolution As stated before, the objective in the event coreference task is two-fold. First, to correctly determine which event mentions do or do not corefer and. Second, to determine the relation (identity/part-whole) between coreferring events. The pipeline setup does this through two binary classification tasks: coreference/non-coreference and identity/part-whole. The joint approach, on the other hand, considers the problem as a multiclass classification task with 3 possible classes: non-coreference, identity and part-whole. Note that for the evaluation of the pipeline setup, we set aside the non-corefering events after the first step in the pipeline and then merge those back with the predicted relation after the second step.

The pipeline system consists of a baseline finetuned algorithm for Dutch ECR (De Langhe et al. 2022b) that extracts corefering events from the raw event mentions. We couple this coreference resolution algorithm with the subtype classification algorithms that were previously trained. The joint models consist of the same algorithms that were used in the previous section, but now finetuned for a multiclass classification task. The results of the pipeline approach can be found in Table 2a, while the results of the joint approach are represented in Table 2b. Note that the ensemble and weighted ensemble architectures are identical to those in Section 5.2.2.1, consisting of the SVM model and finetuned BERTje and RobBERTje models.

Model	F1 score
SVM	0.47
BERTje	**0.56**
RobBERT	0.54
RobBERTje	0.54
XLM-RobBERTa	0.45
Ensemble 1	0.51
Ensemble 2	**0.56**

(a) Results for the pipeline ECR experiments

Model	F1 score
SVM	0.34
BERTje	0.47
RobBERT	0.44
RobBERTje	0.48
XLM-RobBERTa	0.29
Ensemble 1	0.51
Ensemble 2	**0.53**

(b) Results for the joint ECR experiments

Table 2: Results of the event coreference task. Table 2a and 2b detail results for for the subtype integration using a pipeline and joint approach respectively .

A first striking observation here is that most of the pipeline models outperform the joint models by a large margin. Looking at the individual class F1 scores for the pipeline and joint models (Table 3) we can infer that this large difference in performance is mainly due to the much lower scores for the non-coreference and part-whole labels in the joint setup. It is possible that the distinction between non-coreference and part-whole coreference is problematic due to the fact that the classification decision cannot be made on the basis of word-level lexical similarity. The strength of current transformer architectures lies primarily in associations at the lexical-semantic level, rather than large-distance discourse relationships between parts of the text, which is required in order to resolve coreference and detailed subrelations between events.

Model	**F1 (Non-Coreference)**		**F1 (Identity)**		**F1 (Part-whole)**	
	Pipeline	*Joint*	*Pipeline*	*Joint*	*Pipeline*	*Joint*
SVM	0.63	0.30	0.48	0.51	0.32	0.21
BERTje	0.63	0.48	0.57	0.61	**0.48**	0.33
RobBERT	0.63	0.30	0.52	0.58	0.47	0.44
RobBERTje	0.63	0.38	0.58	0.64	0.41	0.42
XLM-RoBERTa	0.63	0.21	0.38	0.48	0.34	0.18
Ensemble 1	0.63	0.40	0.58	0.67	0.32	**0.46**
Ensemble 2	0.63	**0.50**	**0.60**	**0.69**	0.45	0.40

Table 3: Individual class labels for the pipeline and joint setup

Second, it is of course no surprise that in the pipeline setting we observe largely the same trends in model performance that were reported for the coreference relation classification tasks (Section 5.2.2.1). Given the fact that the first model in the pipeline architecture is identical, that the input for the applied relation classification model is also identical.

5.2.3 Analysis and Discussion

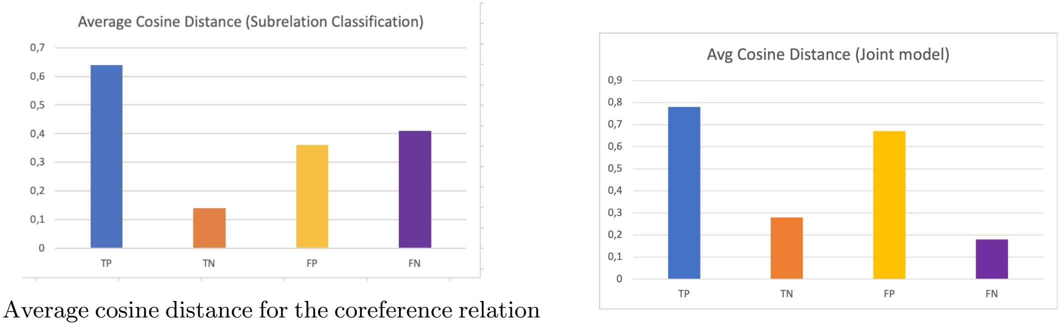
5.2.3.1 Influence of lexical similarity in transformers If it is indeed the case that in the coreference relation classification task the model simply learns that events that are not semantically or outwardly similar should be classified as part-whole events, it would provide a good argument as to why the performance of the joint models is so low. If the models can only learn to classify based on this similarity, rather than the underlying principles governing coreference, the joint models would ultimately fail due to the fact that non-coreferring events and part-whole events would be virtually indistinguishable.

In order to gauge the impact of lexical similarity on the predictions of the models, we measure the cosine distance based on Sentence-BERT embeddings (Reimers and Gurevych 2019) for each event pair in the test set. We obtain these embeddings by feeding individual event mentions into a Dutch sentence-BERT implementation[2], averaging the last 4 hidden layers of the model and then use a mean pooling operation of the individual tokens of the mention in order to compress the representation into a single vector. Cosine similarity was then computed between each pair of possible event mention vector representations. We then plot the average cosine distance for each of the examples classified as True positive (TP), False positive (FP), True negative (TN) and False negative (FN) for the best performing models (weighed ensembles w/ SVM, BERTje and RobBERTje) for both the coreference relation classification and event coreference task.

For the coreference relation classification task, we see that average cosine distance is highest for True Positive cases and lowest for True Negative cases. Then, for the event coreference task we see that average cosine distance is highest between True Positive examples, while being lowest for the False Negatives. From further analysis we also learn that in the ECR task class accuracy for both non-coreference and part-whole coreference is extremely low.

Based on those observations it would be safe to assume that the joint model does not learn the discourse properties underlying event coreference and event-subevent relations, but rather bases

[2] jegormeister/bert-base-dutch-cased-snli



(a) Average cosine distance for the coreference relation classification task

(b) Average cosine distance for the joint ECR task

Figure 6: Average Cosine distances for the best performing model

the classification on outward lexical similarity. In the pipeline approach, however, we observe that average cosine distance for False Negative cases exceeds the average distance for the False Positive cases. While the difference is small, it does indicate that lexical similarity is not the only aspect that plays a role in the classification process. A possible hypothesis that could explain the observed differences between the joint and pipeline approaches is that for the pipeline approach most non-coreferring mentions (i.e pairs with low lexical similarity) have already been identified in the first step of the pipeline. This in turn would lead to less noise in the second step of the pipeline and would both explain the (relatively) higher scores in the coreference relation classification task and the confusion between the part-whole and non-coreference labels in the joint setting.

5.3 Self-Ensembling and Self-Distillation Experiments

As stated in Section 2, past studies in event coreference and classification of event-event relations often tend to make use of knowledge-base structures or complex pipeline architectures (Rospocher et al. 2016). In that regard, a logical next step to improve on the baseline scores in the previous section would thus be to implement such an algorithm for the Dutch language. However, it should be noted that the current state of the art in event coreference is still attained by transformer-based language models by a large margin, rather than by the domain-specific algorithms of the past. Based on the success of the application of specialized transformer architectures in the English language domain (Joshi et al. 2020), we therefore propose to extend existing Dutch language models with techniques that have been known to improve model performance on a variety of NLP tasks.

While training the earlier-mentioned span-based transformer architectures (Joshi et al. 2020) for Dutch would be far beyond the scope of this paper, there exist several methods of extending the fine-tuning process of LMs that can be applied with relative ease and have a proven track record. Concretely, we investigate two techniques proposed by Xu et al. (2020): self-ensembling and self-distillation, which are discussed in more detail in the sections below.

5.3.1 Models

5.3.1.1 Self-Ensembling Xu et al. (2020) posit that the fine-tuning of transformer language models can be improved by accumulating model weights over all training steps of the fine-tuning task and as such create a so-called 'self-ensemble' model. This is based on the observation of Polyak and Juditsky (1992) that averaging model parameters will almost always provide better results in classification models. The authors employ a BERT-based self-ensemble model, where model parameters are accumulated and averaged over the total number of model training steps. We intend to make use of a similar setup using the monolingual models BERTje, RobBERT and RobBERTje as a base for the self-ensemble. We also propose the inclusion of a warmup parameter R, which will ensure that model weights will only be accumulated and averaged after a warmup period of Y training steps. While it is known that warmup measures are often only required in the pretraining phase of a language model, the effects of only averaging later training steps in the fine-tuning process might be beneficial for the model's performance and worth studying. Concretely, the warmup is set as:

$$Y = \frac{\text{total training steps}}{R} \tag{1}$$

where R equals the proportional value of the number of warmup steps. Concretely this means that if R is set as 2, half of the total of number of training steps will be reserved for warmup and only the latter half of training steps will be reserved for accumulation.

5.3.1.2 Self-Distillation Another alternative training method proposed in Xu et al. (2020) is the use of self-distillation. Model distillation is usually done by letting a small model learn from the output logits of a larger teacher model in order to reduce model complexity while still maintaining performance. However, it has been shown by the authors that letting a model learn from its own output logits, without any reduction in model size, can also be beneficial to a variety of NLP text classification tasks improving results over standard transformer-based language models.

Concretely, the standard learning objective is modified by appending a loss term, which is the Mean Square Error (MSE) between output logits of the student and teacher models. The model's standard Cross-entropy loss (CE) and distillation loss (MSE) are balanced by a weight parameter, which is set at the start of the training process. The model weights of the distilled model are obtained by self-ensembling the model's previous K training steps. In other words, the teacher model in this distillation setup is an average of its previous states, with larger values of K increasing the robustness of the overall model due to the aggregation of information throughout its training process.

A notable omission from the formulas detailed above is the temperature parameter (Gou et al. 2021). In knowledge distillation temperature is used for decreasing the absolute values of the teacher model's logits before passing them through the softmax function in the training process. This is typically done by dividing those logits by a fixed parameter T, with larger values of T ensuring a softening of the teacher's predictions, which in turn results in an increase of expressiveness in the Teacher's output distribution. For inference, the temperature parameter is set at 1. The original authors of this model give no explicit reason as to why temperature is removed from the equation. It is possible that they esteem that the aggregation and averaging done by the self-ensembling will provide sufficient smoothing for the output logits. Given the fact that as a rule temperature is included in knowledge distillation, as well as the demonstrated effect of using this parameter on the performance of distilled models (Hinton et al. 2015), we propose including temperature τ in order to smooth softmax probabilities p_i based on the output logits z for our own experiments:

$$p_i = \frac{exp(z_i/\tau)}{\sum_j exp(z_j/\tau)} \tag{2}$$

5.3.2 Results

5.3.2.1 Coreference relation classification Table 4 below details the results for the modified versions of the monolingual language models BERTje, RobBERT and RobBERTje. In the table, we also included the best performing model from our baseline experiments, which was the weighed ensemble of the SVM, BERTje and RobBERTje. Note that for the self-ensemble and self-distillation models, only the best performing models were included. A detailed discussion of the influence of the introduced parameters warmup and temperature will be presented in Section 5.3.3.

From the table above we can infer that it is the self-ensemble monolingual model that seems to work best for the task as a whole. Note also that the BERTje self-distillation model achieves similar

Model	**F1 score**
BERTjeSE	**0.66**
RobBERTSE	0.56
RobBERTjeSE	0.61
BERTjeSD	0.63
RobBERTSD	0.57
RobBERTjeSD	0.59

Table 4: Results for the coreference relation classification using self-ensembling (SE) and self-distillation (SD)

performance (0.66) to the weighed ensemble from the previous experiment section (0.63, cfr. Table 1), with the difference that only one model needs to be trained when self-distilling, rather than 3 for the weighed ensemble.

5.3.2.2 Event coreference resolution Table 5a and Table 5b below indicate the performance of the modified models in pipeline and joint settings respectively. We can infer largely the same trends that were observed in the previous section, with the BERTje self-ensemble model outperforming the other models by a significant margin. As compared to the best results presented with the weighed ensemble in Table 5, we again observe similar performances, i.e., 0.56 versus 0.57 for the pipeline setting and 0.52 versus 0.53 for the joint setting. As before, the self-distillation seems to have a smaller effect on the outcome.

Model	**F1 score**
BERTjeSE	**0.57**
RobBERTSE	0.53
RobBERTjeSE	0.51
BERTjeSD	0.52
RobBERTSD	0.50
RobBERTjeSD	0.48

(a) Results for the pipeline ECR experiments

Model	**F1 score**
BERTjeSE	**0.52**
RobBERTSE	0.49
RobBERTjeSE	0.49
BERTjeSD	0.51
RobBERTSD	0.46
RobBERTjeSD	0.44

(b) Results for the joint ECR experiments

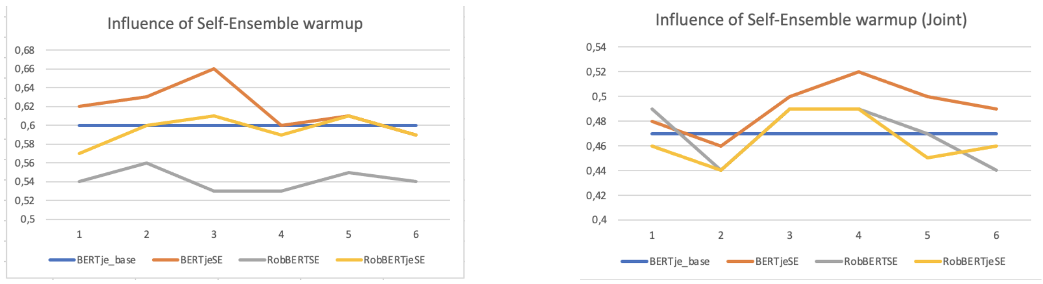
Table 5: Results of the ECR experiments in a pipeline (a) and joint (b) setting using self-ensembling (SE) and self-)distillation (SD).

5.3.3 Analysis and Discussion

5.3.3.1 Temperature and Warmup We also wish to measure the influence of the two new parameters that were introduced for the self-ensemble and self-distillation models, respectively.

In order to gauge the effect of the warmup step parameter W in the self-ensemble models we perform additional experiments where we take the 3 fine-tuned monolingual transformer models (BERTje, RobBERT and RobBERTje) and self-ensemble the models using different values of the warmup parameter W. For each model, we perform 6 training runs using a warmup parameter ranging from 1 to 6. The results for the coreference relation classification and ECR (joint) experiments are found in Figure 7a and Figure 7b.

As can be inferred from Figure 7, adding a warmup parameter to the self-ensemble models can have a favorable effect on the model's end result in some cases. The overall best performance is obtained by taking 1/3 of the total amount of training steps as a warmup period using the BERTje model.

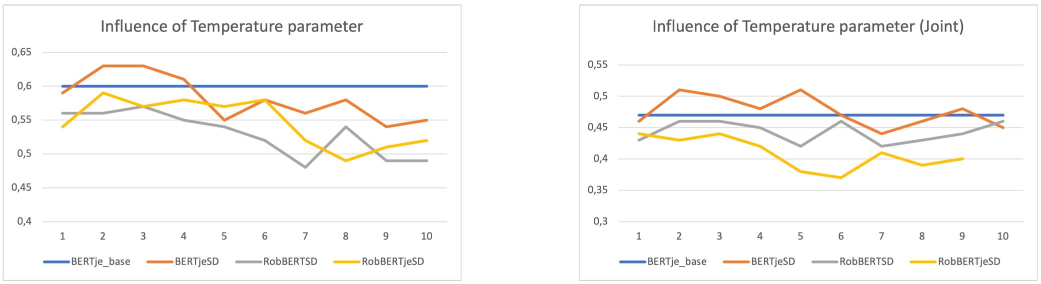


(a) Influence on coreference relation classification

(b) Influence on ECR (joint)

Figure 7: Warmup parameter influence on Self-Ensemble models

We perform a similar experiment for the proposed inclusion of the temperature parameter (τ) in the self-distillation models. We perform 10 training runs for each of the 3 monolingual base models, with the temperature parameter ranging from 1 to 10. Figures 8a and 8b review the effect of adding this parameter in self-distillation for the coreference relation classification and joint ECR, respectively.



(a) Influence on coreference relation classification

(b) influence on ECR (joint)

Figure 8: Temperature parameter influence on Self-Distillation models

From the graphs we can see that including temperature can also have a favourable effect on the classification scores in select cases (using BERTje as the base model), making it a worthwhile inclusion in the self-distillation models as a whole. More specifically, lower values (2-4) seem to consistently outperform the base self-distillation model where $\tau = 1$.

6. Conclusion and Future research

In this paper we have presented our efforts to create the first coreference-based study to automatically detect event-subevent relations in the Dutch language, an important step towards better discourse analysis. In our broad theoretical overview we explained how subevents can easily be integrated into coreference research by extrapolating entity-level coreference relations to event-event relations and by zooming in on two relations in particular, the identity and part-whole (or subevent) relation. To investigate this we relied the Dutch event coreference ENCORE corpus and explained how subevents have been defined in this corpus and some of the peculiarities associated with them. We also provided a more in-depth IAA study on the annotation of the coreference data and revealed a high agreement for both the annotation of coreference between events (0.80) and the actual relation (0.83).

As our main objective was to test the viability of integrating subevent detection as a subsidiary task to general event event coreference resolution we performed two experiments. A coreference relation classification task, in which the nature of the relations between gold-standard coreferring events was tested, and a general ECR task in which the goal was to predict either non-coreference, an identity relation or a part-whole relation between two events in either a pipeline or joint setting. Our baseline experiments consisted of fine-tuning various transformer language models, after which model ensembles were created to gauge the combined performance of these models. Initially, the best results were achieved with these ensemble models. However, in a second step, we also applied self-ensembling and self-distillation techniques to improve the fine-tuning process of the existing monolingual language models. Here we demonstrated that adding a warmup parameter in the self-ensembling process and a temperature in the self-distillation algorithm can have a noticeable effect on model performance, leading to on par or better performance than the ensemble models.

While transformers attain state-of-the-art results in many NLP tasks, including coreference, we note that in our case the actual classification descision was almost always based on lexical similarity rather than on the actual principles underlying coreference in within -and cross-document contexts. We supported this claim by analyzing individual class labels as well as average cosine distance between event mentions. This poses an interesting choice between overall model performance and algorithmic language understanding. In future research we propose to further explore the possibility of bridging the gap between those two choices by combining (older) graph-based and clustering methods with powerful semantic transformer representations. In addition to this, we also aim to integrate the detection of the events themselves, which is notably absent from this paper, into our coreference pipeline. This final component would allow us to create an end-to-end system for fine-grained event coreference resolution and structural text analysis.

References

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

Ahn, David (2006), The stages of event extraction, *Proceedings of the Workshop on Annotating and Reasoning about Time and Events - ARTE '06* (July), pp. 1–8. http://portal.acm.org/citation.cfm?doid=1629235.1629236.

Allein, Liesbeth, Artuur Leeuwenberg, and Marie-Francine Moens (2020), Automatically correcting dutch pronouns" die" and" dat", *Computational Linguistics in the Netherlands Journal* **10**, pp. 19–36.

Altmami, Nouf Ibrahim and Mohamed El Bachir Menai (2020), Automatic summarization of scientific articles: A survey, *Journal of King Saud University-Computer and Information Sciences*, Elsevier.

Araki, Jun (2018), *Extraction of Event Structures from Text*, PhD thesis, Ph. D. thesis, Carnegie Mellon University.

Bagga, Amit and Breck Baldwin (1998), Algorithms for scoring coreference chains, *The first international conference on language resources and evaluation workshop on linguistics coreference*, Vol. 1, Citeseer, pp. 563–566.

Bedi, Harsimran, Sangameshwar Patil, Swapnil Hingmire, and Girish Palshikar (2017), Event timeline generation from history textbooks, *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pp. 69–77.

Bejan, Cosmin and Sanda Harabagiu (2010), Unsupervised Event Coreference Resolution with Rich Linguistic Features, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (July), pp. 1412–1422. http://www.aclweb.org/anthology/P10-1143.

Benedetti, Fabio, Domenico Beneventano, Sonia Bergamaschi, and Giovanni Simonini (2019), Computing inter-document similarity with context semantic analysis, *Information Systems* **80**, pp. 136–147, Elsevier.

Bhyravajjula, Sriharsh, Ujwal Narayan, and Manish Shrivastava (2022), Marcus: An event-centric nlp pipeline that generates character arcs from narratives., *Text2Story@ ECIR*, pp. 67–74.

Bouma, Gosse (2021), Probing for dutch relative pronoun choice, *Computational Linguistics in the Netherlands Journal* **11**, pp. 59–70.

Chen, Chen and Vincent Ng (2014), SinoCoreferencer : An End-to-End Chinese Event Coreference Resolver, *Lrec 2014* pp. 4532–4538. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1144_Paper.pdf.

Chen, Chen and Vincent Ng (2016), Joint Inference over a Lightly Supervised Information Extraction Pipeline: Towards Event Coreference Resolution for Resource-Scarce Languages, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pp. 2913–2920. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12413.

Chen, Zheng and Heng Ji (2010), Graph-based clustering for computational linguistics: A survey, *Proceedings of TextGraphs-5-2010 Workshop on Graph-based Methods for Natural Language Processing*, pp. 1–9.

Choubey, Prafulla Kumar and Ruihong Huang (2017), Event Coreference Resolution by Iteratively Unfolding Inter-dependencies among Events, Association for Computational Linguistics, pp. 2124–2133. http://aclweb.org/anthology/D17-1226.

Chowdhury, Shatadru Roy, Srinka Basu, and Ujjwal Maulik (2022), A survey on event and subevent detection from microblog data towards crisis management, *International Journal of Data Science and Analytics* pp. 1–31, Springer.

Cohen, Jacob (1960), A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20** (1), pp. 37–46, Sage Publications Sage CA: Thousand Oaks, CA.

Cybulska, Agata and Piek Vossen (2014), Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution., *LREC*, pp. 4545–4552.

Cybulska, Agata and Piek Vossen (2015), Translating Granularity of Event Slots into Features for Event Coreference Resolution., *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Association for Computational Linguistics, Denver, Colorado, pp. 1–10. http://aclweb.org/anthology/W15-0801.

Daumé III, Hal (2004), Support vector machines for natural language processing, *Lecture Notes*, Citeseer.

De Langhe, Loic, Orphée De Clercq, and Veronique Hoste (2022a), Constructing a cross-document event coreference corpus for dutch, *Language Resources and Evaluation* pp. 1–30, Springer.

De Langhe, Loic, Orphée De Clercq, and Veronique Hoste (2021), Guidelines for Annotating Events and Event Coreference in Dutch News Articles, *Technical report.*

De Langhe, Loic, Orphée De Clercq, and Veronique Hoste (2022b), Investigating cross-document event coreference for dutch.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), Bertje: A dutch bert model, *arXiv preprint arXiv:1912.09582.*

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), Robbert: a dutch roberta-based language model, *arXiv preprint arXiv:2001.06286.*

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2022), Robbertje: A distilled dutch bert model, *arXiv preprint arXiv:2204.13511.*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805.*

Fillmore, Charles J, Collin F Baker, and Hiroaki Sato (2002), The framenet database and software tools., *LREC.*

Fleiss, Joseph L and Jacob Cohen (1973), The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and psychological measurement* **33** (3), pp. 613–619, Sage Publications Sage CA: Thousand Oaks, CA.

Girish, KK, Jeni Moni, Joel Gee Roy, CP Afreed, S Harikrishnan, and Gokul G Kumar (2022), Extreme event detection and management using twitter data analysis, *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, pp. 917–921.

Glavaš, Goran, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi (2014), HiEve: A Corpus for Extracting Event Hierarchies from News Stories, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (1), pp. 3678–3683. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1023_Paper.pdf.

Gou, Jianping, Baosheng Yu, Stephen J Maybank, and Dacheng Tao (2021), Knowledge distillation: A survey, *International Journal of Computer Vision* **129** (6), pp. 1789–1819, Springer.

Gwet, Kilem Li (2008), Computing inter-rater reliability and its variance in the presence of high agreement, *British Journal of Mathematical and Statistical Psychology* **61** (1), pp. 29–48, Wiley Online Library.

Hinton, Geoffrey, Oriol Vinyals, Jeff Dean, et al. (2015), Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531.*

Hong, Yu, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer (2016), Building a cross-document event-event relation corpus, *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pp. 1–6.

Hoste, Veronique (2005), *Optimization issues in machine learning of coreference resolution*, PhD Thesis, Universiteit Antwerpen. Faculteit Letteren en Wijsbegeerte.

Huang, Lifu et al. (2013), Optimized event storyline generation based on mixture-event-aspect model, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 726–735.

Huang, Yin Jou (2021), Event centric approaches in natural language processing, Kyoto University.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy (2020), Spanbert: Improving pre-training by representing and predicting spans, *Transactions of the Association for Computational Linguistics* **8**, pp. 64–77, MIT Press.

Kingsbury, Paul and Martha Palmer (2003), Propbank: the next level of treebank, *Proceedings of Treebanks and lexical Theories*, Vol. 3, Citeseer.

Lample, Guillaume and Alexis Conneau (2019), Cross-lingual language model pretraining, *arXiv preprint arXiv:1901.07291.*

Lu, Jing and Vincent Ng (2016), Event Coreference Resolution with Multi-Pass Sieves, p. 8.

Lu, Jing and Vincent Ng (2017), Learning antecedent structures for event coreference resolution, *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 113–118.

Lu, Jing and Vincent Ng (2018a), Event Coreference Resolution: A Survey of Two Decades of Research, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, pp. 5479–5486. https://www.ijcai.org/proceedings/2018/773.

Lu, Jing and Vincent Ng (2018b), Event coreference resolution: A survey of two decades of research., *IJCAI*, pp. 5479–5486.

Lu, Jing and Vincent Ng (2021), Conundrums in event coreference resolution: Making sense of the state of the art, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1368–1380.

Luo, Xiaoqiang (2005), On coreference resolution performance metrics, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 25–32.

Miller, George A (1995), Wordnet: a lexical database for english, *Communications of the ACM* **38** (11), pp. 39–41, ACM New York, NY, USA.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Marieke van Erp, Anneleen Schoen, and Chantal van Son (2016), MEANTIME, the NewsReader Multilingual Event and Time Corpus, *Proceedings of the 10th language resources and evaluation conference (LREC 2016)*, European Language Resources Association (ELRA), Portorož, Slovenia, p. 6.

Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy (2015), Overview of TAC KBP 2015 Event Nugget Track, *Kbp Tac 2015* pp. 1–31.

Mitra, Mandar, Amit Singhal, and Chris Buckley (1997), Automatic text summarization by paragraph extraction, *Intelligent Scalable Text Summarization.*

Moosavi, Nafise Sadat and Michael Strube (2016), Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 632–642.

Nguyen, Thien Huu, Adam Meyers, and Ralph Grishman (2016), New york university 2016 system for kbp event nugget: A deep learning approach., *TAC.*

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013), The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch, pp. 219–247.

Ordelman, Roeland J.F., Franciska M.G. de Jong, Adrianus J. van Hessen, and G.H.W. Hondorp (2007), Twnc: a multifaceted dutch news corpus, *ELRA Newsletter.*

O'Gorman, Tim, Kristin Wright-Bettner, and Martha Palmer (2016), Richer event description: Integrating event coreference with temporal, causal and bridging annotation, *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pp. 47–56.

Pohl, Daniela, Abdelhamid Bouchachia, and Hermann Hellwagner (2012), Automatic sub-event detection in emergency management using social media, *Proceedings of the 21st international conference on world wide web*, pp. 683–686.

Polyak, Boris T and Anatoli B Juditsky (1992), Acceleration of stochastic approximation by averaging, *SIAM journal on control and optimization* **30** (4), pp. 838–855, SIAM.

Pradhan, Sameer S., Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla (2007), Unrestricted coreference: Identifying entities and events in ontonotes, *ICSC 2007 International Conference on Semantic Computing* pp. 446–453.

Pustejovsky, James, Jose Castano, Robert Ingria, Roser Saurı, Robert Gaizauskas, Andrea Setzer, and Graham Katz (2003), TimeML: Robust Specification of Event and Temporal Expressions in Text, *New Directions in Question Answering* **3**, pp. 28–34.

Reimers, Nils and Iryna Gurevych (2019), Sentence-bert: Sentence embeddings using siamese bert-networks, *arXiv preprint arXiv:1908.10084*.

Reiter, Nils, Anette Frank, and Oliver Hellwig (2014), An nlp-based cross-document approach to narrative structure discovery, *Literary and Linguistic Computing* **29** (4), pp. 583–605, Oxford University Press.

Rospocher, Marco, Marieke Van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard (2016), Building event-centric knowledge graphs from news, *Journal of Web Semantics* **37**, pp. 132–151, Elsevier.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019), Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108*.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary (2019), Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Leibniz-Institut für Deutsche Sprache.

Vilain, Marc, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995), A model-theoretic coreference scoring scheme, *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Xu, Yige, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang (2020), Improving bert fine-tuning via self-ensemble and self-distillation, *arXiv preprint arXiv:2002.10345*.

Zhou, Zhi-Hua, Jianxin Wu, and Wei Tang (2002), Ensembling neural networks: many could be better than all, *Artificial intelligence* **137** (1-2), pp. 239–263, Elsevier.