Investigating Cross-Document Event Coreference for Dutch

Loic De Langhe, Orphée De Clercq, Veronique Hoste
LT3, Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

In this paper we present baseline results for Event Coreference Resolution (ECR) in Dutch using gold-standard (i.e non-predicted) event mentions. A newly developed benchmark dataset allows us to properly investigate the possibility of creating ECR systems for both within and cross-document coreference. We give an overview of the state of the art for ECR in other languages, as well as a detailed overview of existing ECR resources. Afterwards, we provide a comparative report on our own dataset. We apply a significant number of approaches that have been shown to attain good results for English ECR including feature-based models, monolingual transformer language models and multilingual language models. The best results were obtained using the monolingual BERTje model. Finally, results for all models are thoroughly analysed and visualised, as to provide insight into the inner workings of ECR and long-distance semantic NLP tasks in general.

1 Introduction

With the focus of Natural Language Processing (NLP) applications shifting more towards large-scale discourse-oriented tasks, there is a growing need for systems that can model language not only at the word level, but which can also capture long-distance semantic dependencies. Event coreference resolution (ECR) has been one of the domains within NLP that has been at the forefront of this transition. The ambition in ECR is to determine whether or not two textual events refer to the same real-life or fictional event. For this to be true, two candidate event mentions should have the same event trigger, which denotes the action performed, and non-contradicting event arguments, which include spatio-temporal information and possible participants to the event. Consider the examples below that were taken from two different Dutch (Flemish) newspaper articles:

1. Frankrijk Verslaat België in de halve finales van de FIFA wereldbeker voetbal *EN: France beats Belgium in the semi-final of the FIFA world cup.*

2. België verliest halve finale *EN: Belgium loses semi-final.*

For a human reader, it is perfectly obvious that these two events refer to the same real-world occurrence, even though the event pair has different triggers and the second event mention has no additional argument information. For algorithms, however, this is no trivial task because event mentions are often spread throughout a text, which requires insight into the general discourse structure rather than the local context alone. In addition to this, it is also paramount that coreference can be performed not only at a within-document level, but also across different documents, dramatically increasing the search space of potential event antecedents. In the latter case, the task is possibly further complicated by the fact that the context, target audience and register can inevitably vary between documents. Other than the inherent complexity of creating a (language) model that can accurately resolve long-distance semantic dependencies, ECR research is hindered by the lack of available resources, especially for traditionally lower-resourced languages. In addition to this, data is generally sparse and creating new fully-annotated resources takes considerable time and effort. Despite the challenges, it is important to thoroughly investigate the potential of event coreference resolution because it is a key component of many practical applications such as content-based news recommendation, question answering and contradiction detection. Moreover, researching the links between individual entities and events in texts is paramount to a good understanding of natural language in general.

In this paper, we present baseline results for the task of event coreference resolution on the first

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 88–98
Gyeongju, Republic of Korea, October 16–17, 2022.

large-scale Dutch cross-document ECR corpus using gold-standard event mentions. This new resource allows us to investigate the possibility of performing ECR on languages other than English and to potentially create an effective end-to-end event coreference resolution system for Dutch in the future. As previous research has exclusively focused on English, Chinese and Spanish, we aim to adapt existing methodologies for those languages and apply them to this Dutch dataset. We hope that this paper, combined with the first large-scale Dutch corpus can be an incentive for future research into event coreference resolution and discourse-oriented tasks for both Dutch and lower-resourced languages in general.

2 Related work

2.1 Resources

Existing annotated datasets for event coreference resolution are scarce even for languages that are generally well-resourced. In this section, we briefly discuss the most widely used corpora for event coreference resolution, detailing strengths and weaknesses for each of them.

Among the most popular of event coreference corpora is the EventCorefBank+ (ECB+) dataset (Cybulska and Vossen, 2014b) which is itself an extension of the earlier EventCorefBank (ECB) (Bejan and Harabagiu, 2010) corpus. ECB+ includes both within and cross-document event coreference annotations, as well as extensive annotation of event arguments and linguistic properties. In addition to this, this dataset contains events belonging to a variety of topics, such as financial news, geopolitical events and local news stories, making it particularly fit for simulation of real-word practical scenarios. Another large-scale resource which is often used as a benchmark dataset for ECR is the OntoNotes corpus (Pradhan et al., 2007). In this corpus both entity and event coreference has been annotated in a within-document fashion. However, a notable caveat for this corpus is that no distinction has been made between entities and events in the annotation. Another group of datasets often used to train and evaluate ECR systems are the TAC KBP corpora (Mitamura et al., 2015). This resource is strictly limited to within-document coreference and events are only annotated when belonging to a more strict event typology. In addition to its English component, the corpus includes a more limited set of Chinese and Spanish documents for event coreference resolution. The last large-scale cross-document corpus for English that should be mentioned is the more recently created WEC-Eng dataset (Eirew et al., 2021), which adopts a novel method of leveraging data where both event mentions and coreference links between events are not restricted to pre-defined topics. A final ECR corpus that should be mentioned is the Newsreader Meantime dataset (Minard et al., 2016). While this corpus is very limited in size, it has extensive event annotations and includes both within and cross-document coreference. Moreover, it includes documents in English, Italian, Dutch and Spanish. However, the articles in Dutch, Spanish and Italian were machine-translated from the original English news articles which is arguably a non-optimal way of collecting data. Table 1 presents an overview of the relative size and most important characteristics of the aforementioned corpora.

Corpus	#Documents	Coref	Languages
OntoNotes	600	CD	EN
TAC KBP	1000, 800, 400	WD	EN, SP, CH
ECB	480	CD	EN
ECB+	982	CD	EN
Newsreader Meantime	120	CD	EN, DU, IT, SP

Table 1: Overview of the most popular corpora annotated with event coreference, both within-document (WD) and cross-document (CD).

2.2 Methodology

Following standards set by research in entity coreference resolution (Rahman and Ng, 2009), event coreference resolvers often take the form of mention-pair models. The mention-pair approach reduces the task to a binary decision problem in which two candidate events are presented to a classification algorithm. The task is then to determine whether or not the two candidates refer to the same event, where the event can be either a fictitious or real-world event. The classification algorithms selected for mention-pair models are often traditional feature-based machine-learning approaches such as support vector machines (Chen and Ng, 2014), decision trees (Cybulska and Vossen, 2015) and, more recently, deep neural networks (Nguyen et al., 2016) and transformer architectures. Note that after this pairwise task, an additional step is needed to construct coreference clusters.

A shortcoming of the mention-pair models is their inability to consider an event coreference chain consisting of more than two events collectively, as the algorithm boils down to pairwise deci-

sions and not to a decision based on the document as a whole. A possible solution to this conundrum can be found in the mention-ranking models. In these systems, all possible candidate antecedents are considered simultaneously and a probability distribution over the most likely partition within a given document is generated (Lu and Ng, 2017b).

Note that the algorithms discussed above strictly require events as input. While this is not an issue in optimal settings where all gold-standard events are known to us, it does raise some problems when trying to apply event coreference resolution in real-life practical applications on unseen data. In this case, events first need to be extracted and analyzed in order to make an accurate prediction regarding a possible coreferential relation. To this purpose, recent work in ECR research has primarily focused on end-to-end systems (Lu and Ng, 2018a). These systems often include a mention detection component, which extracts the events from raw text, a component that identifies spatio-temporal information of the event and finally a component that identifies coreference relationships between entities partaking in the event, as logically, knowing which entities participate in the events is a huge step towards resolving the coreference of the events themselves. Until recently, this was primarily done through pipeline architectures, where one component feeds directly into the next one (Choubey and Huang, 2017). While effective, pipelines are inherently prone to error propagation, which complicates matters enormously. In order to circumvent this problem, interest in joint-modelling techniques for end-to-end coreference resolution has been steadily growing (Lu and Ng, 2018a). Joint models have typically focused on performing joint inference over the output of the various tasks contained within the pipeline through the use of integer linear programming (Chen and Ng, 2016) and Markov Logic Networks (Lu and Ng, 2016), where manually defined constraints are used in order for the individual components to improve one another. Alternatively, joint-learning techniques in which interactions between upstream tasks are modelled have also been applied successfully using both traditional probabilistic methods (Lu and Ng, 2017a) and deep learning (Lu et al., 2022).

Finally and perhaps most importantly, advancements in transformer-based language architectures (Vaswani et al., 2017) have had a major impact on both entity and event coreference alike. Transformer-based language embeddings are often used to extend and improve existing ECR systems for both within -and cross-document settings (Cattan et al., 2021a). Additionally, span-based models have been shown to provide massive improvements when integrated in earlier entity pipelines (Joshi et al., 2020). Similarly, span-based architectures attain state-of-the-art results on the benchmark KBP2017 for event coreference resolution, both in pipeline (46,2 F1) and in joint settings (48,0 F1) (Lu and Ng, 2021).

3 The ENCORE Corpus

The recently developed ENCORE corpus (De Langhe et al., 2022) provides us with the opportunity to lay the groundwork for cross-document event coreference in Dutch. As far as we know, the ENCORE corpus is the largest annotated cross-document event coreference corpus in existence, not only for the Dutch language, but also compared to existing English language corpora.

Data for the ENCORE corpus was sourced from a large collection of unannotated Dutch (Flemish) news texts (De Clercq, Orphée and De Bruyne, Luna and Hoste, Veronique, 2020) collected from a variety of online sources during a one-year period. As event coreference data is notoriously sparse, additional measures were taken in order to maximise the total number of coreference links i.e events referring to one another in the corpus. First, named entities were extracted from each of the documents in the aforementioned larger collection. Second, articles containing a given number (>5) of unique overlapping entities were grouped together in so-called "event clusters", as it was hypothesized that news texts containing a high number of overlapping named entities are much more likely to contain overlapping events as well. Finally, the resulting event clusters were (manually) pruned in order to avoid duplicate and irrelevant news texts. After this process, the corpus totalled 91 event clusters, each containing on average 13 - 14 unique documents.

Table 2 provides a side-by-side view of the ENCORE corpus and comparable event coreference corpora. As the ECB+ corpus was considered to be the largest ECR corpus in existence, the newly created corpus is larger than the corpora presented in Table 1, both in terms of actual size (number of documents) and in terms of the total number of event clusters.

Corpus	Doc.	Topics	Events
ECB (ENG)	482	43	1744
ECB+ (ENG)	982	43	14884
MeanTime (DU)	120	4	1510
ENCORE (DU)	**1115**	**91**	**15407**

Table 2: Comparison of various event coreference corpora at the level of the number of annotated documents, topics and events.

3.1 Event annotation

Annotating event data can be a complicated task in itself. There exists a multitude of annotation schemes ranging from concise, in which the main verb alone is considered to be representative of the entire event (NIST, 2005), to extensive fine-grained annotation where participant information, (extra-) linguistic properties and spatio-temporal cues of the events are all annotated. Since the explicit goal of the corpus is to perform event coreference resolution, a rich annotation style was employed based on the aforementioned ECB+ corpus (Cybulska and Vossen, 2014a). Concretely, the ECB+ guidelines specify four types of event arguments: EVENT-PARTICIPANT, EVENT-TIME, EVENT-LOCATION and EVENT-ACTION that are (if present) annotated for each event. The example below illustrates how an event is typically annotated in the ENCORE corpus.

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

3.2 Coreference annotation

Coreference between events was annotated, both on the within and cross-document level. Events were considered to be coreferent when three criteria were fulfilled: events should occur at the same time (i), in the same place (ii) and the same participants should be involved (iii). Note that the cross-document annotation of event coreference was limited to documents within one event cluster, as manual coreference annotation over the entire corpus would be an almost insurmountable task. Subtypes of coreference were also annotated for events. A distinction was made here between identity relations and part-whole relations. Traditionally, studies in event coreference resolution have exclusively focused on the identity relation between events, even though a solid case can be made that other relationships exist between textual events. For instance, one can argue that, given the proper context, an event such as *the opening speech* is a part of *the Oscars ceremony*, a nuance that is currently overlooked in, to the best of our knowledge, virtually all ECR research.

4 Experimental Setup

We present baseline results using gold event mentions on the Dutch ENCORE corpus. The goal is to correctly reconstruct coreference chains for the events in the documents based on the gold mentions and any spatio-temporal, participant and (meta) linguistic information that was annotated. We report experimental results for both a within and a cross-document coreference resolution task using a variety of algorithms that have shown to perform well throughout the years. The algorithms used for this set of baseline experiments includes both traditional feature-based mention-pair and mention-ranking systems, as well as newer monolingual and multilingual transformer models.

4.1 Feature-based approaches

As there is no earlier work regarding Dutch event coreference resolution, we use a combination of traditional Dutch entity coreference features as well as a set of well-performing language-independent features that have been used previously for English and Chinese ECR. For both the mention-pair and mention-ranking approach, features are identical and have been generated for each possible pair of events.

Lexical-semantic features mostly compare events based on outward similarity. Both string-matching and string-similarity features are known to be important for event coreference resolution, despite their apparent simplicity (Lu and Ng, 2018b). Among the lexical features we apply the exact string match of both event action and span for each pair, as well as POS matching of the event actions. In addition, we add a hoist of string similarity features for both spans and actions in event pairs including Levenshtein distance, Dice coefficient, Jaro-Winkler coefficient and cosine distance based on FastText embeddings (Bojanowski et al.,

2017). Finally, synonym-hypernym relations of the event actions are also extracted.

Discourse features are another category of regularly used characteristics for event coreference resolution. These features include sentence distance between two events, event distance and encoded token distance. In addition, we include matching of (meta) linguistic event aspects that have been specifically annotated in the corpus such as the events' prominence, realis and sentiment.

Logical and constraining features are entirely reliant on successful completion of upstream tasks in the ECR pipeline. Among others, possible conflict of event times and locations are modelled through these features, as well as the possible coreference between event participants. Finally, following earlier success with applying distance-based features for event arguments (Lu and Ng, 2018b), we also include the use of Dice coefficient and FastText-based cosine distance between event locations, times and participant head words.

4.1.1 Feature-based Mention-Pair

We use the popular XGBoost algorithm (Chen et al., 2015) for the pairwise classification of event pairs and then reconstruct the event coreference chains from those pairs using agglomerative clustering. The model is trained using 10-fold cross-validation and extensive hyperparameter tuning for both the within and cross-document setting.

4.1.2 Feature-based Mention-Ranking

We use an adapted implementation of the mention-ranking algorithm used in Lu and Ng (2017c). The base algorithm first generates all possible partitions for the events in a given document. In the partition, each event slot can either be the start of a new coreference chain, or can designate the possible anaphora of said event. Concretely, this means that a document with three events (*event 1*, *event 2*, *event 3*) has 6 possible partitions, as shown in Figure 1. In this setting, each event can either be the start of a new coreference chain (i.e *NEW*) or refer to each of its possible antecedents, which would indicate that these events corefer. Logically, some partitions will, in practice, result in the same output coreference chain e.g. [NEW, E1, E1] and [NEW, E1, E2], where *event 1* starts a new coreference chain and both *event 2* and *event 3* refer to that real-life event.

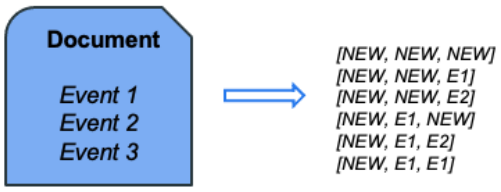


Figure 1: Generated partitions for the mention-ranking model

The original log-linear model defines a distribution over all possible partition vectors a given document d, weights w and feature vector f.

$$p(a|d;w) \propto exp(\sum_{i=1}^{n} w \cdot f(i, a_i, d)) \quad (1)$$

The authors include a task-specific loss function in their original implementation where the weighted sum of three different error types is taken into account.

$$p(a|d;w)' \propto p(a|d;w)l(a, C_d^*) \quad (2)$$

The augmentation for the task-specific loss function $l(a, C_d^*)$ includes the number of non-anaphoric mentions misclassified as anaphoric, anaphoric mentions misclassified as non-anaphoric and incorrectly resolved anaphora based on the gold-standard document partitions C_d^*. Each error type is individually weighed by a floating point parameter, optimized during the training process. For this set of baseline experiments, we test the system using both a general and task-specific loss function and learn the weights that maximise the conditional likelihood of our training data:

$$L(\Theta) = \sum_{d=1}^{t} log \sum_{a \in A(C_d^*)} p(a|d;w)' + \lambda||\Theta||_1 \quad (3)$$

In addition to the two base algorithms described above, we make a series of modifications, as described in the paragraphs below.

For the within-document version, instead of selecting the most likely document partition for each of the documents, we implement a k-majority voting system. We found that in many cases some of the top predicted partitions would result in the correct output chain. By issuing a hard majority vote over the top k predictions we can use this to our advantage and optimally use the probability mass assigned to the resulting output chain.

Additionally, we present two versions of the cross-document algorithm. The original algorithm did not account for the possibility of cross-document coreference and while one can simply concatenate all documents in a given event cluster and generate all cluster partitions similarly to the document partitions, this does pose some scaling issues. First, generating the number of total possible partitions increases almost exponentially when the number of events within a cluster increases, potentially causing memory issues. Second, generating all possible event cluster partitions creates an artificial sparsity problem since, as stated before, the number of total partitions is large. Despite this, the number of correct partitions remains relatively low. While generating all cluster partitions is still feasible with this dataset, we believe that this would be a significant problem in end-to-end settings. We therefore propose an alternative way of performing cross-document coreference using pairwise chain classification. We first determine and extract the coreference chains using the within-document algorithm, then we generate word2vec embeddings for each of the event mentions and average them. Finally, we apply a simple feedforward neural network to determine pairwise coreference between chain representations and reconstruct the final chains using the same clustering algorithm mentioned in section 4.1.1. For the final evaluation, we present cross-document scores using both concatenated cluster partitions (MR) and pairwise document coreference chains (MR Embedding).

4.2 Transformer-based approaches

Fine-tuned transformer language models attain state-of-the-art performance on a multitude of NLP tasks and event coreference resolution is no exception in this regard. The best results are obtained using span-based transformers such as modified versions of SpanBERT-base and SpanBERT-large (Lu and Ng, 2021). It should be noted, however, that results for ECR are still comparatively low (SOTA F1 is 58 on KBP2017).

As no span-based models are available for Dutch, we opt for a series of transformer-based mention-pair models based on the Dutch language models BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020). These models 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 on this dataset. RobBERT on the other hand was pre-trained on 6.6B tokens of Commoncrawl webdata (Suárez et al., 2019). However, 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 on our dataset.

Finally, we also finetune the monolingual RobBERTje model for this task. The RobBERTje models include a series of distilled languague 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 previous language models on coreference-based tasks such as die-dat disambiguation (Allein et al., 2020) and pronoun prediction (Delobelle et al., 2022). In addition to these three monolingual models, we finetune the multilingual models XLM-ROBERTa (Lample and Conneau, 2019) and multilingual BERT (mBERT) (Devlin et al., 2018), as they both contain a substantial amount of Dutch data and have been shown to be quite effective at a number of Dutch NLP tasks (Bouma, 2021).

5 Evaluation

5.1 Evaluation metrics for coreference

Evaluating coreference, much like any cluster-based task, can be a complex affair. Many different evaluation metrics have been proposed throughout the years with some being more robust, while others provide counter-intuitive results in certain situations. Common practice is to evaluate coreference systems by computing the average F1-score of 3 metrics in particular: MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). In addition to this, we also report evaluation using the recently developed LEA metric, a link-based evaluation method that has shown to often produce reliable and highly interpretable results (Moosavi and Strube, 2016). It must be noted that in our evaluation we exclude any singleton event mention, i.e. events that are predicted to form a coreference

	CONLL	LEA
MP XGBoost	0.36	0.21
MR_{base}	0.39	0.25
$MR_{task-specific}$	0.42	0.26
MR $Embedding_{base}$	/	/
MR $Embedding_{task-specific}$	/	/
MP BERTje	**0.52**	**0.33**
MP RobBERT	0.49	0.29
MP RobBERTje	0.48	0.29
MP XLM-RoBERTa	0.17	0.11
MP mBERT	0.14	0.08

(a) Results for within-document ECR

	CONLL	LEA
MP XGBOOST	0.37	0.23
MR_{base}	0.35	0.22
$MR_{task-specific}$	0.38	0.25
MR $Embedding_{base}$	0.36	0.24
MR $Embedding_{task-specific}$	0.40	0.28
MP BERTje	**0.59**	**0.39**
MP RobBERT	0.56	0.38
MP RobBERTje	0.54	0.35
MP XLM-RoBERTa	0.23	0.14
MP mBERT	0.19	0.10

(b) Results for cross-document ECR

Table 3: Results of the baseline ECR experiments in the within (a) and cross-document (b) setting for both the Mention-Pair (MP) and Mention-Ranking (MR) paradigms. Naturally, the Mention-ranking algorithm using chain embeddings is not applicable to the within-document setting.

chain of size one. While the inclusion of singleton clusters can be useful for the evaluation of joint and pipeline systems, it has been shown that singletons can artificially inflate certain metrics. B3 and CEAF are particularly prone to this, but recent work has revealed that also the LEA metric can be distorted by it to some extent (Poot and van Cranenburgh, 2020; Cattan et al., 2021b).

5.2 Results

Tables 3a and 3b show results for the within and cross-document respectively. These are fully in line and proportional to similar research for English and Chinese ECR (Lu and Ng, 2018b). Monolingual transformer language models such as BERTje (0.59 F1) and RobBERT (0.56 F1) produce by far the best results, followed by feature-based mention-ranking (0.40 F1) and mention-pair (0.37 F1) models respectively. Somewhat surprisingly, multilingual transformer models such as XLM-RoBERTa (0.23 F1) and mBERT (0.19 F1) perform rather poorly, especially when considering their potential when it came to other multilingual NLP problems (Li et al., 2021). Finally, we also notice a slight increase in performance for almost all models when comparing the within-document trial to the cross-document setting.

5.3 Analysis and discussion

5.3.1 Feature-based models

Despite the discrepancy in performance between transformer models and more traditional approaches, the inclusion of feature-based models can still be useful going forward, as hybrid models combining transformer-based embeddings with traditional features and encoding feature information within transformer architectures have shown to provide promising results for many NLP applications (van Cranenburgh et al., 2021). In order for such an approach to be explored in closer detail it is important to know which features can exactly be useful.

We explore feature importance for the XGBoost algorithm by calculating the amount that each feature improves the overall performance for each decision tree weighted by the number of observations the feature node is responsible for. The final score for each individual feature is then determined by averaging the aforementioned per-tree score over all trees in the model. For the log-linear mention ranking algorithm we study which feature coefficients it employed in order to determine the weight of each feature in the classification decision. Figures 2a and 2b report feature importance for the 10 most important features in the used mention-pair and mention-ranking models, respectively. The most important features were fairly consistent for the mention-pair and mention-ranking approaches respectively. Our observations generally confirm earlier research in the sense that outward (Dice coefficient) and lexical similarity (cosine similarity) between the two events are paramount when it comes to resolving coreference between them. For the cross-document setting specifically, argument-constraining features also seem to have an (minimal) impact on the task, while discourse-based features seem to have no real contribution.

(a) Top 10 features Mention-Pair

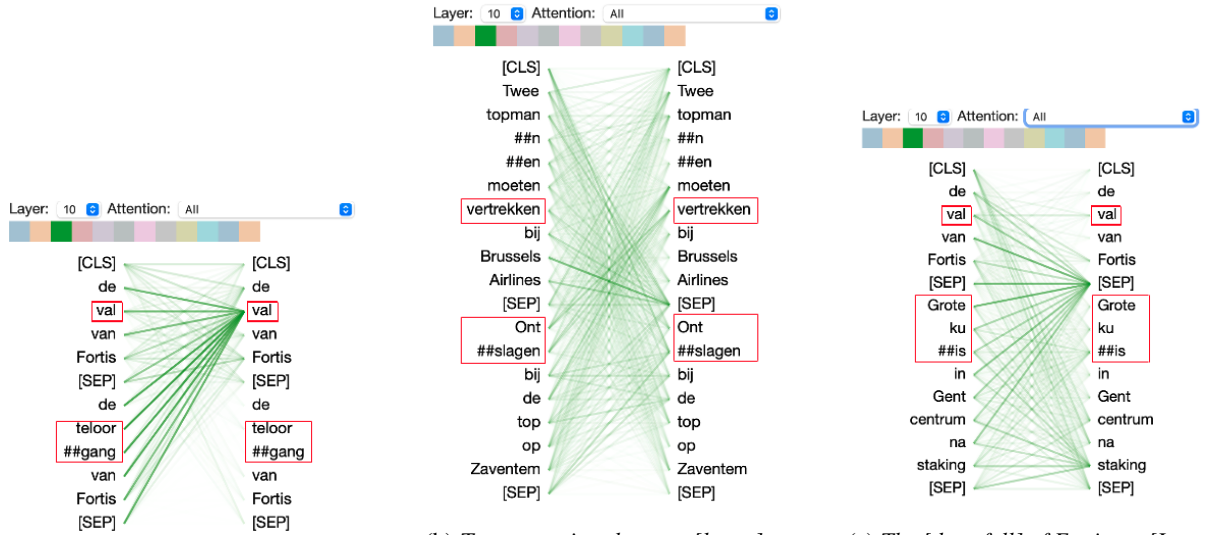
(b) Top 10 features Mention-Ranking

Figure 2: Feature Importances for the Mention-pair and Mention-ranking algorithms

5.3.2 Transformer-based models

As could be observed in Table 3, BERTje performs best. This is most likely due to the thematic overlap of the training corpus (news) and the ENCORE dataset, as well as the fact that the data tends to be less fragmented than RobBERT's. As stated before, successful event coreference resolution is mainly dependent on successfully modelling long-distance semantic dependencies and RobBERT's training data might not be sufficient. Nonetheless, both models perform well, especially when compared to multilingual models XLM-RoBERTa and mBERT. Intuitively, we assumed the task of cross-document coreference to be more difficult than within-document coreference, however, when looking at the results the opposite seems to be true. We assume this is because for the cross-document setting the models had access to significantly increasing training data (1M event pairs compared to 100k for within-document).

Recently, interpretation of transformer-based models has been a hot topic. Vig (2019) and Vig and Belinkov (2019) have revealed that insights regarding syntactic and semantic relations important to a given task can be gained from transformer architectures by visualizing attention heads. We use the Bertviz tool (Vig, 2019) to visualize attention between mention-pairs. We observe that our best performing model (Cross-document BERTje) can consistently model action-to-action relationships for both semantically similar events (figure 3a) and, to a lesser degree, between semantically more distant events (Figure 3b). In addition to this, these aforementioned relationships were absent in the



(a) *The [downfall] of Fortis vs. The [decline] of Fortis*

(b) *Two executives have to [leave] at Brussels Airlines vs. [Dismissals] at the top of Zaventem*

(c) *The [downfall] of Fortis vs. [Large cleanup] in Gent city center after strikes*

Figure 3: Visualisation of the CD BERTje attention heads

same layer and attention head for events that did corefer (Figure 3c).

6 Conclusion

In this paper, we presented baseline results for Dutch ECR on the recently developed ENCORE dataset, which we hope will serve as a benchmark for future investigations into the possibility of developing ECR applications for Dutch. We use a selection of both feature-based and transformer-based models that have shown to work well for English ECR and evaluate these for within-document and cross-document coreference. Our experiments show that monolingual Dutch language models perform best. It should also be noted that multilingual language models perform poorly. This has implications for future work not only in Dutch, but possibly for ECR research in other lower-resourced languages. We also present an analysis of our models, confirming earlier observations that semantic similarity features have a large impact on the task of ECR, while discourse features are less effective. Additionally, by visualising the attention heads we reveal that transformer architectures can specifically model syntactic and semantic relationships that are important in event coreference. In future work we will progress to the development of an end-to-end Dutch ECR system. We will also focus on systems that can accurately model long-distance semantic dependencies, both in context of ECR and language understanding in general.

7 Acknowledgements

This research is part of the ENCORE project which is funded by the Research Foundation–Flanders, Project No. G013820N

References

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

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

Cosmin Bejan 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):1412–1422.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

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

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021a. Cross-document coreference resolution over predicted mentions. *arXiv preprint arXiv:2106.01210*.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021b. Realistic evaluation principles for cross-document coreference resolution. *arXiv preprint arXiv:2106.04192*.

Chen Chen and Vincent Ng. 2014. SinoCoreferencer : An End-to-End Chinese Event Coreference Resolver. *Lrec 2014*, pages 4532–4538.

Chen Chen and Vincent Ng. 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.*, pages 2913–2920.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4).

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event Coreference Resolution by Iteratively Unfolding Inter-dependencies among Events. pages 2124–2133. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ecb+ annotation of events and their coreference. In *Technical Report*. Technical Report NWR-2014-1, VU University Amsterdam.

Agata Cybulska and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552.

Agata Cybulska and Piek Vossen. 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*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

De Clercq, Orphée and De Bruyne, Luna and Hoste, Veronique. 2020. News topic classification as a first step towards diverse news recommendation. *COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS JOURNAL*, 10:37–55.

Loic De Langhe, Orphee De Clercq, and Veronique Hoste. 2022. Constructing a cross-document event coreference corpus for dutch. *Language Recources and Evaluation*, page Accepted for publication.

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

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

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

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

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. Wec: Deriving a large-scale cross-document event coreference dataset from wikipedia. *arXiv preprint arXiv:2104.05022*.

Mandar Joshi, 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*, 8:64–77.

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

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Jing Lu and Vincent Ng. 2016. Event Coreference Resolution with Multi-Pass Sieves. page 8.

Jing Lu and Vincent Ng. 2017a. Joint Learning for Event Coreference Resolution. pages 90–101. Association for Computational Linguistics.

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

Jing Lu and Vincent Ng. 2017c. Learning Antecedent Structures for Event Coreference Resolution. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pages 113–118. IEEE.

Jing Lu and Vincent Ng. 2018a. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486.

Jing Lu and Vincent Ng. 2018b. Event Coreference Resolution: A Survey of Two Decades of Research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5479–5486, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Jing Lu and Vincent Ng. 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*, pages 1368–1380.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. 2022. End-to-end neural event coreference resolution. *Artificial Intelligence*, 303:103632.

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

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

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

Nafise Sadat Moosavi and Michael Strube. 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)*, pages 632–642.

Thien Huu Nguyen, Adam Meyers, and Ralph Grishman. 2016. New York University 2016 System for KBP Event Nugget: A Deep Learning Approach. *Text Analysis Conference*, page 7.

NIST. 2005. The ACE 2005 (ACE 05) Evaluation Plan.

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

Roeland J.F. Ordelman, Franciska M.G. de Jong, Adrianus J. van Hessen, and G.H.W. Hondorp. 2007. Twnc: a multifaceted dutch news corpus. *ELRA Newsletter*, 12(3-4).

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in dutch novels and news. *arXiv preprint arXiv:2011.01615*.

Sameer S. Pradhan, 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*, pages 446–453.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 968–977.

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

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

Andreas van Cranenburgh, Esther Ploeger, Frank van den Berg, and Remi Thüss. 2021. A hybrid rule-based and neural coreference resolution system with an evaluation on dutch literature. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–56.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

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