# What Does BERT actually Learn about Event Coreference? Probing Structural Information in a Fine-Tuned Dutch Language Model

**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 evaluate a fine-tuned BERT model's performance on a set of auxiliary probe tasks to gauge whether the model can indirectly encode discourse properties. The focus is on structural properties that have proven important predictors in feature-based Event Coreference Resolution (ECR). We demonstrate that fine-tuning a language model for ECR also increases performance for event prominence and sentiment matching tasks. This contradicts earlier work where coreference models seemed unable to encode any sort of significant structural or discourse information.

## 1 Introduction

The advent of Large Language Models (LLMs) has drastically improved performance in the field of Natural Language Processing (NLP) on a large variety of tasks that require thorough syntactic and semantic knowledge (Tenney et al., 2019; Koroteev, 2021). However, discourse-based tasks, which typically require a deeper understanding of long-distance semantic relationships and dependencies within a given text, remain a tough nut to crack. One of such tasks, Event Coreference Resolution (ECR), aims to determine whether or not two textual events refer to the same real-life or fictional event. While transformer-based architectures have been moderately successful in tackling this problem (Lu and Ng, 2021; Joshi et al., 2020), much work remains to be done, especially in lower-resourced language domains. Consider the two examples below, which have been taken from a collection of 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.*

Determining that the examples 1 and 2 refer in fact to the same real-world event is fairly straightforward for human readers, owing to their extra-linguistic knowledge. For LLMs however, this task is far from trivial and the mechanisms supporting classification decisions for ECR are currently not well understood. Recent research has suggested that the classification of coreferring mentions in LLMs is entirely dependent on the degree of outward lexical similarity of two candidate events (De Langhe et al., 2023). If true, this is problematic because lexical similarity does not automatically imply a coreferential relation, as illustrated in Examples 3 and 4 below.

3. De Franse president Macron ontmoette de Amerikaanse president voor de eerste keer vandaag *EN: The French president Macron met with the American president for the first time today*

4. Frans President Sarkozy ontmoette de Amerikaanse president *EN: French President Sarkozy met de American president*

Given the high degree of similarity between both examples, most existing classifiers would detect a coreferential relation between the events, despite the fact that they refer to two entirely separate real-world events. Interestingly, earlier work on feature-based classifiers for ECR has shown that discourse and meta-linguistic information surrounding an event are in fact important, to some degree, for the classification of coreference (Lu and Ng, 2018). In this paper, we will devise a series of linguistic probes in order to gauge a Dutch transformer-based coreference model's understanding of certain discourse and meta-linguistic event traits that have been shown to be important for within-document ECR (De Langhe et al., 2022c; Lu and Ng, 2018). Currently, it is assumed that this type of information is implicitly encoded into the transformer's



*The Fourth Workshop on Insights from Negative Results in NLP*, pages 103–108

contextual embeddings, but with this paper we intend to verify this. We believe that if these models do not encode this information, this opens up many possibilities towards extending current models. Moreover, it will allow to further boost our understanding of the linguistic mechanisms behind event coreference.

## 2 Related Work

### 2.1 Linguistic Probing

In recent years, interpretability and explainability of LLMs have been researched through the use of linguistic probes (Conneau et al., 2018). By freezing model weights and training a classifier on a linguistic task such as part-of-speech tagging, subject verb agreement or syntax tree reconstruction, the presence or absence of such basic linguistic capabilities can be evaluated within a model (Adi et al., 2016). Through the use of linguistic probes it has been demonstratively shown that transformer-based encoders such as BERT (Devlin et al., 2018) can successfully encode a hoist of fine-grained syntactic and semantic information (Jawahar et al., 2019). Additionally, research has also been done on the probing of fine-tuned LLMs with applications in conversational recommendation (Penha and Hauff, 2020), reading comprehension (Cai et al., 2020) and question-answering (Van Aken et al., 2019) showing that task-specific knowledge is encoded in such models to a certain degree.

### 2.2 Event Coreference Resolution

There exist several paradigms within ECR research. First, mention-pair approaches reduce the task to a binary decision problem in which two candidate events are presented to a classifier, which has to determine whether or not the two candidates refer to the same event. Past studies often focused on coreference resolution through the use of decision trees (Cybulska and Vossen, 2015), support vector machines (Chen et al., 2015) and standard deep neural networks (Nguyen et al., 2016). More recent work is marked by the use of LLMs and transformer encoders (Cattan et al., 2021a,b), with span-based architectures attaining the best overall results (Joshi et al., 2020; Lu and Ng, 2021). Mention-ranking approaches constitute another paradigm within ECR, in which 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, 2017). Other than the dominant mention-pair and mention-ranking paradigms, studies have also focused on rule-based methods such as multi-pass sieves (Lu and Ng, 2016) and statistical approaches such as Integer Linear Programming (ILP) (Chen and Ng, 2016) and Markov Logic Networks (Lu et al., 2016).

## 3 Experimental Setup

In our experiments we aim to evaluate a fine-tuned BERT model's performance on a set of auxiliary probe tasks in order to gauge whether the model can indirectly encode discourse properties that have proven important predictors in feature-based ECR.

### 3.1 Data

Our data consists of the Dutch ENCORE corpus (De Langhe et al., 2022a), which includes 15,407 events spread over 1,015 documents that were sourced from a Dutch newspaper article collection (Vermeulen, 2018). The corpus is comparable in size to most large-scale English-language ECR datasets. It includes event coreference annotation on both the within- and cross-document level and meta-linguistic information such as the event's prominence (is it a main event or does it provide background information), realis (does the event happen with certainty) and implicit sentiment (positive/negative/neutral). For our probing experiments, we adhere to an identical split of the data as in the original model paper (De Langhe et al., 2022c). We reserve 85% of data for fine-tuning (70% for training and 15% for development) and use the remaining 15% of data for our probing experiments.

### 3.2 Coreference Resolution Model

The ECR model consists of the fine-tuned Dutch BERT model BERTje (de Vries et al., 2019). While this BERTje model has been outperformed by Dutch RobBERTa-based models on most standard NLP tasks (Delobelle et al., 2020, 2022), it is still the model of choice for discourse-type tasks such as coreference resolution, which often require the encoding of long-range semantic and syntactic information (De Langhe et al., 2022c).

As explained in Section 2 there exist two widely used paradigms within the domain of event coreference resolution. For our model, we opt for a mention-pair approach which has demonstratively better results compared to other existing methods

(Lu and Ng, 2018, 2021). Concretely, we obtain pairwise scores for each pair of event mentions in the dataset. First, each possible within-document event pair in the data is encoded by concatenating and tokenizing them and by subsequently feeding them to the BERTje encoder. A special *[SEP]* token is inserted between the two event mentions to indicate where one ends and the other begins. We use the token representation of the classification token *[CLS]* as the aggregate embedding of each event pair, which is subsequently passed to a softmax-activated classification function. Finally, the results of the binary text pair classification are passed through a clustering algorithm in order to obtain output in the form of coreference chains.

### 3.3 Auxiliary Probe Tasks

We define a set of pairwise probes, in which we generate an aggregate embedding of each event pair (as described in Section 3.2) and try to predict whether or not each event mention shares certain structural and discourse properties. The same methodology is applied to the non fine-tuned BERTje language model (de Vries et al., 2019) to serve as a comparable baseline to our coreference model. For the probes we implement the probe classifier as a 2-layer feed-forward network with ReLU activations and layer Normalization (Ba et al., 2016):

$$h_0 = [CLS]$$

$$h_1 = LayerNorm(ReLU(W_1 h_0))$$

$$h_1 = LayerNorm(ReLU(W_2 h_1))$$

Moreover, as previous research has revealed that different BERT encoder layers tend to focus on different linguistic properties (Jawahar et al., 2019), we also extract and classify the encodings for each of the encoder's 12 layers in order to gauge whether the same is true for the coreference BERTje model. Additionally, shifts in layer performance could also provide us with valuable information w.r.t the inner workings of ECR in BERT-based models.

#### 3.3.1 Classification Probe

Meta-information, such as an event's *prominence*, *realis* and *sentiment* (see Section 3.1), can implicitly aid towards the classification of event coreference. With this set of probe tasks, we aim to test whether or not a BERT-based model can implicitly learn these event properties by being fine-tuned on an ECR dataset. Concretely, we set up this probe as a classification task where the classifier's goal is to determine if two events match in their *Prominence*, *Realis* or *Sentiment*, respectively. Our intuition is that if the shared contextual embedding of the two spans encodes this information it is probably an important aspect of the coreferential relation between the events and could be used as a potentially rewarding avenue for future ECR research.

#### 3.3.2 Regression Probe

Feature-based studies for within-document event coreference have shown that two structural features are typically key in the resolution of event mentions (Lu and Ng, 2018): the sentence distance *SD*, where the distance for events in the same sentence is set to 0, and event distance *ED*, where *ED* is equal to the number of events between the events in the pair when traversing the text. The intuition behind this is fairly straightforward: coreferring event mentions are often grouped closely together, resulting in a low sentence and event distance. This corresponds well with general theories on discourse structure where related concepts are usually found within close proximity of each other, be it on the sentence, paragraph or section level (Hoeken and Van Vliet, 2000; Glasbey, 1994). Ideally, if a BERT-based model were able to encode rudimentary discourse information to some extent it would learn that coreferring events are, on average, grouped closer together than non-coreferring events. We define two regression tasks in which we use the shared contextual embeddings for the event pairs to predict the event and sentence distances between them.

## 4 Results and Discussion

Table 1 shows the macro F1 scores (classification tasks) and Root Mean Squared Error (regression tasks) for each of the pairwise probes based on the models' *[CLS]* tokens in each layer, with the baseline scores in between brackets. Our primary interest is in the results of the final layer, as the model's coreference classification decision is entirely dependent on the output of this layer.

For the classification probe tasks we establish that the fine-tuned model outperforms the baseline pre-trained model in both the prominence and sentiment matching tasks, while showing no improvement when it comes to realis matching. This indicates that by fine-tuning, the BERT model does implicitly learn some basic information regarding document structure and can differentiate between the importance of events within a given document

| Layer | Prominence Match | Realis Match | Sentiment Match |
|---|---|---|---|
| 1 | 0.531 (0.523) | 0.537 (0.530) | 0.570 (0.570) |
| 2 | 0.530 (0.526) | 0.547 (0.488) | 0.578 (0.616) |
| 3 | 0.554 (0.522) | 0.535 (0,523) | 0.629 (0.600) |
| 4 | 0.522 (0.531) | 0.545 (0.536) | 0.594 (0.612) |
| 5 | 0.535 (0.530) | 0.558 (0.566) | 0.599 (0.625) |
| 6 | 0.542 (0.535) | 0.543 (0.542) | 0.633 (0.627) |
| 7 | 0.537 (0.514) | 0.561 (0.562) | 0.625 (0.637) |
| 8 | 0.575 (0.512) | 0.544 (0.562) | 0.630 (0.612) |
| 9 | 0.561 (0.561) | 0.556 (0.567) | 0.640 (0.603) |
| 10 | 0.573 (0.562) | 0.570 (0.578) | 0.629 (0.618) |
| 11 | 0.550 (0.541) | 0.568 (0.588) | 0.681 (0.651) |
| 12 | **0.567** (0.493) | 0.564 (**0.570**) | **0.660** (0.649) |

(a) Macro F1 scores for the classification tasks

| Layer | Sentence Distance (SD) | Event Distance (ED) |
|---|---|---|
| 1 | 28.54 (27.99) | 14.4 (14.37) |
| 2 | 34.58 (23.95) | 15.74 (16.52) |
| 3 | 26.17 (26.06) | 23.33 (20.16) |
| 4 | 23.58 (23.58) | 18.32 (14.4) |
| 5 | 27.45 (23.84) | 16.48 (15.83) |
| 6 | 24.78 (27.42) | 17.65 (16.98) |
| 7 | 27.78 (23.59) | 15.94 (16.04) |
| 8 | 23.65 (29.33) | 17.32 (15.88) |
| 9 | 33.29 (45.03) | 16.74 (15.1) |
| 10 | 28.82 (23.83) | 14.36 (16.65) |
| 11 | 28.41 (27.67) | 15.66 (17.48) |
| 12 | 26.05 (**23.83**) | **14.31** (16.72) |

(b) RMSE results for the regression tasks

Table 1: Layer-by-layer comparison of the pairwise probe tasks, with baseline results in between brackets

and use this information for the classification of coreferential relations between events.

While the improvement in the sentiment task is minor, results for prominence show significant improvement over the baseline, showing that the prominence of two events can be a component to consider for future studies in ECR. Conversely, the realis and sentiment properties seem to be not directly related to the correct classification of coreferential events within this model. To get a more complete picture of the models' layer-by-layer performance we also calculate Spearman's correlation coefficients over different layer performances. Correlation coefficients on the promincence (0.146 & 0.720), realis (0.914 & 0.748) and sentiment (0.637 & 0.851) tasks indicate no significant changes in layer performance for the baseline and fine-tuned models as, overall, for all tasks performance increases towards the higher layers.

For the regression tasks we see that final layer performance improves for the Event Distance task in the fine-tuned model, albeit only slightly. It should be noted, though, that the RSME for both tasks is very high, leading us to believe that no significant knowledge regarding event or sentence distance is encoded within the fine-tuned coreference model. Similarly to the classification tasks we also calculate Spearman's correlation coefficients for the performance on both regression tasks over different layers, showing again no different trends for the ED (0.34 & 0.38) and SD (0 & -0.048) tasks for the baseline and fine-tuned models, respectively. Finally, as the raw RMSE result scores from the pairwise distance probes are hard to interpret without context, we also compare the RMSE for the SD and ED tasks on each layer for both coreferring and non-coreferring mentions to see if the fine-tuned model has implicitly learned something about event and sentence distances in within-document contexts for individual class labels. Table 2 shows that on average the RMSE for coreferring mentions is slightly lower than the RMSE for non-coreferring mentions in both the fine-tuned and baseline models in the ED and SD task for the final layer of both models. While these latter results could indicate that both models intrinsically learn that coreferring mentions tend to me grouped closer together, the overall regression scores remain poor. Ultimately, this leads us to conclude that no significant information regarding the closeness of events within a given text is encoded in either model.

| Model | ED (+) | ED (-) | SD (+) | SD (-) |
|---|---|---|---|---|
| Baseline | 16.62 | 16.85 | 23.50 | 23.87 |
| Coreference Model | 14.02 | 14.96 | 25.87 | 26.07 |

Table 2: Average RMSE for coreferring and non-coreferring event pairs for both regression tasks

## 5 Conclusion

In this paper we devised a set of rudimentary probes to determine if a fine-tuned Dutch BERT event coreference model can learn a set of basic characteristics regarding the nature of coreferential relations. We show that the fine-tuned BERT model can in fact encode a limited number of these properties. This goes against previous findings that event coreference resolution in transformer-based models is entirely based on outward lexical similarity, rather than the proper discourse mechanisms governing coreferential relations in natural language (De Langhe et al., 2022b, 2023). In future research, we aim to further investigate and integrate structural and discourse aspects of coreference in LLMs, which will hopefully lead to more stable, interpretable and better performing ECR models.

## References


Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Jie Cai, Zhengzhou Zhu, Ping Nie, and Qian Liu. 2020. A pairwise probe for understanding bert fine-tuning on machine reading comprehension. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1665–1668.

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. 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.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 167–176.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

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.

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

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2022b. Towards fine (r)-grained identification of event coreference resolution types. *Computational Linguistics in the Netherlands Journal*, 12:183–205.

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

Loic De Langhe, Thierry Desot, Orphée De Clercq, and Veronique Hoste. 2023. A benchmark for dutch end-to-end cross-document event coreference resolution. *Electronics*, 12(4).

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*.

Sheila R Glasbey. 1994. *Event structure in natural language discourse*. Ph.D. thesis, University of Edinburgh.

Hans Hoeken and Mario Van Vliet. 2000. Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics*, 27(4):277–286.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

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.

MV Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

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

Jing Lu and Vincent Ng. 2017. 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. 2018. 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.

Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng. 2016. Joint inference for event coreference resolution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3264–3275.

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

Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 388–397.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832.

Judith Vermeulen. 2018. # newsdna: promoting news diversity: an interdisciplinary investigation into algorithmic design, personalization and the public interest (2018-2022). In *ECREA 2018 pre-conference on Information Diversity and Media Pluralism in the Age of Algorithms*.