Leveraging Structural Discourse Information for Event Coreference Resolution in 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

Structural information is known to be important in resolving coreferential relations. We directly embed discourse structure information (subsection, paragraph and text location) in a transformer-based Dutch event coreference resolution model in order to more explicitly provide it with structural information. Results reveal that integrating this type of knowledge leads to a significant improvement in CONLL F1 for within-document settings (+ 8.6%) and a minor improvement for cross-document settings (+ 1.1%).

1 Introduction

Large Language models (LLMs) and transformer-based architectures have significantly changed the domain of Natural Language Processing (NLP) in recent years. Through pre-training and fine-tuning masked language models (MLMs) such as BERT (Devlin et al., 2018), state-of-the-art results can be obtained for tasks requiring deep semantic or syntactic knowledge such as readability assessment (Imperial, 2021), syntactic parsing (He and Choi, 2019) and conversational question-answering (Staliūnaitė and Iacobacci, 2020). However, despite their apparent dominance over other methods, transformer-based language models are still not the 'one-size-fits-all' solution for a subset of NLP tasks. In particular, discourse-based tasks such as Event Coreference Resolution (ECR) still pose a major challenge. Within ECR, the goal is to determine whether or not two textual events refer to the same real-life or fictional event, as is the case in Examples 1 and 2

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.*

Understanding that these two Examples, which have been sourced from a Dutch newspaper article, refer in fact to the same real-world occurrence is straightforward for human readers. Tasks like these typically require understanding of long-distance semantic relationships and dependencies within a given text, or even across multiple texts. While human readers can take advantage of both their extensive extra-linguistic knowledge and structural awareness of the text, AI algorithms typically do not possess such skills. For transformer-based language models in particular long-distance semantic dependencies throughout texts might pose a particular problem. Because MLM pre-training is typically limited to the immediate (sentence) context, the model is unable to learn these dependencies. Additionally, while models such as ALBERT (Lan et al., 2019) have tried to explicitly integrate textual and discourse structure in transformer-based architectures, these models still tend to focus on immediate local context and not on the discourse as a whole.

These limitations pose significant problems for ECR. Recent work has indicated that the correct classification of coreferential links between events in BERT-based models is primarily dependent on the outward lexical similarity of those events (De Langhe et al., 2023b). While logical in principle, this is highly problematic in cases such as Example 3 and Example 4, as there exist many instances in which two events are lexically similar, but that do not corefer.

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*

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 48–53
July 13-14, 2023 ©2023 Association for Computational Linguistics

Vice versa, non-similar event mentions are not necessarily not in a coreferential relation. Although the latter cases are more exceptional, the overall sparseness of ECR results makes that the bulk of training data often consists of similar, but not-coreferring event mentions. Overall, transformer-based approaches have made significant strides within the field of ECR (Joshi et al., 2020). Nonetheless, the over-reliance on lexical similarity between event mentions might impede further improvement. Interestingly, earlier feature-based studies have shown that integrating certain structural features, such as the proximity of two events in a given text can have a positive effect (Lu and Ng, 2018). We aim to improve an existing Dutch transformer-based ECR model (De Langhe et al., 2022b) by enriching it with structural discourse-level information. The goal of this paper is two-fold. First, it is our ambition to illustrate that concepts rooted in general linguistic theory and fundamental to our own understanding of coreferential relations can also improve the performance of LLMs on this task. Second, we wish to address the gap between ECR studies in the English language domain and those in lower-resourced languages. Currently, there exists very little data or available research for languages other than English. In our experiments we show that including discourse-level information leads to a significant and consistent improvement for within-document ECR models. We also note minor improvements in cross-document contexts.

2 Related Work

There exist two important model paradigms within the domain of ECR. First, mention-ranking approaches focus on finding all possible antecedents for a given event and on generating a ranking of those antecedents based on the likelihood of coreference with the event in question. In Lu and Ng (2017a) a feature-based probabilistic model was used for within-document ECR. The authors show that lexical features such as full or partial overlap of events and cosine similarity between event mentions are among the most important information sources of the model. In addition, they revealed that distance-based features such as the number of sentences between two events also have a noticeable positive effect on the classifier's performance. A second and more important series of models are mention-pair approaches. This method generates all possible event pairs and reduces the classification to a binary decision (coreferring or not-coreferring) between each event pair. Earlier models within this paradigm were entirely feature-based and relied on a series of lexical, structural and logical constraining features. A large variety of classical machine learning algorithms has been tested using the mention-pair paradigm such as 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 has focused on 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). It has to be noted that mention-pair approaches relying on LLMs suffer most from the limitations discussed in Section 1. Therefore, recent studies have attempted, with some success, to integrate insights regarding discourse coherence (Held et al., 2021) and domain-specific document discourse information (Choubey et al., 2020) into existing pipelines. Research for comparatively lower-resourced languages has generally followed the paradigms and methods described above and has focused on languages such as Chinese (Mitamura et al., 2015), Arabic (NIST, 2005) and Dutch (Minard et al., 2016).

3 Experimental setup

3.1 Data

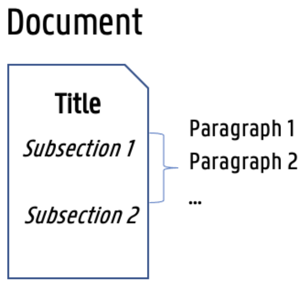

Figure 1: News article structure in the ENCORE dataset

Our data consists of a subset of the Dutch ENCORE corpus (De Langhe et al., 2022a), which in its totality consists of 15,407 annotated events spread over 1,015 documents that were sourced from a collection of Dutch (Flemish) newspaper articles. Coreferential relations between events were annotated at both the within- and cross-document level. For the research presented in this paper a

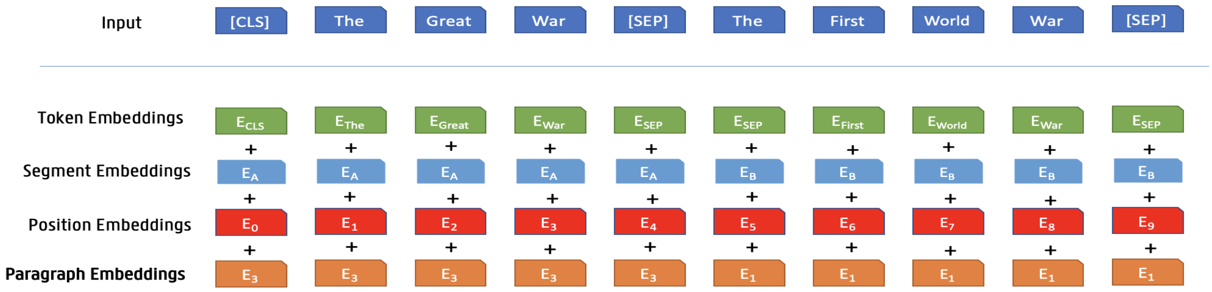

Figure 2: Visualisation of BERT's input embeddings with added discourse-level paragraph embeddings. Event 1 (*The Great War*) is found in paragraph 3 of the document, while potential antecedent Event 2 (*The First World War*) is found in the first paragraph of the document

subset of 8,794 events was selected which all come from documents for which a detailed discourse structure was available. Each of these documents can be broken down into subsections, which in turn consist of a number of paragraphs. Subsections are preceded by a subtitle in bold and typically group together a piece of related information. Figure 1 visualizes the general structure of a document in the ENCORE corpus.

For each event in our dataset we thus know at which subsection and paragraph level it is located within a given article. Additionally, for each event we can derive its *Text Location*, depending on where in the article the event is located. There are 5 possible locations, being the *Article Header*, *Article Subheader*, *Article Introduction*, *Subsection Title* and *Paragraph*.

3.2 Experiments

In our experiments we draw inspiration from earlier work on feature-based models (Lu and Ng, 2018) and integrate specific event proximity and structural information into a state-of-the-art Dutch transformer-based ECR model (De Langhe et al., 2023b). We focus on the usage of the readily available discourse and document-level information which was described in Section 3.1.

3.2.1 Baseline coreference algorithm

The ECR model consists of the fine-tuned Dutch BERT model BERTje (de Vries et al., 2019). For this model, the mention-pair approach has demonstratively better results compared to other existing methods (Lu and Ng, 2018, 2021). Concretely, pairwise scores for each pair of event mentions in the dataset are obtained. First, each possible event pair in the data is encoded by concatenating and tokenizing the two events and by subsequently feeding these to the BERTje encoder. A special *[SEP]* token is inserted between the two event mentions to indicate where one event 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.

3.2.2 Discourse Embeddings Model

In our proposed algorithm discourse-level positional information (paragraph, subsection and text location) is passed to BERT's first encoder layer for each individual event during the fine-tuning process. This is done in a similar way as how the positional, segment and token embeddings are used in the original BERT implementation. We believe that this structural information corresponds well with established general theories on discourse structure where related concepts are usually found within close proximity of each other, be it at the sentence, paragraph or section level (Hoeken and Van Vliet, 2000; Glasbey, 1994). By directly integrating this knowledge into the model it would ideally learn that, overall, coreferring mentions are grouped closer together compared to non-coreferring mentions at the document level. As mentioned before in Section 2, it has already been shown that knowledge regarding the proximity of two events can have a positive impact on the classification decision in feature-based models (Lu and Ng, 2017b, 2018). Earlier research has also shown that currently this knowledge is not encoded by BERT-like models (De Langhe et al., 2023a). These findings led us to believe that this specific knowledge can be leveraged by the model to learn about a fundamental aspect of coreferential relations, as well as

Model	**CONLL**
BERTje$_{Baseline}$	0.432
BERTje$_{Paragraph}$	0.466 ± 0.012
BERTje$_{Subsection}$	0.517 ± 0.008
BERTje$_{Text\ Location}$	0.424 ± 0.032
BERTje$_{Paragraph\ +\ Subsection}$	**0.518** ± 0.009
BERTje$_{Paragraph\ +\ TextLocation}$	0.434 ± 0.028
BERTje$_{Subsection\ +\ TextLocation}$	0.437 ± 0.019
BERTje$_{Paragraph\ +\ Subsection\ +\ Text\ Location}$	0.516 ± 0.022

(a) Results for the Within-document setting

Model	**CONLL**
BERTje$_{Baseline}$	0.519
BERTje$_{Paragraph}$	**0.530** ± 0.014
BERTje$_{Subsection}$	0.517 ± 0.011
BERTje$_{Text\ Location}$	0.460 ± 0.009
BERTje$_{Paragraph\ +\ Subsection}$	0.481 ± 0.032
BERTje$_{Paragraph\ +\ TextLocation}$	0.476 ± 0.048
BERTje$_{Subsection\ +\ TextLocation}$	0.468 ± 0.064
BERTje$_{Paragraph\ +\ Subsection+\ Text\ Location}$	0.472 ± 0.026

(b) Results for the Cross-document setting

Table 1: Subtables report average CONLL results and standard deviation over 3 trials using different random seeds for the discourse-level embeddings in a within-document and cross-document setting respectively. All results in the cross-document table, except the baseline model, automatically include document-level embeddings

break away from its aforementioned dependency on outward lexical similarity of events.

In our implementation, all possible subsection, paragraph and text location levels are encoded using a tokenizer-like mechanism where each level of the respective subsection, paragraph or text location is assigned a unique ID, much like individual tokens are encoded using BERT's own tokenizer. Then, an input embedding matrix of size *A x 768* is randomly initiated for each type of segment information (subsection, paragraph and text location), where *A* is the maximum depth level of a given segment across the dataset and 768 is the standard embedding length for a BERT$_{Base}$ model. Concretely, the maximum depth at the paragraph level is 10 if the longest document across the dataset (in number of paragraphs) has 10 paragraphs. The resulting input embedding matrix will then be of dimension *10x768* and a total of 7680 trainable parameters (*A x 768*) will be added to the model. The first paragraph in each document is encoded by the same unique ID (i.e., 1) and the paragraph-level embedding for each individual token is obtained by embedding the unique ID through the generated input embedding matrix. The same process is followed for the subsection and text location embeddings. Finally, the resulting discourse-level embeddings are summed with the token, segment and positional embeddings to obtain the input for the first encoder layer. As is the case in the original BERT implementation, the weights of our custom discourse input embedding matrices are also optimized during the fine-tuning process. A high-level visualization of the integration of a paragraph embedding can be found in Figure 2. Subsection and text location embeddings are implemented in an analogous manner.

While, intuitively, our proposed structural embeddings would most likely be most useful in a within-document setting, we also include results for a cross-document setting in order to gauge the effectiveness of discourse-level features in those contexts specifically. Our setup for cross-document ECR is identical to the one described above with the notable exception that we add a fourth type of discourse-level embedding, namely a document embedding. When events are found within the same document this embedding is identical.

4 Results and discussion

Tables 1a and 1b show the results of testing various discourse-level embeddings in a within-document and cross-document context, respectively. We evaluate our results using the established CONLL metric, which is an average of 3 commonly used metrics for coreference evaluation: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). We report the average and standard deviations of 3 runs of experiments with different random seeds for the discourse-level input embedding matrices. For the within-document experiments, we see a significant impact on overall performance when including paragraph- and subsection-level information in the fine-tuning process. A combination of paragraph embeddings and subsection embeddings provides the best results. Conversely, we find that the inclusion of Text Location embeddings does not have any noticeable impact on the classification of within-document event coreference.

In the cross-document setting, we find that including structural discourse information does not have a significant impact on classifier performance. While including document and paragraph-level embeddings results in a minor improvement over the baseline coreference model, we find that in general including discourse-specific embeddings does not help with cross-document event coreference.

5 Conclusion and Future Work

In this paper we explored the potential of using discourse-level embeddings in transformer-based models for event coreference resolution. Motivated by general linguistic theory on the overall structure of language we integrate paragraph, subsection and text location information in a Dutch BERT-based mention-pair event coreference algorithm. We find that in within-document contexts the inclusion of discourse-level information has a significant positive effect on overall classifier performance. In particular, the inclusion of paragraph and subsection information consistently leads to better results. Results for the cross-document setting show only minimal improvement over the baseline model. In future work, we aim to further develop structurally informed models for event coreference resolution as well as look into improving the existing cross-document setup.

Acknowledgements

This work was supported by Ghent University under grant BOFGOA2018000601 and by the Research Foundation–Flanders under project grant number FWO.OPR.2020.0014.01.

References

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.

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*.

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.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

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, Orphee De Clercq, and Veronique Hoste. 2023a. What does bert actually learn about event coreference? probing structural information in a fine-tuned dutch language model. *Accepted*.

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

Loic De Langhe, Thierry Desot, Orphée De Clercq, and Veronique Hoste. 2023b. 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*.

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.

Han He and Jinho D Choi. 2019. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with bert. *arXiv preprint arXiv:1908.04943*.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. *arXiv preprint arXiv:2110.05362*.

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.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jing Lu and Vincent Ng. 2017a. 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. 2017b. 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.

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, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76, Denver, Colorado. Association for Computational Linguistics.

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

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

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in roberta, bert and distilbert: a case study on coqa. *arXiv preprint arXiv:2009.08257*.

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*.