Filling in the Gaps: Efficient Event Coreference Resolution using Graph Autoencoder Networks

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

We introduce a novel and efficient method for Event Coreference Resolution (ECR) applied to a lower-resourced language domain. By framing ECR as a graph reconstruction task, we are able to combine deep semantic embeddings with structural coreference chain knowledge to create a parameter-efficient family of Graph Autoencoder models (GAE). Our method significantly outperforms classical mention-pair methods on a large Dutch event coreference corpus in terms of overall score, efficiency and training speed. Additionally, we show that our models are consistently able to classify more difficult coreference links and are far more robust in low-data settings when compared to transformer-based mention-pair coreference algorithms.

1 Introduction

Event coreference resolution (ECR) is a discourse-centered NLP task in which the goal is to determine whether or not two textual events refer to the same real-life or fictional event. While this is a fairly easy task for human readers, it is far more complicated for AI algorithms, which often do not have access to the extra-linguistic knowledge or discourse structure overview that is required to successfully connect these events. Nonetheless ECR, especially when considering cross-documents settings, holds interesting potential for a large variety of practical NLP applications such as summarization (Liu and Lapata, 2019), information extraction (Humphreys et al., 1997) and content-based news recommendation (Vermeulen, 2018).

However, despite the many potential avenues for ECR, the task remains highly understudied for comparatively lower-resourced languages. Furthermore, in spite of significant strides made since the advent of transformer-based coreference systems, a growing number of studies has questioned the effectiveness of such models. It has been suggested that classification decisions are still primarily based on the surface-level lexical similarity between the textual spans of event mentions (Ahmed et al., 2023; De Langhe et al., 2023), while this is far from the only aspect that should be considered in the classification decision. Concretely, in many models coreferential links are assigned between similar mentions even when they are not coreferent, leading to a significant number of false positive classifications, such as between Examples 1 and 2.

1. The French president Macron met with the American president for the first time today

2. French President Sarkozy met the American president

We believe that the fundamental problem with this method stems from the fact that in most cases events are only compared in a pairwise manner and not as part of a larger coreference chain. The evidence that transformer-based coreference resolution is primarily based on superficial similarity leads us to believe that the current pairwise classification paradigm for transformer-based event coreference is highly inefficient, especially for studies in lower-resourced languages where the state of the art still often relies on the costly process of fine-tuning large monolingual BERT-like models (De Langhe et al., 2022b).

In this paper we aim to both address the lack of studies in comparatively lower-resourced languages, as well as the more fundamental concerns w.r.t. the task outlined above. We frame ECR as a graph reconstruction task and introduce a family of graph autoencoder models which consistently outperforms the traditional transformer-based methods on a large Dutch ECR corpus, both in terms of accuracy and efficiency. Additionally, we introduce a language-agnostic model variant which disregards the use of semantic features entirely and even outperforms transformer-based classification in some

Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023), pages 1–7
December 6–7, 2023

situations. Quantitative analysis reveals that the lightweight autoencoder models can consistently classify more difficult mentions (cfr. Examples 1 and 2) and are far more robust in low-data settings compared to traditional mention-pair algorithms.

2 Related Work

2.1 Event Coreference Resolution

The primary paradigm for event coreference resolution takes the form of a binary mention-pair approach. This method generates all possible event pairs and reduces the classification to a binary decision (coreferent or not) between each event pair. 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. In an effort to mitigate these issues some studies have sought to move away from the pairwise computation of coreference by modelling coreference chains as graphs instead. These methods' primary goal is to create a structurally-informed representation of the coreference chains by integrating the overall document (Fan et al., 2022; Tran et al., 2021) or discourse (Huang et al., 2022) structure. Other graph-based methods have focused on commonsense reasoning (Wu et al., 2022).

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).

2.2 Graph Autoencoders

Graph Autoencoder models were introduced by Kipf and Welling (2016b) as an efficient method for graph reconstruction tasks. The original paper introduces both variational graph autoencoders (VGAE) and non-probabilistic graph autoencoders (GAE) networks. The models are parameterized by a 2-layer graph-convolutional network (GCN) (Kipf and Welling, 2016a) encoder and a generative inner-product decoder between the latent variables. While initially conceived as lightweight models for citation network prediction tasks, both the VGAE and GAE have been successfully applied to a wide variety of applications such as molecule design (Liu et al., 2018), social network relational learning (Yang et al., 2020) and 3D scene generation (Chattopadhyay et al., 2023). Despite their apparent potential for effectively processing large amounts of graph-structured data, application within the field of NLP has been limited to a number of studies in unsupervised relational learning (Li et al., 2020).

3 Experiments

3.1 Data

Our data consists of the Dutch ENCORE corpus (De Langhe et al., 2022a), which in its totality consists of 12,875 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 the within-document and cross-document level.

3.2 Experimental Setup

3.2.1 Baseline Coreference Model

Our baseline model consists of the Dutch monolingual BERTje model (de Vries et al., 2019) fine-tuned for cross-document ECR. First, each possible event pair in the data is encoded by concatenating the two events and by subsequently feeding these to the BERTje encoder. 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 text pair classification are passed through a standard agglomerative clustering algorithm (Kenyon-Dean et al., 2018; Barhom et al., 2019) in order to obtain output in the form of coreference chains.

We also train two parameter-efficient versions of this baseline model using the distilled Dutch Language model RobBERTje (Delobelle et al., 2022) and a standard BERTje model trained with bottleneck adapters (Pfeiffer et al., 2020).

3.2.2 Graph Autoencoder Model

We make the assumption that a coreference chain can be represented by an undirected, unweighted graph $\mathcal{G} = (V, E)$ with $|V|$ nodes, where each node represents an event and each edge $e \in E$ between

Model	CONLL F1	Training Runtime (s)	Inference Runtime (s)	Trainable Parameters	Disk Space (MB)
MP RobBERTje	0.767	7962	16.31	74M	297
MP BERTje$_{ADPT}$	0.780	12 206	20.61	0.9M	3.5
MP BERTje	0.799	9737	21.78	110M	426
GAE NoFeatures	0.832 ± 0.008	1006	0.134	825856	3.2
GAE BERTje$_{768}$	0.835 ± 0.010	975	0.263	51200	0.204
GAE BERTje$_{3072}$	**0.852 ± 0.006**	1055	0.294	198656	0.780
GAE RobBERT$_{768}$	0.838 ± 0.004	1006	0.273	51200	0.204
GAE RobBERT$_{3072}$	0.841 ± 0.007	1204	0.292	198656	0.780
GAE SBERT	0.801 ± 0.002	982	0.291	51200	0.204
VGAE NoFeatures	0.824 ± 0.009	1053	0.139	827904	3.2
VGAE BERTje$_{768}$	0.822 ± 0.011	1233	0.282	53248	0.212
VGAE BERTje$_{3072}$	0.842 ± 0.009	1146	0.324	200704	0.788
VGAE RobBERT$_{768}$	0.828 ± 0.0021	1141	0.288	53248	0.212
VGAE RobBERT$_{3072}$	0.831 ± 0.004	1209	0.301	200704	0.788
VGAE SBERT	0.773 ± 0.012	1185	0.295	53248	0.212

Table 1: Results for the cross-document event coreference task. We report the average CONLL score and standard deviation over 3 training runs with different random seed initialization for the GCN weight matrices (GAE/VAE) and classification heads (Mention-Pair models). Inference runtime is reported for the entire test set.

two nodes denotes a coreferential link between those events. We frame ECR as a graph reconstruction task where a partially masked adjacency matrix A and a node-feature matrix X are used to predict all original edges in the graph. We employ both the VGAE and GAE models discussed in Section 2.2. In a non-probabilistic setting (GAE) the coreference graph is obtained by passing the adjacency matrix A and node-feature matrix X through a Graph Convolutional Neural Network (GCN) encoder and then computing the reconstructed matrix $\hat{A}$ from the latent embeddings Z:

$$Z = GCN(X, A) \quad (1)$$

$$\hat{A} = \sigma(ZZ^{\tau}) \quad (2)$$

For a detailed overview of the (probabilistic) variational graph autoencoder we refer the reader to the original paper by Kipf and Welling (2016b).

Our experiments are performed in a cross-document setting, meaning that the input adjacency matrix A contains all events in the ENCORE dataset. Following the original approach by Kipf and Welling (2016b) we mask 15% of the edges, 5% to be used for validation and the remaining 10% for testing. An equal amount of non-edges is randomly sampled from A to balance the validation and test data.

We extract masked edges and non-edges and use them to build the training, validation and test sets for the mention-pair baseline models detailed above, ensuring that both the mention-pair and graph autoencoder models have access to exactly the same data for training, validation and testing. We define the encoder network with a 64-dimension hidden layer and 32-dimension latent variables. For all experiments we train for a total duration of 200 epochs using an Adam optimizer (Kingma and Ba, 2014) and a learning rate of 0.001.

We construct node features through Dutch monolingual transformer models by average-pooling token representations for each token in the event span in the models' final hidden layer, resulting in a 768-dimensional feature vector for each node in the graph. For this we use the Dutch BERTje model (de Vries et al., 2019), a Dutch sentence-BERT model (Reimers and Gurevych, 2019) and the Dutch RoBERTa-based RobBERT model (Delobelle et al., 2020). Additionally, we create a second feature set for the BERTje and RobBERT models where each event is represented by the concatenation of the last 4 layers' average-pooled token representations Devlin et al. (2018). This in turn results in a 3072-dimensional feature vector.

Finally, we also evaluate a language-agnostic featureless model where X is represented by the identity matrix of A.

3.2.3 Hardware Specifications

The baseline coreference algorithms were trained and evaluated on 2 Tesla V100-SXM2-16GB GPUs. Due to GPU memory constraints, the Graph encoder models were all trained and evaluated on a single 2.6 GHz 6-Core Intel Core i7 CPU.

4 Results and Discussion

Results from our experiments are disclosed in Table 1. Results are reported through the CONLL F1 metric, 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 find that the graph autoencoder models consistently outperform the traditional mention-pair approach. Moreover, we find the autoencoder approach significantly reduces model size, training time and inference speed even when compared to parameter-efficient transformer-based methods. We note that the VGAE models perform slightly worse compared to their non-probabilistic counterparts, which is contrary to the findings in Kipf and Welling (2016b). This can be explained by the use of more complex acyclic graph data in the original paper. In this more uncertain context, probabilistic models would likely perform better.

As a means of quantitative error analysis, we report the average Levenshtein distance between two event spans for the True Positive (TP) pairs in our test set in Figure 1. Logically, if graph-based models are able to better classify harder (i.e non-similar) edges, the average Levenstein distance for predicted TP edges should be higher than for the mention-pair models. For readability's sake we only include results for the best performing GAE-class models. A more detailed table can be found in the Appendix. We find that the average distance between TP pairs increases for our introduced graph models, indicating that graph-based models can, to some extent, mitigate the pitfalls of mention-pair methodologies as discussed in Section 1.

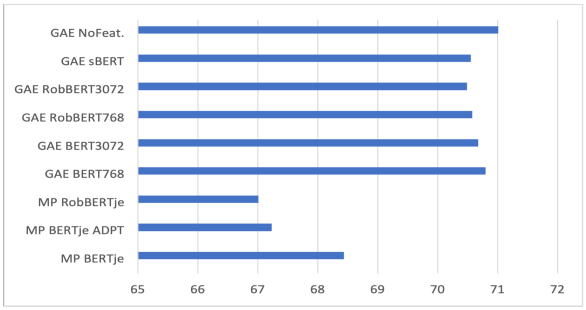

Figure 1: Average Levenshtein distance for True Positive (TP) classifications across all models

5 Ablation Studies

We gauge the robustness of the graph-based models in low-data settings by re-running the original experiment and continually reducing the available training data by increments of 10%. Figure 2 shows the CONLL F1 score for each of the models with respect to the available training data size. Also here, only the best-performing GAE-class models are visualized and an overview of all models' performance can be found in the Appendix. Surprisingly, we find that training the model on as little as 5% of the total amount of edges in the dataset can already lead to satisfactory results. Logically, feature-less models suffer from a significant drop in performance when available training data is reduced. We also find that the overall drop in performance is far greater for the traditional mention-pair model than it is for the feature-based GAE-class models in low-data settings. Overall, we conclude that the introduced family of models can be a lightweight and stable alternative to traditional mention-pair coreference models, even in settings with little to no available training data.

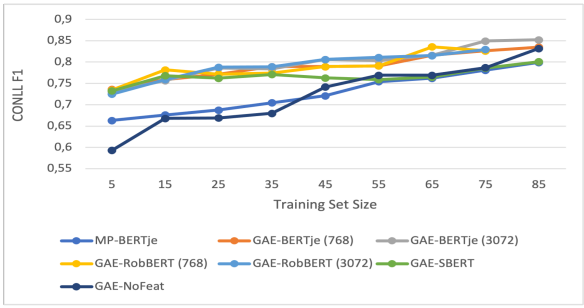

Figure 2: CONLL F1 performance with respect to the available training data.

6 Conclusion

We show that ECR through graph autoencoders significantly outperforms traditional mention-pair approaches in terms of performance, speed and model size in settings where coreference chains are at least partially known. Our method provides a fast and lightweight approach for processing large cross-document collections of event data. Additionally, our analysis shows that combining BERT-like embeddings and structural knowledge of coreference chains mitigates the issues in mention-pair classification w.r.t the dependence on surface-form lexical similarity. Our ablation experiments reveal that only a very small number of training edges is needed to obtain satisfactory performance.

Future work will explore the possibility of combining mention-pair models with the proposed graph autoencoder approach in a pipeline setting in order to make it possible to employ graph reconstruction models in settings where initially all edges in the graph are unknown. Additionally, we aim to perform more fine-grained analyses, both quantitative and qualitative, regarding the type of errors made by graph-based coreference models.

7 Limitations

We identify two possible limitations with the work presented above. First, by framing coreference resolution as a graph reconstruction task we assume that at least some coreference links in the cross-document graph are available to train on. However, we note that this issue can in part be mitigated by a simple exact match heuristic for event spans on unlabeled data. Moreover, in most application settings it is not inconceivable that at least a partial graph is available.

A second limitation stems from the fact that we modelled coreference chains as undirected graphs. It could be argued that some coreferential relationships such as pronominal anaphora could be more accurately modelled using directed graphs instead.

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

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H Martin, and Nikhil Krishnaswamy. 2023. $2 * n$ is better than n^2: Decomposing event coreference resolution into two tractable problems. *arXiv preprint arXiv:2305.05672*.

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.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. *arXiv preprint arXiv:1906.01753*.

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*.

Aditya Chattopadhyay, Xi Zhang, David Paul Wipf, Himanshu Arora, and René Vidal. 2023. Learning graph variational autoencoders with constraints and structured priors for conditional indoor 3d scene generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 785–794.

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.

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. 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*.

Chuang Fan, Jiaming Li, Xuan Luo, and Ruifeng Xu. 2022. Enhancing structure preservation in coreference resolution by constrained graph encoding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2557–2567.

Congcheng Huang, Sheng Xu, Longwang He, Peifeng Li, and Qiaoming Zhu. 2022. Incorporating generation method and discourse structure to event coreference resolution. In *International Conference on Neural Information Processing*, pages 73–84. Springer.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robus Anaphora Resolution for Unrestricted Texts*, pages 75–81.

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.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving Event Coreference with Supervised Representation Learning and Clustering-Oriented Regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2016a. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Thomas N Kipf and Max Welling. 2016b. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Irene Li, Alexander Fabbri, Swapnil Hingmire, and Dragomir Radev. 2020. R-vgae: Relational-variational graph autoencoder for unsupervised prerequisite chain learning. *arXiv preprint arXiv:2004.10610*.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. 2018. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

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.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

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

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4840–4850.

Judith Vermeulen. 2018. newsdna : promoting news diversity : an interdisciplinary investigation into algorithmic design, personalization and the public interest (2018-2022).

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*.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558.

Carl Yang, Jieyu Zhang, Haonan Wang, Sha Li, Myungwan Kim, Matt Walker, Yiou Xiao, and Jiawei Han. 2020. Relation learning on social networks with multi-modal graph edge variational autoencoders. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 699–707.

A Appendix

Model	Levenshtein Distance (TP)
MP RobBERTje	67.01
MP BERTje (ADPT)	67.23
MP BERTje	68.44
GAE NOFEAT	71.01
GAE BERTje (768)	70.8
GAE BERTje (3072)	70.68
GAE RobBERT (768)	70.57
GAE RobBERT (3072)	70.49
GAE SBERT	70.55
VGAE NOFEAT	69.95
VGAE BERTje (768)	68.71
VGAE BERTje (3072)	70.04
VGAE RobBERT (768)	70.21
VGAE RobBERT (3072)	70.15
VGAE SBERT	70.04

Table 2: Average Levenshtein distance for each True Positive (TP) pair in the test set indicating how well each model predicts comparatively more difficult coreference links.

Model	5	15	25	35	45	55	65	75
MP RobBERTje	0.627	0.631	0.667	0.683	0.701	0.736	0.753	0.766
MP BERTje (ADPT)	0.638	0.640	0.662	0.685	0.692	0.724	0.729	0.754
MP BERTje	0.663	0.675	0.687	0.704	0.721	0.754	0.762	0.781
GAE NOFEAT	0.593	0.667	0.669	0.679	0.747	0.769	0.769	0.786
GAE BERTje (768)	0.736	0.759	0.771	0.789	0.789	0.791	0.815	0.826
GAE BERTje (3072)	0.730	0.756	0.786	0.784	0.805	0.803	0.815	0.849
GAE RobBERT (768)	0.734	0.781	0.771	0.774	0.783	0.791	0.835	0.826
GAE RobBERT (3072)	0.725	0.759	0.788	0.788	0.806	0.810	0.815	0.829
GAE SBERT	0.732	0.768	0.762	0.770	0.762	0.759	0.765	0.786
VGAE NOFEAT	0.632	0.653	0.742	0.752	0.747	0.766	0.781	0.786
VGAE BERTje (768)	0.672	0.747	0.753	0.758	0.758	0.773	0.795	0.809
VGAE BERTje (3072)	0.712	0.769	0.781	0.780	0.776	0.818	0.802	0.818
VGAE RobBERT (768)	0.672	0.745	0.757	0.758	0.759	0.770	0.791	0.799
VGAE RobBERT (3072)	0.691	0.753	0.762	0.764	0.761	0.791	0.800	0.801
VGAE SBERT	0.651	0.681	0.735	0.738	0.726	0.711	0.745	0.735

Table 3: Results (CONLL F1) for the ablation experiments for each individual model. Columns indicate the percentagewise amount of available training data w.r.t the overall size of the ENCORE dataset.