Enhancing Unrestricted Cross-Document Event Coreference with Graph Reconstruction 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

Event Coreference Resolution remains a challenging discourse-oriented task within the domain of Natural Language Processing. In this paper we propose a methodology where we combine traditional mention-pair coreference models with a lightweight and modular graph reconstruction algorithm. We show that building graph models on top of existing mention-pair models leads to improved performance for both a wide range of baseline mention-pair algorithms as well as a recently developed state-of-the-art model and this at virtually no added computational cost. Moreover, additional experiments seem to indicate that our method is highly robust in low-data settings and that its performance scales with increases in performance for the underlying mention-pair models.

Keywords: Event Coreference Resolution, Graph Variational Auto-Encoder

1. Introduction

Event Coreference Resolution (ECR) is a discourse-oriented NLP task in which the primary goal is to find textual references that refer to the same happening, be it a fictional or real-world event. Typically, the textual representation of an event is designated as an *event mention*. Consider the following:

1. Elon Musk [completes]$_{Event}$ $44 billion deal to own Twitter

2. Elon Musk's contested and tumultuous Twitter [acquisition]$_{Event}$

While human readers can easily call on extra-linguistic knowledge to determine that Examples 1 and 2 do indeed refer to the same real-word event, this is no trivial task for most AI algorithms. Many state-of-the-art Large Language Models (LLMs) possess powerful commonsense reasoning abilities and are, to a certain degree, able to resolve local instances of coreference given sufficient document-level context (Ravi et al., 2023; Liu et al., 2023; Zhang et al., 2023). However, in application settings where wider context is often lacking and coreference resolution needs to be performed on a very large scale such methods currently fall short, even more so in languages other than English (Lu and Ng, 2018). Nonetheless ECR, especially when considering cross-documents settings, holds great potential for a large variety of practical NLP applications such as large-scale document summarization (Liu and Lapata, 2019), information extraction (Humphreys et al., 1997) and content-based news recommendation (Vermeulen, 2018).

Recent work has pointed out the potential of using small and efficient graph models, in which coreference resolution is framed as a graph reconstruction task as a viable strategy for large-scale cross-document ECR (De Langhe et al., 2023b). In this setting, one assumes that a number of coreferential links between events are given and a graph-based auto-encoder is then used to predict the missing links. However, a major problem emerges when evaluating the applicability of these methods in practical settings. For a graph reconstruction algorithm to work, at least part of the coreference graph should be known. As such, from-scratch prediction of coreferential chains on large collections of unstructured data seems impossible.

In this paper, we overcome this problem by outlining a methodology in which traditional event coreference resolution algorithms are extended by building a modular and lightweight graph reconstruction model on top, allowing for both from-scratch coreference link prediction, as well as fast generalisation across large document collections. To this purpose we perform a large number of experiments using different baseline encoder ECR models (De Langhe et al., 2023b), as well as a recently developed state-of-the-art ECR model (Yao et al., 2023). Our results reveal that performance increases significantly across the board by using simple graph auto-encoder networks (Kipf and Welling, 2016b) as a supplement to LLM-based coreference classification at virtually no added cost with respect to both training time and number of model parameters. Moreover, we show that the proposed methodology involving modular graph reconstruction algorithms is robust against incorrect classifications made by the underlying mention-pair models whilst being highly data efficient at the same time.

The remainder of this paper is organized as follows: first, we give an overview of available corpora

LREC-COLING 2024, pages 6122–6133
20-25 May, 2024.

and state-of-the-art methods for ECR (Section 2). Then, we describe the different components of our experimental pipeline (Section 3). Finally, we analyze the obtained results (Section 4) and perform a series of additional experiments highlighting certain properties of the newly developed approach (Section 5).

2. Related Work

2.1. Data

Large-scale cross-document ECR corpora are notoriously difficult to create due to their extensive annotation process involving the annotation of each mention and its potential arguments, as well as cross-linking these mentions over an entire document collection which often comprises hundreds or thousands of documents. Although some research has attempted to create large-scale corpora in a semi-supervised manner (Eirew et al., 2021), issues with respect to more fine-grained event annotation persist. As such, only a few established corpora currently exist, with most of them being entirely composed of English language data.

We can distinguish two broad categories of ECR corpora. First, there are corpora adhering to a strict event taxonomy, where event mentions are only included if they fall within certain predefined event types and subtypes such as *Life-BeBorn* or *Business-StartOrganization* (NIST, 2005). This category includes datasets such as the ACE corpora (*English/Chinese*) (NIST, 2005), TAC-KBP (*English/Chinese/Spanish*) (Mitamura et al., 2015), ECB+ (*English*) (Cybulska and Vossen, 2014b) and the MEANTIME NewsReader corpus (*Dutch*) (Minard et al., 2016). Second, some corpora forgo the aforementioned event taxonomy and focus instead on unrestricted events. In this case, event mentions do not have to belong to a predefined list of possible event types. Unrestricted ECR corpora include the OntoNotes dataset (*English*) (Pradhan et al., 2007), which does not distinguish between entity and event coreference chains, and the ENCORE corpus (*Dutch*) (De Langhe et al., 2022).

2.2. ECR

Methods in ECR research are primarily based on earlier work in entity coreference studies (Rahman and Ng, 2009) where instead of finding links between events, the task is to link certain entities which are in a coreferential relation. The primary paradigm for the task of coreference resolution, be it entity or event resolution, takes the form of a binary mention-pair approach. This method generates all possible mention pairs and reduces the task to a binary classification decision (coreferent or not) for each mention pair.

For event coreference resolution specifically, a large variety of classical machine learning algorithms have been tested using the mention-pair paradigm, including 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 adopted 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 here that while traditional feature-based machine learning has been entirely replaced by neural encoder-based methods, many of the core tenets and observations remain deeply embedded withing present-day ECR research (Lu and Ng, 2021). Features such as lexical closeness, often in the form of dice, cosine and element-wise similarity, as well as structural discourse properties such as sentence and event distance are still explicitly modelled and integrated in many state-of-the-art coreference resolvers today (Yao et al., 2023). Additionally, in restricted settings (i.e event mentions always belonging to a predefined taxonomy), type-based decoding and pruning methods have been shown to significantly improve results for many of the established coreference benchmark datasets (Lu et al., 2022; Yao et al., 2023).

While the mention-pair paradigm has been the preferred setup for most studies in coreference resolution, both entity and event alike, there exist some notable caveats to this approach. First, mention-pair models notoriously suffer from an over-reliance on superficial lexical similarity (Ahmed et al., 2023), meaning that lexically similar mentions are consistently classified as being coreferent and that non-lexically similar mentions are often deemed non-coreferent. While similarity between mentions, as previously discussed, is indeed one of the primary predictors of a coreferential relation, many mentions will be wrongfully designated as coreferent for merely belonging to the same domain. A second problem is the number of possible mention-pairs to compute in large-scale settings, which is defined as:

$$\frac{n!}{2!(n-2)!}$$

where *n* is the number of total events in the collection. Previous work has proposed intervention strategies either at the annotation-level (Cybulska and Vossen, 2014b,a) or by inserting pruning algorithms in the classification pipelines (Cattan et al., 2021a) in controlled settings, however, some of these methods cannot be fully extrapolated to application settings.

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). More recently, graph reconstruction algorithms have also been proposed for efficient large-scale ECR (De Langhe et al., 2023b). However, as discussed in Section 1, these algorithms suffer from a fundamental flaw in that they require a number of coreferential links to accurately predict the remaining connections, meaning that their deployment in from-scratch settings remains difficult.

2.3. (V)GAE

Graph auto-encoder models were introduced by Kipf and Welling (2016b) as an efficient method for graph reconstruction tasks. In this original paper, both variational graph auto-encoder (VGAE) and non-probabilistic graph auto-encoder (GAE) networks were introduced. 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. Both the VGAE and GAE have been successfully applied to a wide variety of applications such as molecule design (Liu et al., 2018) and social network relational learning (Yang et al., 2020). 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

Our proposed methodology, which is illustrated in Figure 1, consists of a standard training, validation and testing pipeline using mention-pair coreference models on top of which a small auto-encoder is built. This auto-encoder uses the predicted mentions of the first model as its own training data in order to reconstruct missing coreferential links.

Please note that for the research presented in this paper, we are mainly interested in enhancing the actual resolution of the coreferential links, which is why we start from gold-standard event mentions.

In the following sections, we describe and motivate the data that has been used for the experiments (Section 3.1) and explain which design choices were made for the proposed architecture. In order to illustrate the broad applicability of our pipeline we evaluate with both a variety of baseline models, each with a different encoder (Section 3.2.1), as well as with a modified state-of-the-art neural ECR model (Section 3.2.2).

3.1. Data

Our data consists of the Dutch ENCORE corpus (De Langhe et al., 2022), which in total comprises 15,546 annotated events spread over 1,087 documents that were sourced from a collection of Dutch (Flemish) newspaper articles collected in the calendar year 2019. Coreferential relations between events were annotated at the within-document and cross-document level. Note that for the ENCORE corpus during corpus creation raw documents were grouped based on their approximate content to create 'topic clusters' similar to the ones in the ECB+ corpus (Cybulska and Vossen, 2014b). Cross-document coreference links were annotated for documents belonging to the same topic clusters in order to maximize annotator efficiency. In addition to coreferential links 3 key event properties were annotated for each event mention: the prominence/importance of the event in a given news article (main event/background event), its realis (certainly happened/may not happen) and the general sentiment (positive/negative/neutral) of the event.

Our motivation to rely on the ENCORE corpus over other available corpora is two-fold. First, we aimed to focus on unrestricted ECR, as we believe that in application settings events will often not be restricted to a certain taxonomy, but rather exhibit a large topical variety. Second, while the English-language OntoNotes corpus is also unrestricted, it contains both entity and event mentions, which we believe could incur certain difficulties for traditional event conference resolution algorithms. This primarily includes the presence of many anaphorical pronominal mention-pairs which are more common to corefer with entities, but are often absent for events.

For our experimental setup we reserved 70 % of the data for training, 15 % for validation and a final 15 % for testing. As the ENCORE corpus is split up in several aforementioned topical clusters with coreference links annotated intra-topic, we ensure that there is no overlap between topics in training, validation and test sets respectively. In accordance with earlier studies on event coreference (Cattan et al., 2021a), we sampled the number of instances in the training set at a ratio of 20 negative samples to each positive mention-pair.

3.2. Mention-Pair Coreference Models

3.2.1. Baseline Mention-Pair Models

The baseline mention-pair event coreference models we developed consist of a number of fine-tuned BERT-based transformer models. First, each possible event pair in the training data is encoded

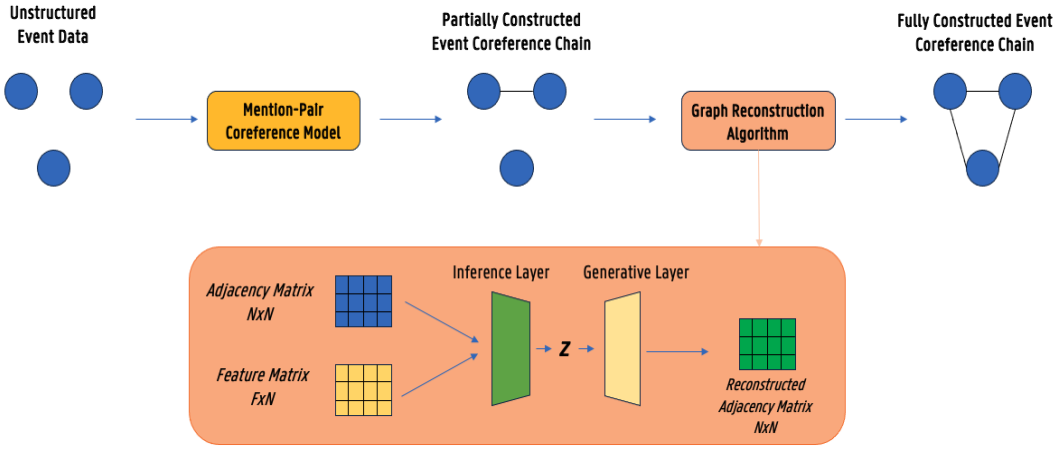

Figure 1: Visualisation of the proposed architecture. A standard mention-pair coreference model is used to predict coreference links between mentions. Then, a reconstruction algorithm is applied to find the missing links between all mentions.

by concatenating the two events and by subsequently feeding these to a BERT-based 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. Each of these models was trained for 3 epochs using an ADAM optimizer (Kingma and Ba, 2014) and a learning rate of 2e-5 with a batch size of 64.

We employ a variety of established transformer encoders including the Dutch BERT-based BERTje (de Vries et al., 2019) and RoBERTa-based RobBERT (Delobelle et al., 2020) models. Additionally, we also fine-tune on the multilingual models XLM-RoBERTa (Conneau et al., 2019) and mDeBERTa-v3 (He et al., 2021), the latter improving on the earlier mentioned encoder models by using a variety of highly effective pre-training techniques such as gradient Disentangled Embedding Sharing, disentangled attention and the use of an enhanced mask decoder.

None of the models described above have been explicitly trained on newspaper data. Pre-training for XLM-RoBERTa, RobBERT and mDeBERTa-v3 was all performed on cleaned multilingual (XLM-R, mDeBERTa-v3) and monolingual Dutch (RobBERT) CommonCrawl data (Wenzek et al., 2020; Conneau et al., 2019). While the BERTje model does contain some (online) Dutch news articles up to 2019 (de Vries et al., 2019; Ordelman et al., 2007), the majority of its pre-training data still comes from other sources such as books and crawled Wikipedia data. As the ENCORE corpus we are working with only comprises news data and studies have shown that transformer models which have been domain-adapted by continually pre-training them often outperform general models in domain-specific tasks (Gururangan et al., 2020), we believe that we can present a stronger baseline encoder algorithm by supplementing the monolingual BERTje transformer model with additional news data. The choice for BERTje as a starting point is motivated both by the composition of its original training data as well as its demonstratively better results on news-based tasks, including ECR (de Vries et al., 2023). In order to create the domain-adapted BERTje model we continued its pre-training on a (filtered) collection of Dutch online news articles spanning the years 2020-2023[1]. In total, this additional pre-training corpus contains around 20 million tokens of Dutch news articles, including headlines. The model was trained using the original pre-training objectives which are a Masked Language Modelling of consecutive word pieces (MLM) and a Sentence order prediction (SOP) task where each second sentence in a training example is either the next or previous sentence (de Vries et al., 2019). We train the model with the supplemental corpus for a total of 4 epochs keeping the same hyperparameter configuration that was used in the model's original pre-training. Throughout the rest of this paper we will refer to this domain-adapted BERTje model as *NewsBERTje*. This model is freely available[2] through the HuggingFace Transformer framework (Wolf et al., 2019).

[1] https://www.kaggle.com/datasets/maxscheijen/dutch-news-articles

[2] https://huggingface.co/LoicDL/NewsBERTje-base

3.2.2. State-of-the-art Mention-Pair Model

In addition to the baselines we also couple our proposed graph reconstruction model with the recently developed event-aware ECR model by Yao et al. (2023) which attains state-of-the-art performance on the English ACE-05 dataset (NIST, 2005).

This model integrates multiple linguistically-motivated features, as well as BERT-encoded event spans for a more fine-grained event representation. We did, however, make several modifications to the model's original implementation: (1) as we work within an unrestricted event setup, we removed those feature representations explicitly related to event type, (2) we replaced the model's SpanBERT (Joshi et al., 2020) encoder for event mention spans by a standard BERTje encoder, since there are currently no multilingual or monolingual Dutch SpanBERT models available, (3) we also decided to add three domain-specific news features (cfr. infra) to the model as it has been shown that these features potentially play a role in the coreference classification decision (De Langhe et al., 2023a). Finally, (4) in the original implementation the authors first fuse the events' feature representations with similarity embeddings and then concatenate those similarity embeddings again to the fused result. We did not fuse the feature encodings with the similarity encodings through a feedforward neural network prior to computing the final coreference score, but just concatenated these two information vectors and used this as the input for a classification network instead. We did this in order to shy away from the aforementioned over-reliance on lexical similarity features (Section 2.2). The paragraphs below briefly describe how we encode both the similarity between events in a mention pair as well as the individual event features prior to applying the coreference classification algorithm.

Similarity Encoding For each event pair *i*, *j* in a given mention pair we create a pooled representation *r* of its span by passing it through a BERTje encoder and averaging token representations for each token in the span from the encoder's hidden layer. Then, we calculate cosine and element-wise similarity between the obtained embeddings and concatenate the pooled representations and similarity measures in a pair-encoding vector:

$$V_{sim} = [r_i; r_j; Sim_{cos}(r_i, r_j); Sim_{ele}(r_i, r_j)]$$

This pair representation is subsequently passed through a fusion layer parameterized by a feedforward neural network to obtain the fused similarity pair representation $S_{i,j}$:

$$S_{i,j} = \textbf{FFNN}(V_{sim})$$

Feature Encoding We integrate three domain-specific event features for each event. First, the event prominence (*Background/main*) to indicate whether an event is the key event reported on in a given document, or whether it is used as background information. Second, the event realis (*Certain/Uncertain*) which indicates whether an event happened/will happen with absolute certainty or not. Third, the event sentiment (*Positive/negative/neutral*) which expresses the overall sentiment of the event. For each individual event, we create a representation E_s of its span by feeding it through a BERTje encoder and average-pooling the 768-dimensional token representations for each token *t* in the event span *X*. We then use a 2-layer feedforward Neural Network to determine the events' label for the feature:

$$E_s = Pooler(BERTjeEncoder(X))$$

$$P(y|X) = \textbf{FFNN}(E_s)$$

For each event *e* and feature *k*, we then use an embedding layer (Lai et al., 2021) to extract the feature vector F_s^k. Finally, for each mention-pair *i,j* an individual feature is concatenated and passed through a feature filtering layer parameterized by another simple feedforward neural network:

$$F_{i,j}^k = \textbf{FFNN}([F_i^k; F_j^k])$$

Coreference Classification In a final step, the pair representation $S_{i,j}$ and each of the individual filtered feature vectors are concatenated and passed to a coreference classification layer to obtain a label for the mention pairs:

$$Score = \textbf{FFNN}([S_{i,j}; F_{i,j}^1; ...; F_{i,j}^k])$$

3.3. Graph Auto-Encoder Model

On top of the mention-pair models described in Section 3.2 we built a small modular graph reconstruction network that uses known edges in the graph to predict missing links. By supplementing the original model with a graph-based model we are able to integrate information for each event pair as well as the relation of these events to other events in the data. 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 two nodes denotes a coreferential link between those events.

We frame ECR as a graph reconstruction task where a partially masked adjacency matrix *A*, of dimension *nxn*, and a node-feature matrix *X*, of dimension *fxn* are used to predict all original edges in the graph, where *n* denotes the total number of

Base Model	Extension Type	Features	CONLL F1	Num Parameters	Training Time (S)	Disk Space (MB)
BERTje	-	-	0.635	110M	815.28	418
	GAE	*BERTje*	0.691	51200	7.35	0.204
	VGAE	*BERTje*	0.689	53248	8.57	0.212
RobBERT	-	-	0.597	117M	811.35	449
	GAE	*RobBERT*	0.611	51200	6.89	0.204
	VGAE	*RobBERT*	0.607	53248	8.26	0.212
mDeBERTa-v3	-	-	0.669	276M	1232.11	1100
	GAE	*mDeBERTa*	0.700	51200	9.10	0.204
	VGAE	*mDeBERTa*	0.695	53248	10.41	0.212
XLM-RoBERTa	-	-	0.613	123M	940.07	1100
	GAE	*XLM-RoBERTa*	0.650	51200	7.90	0.204
	VGAE	*XLM-RoBERTa*	0.632	53248	9.56	0.212
NewsBERTje	-	-	0.641	110M	835.16	418
	GAE	*NewsBERTje*	0.698	51200	7.57	0.204
	VGAE	*NewsBERTje*	0.675	53248	9.16	0.212
(Adapted) SOTA (Yao et al., 2023)	-	-	0.722	112M	1459.36	532
	GAE	*BERTje*	**0.746**	51200	7.14	0.204
	VGAE	*BERTje*	0.738	53248	8.46	0.212

Table 1: Results for the unrestricted ECR task using (V)GAE extension models. For each model component we additionally report the total number of (trainable) parameters, the training time and the allocated disk space for the saved model for each of their components.

events in the test set and *f* denotes feature length. A high-level visualisation of this process can be found at the bottom of Figure 1. We employ both the probabilistic VGAE and non-probabilistic GAE models introduced in 2.3. 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 compute the reconstructed matrix $\hat{A}$ from the latent embeddings *Z*:

$$Z = GCN(X, A)$$

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

In its probabilistic setting (VGAE) a set of latent stochastic variables z_i is introduced, which is summarized in matrix **Z**. The two-layer GCN encoder is defined as GCN(X,A) = $\hat{A}$ ReLU($\hat{A}XW_0)W_1$. In learning, the following variational lower bound $\mathcal{L}$ is optimised with respect to weights W_i:

$$\mathcal{L} = \mathcal{E}_{q(Z|X,A)}[logp(A|Z)] - KL[q(Z|X,A)||p(Z)]$$

As shown in Figure 1 the (V)GAE models are trained, validated and tested on the output of the mention-pair models. First, the fine-tuned model predicts presence or absence of a coreferential link (*e*) between each of the test set instances. Then, the adjacency matrix *A* is constructed from these predictions. Matrix *A* is of dimension *nxn*, where *n* is the amount of unique events in the test set.

We create training and validation sets for the graph reconstruction task by setting aside 90 % of predicted edges for training and 10 % for validation. In accordance with the original setup described by Kipf and Welling (2016b) we then sample an equal amount of predicted non-edges to balance the validation data. For testing, the model is tasked with predicting each cell in the adjacency matrix, which corresponds to the original test set used to evaluate the mention-pair models.

The node-feature matrix *X* is constructed by encoding each individual event span through a BERT-based model. We do this 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/event in the graph. For each mention-pair model we create 5 separate node-feature matrices, one for each of the models described in Section 3.2, resulting in a total of 16 possible experiment configurations for both the GAE and VGAE models.

Finally, 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 300 epochs using an Adam optimizer (Kingma and Ba, 2014) and a learning rate of 0.001.

3.4. Hardware Specifications

The baseline coreference algorithms were trained and evaluated on 1 Tesla V100-SXM2-16GB GPU. The graph encoder models were all trained and evaluated on a 16 Intel(R) Xeon(R) Gold 6142 2.60GHz CPUs. Finally, the domain-adapted BERT model (NewsBERTje) was trained on 4 Tesla V100-SXM2-16GB GPUs.

4. Results and Discussion

Results from our experiments are detailed in Table 1. As is standard in coreference resolution studies, evaluation is done 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). For each of the base mention-pair models, we report scores with and without our proposed

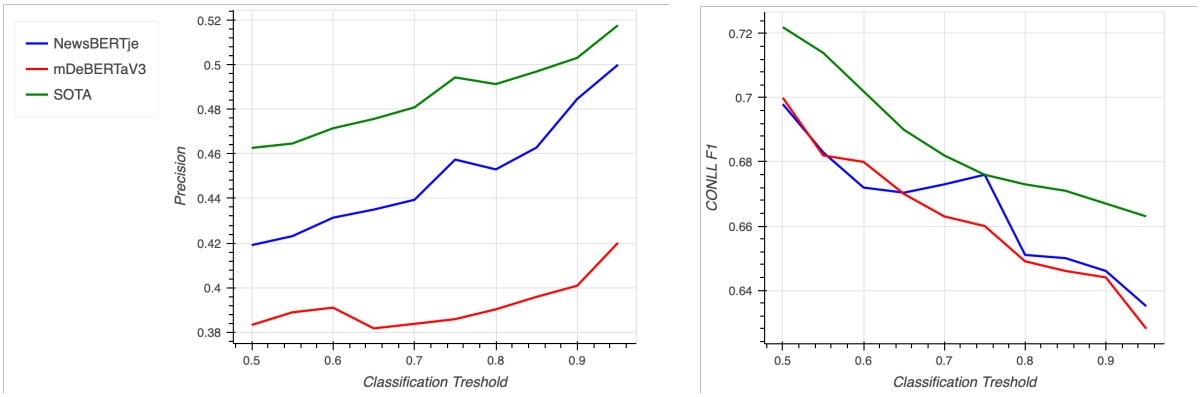

Figure 2: Effects of raising the classification threshold p on mention-pair precision (left) and CONLL F1 for the extension models (right) for the best performing mono- and multilingual models as well as the SOTA model.

(V)GAE extension algorithms. As reporting all possible base model - embedding combinations would take up a significant amount of space and hinder readability, we limit the results in Table 1 to base-(V)GAE combinations that use the same (feature) encoder.

Overall, we find that the multilingual DeBERTa-v3 model performs best out of the base mention-pair models that were tested, while our domain-adapted NewsBERTje model was the best performing baseline monolingual model. As shown in Table 1, extending the base mention-pair models with the (V)GAE graph reconstruction models leads to an increase in performance across the board in exchange for only a minimal increase in the number of parameters and overall training time. In addition, the implemented SOTA model also benefits greatly from our proposed methodology, indicating that our method generalizes well over different mention-pair classifiers and can thus potentially be more widely applied and integrated into many ECR pipelines. Consistent with earlier results (De Langhe et al., 2023b) we find that on average the non-probabilistic GAE models are superior to the probabilistic VGAE models. We hypothesize that as (event) coreference resolution graphs are in essence always complete graphs, non-probabilistic models will by default perform better.

5. Ablation Studies

The results of applying reconstruction models to a large-scale coreference resolution task seem promising. However, it should be noted that by default some false positive coreference links will be present in the output of the base mention-pair models. Naturally, this means that the (V)GAE extension models will be trained on partially incorrect and inconsistent data. In the following sections we aim to explore the effects of training the extension models on such faulty data, as well as try out possible mitigation strategies.

5.1. Modulating Precision as a Parameter

A first mitigation strategy to avoid the (V)GAE models' training on faulty predicted coreference links could be to simply maximize the precision of the underlying mention-pair models by regulating the decision boundary p in their classification layer. For all experiments described in Section 3 we assumed a default decision boundary of $p = 0.5$, where all mention-pairs with a classification output o are classified as coreferent if $o > 0.5$ and as non-coreferent if $o < 0.5$.

For each of the individual mention-pair models we raised p by increments of 0.05 and used the new set of predicted coreference links to train and evaluate the respective (V)GAE extension models. Figure 2 plots the effects of raising treshhold p for the best performing (GAE) monolingual (NewsBERTje), multilingual (mDeBERTaV3) and SOTA models respectively. The left graph depicts the changes in precision for each of the mention-pair models when raising p. Additionally, the right one illustrates the impact of raising p in the mention-pair step on the resulting CONLL F1 score. Interestingly, while it is indeed confirmed that raising p increases precision in the mention-pair step (and thus reduces the number of faulty coreference links for the (V)GAE models to be trained on), overall CONLL F1 decreases again for higher values of the decision boundary p for each of the models. While this is a counter-intuitive observation, a possible explanation might be found when considering that as precision (and decision boundary p) in this setting increase, the absolute number of True Positive coreferential links will naturally decrease. This in turn results into less available edge training data, which may indicate that the proposed (V)GAE models are far more robust against faulty or inconsistent training data

and that the primary strategy of this methodology should be to provide those models with the highest amount of True Positive (TP) coreference links rather than avoiding False Positive (FP) links. We also note a more pronounced downward trend for the multilingual model (the green line), possibly due to multilingual models as a whole needing more training data for monolingual tasks as a result of an inherent skew in their internal language distribution (Wu and Dredze, 2020).

5.2. Influence of Graph Corruption

In this section, we investigate the potential of graph reconstruction models in optimal conditions by gradually removing false positive links from the individual mention-pair models' output. We designate the number of false positive edges in the mention-pair models' output as the degree of corruption (DoC) and then remove incorrect edges from the extension models' training data by increments of 10 %. Figure 3 plots the performance for best performing (GAE) mono- and multilingual models (in blue and red respectively) as well as the SOTA model (green) detailed in Table 1 relative to the DoC in the training data. Results show that if faulty links can be accurately pruned from the mention-pair output, whilst keeping the number of True Positive links as is, applying the extension models can dramatically increase performance in cross-document coreference tasks.

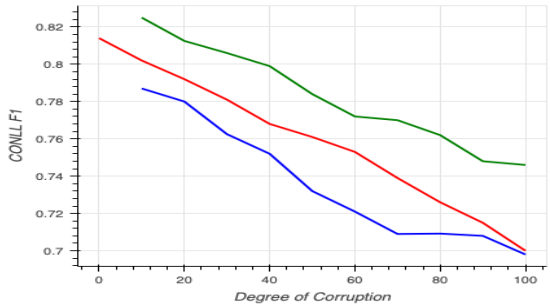

Figure 3: CONLL F1 scores relative to the % of False Positive links in the training data

5.3. Robustness in Low-Data Settings

Earlier studies have shown that graph reconstruction methods such as GAE and GVAE only need a fraction of available in-domain training edges to obtain on-par or better results than their traditional mention-pair counterparts (De Langhe et al., 2023b). In order to validate the extension models' performance in situations where only a limited amount of edges are available to train on, we remove all false positive edges from the extension models' training data and then gradually remove the number of available true positive training edges. In this way, we aim to estimate the number of correct coreferential links needed to achieve on-par or better performance compared to the baseline models.

In Figure 4 the graph shows the performance of the best (GAE) mono- and multilingual models (in blue and red respectively) as well as the SOTA model (green) that were reported in Table 1. The available training edges were removed by increments of 10 %, relative to all predicted true positive edges. Interestingly, we find that monolingual extension models only require 10 to 20% of the correctly predicted links to exceed the original mention-pair models' performance. This might imply that in future research it might be beneficial to focus on mention-pair models that emphasize precision above everything else. Additionally, we observe that the multilingual model suffer from the largest drop in performance relative to the results in Table 1 (-22% for mDeBERTa-v3), indicating that as a whole monolingual models are more robust in low-data settings.

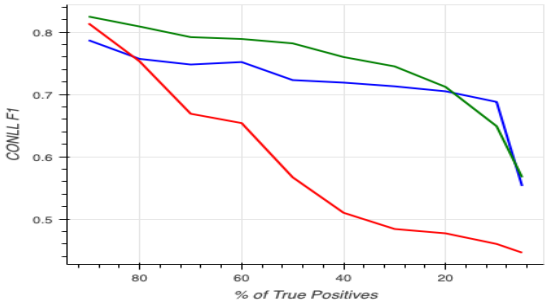

Figure 4: CONLL F1 scores relative to the amount of True Positive edges in the extension models' training data

5.4. Applicability to Other Languages

In order to demonstrate that the proposed (V)GAE method can be easily extrapolated to other languages and corpora, we present a set of rudimentary experiments on the widely used English ECB+ corpus. Given the similarity in size and overall design between the ENCORE and ECB+ corpora (i.e. the use of topical clusters), we copy the experimental setup that was described in Sections 3.1 and 3.2.1. For the encoder models used in these experiments we select standard BERT (Devlin et al., 2018) and RoBERTa (Conneau et al., 2019) models, which are simply the English Language equivalents of the monolingual Dutch models used in Section 3. Table 2 displays the results for extending a set of basic coreference models with (V)GAE networks for the ECB+ dataset. Overall, the results are in line with those reported in Table 1 and show that the use of the extension models leads to a

Base Model	Extension Type	Features	CONLL F1	Num Parameters	Training Time (S)	Disk Space (MB)
BERT	-	-	0.621	110M	807.4	418
	GAE	*BERT*	0.659	48360	7.15	0.200
	VGAE	*BERT*	0.654	49300	8.32	0.206
RoBERTa	-	-	0.641	117M	889.45	449
	GAE	*RoBERTa*	0.667	48360	6.41	0.200
	VGAE	*RoBERTa*	0.667	49300	8.29	0.206
mDeBERTa-v3	-	-	0.716	276M	1295.45	1100
	GAE	*mDeBERTa*	0.754	48360	9.84	0.200
	VGAE	*mDeBERTa*	0.748	49300	11.01	0.206
XLM-RoBERTa	-	-	0.748	123M	960.08	1100
	GAE	*XLM-RoBERTa*	0.650	48360	7.62	0.200
	VGAE	*XLM-RoBERTa*	0.632	49300	9.48	0.206

Table 2: Results for the unrestricted ECR task using (V)GAE extension models on the English language ECB+ corpus. For each model component we additionally report the total number of (trainable) parameters, the training time and the allocated disk space for the saved model for each of their components.

steady improvement over the baseline coreference models at a relatively increases in training time, parameter and disk space.

6. Conclusion

We propose a new methodology in which we combine the strengths of traditional mention-pair event coreference resolution (ECR) algorithms with a series of lightweight graph reconstruction algorithms. The pairwise output of the initial model is used to create an adjacency matrix which forms the input of the reconstruction model. Then, likelihood of individual edges is predicted based on a node-feature matrix and the already established edges in the graph. We show that directly training (variational) graph auto-encoders on the output of ECR mention-pair models can enhance their performance, irrespective of the underlying encoder/model. We believe that due to their little training time and computational cost graph reconstruction models can be easily inserted to most mention-pair pipelines. Our ablation experiments show that our lightweight models remain robust even in low-data settings and that inserting extra pruning steps in the proposed pipeline might even further increase performance down the line.

7. Limitations

While in theory our methodology using a graph reconstruction algorithm can be readily applied to other coreference-based tasks such as entity coreference and bridging relationships we do note one potential caveat. For our proposed Graph reconstruction model we made the simplifying assumption that each event coreference chain could be modelled as an undirected graph. Consider the following coreference chain between entity mentions: *Barack Obama - He*. In this case it can be argued that the relationship between these mentions is in fact directed, as the meaning of the mention *He* depends entirely on its antecedent *Barack Obama*. Unlike in ECR, such pronominal relationships between mentions are far more common for entity coreference tasks. Indeed, many state-of-the-art entity coreference resolution algorithms throughout the years have explicitly modelled noun-pronoun relationships. As of now, this is not possible with our proposed method. The same argument can be made for the more complex bridging relationships such as: *the student - (part-of) - the class*, where there is always a certain unweighted dependence between mentions.

8. Acknowledgements

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

9. Bibliographical 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*.

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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ECB+ Annotation of Events and their Coreference. 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.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2022. 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. 2023b. Filling in the gaps: Efficient event coreference resolution using graph autoencoder networks. *arXiv preprint 2310.11965*.

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*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2023. Dumb: A benchmark for smart evaluation of dutch models. *arXiv preprint arXiv:2305.13026*.

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

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*.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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*.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. *arXiv preprint arXiv:2104.01697*.

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.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, pages 1–43.

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

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.

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, 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.

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).

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.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwarz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. *arXiv preprint arXiv:2302.09715*.

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*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

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.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Learning event-aware measures for event coreference resolution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13542–13556.

Hang Zhang, Wenjun Ke, Jianwei Zhang, Zhizhao Luo, Hewen Ma, Zhen Luan, and Peng Wang. 2023. Prompt-based event relation identification with constrained prefix attention mechanism. *Knowledge-Based Systems*, page 111072.