NUS-IDS at CASE 2021 Task 1: Improving Multilingual Event Sentence Coreference Identification With Linguistic Information

Fiona Anting Tan, Sujatha Das Gollapalli, See-Kiong Ng
Institute of Data Science
National University of Singapore, Singapore
tan.f@u.nus.edu, idssdg@nus.edu.sg, seekiong@nus.edu.sg

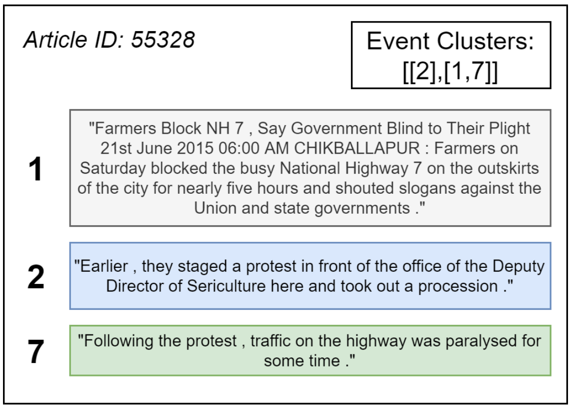
Abstract

Event Sentence Coreference Identification (ESCI) aims to cluster event sentences that refer to the same event together for information extraction. We describe our ESCI solution developed for the ACL-CASE 2021 shared tasks on the detection and classification of socio-political and crisis event information in a multilingual setting. For a given article, our proposed pipeline comprises of an accurate sentence pair classifier that identifies coreferent sentence pairs and subsequently uses these predicted probabilities to cluster sentences into groups. Sentence pair representations are constructed from fine-tuned BERT embeddings plus POS embeddings fed through a BiLSTM model, and combined with linguistic-based lexical and semantic similarities between sentences. Our best models ranked 2^{nd}, 1^{st} and 2^{nd} and obtained CoNLL F_1 scores of 81.20%, 93.03%, 83.15% for the English, Portuguese and Spanish test sets respectively in the ACL-CASE 2021 competition.

1 Introduction

The ability to automatically extract sentences that refer to the same event from any given document is useful for downstream information extraction tasks like event extraction and summarization, timeline extraction or cause and effect extraction (Örs et al., 2020). Event Sentence Coreference Identification (ESCI) aims to cluster sentences with event mentions such that each cluster comprises of sentences that refer to the same specific event.

We address ESCI for news articles referring to socio-political and crisis event information in a multilingual setting, introduced as one of the ACL-CASE 2021's shared tasks (Hürriyetoğlu et al., 2021). Given that news articles comprise of multiple events spread across a few sentences, and the syntax referring to the same event differs in different contexts, ESCI for news articles is a challenging NLP problem (Hürriyetoğlu et al., 2020). Furthermore, considering the availability of news in various languages, ESCI techniques that are applicable beyond English and robust across different languages are desirable.



"Farmers Block NH 7 , Say Government Blind to Their Plight 21st June 2015 06:00 AM CHIKBALLAPUR : Farmers on Saturday blocked the busy National Highway 7 on the outskirts of the city for nearly five hours and shouted slogans against the Union and state governments ."

"Earlier , they staged a protest in front of the office of the Deputy Director of Sericulture here and took out a procession ."

"Following the protest , traffic on the highway was paralysed for some time ."

Figure 1: Example English article from training dataset from ACL-CASE 2021. The sentences $1, 2, 7$ with event mentions as well as the target clustering $\{[2], [1,7]\}$ are highlighted.

The ESCI task is illustrated using an example article shown in Figure 1. As shown in this figure, ESCI involves the identification of the event clusters (e.g. $\{[2], [1,7]\}$ in the figure) based on the content of the individual sentences.

Contributions: We propose a two-step solution for the ESCI task. In Step-1, we obtain sentence pair embeddings by fine-tuning Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) embeddings combined with parts-of-speech (POS) embeddings that are fed through a bi-directional long short-term memory (BiLSTM) model. Next, these sentence pair embeddings are combined with novel features based on lexical and semantic similarities to train a classifier that predicts if the sentence pair is coreferent.

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 105–112
August 5–6, 2021, ©2021 Association for Computational Linguistics

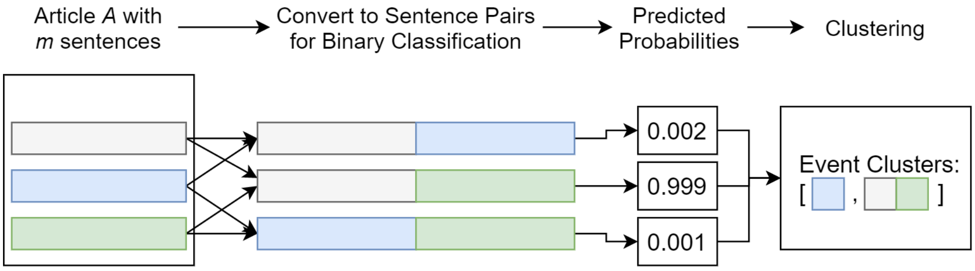

Article *A* with *m* sentences
Convert to Sentence Pairs for Binary Classification
Predicted Probabilities
Clustering
0.002
0.999
0.001
Event Clusters:
[,]

Figure 2: Overall model pipeline. *Notes.* (1) Convert article into sentence pairs for binary classification, and (2) Taking predicted probabilities to perform article level clustering of sentences.

Step-2 involves the clustering of sentences using sentence pair probabilities predicted from Step-1. We apply the clustering algorithm from Örs et al. (2020) to obtain a variable number of clusters for each article.

We illustrate the effectiveness of our proposed solution via detailed validation experiments on the training datasets from ACL-CASE 2021. We show that our features are effective on documents from all three languages studied in the competition, viz, English, Portuguese, and Spanish. Indeed, on the ACL-CASE 2021 Shared Task 1 Subtask 3, our best-performing models ranked 2nd, 1st and 2nd and obtained CoNLL F_1 scores of 81.20%, 93.03%, 83.15% for the English, Portuguese and Spanish test sets respectively.

Organization: In the next section, we present closely related work on ESCI. Subsequently, Section 3 introduces our features and classification model while Section 4 discusses our dataset, experimental setup, results, and findings. In Section 5, we conclude the paper with some future directions.

2 Related Work

Most end-to-end event coreference systems approach the task in a two-stage manner: (1) To detect the mention or event of interest, and (2) To resolve the given mentions or events and cluster if coreferent (Zhang et al., 2018). In our work, we focus only on latter task of coreference resolution and have direct access to identified event sentences.

Early works of ESCI adopted linguistic (Bejan and Harabagiu, 2010) or template-based features (Choubey and Huang, 2017). Subsequently, neural network methods to encode textual events and contexts became increasingly popular (Krause et al., 2016). The combination of the two methods have also proved to be effective in recent works (Zeng et al., 2020; Barhom et al., 2019).

In the previous run of ESCI by the same organisers (Hürriyetoğlu et al., 2019, 2020), the best-performing team (Örs et al., 2020) deconstructed the task into two steps: (1) To predict if a sentence pair is coreferent or not, and (2) Use the predictions as scores for clustering. This approach is common amongst other event coreference resolution methods too (Barhom et al., 2019). We employ this general approach and focus on enriching the feature space using linguistic-based similarity measures along with richer text embeddings based on BERT with POS embeddings.

3 Our Approach

Figure 2 summarizes our proposed pipeline. In this section, we describe our approach in detail.[1]

Let A be an article containing m sentences with event mentions $\{s_1, s_2, ..., s_m\}$. To produce a list of c clusters that group these m sentences, we adopt the approach of Örs et al. (2020) and first extract sentence pairs from A. Next, binary classification is performed to identify if a given pair is coreferent. Let (h, t) represent a sentence pair, with h and t referring to the lower and higher sentence numbers in A, respectively. The features employed for training a binary classifier that identifies coreferent sentences are described next.

3.1 Features for Sentence Pair Classification

BERT Embeddings: We utilize BERT, the bidirectional encoder transformer architecture (Devlin et al., 2019), to obtain sentence pair representations for our task. The models were pretrained with masked language modeling (MLM) and next sentence prediction (NSP) objectives. Our sentence pair input is encoded in this order: the special starting token "[CLS]", the head sentence, the separator "[SEP]" token, the tail sentence, another separator

[1] Our code and supplementary materials can be found on Github at https://github.com/NUS-IDS/EventSentenceCoref

token, and padding up to a fixed maximum length.

The encoded inputs, alongside attention mask and token type indexes, are fed into the BERT model. BERT acts as an encoder by producing sentence pair representations, which is later passed on to the BiLSTM model along with other features to train our classifier. As BERT is exposed to label information in the downstream layers, we are able to obtain fined-tuned representations for our task.

POS Embeddings: For each sentence, we obtain parts-of-speech (POS) tags for each word token to represent grammatical structure. To align with tokens of BERT embeddings, we similarly concatenate a starting token, POS tags of the head sentence plus a separator token, POS tags of the tail sentence plus a separator token, followed by padding. These POS tags are subsequently encoded as one-hot vectors and combined with the BERT embeddings per word before feeding them through a BiLSTM.

Lexical and Semantic Similarities: Event mentions in a sentence often correspond to specific POS and Named-Entity (NE) tags. Thus, similarity values capturing the overlap of these token types between the two sentences are indicative of whether they are coreferent. We incorporated lexical similarity based on surface-form overlap of POS and NE tags of sentences and semantic similarity based on sentence embeddings and overlap of the dependence trees of the two sentences. We represent the counts of verb, nouns, and entities occurring in both head and tail sentences using two similarity functions: raw counts and Jaccard overlap. These six features are referred to as "Basic Similarities" in our experiments.

For an "extended" set of similarities, we also computed the cosine similarity based on words of the sentences after stopword removal, and normalized dot product of vectors corresponding to words with POS tags pertaining to nouns, adjectives, verbs, and adverbs, and NER tags corresponding to tangible types such as person, organizations, products, geopolitical location. That is, named-entity tags corresponding to concepts such as money, quantities, and dates as well as POS tags corresponding to punctuation, and pronouns were ignored since they are unlikely to refer to event mentions.

For *semantic similarity* we use cosine similarity between the average word vectors from GloVE[2] for the two sentences. Ozates, et al.(2016) proposed incorporating the type information of the dependency relations for sentence similarity calculation in context of sentence summarization for better capturing the syntactic and semantic similarity between two sentences. We use similarity between two sentences computed using their proposed "Simple Bigram Approximate Kernel" as an additional feature.

Overall, the set of "Extended Similarities", correspond to a total of 27 features.

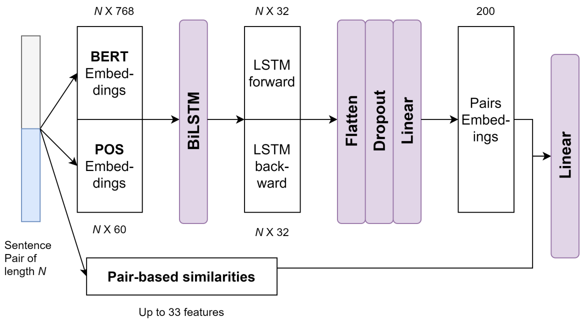


Figure 3: Overview of the sentence pair classification model. BERT embeddings, POS embeddings and similarity features are used to train a BiLSTM-based deep learning model.

3.2 Sentence Pair Classifier

Our deep learning setup for learning sentence pair classification is shown in Figure 3. We use the features described in the previous section for training our classifier. The BERT with POS embeddings are first fed into a BiLSTM layer with an output dimension of 64. Next, we flatten the $n \times 64$ matrix into a $n * 64$ vector and run it through a dropout layer with 0.3 dropout rate. Another linear layer is applied to convert the representation into a vector with length 200. From here, we concatenate our similarity features and send them through a linear layer to obtain class probabilities representing the coreferent (label = 1) and non-coreferent (label = 0) classes.

3.3 Article-level Clustering

Given labels corresponding to each pair of sentences obtained from our classification module, we employ the clustering algorithm from Örs et al. (2020) for grouping the sentences in the document. This algorithm, similar to hierarchical clustering, creates clusters in a bottom-up fashion using maximum scores instead of the minimum distance to

[2] `http://nlp.stanford.edu/data/glove.6B.zip`

Obs Unit	English	Portuguese	Spanish
Train			
Articles	596	21	11
Sentences	2581	88	45
Pairs	6241	235	86
Test			
Articles	100	40	40
Sentences	486	144	188
Pairs	1554	257	549

Table 1: Number of observations at different unit levels for train and test set

group two points into the same cluster. For us, score of a pair refers to the probability of the sentences being coreferent with. We refer the interested reader to Algorithm 2 in Örs et al. (2020) for the pseudo-code. In contrast, with algorithms such as k-medoids, the algorithm employed in our solution has the advantage of determining a different number of clusters for each article in a flexible manner.

4 Experiments and Results

4.1 Dataset and Evaluation

We used the data from ACL-CASE 2021 (Hürriyetoğlu et al., 2021) (Task 1 Subtask 3) for training and testing our models. The dataset comprises of news articles referring to socio-political and crisis event in three languages: English, Portuguese, and Spanish. We refer the interested reader to the overview paper (Hürriyetoğlu et al., 2021) and the task websites[3] for details of this dataset. We summarize the train and test sizes of the dataset in Table 1. For the train set, we were provided with 596 English news articles, 21 Portuguese articles, and 11 Spanish articles. For each article, only sentences with event mentions are included in the dataset instead of all sentences.

The test performance was evaluated using the CoNLL-2012 average F_1 scores obtained by averaging across MUC, B^3 and $CEAF_e$ F_1 scores (Pradhan et al., 2012) and was computed on the setup provided by the organizers on Codalab.

4.2 Experimental Setup

Training Datasets: To handle the low number of examples available with Portuguese and Spanish, we create two datasets for training our models: (1) The "Multilingual train set" is obtained by simply putting the examples from all languages together whereas (2) the "English train set" is obtained by first employing the Google Translate API[4] and translating all available non-English training examples to English and combining with the English training data. The multilingual dataset can be used directly for training language-agnostic models, for example using cross-lingual embeddings (Conneau et al., 2017) and Multilingual BERT.

Feature Extraction: We experimented with two BERT implementations from Huggingface (Wolf et al., 2020). The first model, `bert-base-cased`, was pretrained on English text and has 12 layers, 768 hidden, 12 heads and 109M parameters. We fine-tuned this model using our "English train set". Our second model, `bert-base-multilingual-cased`, was pretrained on the top 104 languages in Wikipedia and has 12 layers, 768 hidden, 12 heads and 179M parameters. We fine-tuned this model using our "Multilingual train set".

We used Stanford's `Stanza` package (Qi et al., 2020) for obtaining POS, NER, and dependency tree tags. The "Universal POS tags" (`upos` scheme) with 17 POS tags and the NER tags referring to PERSON, NORP (Nationalities/religious/political group), FAC (Facility), ORG (Organization), GPE (Countries/cities/states), LOC (Location), PRODUCT, EVENT, WORK_OF_ART, and LANGUAGE were used in experiments.[5]

When constructing the "Basic Similarities", all words are lemmatised before we compare their surface-form overlap. For entities, we use `token_sort_ratio`[6] score of more than 90% to define a positive overap occurence instead of an exact match to allow for some small discrepancy in NEs (e.g. "Sohrabuddin Sheikh" and "Sohrabuddin Sheikh 's" refer to the same entity).

Classifier Settings: To train our classifier, we used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of $2e - 5$ with linear decay. Cross Entropy Loss was used with class weights computed from the training sample. Each

[3] https://emw.ku.edu.tr/case-2021/, https://github.com/emerging-welfare/case-2021-shared-task

[4] https://pypi.org/project/google-trans-new

[5] At present, NER models are only available for Spanish and English in Stanza.

[6] https://github.com/seatgeek/fuzzywuzzy

	CoNLL F_1	ARI		F_1	
		Macro	Micro	Macro	Micro
BERT	84.46	64.73	54.76	68.76	60.82
+ POS embeddings	83.15	56.81	48.94	60.58	54.44
+ Basic similarities	84.31	64.63	55.98	67.54	60.35
+ Extended similarities	**84.92**	**66.78**	**57.66**	**70.68**	**62.94**
Multilingual BERT	82.56	59.97	52.64	62.00	55.41
+ POS embeddings	83.98	61.79	52.89	65.59	58.33
+ Basic similarities	82.83	60.62	50.09	64.47	56.20
+ Extended similarities	81.80	57.53	48.74	61.71	54.70

Table 2: Evaluation results over validation sets from 5 folds. *Notes.* Scores are reported in percentages (%) and averaged across the folds. Best score per column is bolded.

iteration was of batch size 16 and all experiments were ran on Tesla V100 SXM2 32GB GPU device.

Five-fold cross-validation (5-CV) experiments were used for parameter tuning. We also report macro and micro Adjusted Rand Index (ARI) and F_1 scores in addition to ConLL F_1 since they were used for selecting the top-3 runs for the test set in line with the measures employed in the previous rounds of the competition (Hürriyetoğlu et al., 2019, 2020). Since the test labels were not released and evaluation is performed on the competition setup, only CoNLL F_1 scores are reported for the test data. Other details, such as hyperparameter settings and run times, are included in Appendix A.1.

4.3 Results and Analysis

Table 2 reports the average scores for our 5-CV setup across the five scoring metrics (CoNLL F_1, Macro ARI, Micro ARI, Macro F_1 and Micro F_1). Table 3 reports the CoNLL F_1 score on the test data for the winning system and our models at the ACL-CASE 2021 Shared Task 1 Subtask 3 across the three languages – English, Portuguese, and Spanish.

Based on CV experiments, our best model for all scoring measures is the English BERT model with all features included, achieving 84.92% CoNLL F_1 score. The same model also performed the best on the English test set with 81.20% CoNLL F_1 score and was ranked 2nd among fellow competitors on this shared task.

For non-English test sets, our best performing model is the Multilingual BERT model with all features excluding "Extended similarities". This model achieved 93.03% CoNLL F_1 score and ranked 1st for Portuguese. For Spanish, we obtained a CoNLL F_1 score of 83.15% and ranked 2nd among competitors.

4.3.1 BERT versus Multilingual BERT

For the English test set, the BERT model performs better than the Multilingual BERT model on average (79.19% versus 78.01%). Additionally, because the train/validation splits are predominantly comprised of English articles (596/628 = 94.90%), the fluctuations in performance on validation splits largely tally with the fluctuations in performance on the English portion of the data. Therefore, unsurprisingly, BERT (English) performs better than Multilingual BERT for English data.

For non-English test sets, we obtained best performance using the Multilingual BERT model. We hypothesize that the translation of non-English examples to English might have caused some loss of inherent signals present in other languages that are useful for the ESCI task. These signals are possibly better harnessed by retaining the language and using language-specific Stanza taggers along with Multilingual BERT.

Overall, we find that combining BERT embeddings, POS embeddings and basic similarity features achieve the best validation performance across all measures. We observe that "Extended similarities" do not show uniform improvement in performance in multilingual settings. Arguably, there is redundancy among our lexical similarity features and semantic similarities were not found to improve performance on the ESCI task. However, considering the small-scale of our working datasets, these features need further study.

5 Conclusions and Future Work

We presented a two-step solution for the ESCI task of ACL-CASE 2021. Our solution based on

	CoNLL F_1		
	English	Portuguese	Spanish
Best Score in Competition	84.44	93.03	84.23
BERT	80.54	88.85	80.18
+ POS embeddings	77.79	90.58	82.17
+ Basic similarities	77.23	90.21	80.11
+ Extended similarities	**81.20**	92.18	80.91
Multilingual BERT	77.89	87.22	76.55
+ POS embeddings	79.33	90.92	81.43
+ Basic similarities	78.13	**93.03**	**83.15**
+ Extended similarities	76.68	90.36	81.52

Table 3: Evaluation results over test sets submitted to Codalab. *Notes.* Scores are reported in percentages (%). Our best score per column is bolded.

sentence pair classification effectively harnesses BERT and POS embeddings and combines them with linguistic similarity features for accurate sentence coreference resolution across languages. Indeed, our models ranked 2$^{\text{nd}}$, 1$^{\text{st}}$ and 2$^{\text{nd}}$ obtaining CoNLL F_1 scores of 81.20%, 93.03%, 83.15% for the English, Portuguese and Spanish test sets, respectively, in the competition.

In this paper, we focused on within-document coreference sentences. It is common for coreference resolution tasks to also focus on cross-document settings (i.e. identify coreferent event mentions across multiple documents) (Zeng et al., 2020) as such models can better aid downstream tasks like contradiction detection or identification of “fake news”. In future, we hope to extend our models to work across documents. Additionally, multiple events might be presented in a sentence. The shared task focuses on hard clustering (i.e. each sentence can only belong to one cluster). However, we believe it is valuable to also investigate cases where the event clusters overlap.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund - Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*

(*CASE 2021*), online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, and Osman Mutlu. 2019. A task set proposal for automatic protest information collection across multiple countries. In *Advances in Information Retrieval*, pages 316–323, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. Event linking with sentential features from convolutional neural networks. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 239–249, Berlin, Germany. Association for Computational Linguistics.

Faik Kerem Örs, Süveyda Yeniterzi, and Reyyan Yeniterzi. 2020. Event clustering within news articles. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 63–68, Marseille, France. European Language Resources Association (ELRA).

Şaziye Betül Özateş, Arzucan Özgür, and Dragomir Radev. 2016. Sentence similarity based on dependency tree kernels for multi-document summarization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2833–2838, Portorož, Slovenia. European Language Resources Association (ELRA).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Replication Checklist

- Hyperparameters: Apart from hyperparameters mentioned in Section 4.2, our BERT models take the default configuration from Huggingface (Wolf et al., 2020).

- Time taken: For 5 folds over 10 epochs each, our code takes on average $5hours : 27minutes : 48seconds$ to train, validate and predict. For a single run over 10 epochs, our code takes on average $43minutes : 49seconds$ to train and predict.