Exploring Pre-Trained Transformers and Bilingual Transfer Learning for Arabic Coreference Resolution

Bonan Min
Raytheon BBN Technologies
10 Moulton Street, Cambridge, MA 02138
bonan.min@raytheon.com

Abstract

In this paper, we develop bilingual transfer learning approaches to improve Arabic coreference resolution by leveraging additional English annotation via bilingual or multilingual pre-trained transformers. We show that bilingual transfer learning improves the strong transformer-based neural coreference models by 2-4 F1. We also systemically investigate the effectiveness of several pre-trained transformer models that differ in training corpora, languages covered, and model capacity. Our best model achieves a new state-of-the-art performance of 64.55 F1 on the Arabic OntoNotes dataset. Our code is publicly available at https://github.com/bnmin/arabic_coref.

1 Introduction

Coreference resolution is the task of identifying all mentions which refers to the same discourse entity. It is a crucial step for many NLP tasks, and has been well-studied (Ng, 2010) for English. Recently, neural coreference resolution models based on pre-trained tranformers, e.g., BERT (Devlin et al., 2019) and SpanBERT (Joshi et al., 2020), have shown to be highly effective for English coreference resolution, achieving 80 F1 on the OntoNotes (Pradhan et al., 2012) dataset. However, coreference resolution in Arabic is still a challenging problem with previous systems (Björkelund and Kuhn, 2014; Chen and Ng, 2012; Fernandes et al., 2012; Zhekova et al., 2012; Stamborg et al., 2012; Xiong and Liu, 2012; Aloraini et al., 2020) performing significantly lower than the English systems.

A main problem is that Arabic coreference resolution datasets are much smaller than their English counterparts. For example, the widely useful benchmark dataset OntoNotes contains 7 times more documents in English than Arabic (Table 2).

Based on this observation, we develop bilingual transfer learning approaches to improve Arabic coreference resolution by leveraging additional English coreference annotation via bilingual or multilingual pre-trained transformers, e.g., multilingual BERT (Devlin et al., 2019). We summarize our contributions below:

- We develop several transformer-based models for Arabic coreference resolution. Our best model achieves a new state-of-the-art of 64.55 F1 in Arabic coreference resolution. Our source code are made publicly available.
- We develop two bilingual transfer learning approaches to improve Arabic coreference resolution. We show bilingual transfer learning, that combines English and Arabic coreference annotation via multilingual pre-trained transformers, leads to significant improvements in Arabic coreference resolution.
- We systemically investigate the effectiveness of several pre-trained transformer models that differ in training corpora, languages covered, and model capacity. We show that language-specific pre-training is a crucial success factor.

2 Related Work

Neural models for coreference resolution: State-of-the-art models for coreference resolution are based on neural networks (NN). E2E-COREF (Lee et al., 2018) is an NN model that performs span extraction and coreference classification in a single end-to-end model. BERT-based models (Joshi et al., 2019a, 2020) further improve E2E-COREF by replacing the Bi-LSTM text encoder with BERT (Devlin et al., 2019) or SPANBERT (Joshi et al., 2020). COREFQA (Wu et al., 2020) formulates coreference resolution as a QA task leveraging pre-trained transformers.

Arabic coreference resolution: Björkelund and Kuhn (2014) investigates structured perceptron models for Arabic coreference resolution. Chen

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 94–99
Punta Cana, Dominican Republic, November 10–11, 2021.

and Ng (2012) uses a hybrid approach combining rule-based methods and learning-based methods. Fernandes et al. (2012) is a latent structure perceptron model along with entropy-guided feature induction. Zhekova et al. (2012) is a memory-based coreference resolution system. Stamborg et al. (2012) uses a mention classifier with features extracted from the dependency graphs of the sentences. Uryupina and Moschitti (2013) proposed an algorithm for multilingual mention detection by extracting mentions from parse trees via kernel-based SVM learning. Haponchyk and Moschitti (2017) improves structured prediction models for Arabic coreference resolution by using more expressive loss functions. Xiong and Liu (2012) is a projection-based model in which they first translate Arabic into English, run a coreference resolution system, and then map the coreference clusters back into Arabic. Aloraini and Poesio (2020) developed a BERT-based cross-lingual models for zero pronoun resolution in Arabic and Chinese.

Aloraini et al (2020) developed an Arabic neural coreference system, and showed that data processing and mention detection improve the performance by about 7.8 F1 points over the pure neural coreference system, resulting in 63.9 F1 that is higher than previous approaches but lower than ours (Table 3). Our systems are pure neural systems and we focus on a different problem (bilingual transfer learning). Our approaches and Aloraini et al (2020) are complementary, so it is possible that combining ours and Aloraini et al (2020) can achieve further improvements. We leave this for future work.

Pre-trained transformers: Pre-trained transformer based language models such as BERT has been shown to be effective for many NLP tasks. ARABERT (Antoun et al., 2020) is a BERT-like model trained with Arabic text. To enable cross-lingual transfer learning, multilingual pre-trained transformers such as MBERT (Devlin et al., 2019) are trained with large corpora with many languages to encode similar words (words with similar meanings) from different languages into nearby high-dimensional vector spaces. GIGABERT (Lan et al., 2020) is an Arabic-English bilingual language model trained with Arabic and English text.

3 Neural Coreference Models for Arabic

Our coreference resolution model is based on an end-to-end neural model BERT-COREF (Joshi et al., 2019b), which is built on top of E2E-COREF (Lee et al., 2018). Below we give a brief overview of BERT-COREF , and then describe our extensions to this model for Arabic coreference resolution.

3.1 Overview of BERT-COREF

BERT-COREF is an neural model that performs span extraction and coreference relation classification in a single end-to-end model. It first enumerates mention spans, and then for pairs of mention spans, scores their likelihoods of referring to the same entity. The training signal is back-propagated to allow the model to learn to perform span extraction and coreference resolution simultaneously, leading to impressive performance on the English coreference resolution task on the OntoNotes dataset (Pradhan et al., 2012).

For each mentions span x, BERT-COREF learns a distribution P over possible antecedent spans Y:

$$P(y) = \frac{e^{s(x,y)}}{\sum_{y' \in Y} e^{s(x,y')}}$$

The scoring function $s(x, y)$ between spans x and y uses span representations $\mathbf{g}_x$ and $\mathbf{g}_y$ to represent its inputs. The span representation are generated with BERT (Devlin et al., 2019). The model computes $s(x, y)$ as the sum of mention scores $s_m(x)$ and $s_m(y)$ for x and y respectively, and a compatibility score $s_c(x, y)$ of x and y that indicates how likely the two spans refer to the same entity:

$$s(x, y) = s_m(x) + s_m(y) + s_c(x, y)$$

$$s_m(x) = \text{FF}_m(\mathbf{g}_x)$$

$$s_c(x, y) = \text{FF}_c(\mathbf{g}_x, \mathbf{g}_y, \phi(x, y))$$

where FF represents a feed-forward (FF) network and $\phi(x, y)$ represents additional features such as speakers and metadata.

For more details, we refer readers to Lee et al. (2018) and Joshi et al. (2019b).

3.2 Arabic Coreference Resolution with Pre-Trained Transformers

We use BERT-COREF to train end-to-end neural coreference models. To encode Arabic text, we replace the English-only BERT text encoder with pre-trained transformers that are trained with Arabic corpora or multilingual corpora including Arabic.

Table 1 summarizes the configurations of the following three pre-trained transformers that we use to replace BERT to encode words in Arabic (and also English words for MBERT and GIGABERT):

Model	Training source	Training tokens (all/en/ar)	# Model parameters
ARABERT	newswire	2.5B/ – /2.5B	136M
MBERT	Wikipedia	21.9B/2.5B/153M	172M
GIGABERT	Gigaword, Wikipedia, Oscar	10.4B/6.1B/4.3B	125M

Table 1: Model configurations for the pre-trained transformers ARABERT (Antoun et al., 2020), MBERT (Devlin et al., 2019), and GIGABERT (Lan et al., 2020). Wikipedia refers to the English and Arabic portions of the Wikipedia. en and ar refers to English and Arabic respectively.

	Train	Development	Test
English	2,802	343	348
Arabic	359	44	44

Table 2: Number of documents in the *train*, *dev* and *test* sections of the English and Arabic OntoNotes datasets.

- ARABERT: An Arabic BERT-based model trained from 2.5 billion words of publicly available Arabic corpora, including El-khair (2016) that includes more than 5 million articles from ten major news sources covering 8 countries and the Open Source International Arabic News Corpus (Zeroual et al., 2019) that consists of 3.5 million articles from 31 news sources in 24 Arab countries.

- MBERT: Multilingual BERT is a multilingual BERT model that was trained with the Wikipedia dump for the top 104 languages with the largest Wikipedias.

- GIGABERT: an English-Arabic bilingual language model pre-trained from the fifth edition of English and Arabic Gigaword corpora [1] which consists of 13 million news articles, English and Arabic portions of Wikipedia [2], and the Arabic section of the Oscar corpus (Ortiz Suárez et al., 2019), a large-scale multilingual dataset filtered from the Common Crawl.

Table 1 summarizes training data and model configurations for the three models above.

3.3 Bilingual Transfer Learning

Pires et al. (2019) show that MBERT is surprisingly good at cross-lingual transfer learning for several NLP tasks, in which the task-specific training data in one language is used to fine-tune the model and then evaluate in a zero-shot setting in a different language. However, its effectiveness in Arabic coreference resolution has not been studied.

Inspired by the prior work, we developed two approaches for bilingual transfer learning to leverage OntoNotes' English coreference resolution annotation to improve Arabic coreference resolution:

Pre-training on English + fine-tuning on Arabic: We pre-train BERT-COREF for m epochs with the English training dataset, and then fine-tune it with n epochs using the Arabic training dataset. We replace the BERT encoder with MBERT or GIGABERT to allow cross-lingual transfer learning. We call this the PIPELINE approach for bilingual transfer learning (or PIPELINE for short).

Joint training on English and Arabic: We combine the English and Arabic training dataset to jointly train the BERT-COREF model. Similarly, the model uses MBERT or GIGABERT instead of the monolingual BERT as the text encoder. For each epoch, we shuffle the combined set of English and Arabic documents and perform stochastic gradient descent (SGD) on the randomized mixture of mini-batches from both English and Arabic. We call this approach the JOINT training approach for bilingual transfer learning (or JOINT for short).

4 Experiments

We evaluate the models on OntoNotes 5.0 (Pradhan et al., 2012), which is widely used as a benchmark dataset for coreference resolution. OntoNotes also contains both Arabic and English documents annotated with coreference, which makes bilingual transfer learning (Section 3.3) experiments feasible. Table 2 shows the numbers of English and Arabic documents in OntoNotes. To not overwhelm the much smaller Arabic dataset with the English dataset, we replicate the Arabic dataset $k = 6$ times [3] so that the amount of English and Arabic data is similar. We use the official *train/dev/test* split from the CoNLL 2012 shared task on multi-

[1] https://catalog.ldc.upenn.edu/LDC2011T07 and https://catalog.ldc.upenn.edu/LDC2011T11

[2] https://www.wikipedia.org/

[3] We found that $k = 6$ produces the best result on the *dev* dataset.

Model	MUC			B^3			CEAF			
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Avg. F1**
(1) Coreference resolution models trained with Arabic data										
MBERT	61.71	43.81	51.24	56.72	36.34	44.3	53.79	38.48	44.86	46.8 (47.44)
ARABERT	70.07	51.61	59.44	65.87	45.32	53.69	64.5	49.62	56.09	56.41 (56.74)
GIGABERT	73.08	59.24	65.44	69.49	52.63	59.9	64.9	57.42	60.93	62.09 (61.91)
(2) Coreference resolution models trained with both English and Arabic data										
MBERT-PIPELINE	61.2	48.2	53.93	56.35	40.84	47.36	53.52	43.8	48.17	49.82 (49.79)
MBERT-JOINT	58.83	46.07	51.67	54.11	39.59	45.73	52.86	43.05	47.45	48.28 (49.13)
GIGABERT-PIPELINE	71.46	63.19	67.07	66.93	57.62	61.92	65	61.48	63.19	64.06 (64.72)
GIGABERT-JOINT	73.63	61.81	**67.21**	70.74	55.92	**62.46**	66.07	62.03	**63.99**	**64.55** (64.67)

Table 3: Performance of the models on OntoNotes Arabic *test* set. For validation, we additionally show the Avg. F1 scores on the OnteNotes Arabic *dev* set in the parentheses in the last column.

lingual coreference (Pradhan et al., 2012).

We use BERT-COREF as the end-to-end neural coreference model and then replace the BERT encoder with other encoders. We implemented three models: MBERT, ARABERT, and GIGABERT, named after the coresponding pre-trained tranformers that are used as the text encoder. We train each model with only the Arabic training data. Results are shown in Table 3 (1).

To leverage English annotation in bilingual transfer learning, we also implemented the following models (shown in Table 3 (2)) that are trained with both Arabic and English training data.

- MBERT-PIPELINE: We use MBERT as the text encoder. We train the model with the PIPELINE approach (Section 3.3). We set the number of pre-training epochs $m = 20$ and the number of fine-tuning epochs $n = 20$.
- GIGABERT-PIPELINE: Same as above except that we use GIGABERT as the text encoder.
- MBERT-JOINT: We use MBERT as the text encoder. We train the model with the JOINT approach (Section 3.3).
- GIGABERT-JOINT: Same as MBERT-JOINT except that GIGABERT is used as the encoder.

For all models, we set the pre-trained transformers' learning rate to $1e-5$, the task learning rate to $2e-4$, the number of epochs to 30 except for the PIPELINE models, and use the default values for the remaining parameters. We use the performance on the *dev* set to choose the best hyperparameter for all models, and use early stopping to prevent overfitting. All models are trained with an NVIDIA QUADRO RTX 8000 GPU. Each model trains for 3 to 12 hours depending on the size of the data (Arabic or the combined set).

Table 3 shows the performance numbers reported by the CoNLL 2012 scorer. GIGABERT-JOINT achieves the highest performance with an average F1 of 64.55. establishing a new state-of-the-art model for Arabic coreference resolution.

On pre-trained tranformers for Arabic: Comparing MBERT, ARABERT and GIGABERT, Table 3 (1) shows a stunning difference in their performance in Arabic coreference resolution. GIGABERT shows the best F1 at 62.09, ARABERT comes second at 56.41 F1, and MBERT performs the worst at 46.8 F1. There are two important factors that contribute to the difference:

First, the size of Arabic text used for pre-training significantly affects the pre-trained transformer' ability to encode text for Arabic coreference resolution. As shown in Table 1, MBERT, ARABERT and GIGABERT use drastically different amounts of Arabic training tokens (153M, 2.5B and 4.3B for MBERT, ARABERT and GIGABERT, respectively). The more data that the transformer is pre-trained on, the higher the performance.

Second, ARABERT and GIGABERT are trained specifically for Arabic, while MBERT is trained for 104 languages and has not been tuned for Arabic. The language-specific training makes ARABERT and GIGABERT models specialized to capture the intricacies of Arabic, leading to better performance.

On bilingual transfer learning: Comparing the bilingual transfer learning models (Table3 (2)) to the Arabic-only models (Table3 (1)), we observe 3 and 2.5 average F1 improvements when using MBERT and GIGABERT respectively. This shows the effectiveness of the bilingual transfer learning. The PIPELINE and JOINT approach shows similar

gains. This indicates that it is the additional training data rather than the learning model that leads to the gain. Interestingly, using MBERT and GIGABERT gives similar gains, which shows that the transferability of the these two models in English to Arabic is about the same (though with MBERT lagging in baseline Arabic performance by 15.29 F1). It is surprising given that GIGABERT is a bilingual model trained specifically for English-Arabic. We plan to investigate this as an immediate next step.

5 Conclusion

We present bilingual tranfer learning for Arabic coreference resolution, that achieved new state-of-the-art performance as evaluated on OntoNotes.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600006 under the IARPA BETTER program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes not withstanding any copyright annotation therein.

References

Abdulrahman Aloraini and Massimo Poesio. 2020. Cross-lingual Zero Pronoun Resolution. In *Proceedings of LREC 2020*.

Abdulrahman Aloraini, Juntao Yu, and Massimo Poesio. 2020. Neural coreference resolution for Arabic. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Barcelona, Spain (online). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Anders Björkelund and Jonas Kuhn. 2014. Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. pages 47–57.

Chen Chen and Vincent Ng. 2012. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*.

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.

Ibrahim Abu El-khair. 2016. 1.5 billion words arabic corpus.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*.

Iryna Haponchyk and Alessandro Moschitti. 2017. Don't understand a measure? learn it: Structured prediction for coreference resolution optimizing its measures. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1028, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019a. BERT for Coreference Resolution: Baselines and Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, arXiv.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019b. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, arXiv.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Confer-*

ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

Marcus Stamborg, Dennis Medved, Peter Exner, and Pierre Nugues. 2012. Using Syntactic Dependencies to Solve Coreferences. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*.

Olga Uryupina and Alessandro Moschitti. 2013. Multilingual mention detection for coreference resolution. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 100–108, Nagoya, Japan. Asian Federation of Natural Language Processing.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Hao Xiong and Qun Liu. 2012. ICT: System Description for CoNLL-2012. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

Desislava Zhekova, Sandra Kübler, Joshua Bonner, Marwa Ragheb, and Yu-Yin Hsu. 2012. UBIU for Multilingual Coreference Resolution in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*.