A Semi-Supervised BERT Approach for Arabic Named Entity Recognition

Chadi Helwe, Ghassan Dib, Mohsen Shamas, Shady Elbassuoni
Télécom Paris, Institut Polytechnique de Paris
Department of Computer Science, American University of Beirut
chadi.helwe@telecom-paris.fr
{gid01, mys12, se58}@aub.edu.lb

Outline

- Introduction
- Related Work
- Approach
- Evaluation
- Conclusion

Introduction

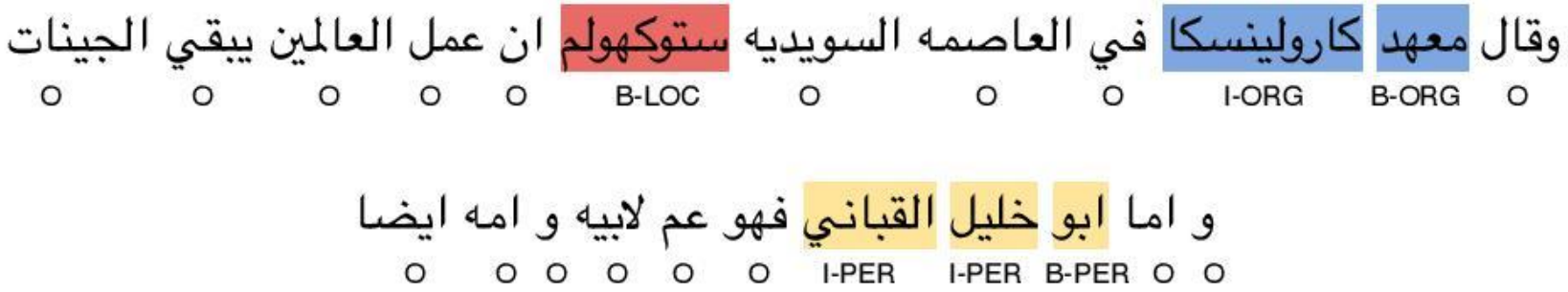
Introduction

In this paper, we proposed a semi-supervised BERT approach to detect and classify named entities in Arabic

Named Entity Recognition (NER):

- It is the task of extracting and locating, and classifying named entities in a given text
- A named entity can be: a proper noun, a numerical expression representing type unit or monetary value, or a temporal value that represents time

Example of NER



Arabic NER

NER in Arabic is considered a particularly difficult task:

- There is no capitalization in the Arabic script
- Arabic can be ambiguous
- Lack of sufficient resources

Solution ?

We proposed a BERT model trained in a semi-supervised fashion using two datasets: a small labeled dataset and a large semi-labeled dataset

Related Work

Related Work

Machine Learning Approaches	Rule Based Approaches	Hybrid Approaches
CRF models (Abdul-Hamid and Darwish [1], Benajiba and Rosso [2], Benajiba et al. [3])	Lexical Triggers (Abuleil [14], Al-Shalabi et al. [15])	Combination of machine learning models and rule based approaches (Abdallah et al. [21], Oudah and Shaalan [22], Shaalan and Raza [23])
SVM models (Abdelali et al. [4], Benjiba et al. [5], Koulali and Meziane [6], Pasha et al. [7])	Morphological Analysers (Elsebai et al. [16], Maloney and Niv [17], Mesfar [18])	
Approaches relied on meta-classifiers (AbdelRahman et al. [8], Benajiba et al. [9], Benajiba et al. [10])	Regular expressions and gazetteers (Shalan and Raza [19])	
Deep learning approaches (Gridach [11], Helwe and Elbassuoni [12], Antoun et al. [13])	Transliteration (Samy et al. [20])	

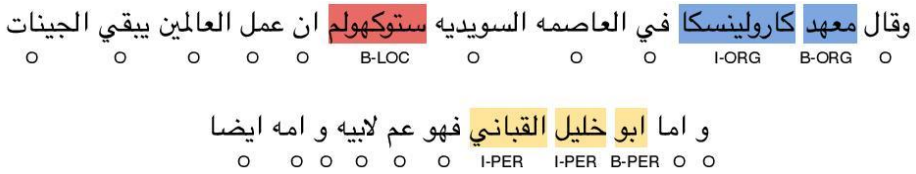
Approach

AraBERT Model

AraBERT model:

- It is a pre-trained Arabic language model developed by Antoun et al. [13]
- It has the same architecture of the BERT model developed by Devlin et al. [24]
- It was pretrained on two tasks:
  - Masked Language Modeling (MLM)
  - Next Sentence Prediction (NSP)
- The datasets used:
  - Arabic Wikidumps
  - 1.5B words Arabic Corpus
  - OSIAN corpus
  - Assafir, Al Akhbar, Annahar, AL-Ahram, and AL-Wafd (news websites)

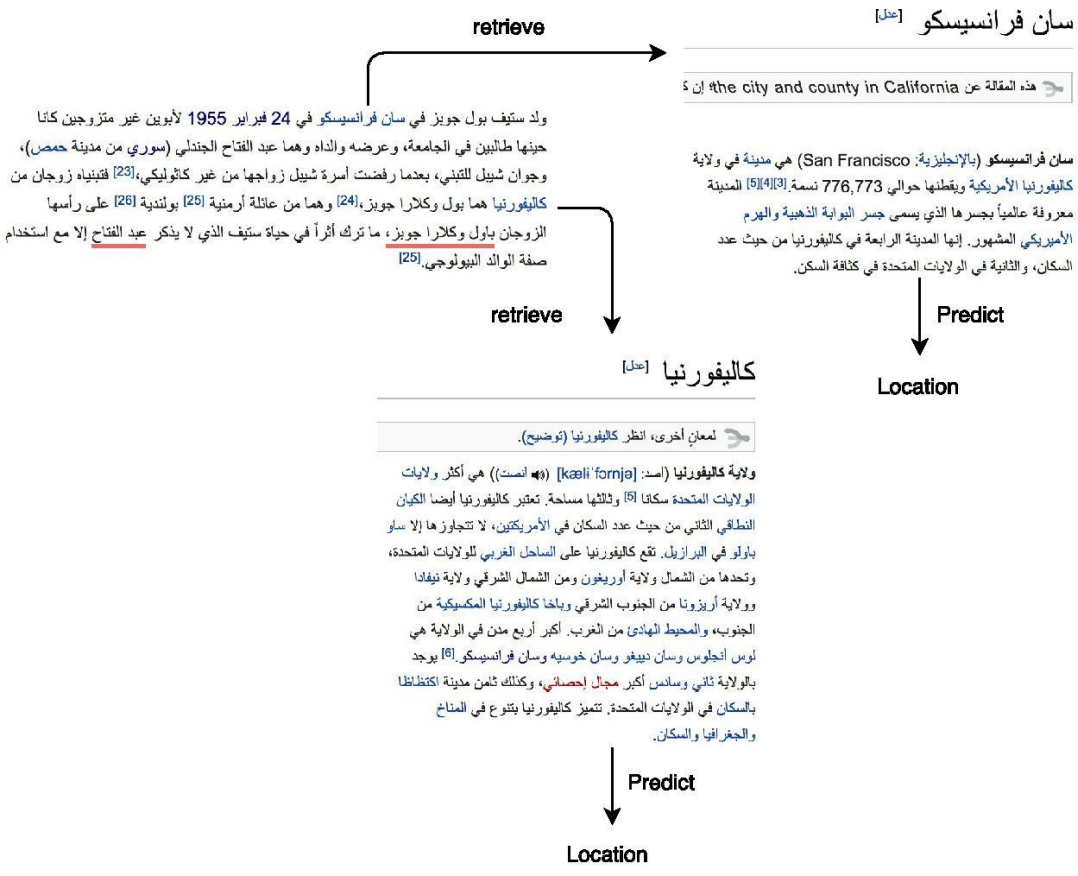
Examples of the Datasets



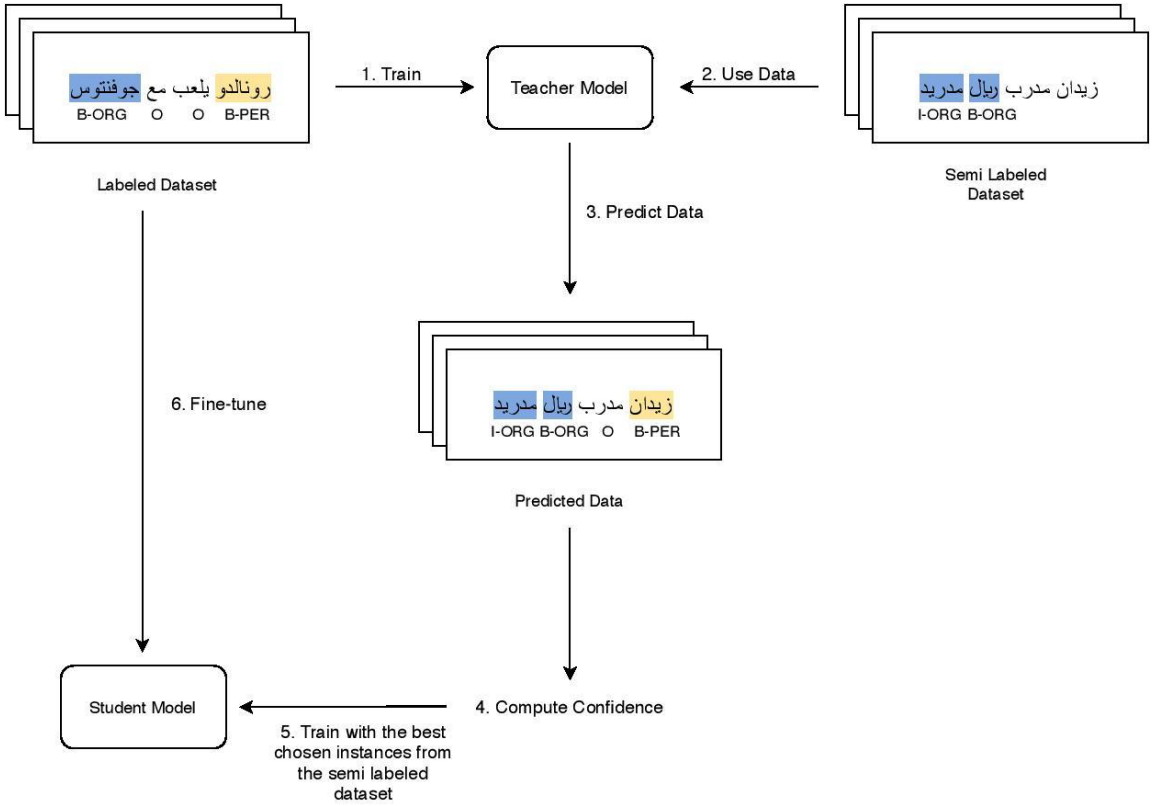
Instances of the Labeled Dataset

Instances of the Semi Labeled Dataset

Semi-labeled Dataset



Semi-Supervised Learning Model for Arabic NER



Evaluation

Datasets

Training and Validation:

- ANERcorp Dataset (Benajiba et al. [3]):
  - Training dataset 80%: 3,686 sentences
  - Validation dataset 20%: 461 sentences
- Semi-labeled Dataset (Helwe and Elbassuoni [12]):
  - 1,617,184 labeled and unlabeled tokens

Testing:

- AQMAR Dataset (Mohit et al. [25]):
  - Corpus of 28 Arabic Wikipedia Articles
  - 2,456 sentences
- NEWS Dataset (Darwish [24]):
  - 292 sentences
- TWEETS Dataset (Darwish [24]):
  - 982 tweets

AQMAR Benchmark

Model	LOC	ORG	PER	AVG
MADAMIRA [7]	39.4	15.1	22.3	29.2
FARASA [4]	60.1	30.6	52.5	52.9
Deep Co-learning [12]	67.0	38.2	65.1	61.8
AraBERT Fully Supervised	63.6	31.0	70.9	61.5
AraBERT Semi-Supervised	68.4	34.6	74.4	**65.5**

NEWS Benchmark

Model	LOC	ORG	PER	AVG
MADAMIRA [7]	38.8	12.6	29.4	28.4
FARASA [4]	73.1	42.1	69.5	63.9
Deep Co-learning [12]	81.6	52.7	82.4	74.1
AraBERT Fully Supervised	74.2	54.2	85.1	73.2
AraBERT Semi-Supervised	80.5	60.8	89.5	**78.6**

TWEETS Benchmark

Model	LOC	ORG	PER	AVG
MADAMIRA [7]	40.3	8.9	18.4	24.6
FARASA [4]	47.5	24.7	39.8	39.9
Deep Co-learning [12]	65.3	39.7	61.3	**59.2**
AraBERT Fully Supervised	57.9	30.7	60.9	54.0
AraBERT Semi-Supervised	63.3	42.1	59.4	57.3

Conclusion

Conclusion and Future Work

Conclusion:

- We proposed a semi-supervised BERT approach to detect and classify named entities in Arabic text
- Our model outperforms all other Arabic NER tools and approaches on two testing datasets: AQMAR and NEWS datasets

Future Work:

- We plan to pre-train the BERT model on tweets to make it more suitable for text that could contain misspelling and mistakes
- We plan to apply our approach on other NLP tasks such as part-of-speech tagging and dependency parsing

Thank You !!

References

References

[1] Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115. Association for Computational Linguistics.

[2] Yassine Benajiba and Paolo Rosso. 2007. Anersys 2.0: Conquering the ner task for the arabic language by combining the maximum entropy with pos-tag information. In *IICAI*, pages 1814–1823.

[3] Yassine Benajiba, Paolo Rosso, and Jose´ Miguel Bened´ıruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computa- tional Linguistics*, pages 143–153. Springer.

[4] Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16. Association for Computational Linguistics, San Diego, California.

[5]Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008a. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293. Association for Computational Linguistics.

[6] Rim Koulali and Abdelouafi Meziane. 2012. A contribution to arabic named entity recognition. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on*, pages 46– 52. IEEE.

[7] Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

[8] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy, and Aly Fahmy. 2010. Integrated machine learning techniques for arabic named entity recognition. *IJCSI*, 7:27–36.

[9] Yassine Benajiba, Mona Diab, Paolo Rosso, et al. 2008b. Arabic named entity recognition: An svm-based approach. In *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*, pages 16–18.

[10] Yassine Benajiba, Imed Zitouni, Mona Diab, and Paolo Rosso. 2010. Arabic named entity recognition: using features extracted from noisy data. In *Proceedings of the ACL 2010 conference short papers*, pages 281–285. Association for Computational Linguistics.

[11] Mourad Gridach. 2016. Character-aware neural networks for arabic named entity recognition for social media. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 23–32.

References

[12] Chadi Helwe and Shady Elbassuoni. 2019. Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, 52(1):197–215.

[13] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

[14] Saleem Abuleil. 2004. Extracting names from arabic text for question-answering systems. In *Coupling ap- proaches, coupling media and coupling languages for information retrieval*, pages 638–647. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[15] Riyad Al-Shalabi, Ghassan Kanaan, Bashar Al-Sarayreh, Khalid Khanfar, Ali Al-Ghonmein, Hamed Talhouni, and Salem Al-Azazmeh. 2009. Proper noun extracting algorithm for arabic language. In *International conference on IT, Thailand*.

[16] Ali Elsebai, Farid Meziane, and Fatma Zohra Belkredim. 2009. A rule based persons names arabic extraction system. *Communications of the IBIMA*, 11(6):53–59.

[17] John Maloney and Michael Niv. 1998. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pages 8–15. Association for Computational Linguistics.

[18] Slim Mesfar. 2007. Named entity recognition for arabic using syntactic grammars. In *Natural Language Processing and Information Systems*, pages 305–316. Springer.

[19] Khaled Shaalan and Hafsa Raza. 2007. Person name entity recognition for arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 17–24. Association for Computational Linguistics.

[20] Doaa Samy, Antonio Moreno, and Jose M Guirao. 2005. A proposal for an arabic named entity tagger leveraging a parallel corpus. In *International Conference RANLP, Borovets, Bulgaria*, pages 459–465.

[21] Sherief Abdallah, Khaled Shaalan, and Muhammad Shoaib. 2012. Integrating rule-based system with classi- fication for arabic named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 311–322. Springer.

[22] Mai Oudah and Khaled F Shaalan. 2012. A pipeline arabic named entity recognition using a hybrid approach. In *COLING*, pages 2159–2176.

[23] Khaled Shaalan and Hafsa Raza. 2009. Nera: Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(8):1652–1663.

References

[24] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[25] Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *ACL (1)*, pages 1558–1567

[26] Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173. Association for Computational Linguistics.