Enhancing Deep Learning with Embedded Features

for Arabic Named Entity Recognition

Ali Lotfy, Caroline Sabty, Slim Abdennadher

LREC 2022

June 21, 2022

Overview

Overview

1. Introduction
    - NER Task for Arabic
    - Related Work
    - The Arabic Language
2. Proposed model
3. Experiments
4. Error Analysis
5. Conclusion

NER Task for Arabic

The most used corpora

- ANERcorp: 150,268 tokens
  Benajiba et al. (2007)
- AQMAR: 73,910 tokens
  Mohit et al. (2012)

Class	Train	Test
Person	2724	882
Location	3778	656
Organization	1579	450
Miscellaneous	889	229
All	8970	2217

Table 1: Number of Entities in the ANERcorp

Class	Train	Test
Person	1048	424
Location	1117	325
Organization	354	102
Miscellaneous	1763	721
All	4282	1572

Table 2: Number of Entities in AQMAR corpus

Related Work

Arabic NER approaches:

Related Work

Arabic NER approaches:

- Rule based approaches.
  Mesfar (2007)

Related Work

Arabic NER approaches:

- Rule based approaches.
  Mesfar (2007)
- Machine learning approaches.
  Benajiba and Rosso (2008)

Related Work

Arabic NER approaches:

- Rule based approaches.
  Mesfar (2007)
- Machine learning approaches.
  Benajiba and Rosso (2008)
- Hybrid of machine learning and rule based.
  Shaalan and Oudah (2014)

Related Work

Arabic NER approaches:

- Rule based approaches.
  Mesfar (2007)
- Machine learning approaches.
  Benajiba and Rosso (2008)
- Hybrid of machine learning and rule based.
  Shaalan and Oudah (2014)
- Deep learning using word embeddings and Neural Networks.
  Youssef et al. (2020)

The Arabic Language

Arabic is highly complex

- Complex morphology.
  .e.g استزرناكم : We asked you to visit

The Arabic Language

Arabic is highly complex

- Complex morphology.
  .e.g استزرناكم : We asked you to visit
- Lack of capitalization: letters has one case.

The Arabic Language

Arabic is highly complex

- Complex morphology.
  .e.g استزرناكم : We asked you to visit
- Lack of capitalization: letters has one case.
- Unwritten diacritics.

  مصر

  مِصْر: Egypt

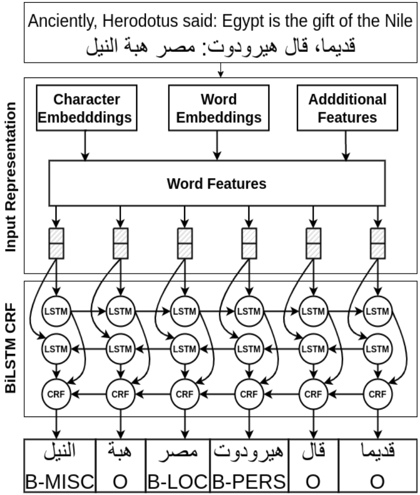
  مُصِرّ: insisting

  مَصَّرَ: divide to provinces

Overview

Model Architecture

- Input is fed to one layer of Bi-LSTM followed by a layer of CRFs.
- Input combines the advantages of deep learning and traditional machine learning by combining word embedding, character embedding, and additional features.



Input Representation

Input consists of three parts:

- Word embeddings are : Glove, Word2Vec, Fastext, AraBert.
- Character embeddings using Convolutional Neural Networks.
- Additional features By Madamira
  - Part of speech.
  - Capitalization from the English translation.
  - Word analysis feature.
    - Suffixes and prefixes.
    - Otehr morphological and syntactical features.

Overview

Model Performance

Training the model using different word embeddings along with features and character embeddings.

Embedding	Alone	+Features	+CE	+Features & CE
One Hot	54.90	59.70	60.00	61.80
GloVe	61.52	72.33*	72.27*	75.22*
AraVec	73.46	77.668*	77.09*	78.10*
FastText	78.34	79.85*	78.87 *	79.44*
AraBert	85.97	**86.47***	85.89	86.00

Table 3: Average F1-scores of 6 runs on ANERcorp in (%)

Embedding	Alone	+Features	+CE	+Features & CE
FastText	70.66	71.77*	71.63	71.95*
AraBert	77.82	**78.51***	77.42	77.56

Table 4: Average F1-scores of 6 runs on AQMAR in (%)

Features Analysis

Training the model using different word embeddings along with every feature.

Embedding	None	+POS	+Capitalization	+Word Analysis	+Quote
GloVe	61.52	67.97*	64.97 *	69.82 *	61.43
AraVec	73.46	76.31*	75.01*	77.46*	73.90*
FastText	78.34	79.29*	79.13*	80.16*	78.23
AraBert	85.97	86.17	85.87	86.26	86.04

Table 5: Average F1-score of 3 runs in (%) after adding features one at a time

* indicates statistical significance on the test set against mere embeddings by a paired sample t-test at level $\alpha = 0.05$.

Performance of Best Model

Class	Recall	Precision	F1
Location	95.59	89.08	92.22
Miscellaneous	64.83	74.63	69.39
Organization	74.00	83.04	78.26
Person	89.75	92.03	90.88
All	85.66	87.77	86.71

Table 6: Performance measures in (%) of the best model on the ANERcorp

Class	Recall	Precision	F1
Location	90.49	84.05	87.15
Miscellaneous	66.90	72.63	69.65
Organization	75.25	71.70	73.43
Person	91.96	89.63	90.78
All	79.07	79.88	79.48

Table 7: Performance measures in (%) of the best model on AQMAR

Comaprison with Previous Models

Model	F1-Score
Benajiba and Rosso (2008)	79.21
Antoun et al. (2020)	83.10
Obeid et al. (2020)	83.00
Our model	**86.71**

Table 8: Comparison with previous taggers results in (%) on the ANERcorp

Model	F1-Score
Helwe and Elbassuoni (2019)	67.22
Bazi and Laachfoubi (2018)	61.80
Liu et al. (2019)	75.82
Youssef et al. (2020)	77.62
Our model	**79.48**

Table 9: Comparison with previous taggers results in (%) on the AQMAR

Overview

Error Analysis

Sources of errors in the test set of ANERcorp:

- Wrong Tags: 44.96%

Error Analysis

Sources of errors in the test set of ANERcorp:

- Wrong Tags: 44.96%
- Unclear Language: 8.50%
  - entities written in foreign languages.

Error Analysis

Sources of errors in the test set of ANERcorp:

- Wrong Tags: 44.96%
- Unclear Language: 8.50%
  - entities written in foreign languages.
- Unclear context: 10.41%
  - information in the context does not accentuate the entity.

Error Analysis

Sources of errors in the test set of ANERcorp:

- Wrong Tags: 44.96%
- Unclear Language: 8.50%
  - entities written in foreign languages.
- Unclear context: 10.41%
  - information in the context does not accentuate the entity.
- Tagging Convention: 13.88%
  - entity identifiers .e.g the word festival in venice film festival مهرجان البندقية .
  university, institute, neighbourhood.

Error Analysis

Sources of errors in the test set of ANERcorp:

- Wrong Tags: 44.96%
- Unclear Language: 8.50%
  - entities written in foreign languages.
- Unclear context: 10.41%
  - information in the context does not accentuate the entity.
- Tagging Convention: 13.88%
  - entity identifiers .e.g the word festival in venice film festival مهرجان البندقية .
  university, institute, neighbourhood.
- Nested Entities: 7.11%
  - Bank of France بنك فرنسا.

Error Analysis

- Remianing Failures: 15.14%
  - Miscellaneous class is too broad
  - Entities with conflicting tags
  - Entities that require deep understanding or prior knowledge.

We have published the ANERcorp with correct annotation at the repository: https://github.com/CSabty/ANERcorp-Correction

Overview

Conclusion

- Our model outperforms previous models by:
  3.6 % for ANERcorp.
  1.86% for AQMAR.
- Additional knowledge can improve deep learning by utilizing it in the input layer.
- Data of higher quality can enhance NER Performance.

References I

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

Ismail El Bazi and Nabil Laachfoubi. Arabic named entity recognition using word representations. *arXiv preprint arXiv:1804.05630*, 2018.

Yassine Benajiba and Paolo Rosso. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer, 2008.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer, 2007.

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

References II

Liyuan Liu, Jingbo Shang, and Jiawei Han. Arabic named entity recognition: what works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67, 2019.

Slim Mesfar. Named entity recognition for arabic using syntactic grammars. In *International Conference on Application of Natural Language to Information Systems*, pages 305–316. Springer, 2007.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 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, 2012.

References III

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032, 2020.

Khaled Shaalan and Mai Oudah. A hybrid approach to arabic named entity recognition. *Journal of Information Science*, 40(1):67–87, 2014.

Abeer Youssef, Mustafa Elattar, and Samhaa R El-Beltagy. A multi-embeddings approach coupled with deep learning for arabic named entity recognition. In *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 456–460, 2020.