WEAK SUPERVISION APPROACH FOR ARABIC NAMED ENTITY RECOGNITION

Olga Simek and Courtland VanDam

MIT Lincoln Laboratory, Wood Street, Lexington, Massachusetts, USA

ABSTRACT

Arabic named entity recognition (NER) is a challenging problem, especially in conversational data such as social media posts. To address this problem, we propose an Arabic weak learner NER model called ANER-HMM, which leverages low quality predictions that provide partial recognition of entities. By combining these predictions, we achieve state of the art NER accuracy for cases for out-of-domain predictions. ANER-HMM leverages a hidden markov model to combine multiple predictions from weak learners and gazetteers. We demonstrate that ANER-HMM outperforms the state-of-the-art Arabic NER methods without requiring any labeled data or training deep learning models which often require large computing resources.

KEYWORDS

Named entity recognition, Arabic, weak learning

1. INTRODUCTION

Named entity recognition (NER) is the process of extracting proper nouns from text and identifying their type (e.g. person, organization) and span (starting and ending characters). NER is an important step in information extraction, network and knowledge graph construction, trend analysis and other tasks.

Conversations pose a unique challenge for NER. The syntax varies both in spelling and sentence structure compared to more formal documents, like news or Wikipedia articles. In this work, we focus on Twitter as a proxy for other types of chat data. Tweets are usually short, informal, and contain non-standard language, making them difficult to analyze using traditional natural language processing techniques.

Arabic poses additional challenges for NER. First, in terms of NER research, Arabic is a low resource language. Annotated datasets, especially conversational Arabic, are limited. Most of the focus has been done on classical and modern standard Arabic NER, but the work on social media or conversational Arabic NER is limited. Second, Arabic does not contain the syntax features, like capitalization, which are effective for NER in other languages like English. Third, Arabic is a very varied language. It consists of three classes: Classical, Modern Standard, and Colloquial. Classical Arabic is used for religious texts like Quran. Modern Standard Arabic is the formal Arabic that is taught at schools and used in media, corporate communications, legal texts and translations. Colloquial Arabic is the spoken form of Arabic and it differs from region to region. The difference between dialects can be very significant. Informal conversations and majority of social media posts are in Colloquial Arabic. These forms of Arabic vary enough that an NER model trained on one form performs poorly on other forms [7].

David C. Wyld et al. (Eds): MLNLP, NWCOM, DTMN, ASOFT, SIGPRO, AIFZ, CSITY, CLSB - 2023
pp. 15-22, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131602

To address these challenges, we propose an Arabic weak learner NER model called ANER-HMM, which leverages low quality predictions to provide partial recognition of entities. By combining these predictions, we achieve higher NER accuracy. ANER-HMM leverages a hidden markov model (HMM) similar to Lison et al. [12] to combine multiple predictions from weak learners and gazetteers. We demonstrate that ANER-HMM outperforms the state-of-the-art Arabic NER methods without requiring any labeled data or training deep learning models which often require large computing resources. To the best of our knowledge, the weak supervision has not been applied to the Arabic NER problem.

The remainder of the paper is as follows. In Section 2, we present related work. In Section 3, we propose ANER-HMM. Our results are presented and discussed in Section 4, and we conclude in Section 5.

2. RELATED WORK

2.1. Weak Learners

Weak supervision is a machine learning approach that combines noisy labels with limited overlap to label a larger training set. Rather than labeling training data by hand, researchers and analysts write labeling functions, which use patterns and external knowledge sources to label the data. Each labeling function will have different levels of coverage and precision. Some will overlap (i.e., label the same phrase) while others may be disjoint. Next, each labeling function is scored and weighted by its roughly estimated performance. Finally, a machine learning model is trained to combine predictions (based on their probability scores rather than the classes) to produce a more accurate model.

Snorkel [14] is a popular weak learning algorithm. Users write labeling functions and provide them to Snorkel. Snorkel learns a generative model over the labeling functions and uses it to estimate their quality and correlations. Snorkel outputs a set of probabilistic labels, which then can be passed to discriminative models, including deep learning models. Ratner et al. improve the recovered accuracy of the weak learners by using a matrix completion-style optimization function, and model the complexity of the estimator by the amount of unlabeled data it can estimate [15].

Lison et al. [12] and Safranchik et al. [16] combine scores using a Hidden Markov Model (HMM). Weak learners include weak classifiers (e.g., out-of-domain NER models), gazetteers, casing, part-of-speech tags, and document-level relations. They show that HMMs constitute an effective approach to weakly supervised sequence tagging, outperforming Snorkel. The general approach can be applied to Arabic, however virtually none of the labeling functions and rules apply. For example, Arabic does not contain capitalization.

2.2. Arabic NER

Arabic named entity recognition research began in 2005. NER approaches fall into two broad categories, rule-based and learned-based [3]. In rule-based approaches, human analysts use gazetteers and write rules in the form of regular expressions to identify entities in text. Learned-based approaches utilize machine learning to extract named entities. Machine learning NER algorithms cluster into two groups, classical machine learning (e.g. Support Vector Machines, decision tress, etc.) and deep neural networks. For full summary of Arabic NER research, see [17] and [3]. We will focus on the current state of the art approaches.

The current state of the art (SOTA) comes from machine learning, especially deep learning. Algorithms use Bidirectional Long Short-Term Memory (BiLSTM), Conditional Random Fields (CRF), or transformers. CRF is a graphical model where each word is linked to its entity type. BiLSTM learns additional context from past and future inputs [4]. BiLSTM-CRF adds a CRF layer to BiLSTM [4].

Transformer models are based on Google's Bidirectional Encoder Representations from Transformers (BERT) model. BERT models are trained either on only Arabic texts [5, 10], or multiple languages [1, 19]. Current SOTA include AraBERT [5], ARBERT [1], MARBERT [1], and mBERT [19].

Arabic NER datasets are not nearly as prevalent as English NER datasets, so previous work combined machine learning models with dictionaries or gazetteers to improve performance. Farasa combines predictions from a CRF model and information on whether an Arabic term's English translation is a named entity [8]. Liu et al. combine a CRF with a gazetteer [13]. Helwe et al. took a semi-supervised approach [11]. They trained an AraBERT model on ANERCorp, applied that model to a partially labeled Wikipedia dataset, then retrained the model on the combined ANERCorp and Wikipedia labels.

3. METHODOLOGY

Weak supervision is related to ensemble learning. We apply a number of labeling functions to the data, these labeling functions have varying levels of accuracy and confidence. For each word, the labeling function will predict that a word is either a specific entity types or is not an entity, or refuse to make a prediction. We combine the predictions to form a model that is more accurate than any of the individual labeling functions.

Labeling functions come in several forms. Out-of-domain NER models are classifiers and neural networks trained on large corpora from different domains, e.g., news articles. Gazetteers are dictionaries of entities. They act as a lookup table for names of people, countries, organizations, etc.

An aggregation model combines the output of the labeling functions to a single layer of annotation [12]. Similar to Lison et al. [12], we use a Hidden Markov Model (HMM) for aggregation. HMM does not require labelled data, which makes it advantageous for settings where labeled data is not available or very limited.

3.1. Weak Learners

We incorporate the following 5 weak learners into our pipeline.

- **Farasa**: Farasa is a combination of a conditional random field (CRF) trained mostly on ANERCorp and crosslingual features (e.g. phrase capitalized in English, translated phrase is entity in English) [8]
- **Hatmimoha (**https://github.com/hatmimoha/arabic-ner): Hatmimoha pretrained an NER transformer model on top of a BERT model on 14,000 sentences collected from the internet. Hatmimoha recognizes 9 entity types, including event and disease. We only consider performance on person, organization, and location entity recognition.
- **EmnamoR (**https://github.com/EmnamoR/Arabic-named-entity-recognition): Emna Amor trained a linear SVM model on ANERCorp.

- **AraBERT**: AraBERT is a BERT model trained on Arabic news corpus. A feed forward output layer is added for NER [5].
- **MultiBERT**: A multilingual BERT model trained on the Wikipedia articles of the top 100 languages with the largest Wikipedias. Similarly to AraBERT, an additional feed forward layer is trained on top of MultiBERT for NER [9, 19].

3.2. Gazetteers

We use the following gazetteers.

- **WikiFANE (**https://sourceforge.net/projects/arabic-named-entity-gazetteer/) is an Arabic named entity recognition gazetteer compiled from Wikipedia. It contains 68,343 entities from 50 classes.
- **NETLexicon (**http://nlp.qatar.cmu.edu/resources/NETLexicon/) Azab et al. [6] automatically construct a bilingual lexicon of NEs paired with the transliteration/translation decisions in two domains.
- **ArabicNEs (**https://sourceforge.net/projects/arabicnes/) Named Entities resource for Arabic, totalling 45,202 NEs. These NEs are extracted from the Arabic Wikipedia, and provided with English translation and ontological information.
- **JRC**-Names is a large multilingual list of names and their spelling variants, collected from 220,000 news reports per day by the Europe Media Monitor (EMM) [18].
- **GeoNames** (http://www.geonames.org/) is a geography database of locations including country names, cities, and towns. We select the Arabic translations of locations.
- **NileAPgazet**} is Arabic persons' names gazetteer, consisting of a list of about 19K full names collected from public resources in addition to lists of first, male, female and family names [20].

3.3. Hidden Markov Model (HMM)

Next, we aggregate the predictions from the weak learners and gazetteers using a Hidden Markov Model (HMM), proposed by Lison et al. [12]. This model is suited for sequence labelling tasks and able to include probabilistic labelling predictions, with the states corresponding to the output labels. The probability of being in a particular state given a token i and the labeling function λ_j is drawn from a Dirichlet distribution with parameter α_j^S. The transition matrix is an inverse logit function of the parameters to the transition probability matrix. The transition matrix and α vectors are estimated using the Baum-Welch algorithm. The initial distribution of the latent states and the initial transition probabilities are drawn from Dirichlet distributions based on counts from the most reliable labelling function. The α vectors are initialized based on the precision and recall of each of the labelling functions j for a given label k. For more detail, see [12].

4. EVALUATION

4.1. Data

4.1.1. Training Data

- **CALC2018** [2] dataset is from the CALCS 2018 task. Tweets are either in modern standard Arabic or Egyptian Arabic by 12 Egyptian political figures. We split the training set from this task into training and validation sets. The training set contains 10,102 tweets. The gold data labels provided with this dataset are not used for training.

4.1.2. Test Data

- **Darwish** [7] scraped Arabic language tweets from Twitter from November 23, 2011 to November 27, 2011. This dataset contains 1,423 tweets and 26k tokens.
- **CALC2018** [2] is dataset from the CALCS 2018 task. We use the development set from this shared task as a test set. Labels are publicly available for the development set but not for the testing set. The development set contains 1,122 tweets.

4.2. Metrics

We evaluate our named entity recognition on the most common entities (person, organization, and location) due to their prevalence across all data sets (including those our weak learners were pre-trained on). We calculate precision, recall and F1 score for each of the three entity types.

4.3. Baselines

We compare the performance of HMM to Snorkel [14] and majority vote.

Snorkel is a weak learning platform which statistically combines the prediction of labeling functions with varying degrees of accuracy to label unlabeled data. A machine learning algorithm is trained from the labels Snorkel produces [14].

Majority Vote (MV) predicts the label with the highest frequency. The labeling functions predict most words to be a non-entity, so at least a threshold *T* algorithms must predict the term or phrase as an entity before MV is applied.

4.4. Results

Table 1: Evaluation results on CALC2018 and Darwish datasets for entity type PERSON; F1 score is followed by precision and recall in parentheses

Model	**CALC2018**	**Darwish**
Majority Rule	0.15(0.08/0.63)	0.10(0.05/0.77)
Farasa	0.46(0.56/0.39)	0.94(0.90/0.99)*
Hatmimoha	0.62(0.65/0.59)	0.62(0.58/0.67)
EmnamoR	0.12(0.16/0.10)	0.13(0.13/0.14)
AraBERT	0.66(0.74/0.60)	0.59(0.56/0.61)
MultiBERT	0.47(0.56/0.41)	0.45(0.42/0.50)
Snorkel	0.15(0.09/0.61)	0.10(0.05/0.71)
HelweFS	0.61(0.64/0.58)	0.62(0.55/0.71)
ANER-HMM	**0.70**(0.70/0.70)	**0.64**(0.63/0.64)

Table 2: Evaluation results on CALC2018 and Darwish datasets for entity type LOCATION; F1 score is followed by precision and recall in parentheses

Model	**CALC2018**	**Darwish**
Majority Rule	0.35(0.24/0.69)	0.31(0.22/0.51)
Farasa	0.64(0.65/0.62)	0.97(0.97/0.98)*
Hatmimoha	0.50(0.64/0.41)	0.46(0.74/0.34)
EmnamoR	0.60(0.72/0.51)	0.48(0.79/0.34)
AraBERT	0.69(0.72/0.67)	0.55(0.74/0.43)
MultiBERT	0.62(0.65/0.60)	0.55(0.80/0.41)
Snorkel	0.37(0.26/0.59)	0.37(0.28/0.53)
HelweFS	0.61(0.58/0.64)	**0.61**(0.70/0.54)
ANER-HMM	**0.72**(0.76/0.68)	0.59(0.89/0.44)

Our performance is summarized in Tables 1, 2, and 3. We include performance of each of the weak learners as well as our baselines. Farasa* has very high performance on Darwish dataset and we hypothesize that Twitter data must have been part of the training data for the Farasa version we used. We retrain our model without Farasa before we run it on Darwish test set. We also compare ANER-HMM to Helwe's fully supervised approach [11]. Helwe presented several state-of-the-art approaches that did not require labeled in-domain data for training. We chose his fully supervised model (HelweFS) since we had access to its code, and ran it on both of our test datasets.

Table 3: Evaluation results on CALC2018 and Darwish datasets for entity type ORGANIZATION; F1 score is followed by precision and recall in parentheses

Model	**CALC2018**	**Darwish**
Majority Rule	0.09(0.05/0.60)	0.09(0.05/0.42)
Farasa	0.19(0.17/0.21)	0.92(0.88/0.98)*
Hatmimoha	0.26(0.17/0.54)	0.37(0.31/0.44)
EmnamoR	0.01(0.01/0.01)	0.02(0.03/0.02)
AraBERT	0.16(0.13/0.19)	0.30(0.41/0.24)
MultiBERT	0.18(0.19/0.17)	0.18(0.24/0.15)
Snorkel	0.12(0.07/0.60)	0.51(0.09/0.51)
HelweFS	0.11(0.11/0.11)	**0.37**(0.41/0.33)
ANER-HMM	**0.29**(0.25/0.34)	0.32(0.39/0.27)

ANER-HMM strongly outperforms both baselines. It significantly outperforms HelweFS on the CALC2018 dataset for all entity types and it also outperforms Helwe on entities of type PERSON in Darwish dataset.

5. CONCLUSION

We have presented Arabic weak learner NER model called ANER-HMM, which leverages low quality predictions that provide partial recognition of entities. By using hidden markov model to combine these predictions, we achieve state of the art NER accuracy for cases of out-of-domain predictions. We have demonstrated that ANER-HMM outperforms the state-of-the-art Arabic NER methods without requiring any labeled data or training deep learning models. For future work we plan to experiment with Arabic and multilingual large language models and utilize them

as potentially very powerful labeling functions. We also plan to train a model that would improve gazetteer annotation precision as well as include additional weak learners and gazetteers.

REFERENCES

[1] Muhammad Abdul-Mageed, Abdel Rahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105, Online. Association for Computational Linguistics.

[2] Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

[3] Brahim Ait Ben Ali, Soukaina Mihi, Ismail El Bazi, and Nabil Laachfoubi. 2020. A recent survey of arabic named entity recognition on social media. Rev. d'Intelligence Artif., 34:125–135.

[4] Sa'a D A. Alzboun, Saia Khaled Tawalbeh, Mohammad Al-Smadi, and Yaser Jararweh. 2018. Using bidirectional long short-term memory and conditional random fields for labeling arabic named entities: A comparative study. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), pages 135–140.

[5] Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer based model for arabic language understanding. CoRR, abs/2003.00104.

[6] Mahmoud Azab, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2013. Dudley North visits North London: Learning when to transliterate to Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 439–444, Atlanta, Georgia. Association for Computational Linguistics.

[7] Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.

[8] Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1070–1074, Portorož, Slovenia. European Language Resources Association (ELRA).

[9] 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.

[10] Hatmimoha. 2020. Hatmimoha/arabic-ner: Named entity recognition system for arabic.

[11] Chadi Helwe, Ghassan Dib, Mohsen Shamas, and Shady Elbassuoni. 2020. A semisupervised BERT approach for Arabic named entity recognition. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 49–57, Barcelona, Spain (Online). Association for Computational Linguistics.

[12] Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. ArXiv, abs/2004.14723.

[13] Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019. Arabic named entity recognition: What works and what's next. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 60–67, Florence, Italy. Association for Computational Linguistics.

[14] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation withweak supervision. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, volume 11, page 269. NIH Public Access.

[15] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher RAl'. 2019. Training complex models with multi-task weak supervision. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):4763–4771.

[16] Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04):5570–5578.

[17] Khaled Shaalan. 2014. A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40(2):469–510.

[18] Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. JRC-NAMES: A freely available, highly multilingual named entity resource. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pages 104–110, Hissar, Bulgaria. Association for Computational Linguistics.

[19] ShijieWu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 120–130, Online. Association for Computational Linguistics.

[20] Omnia Zayed and Samhaa R El-Beltagy. 2015. Named entity recognition of personsâĂŹ names in arabic tweets. In Proceedings of the international conference recent advances in natural language processing, pages 731–738.

AUTHORS

Dr. Olga Simek is a technical staff member in the Artificial Intelligence Technology and Systems Group at Lincoln Laboratory. For the past ten years, she has been leading projects, conducting applied research, and publishing in the area of text analytics, social networks analysis and crowdsourcing.

Prior to joining Lincoln Laboratory, she worked in industry for a number of years. She held leading positions at several companies where she used machine learning to design and build state-of-the-art predictive risk models for global financial risk, market demand estimators, and fraud detection. She also held positions at Lawrence Livermore and Los Alamos National Laboratories. Simek received a BA degree in mathematics and computer science from Eastern Washington University, an MA degree in mathematics from Western Washington University, and a PhD degree in mathematics from the University of Arizona.

Dr. Courtland VanDam is a technical staff member in the Artificial Intelligence Technology and Systems Group at MIT Lincoln Laboratory. Her primary interests are in natural language processing, counter–influence operations, and applications of machine learning to cybersecurity.

Prior to joining Lincoln Laboratory in 2019, VanDam researched disinformation detection at Michigan State University. In 2018, she received the Best Paper Award at the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining for her paper on detecting compromised accounts on Twitter.

VanDam earned a BA degree in Spanish literature from Kalamazoo College in 2008, and an MS degree and a PhD degree in computer science from Michigan State University in 2012 and 2019, respectively.