MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220048

Knowledge-Driven Drug-Use NamedEntity Recognition with Distant Supervision

Goonmeet Bajaj[a], Ugur Kursuncu[b], Manas Gaur[b], Usha Lokala[b], Ayaz Hyder[c], Srinivasan Parthasarathy[a] and Amit Sheth[b]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA
[b] Division of Environmental Health Sciences, College of Public Health, The Ohio State University, Columbus, Ohio, USA
[c] AI Institute, University of South Carolina, Columbia, City Two, South Carolina, USA

Abstract

As Named Entity Recognition (NER) has been essential in identifying critical elements of unstructured content, generic NER tools remain limited in recognizing entities specific to a domain, such as drug use and public health. For such high-impact areas, accurately capturing relevant entities at a more granular level is critical, as this information influences real-world processes. On the other hand, training NER models for a specific domain without handcrafted features requires an extensive amount of labeled data, which is expensive in human effort and time. In this study, we employ distant supervision utilizing a domain-specific ontology to reduce the need for human labor and train models incorporating domain-specific (e.g., drug use) external knowledge to recognize domain specific entities. We capture entities related the drug use and their trends in government epidemiology reports, with an improvement of 8% in F1-score.

Keywords:

Natural Language Processing, Deep Learning, Information Storage and Retrieval

Introduction

Named Entity Recognition (NER) is a fundamental Natural Language Processing (NLP) task used to detect important entities mentioned in text and classify them using predefined types. NER has critical utility for a variety of downstream tasks in pipelines, including knowledge extraction, targeted sentiment analysis [1], question answering [2], information retrieval [2], and dialogue systems [3]. Existing (i.e., off-the-shelf) NER approaches and tools often cover an inadequate set of named entities, which can be incomplete for a specific domain. Many off-the-shelf NER tools tag a limited number of named entities (up to 18) that are often not domain-specific. The lack of domain-specific NER models becomes a significant bottleneck as it can lead to potentially inaccurate recognition and/or inadequate coverage of domain-specific information for high-impact domains, such as healthcare and public health. It further magnifies such impact when such models influence important decision-making processes, e.g., policy-making.

Historically, NER models have been trained using sequential models (e.g., Hidden Markov Model (HMM)) [4] and Conditional Random Fields (CRFs) using handcrafted features [5]. However, designing handcrafted features is a non-trivial task and requires significant time, effort, and knowledge of domain experts. More recently, neural networks have become state-of-the-art models for NER achieving higher performance [6, 7]. However, these models still require an extensive collection of labeled documents which is time-consuming, tedious, and expensive to acquire. Recent approaches employ distant supervision to automatically create labeled data [9, 10, 8] by utilizing knowledge bases (e.g., Wikipedia, YAGO, MeSH) to reduce this need for extensive human expert involvement.

Shang et al. developed a framework to handle noisy distant labels applied using a dictionary [8]. They modify a fuzzy CRF layer to manage tokens with multiple possible labels and use a neural model, AutoNER, with a Tie or Break scheme to determine if adjacent tokens belong to the same entity type. [9] develop a weekly supervised approach, Trove, using medical ontologies for entity classification. Trove requires a user-defined mapping of an ontology's class taxonomy to entity classes, a set of label sources (for weak supervision), regular expressions, and a collection of unlabeled documents (used to build a training set). Similarly, Laing et al. develop a two-stage training algorithm, BOND, that uses pre-trained language models to improve open-domain NER [15]. During the first stage, they adapt a pre-trained language model with distant labels. In the second stage, they apply a self-training approach and drop the distant labels using handcrafted rules and knowledge bases.

In this work, we train a NER model for drug trend reports using the Drug Abuse Ontology (DAO) [11, 12] and spaCy1 (token2vec) that can tag domain-specific (e.g., drug use) and generic named entities. As our approach does not require handcrafted rules, we use the ontology to semantically annotate the data via distant supervision. We employ both generic entities through spaCy and domain-specific entities identified from DAO by our collaborator domain expert. Then, we use the semantically annotated data to fine-tune transformer-based language models that have achieved state-of-the-art results on NER tasks. We find that this framework allows us to extract both domain-specific and generic named entities with high precision, recall, and F1-score without the need for handcrafted regular expressions and rules.

In this study, we make the following contributions utilizing an ontology: i) we automatically label a large amount of data using distant supervision, which is critical for the state-of-the-art deep learning language models and ii) we develop a domain-specific NER model for epidemiology reports that tags both generic named entities and domain-specific entities. We are happy to share the annotated data and trained models.

[1] https://spacy.io

Methods

Datasets

Drug Abuse Ontology: An ontology represents knowledge of a particular domain as a set of concepts and relationships among the concepts. Ontologies describe a domain and are used to reason about the entities within that domain. Precisely, DAO consists of 315 classes, 31 relationships, and 814 instances of the classes [12]. DAO formalizes concepts and entities that are relevant to addiction and mental health. The ontology includes representation of drugs, including slang and brand names, chemical designation, dosage and unit, purity, the form of preparation. We direct interested readers to [11, 12] to learn more about DAO.

Drug Trends Reports from the Ohio Substance Abuse Monitoring (OSAM) Network: The OSAM Network[2] consists of epidemiologists from eight regions of Ohio. The epidemiologists conduct focus groups and individual qualitative interviews with active and recovering drug users and community professionals (treatment providers, law enforcement officials, etc.) to produce epidemiological descriptions of local substance abuse trends. Additionally, the epidemiologists report qualitative findings from coroner's reports and crime laboratory data, provide information related to substance abuse trends from mass media sources. These reports offer policy-makers real-time accurate epidemiological information to plan prevention and intervention strategies. Such information describes substance abuse trends across the state, focusing on drug availability, prices, quality, and abuse patterns. The OSAM network publishes the reports on a semiannual basis in January and June of each year.

In public health settings, the OSAM drug trend reports serve two primary purposes: surveillance and planning. For surveillance, the primary data points of interest are: i) date and time, ii) the drug types, quantities, amounts, units and percent change in quantities of a drug in a community, iii) health conditions affected in the population, and iv) organizations involved in drug-related activities. It is critical to identify a combination of named- and domain-specific entities to capture these data points. For planning purposes, such as making policy changes based on these reports or planning an intervention in a specific region, recognizing similar entities would be of significant interest to public health departments. These entities would include: i) drug type, substance use, and changes in route of administration, ii) information on ethnic groups, iii) health conditions, individual and societal, and economic factors, and iv) information on organizations. We obtained 102 OSAM Drug Trend reports from January 2014 to January 2020, covering the eight Ohio regions. We use the 2020 reports as our validation dataset.

OntoNotes 5.0 datasets: OntoNotes 5.0 dataset [13] contains 18 generic entity types. This dataset provides labeled corpus from many genres (news, conversational telephone speech, weblogs, newsgroups, broadcast, talk shows, etc.) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate-argument structure) and shallow semantics (word sense linked to an ontology and co-reference). We used the OntoNotes 5.0 dataset to provide distant supervision for generic entity types because of the diversity of information and a higher number of entity types when compared to other benchmarking NER datasets.

Final Experimental Datasets: We use three datasets to benchmark the performance of our model:

1. **OntoNotes (ON)** – OntoNotes 5.0 training and validation splits with 18 NER types
2. **OntoNotes + DrugTrends (ON + DT)** – For this dataset, we combine the training splits from the OntoNotes 5.0 and the OSAM Drug Trend Reports from 2014 to 2019. However, we only use the 18 generic NER labels mentioned above. We do not add the labels from DAO and do not modify the validation set from the OntoNotes 5.0 dataset.
3. **OntoNotes + DrugTrends + DAO (ON + DT + DAO)** – For this dataset, we combine the training splits from the OntoNotes 5.0 and the OSAM Drug Trend Reports from 2014 to 2019. We combine the OntoNotes 5.0 validation set with the OSAM Drug Trend Reports from 2020 to create a validation set. Here, we use both the 18 generic NER labels and the domain-specific labels selected from DAO.

Semantic Data Annotation with Distant Supervision: The OSAM Drug Trend reports contain generic named entities (e.g., DATE, LOCATION, GPE) and domain-specific entities such as DRUG and HEALTH-CONDITIONS. Therefore, we employ two different approaches to label our training data. We label our dataset with entity types using the Inside, Outside, Beginning (IBO/BIO) scheme. The IBO format is commonly used to label entities in a chunking task in computational linguistics such as NER. Specifically, we recognize domain-specific named entities in the drug trends reports by mapping words in the text to the instances in DAO and label them with their respective entity types.

Generic Entity Types: We use the en_core_web_lg spaCy (token2vec) model[3] to label OSAM reports with one of the following 18 entity types: CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, and WORK OF ART (WoA). The en_core_web_lg spaCy model is trained using the OntoNotes 5.0 dataset [13]. We experimented with fine-tuning BERT [20] using the OntoNotes 5.0 dataset [13] but found the spaCy model outperforming BERT on the OntoNotes 5.0 validation dataset.

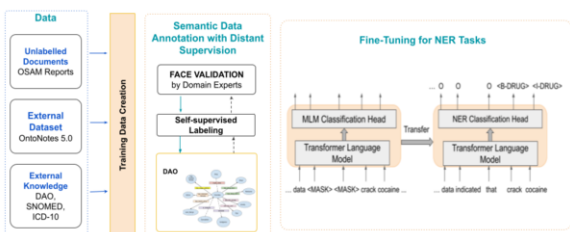

Figure 1 - Framework Overview

Domain-Specific Entity Types: To provide distant supervision for entities in the drug trends reports, we utilize DAO that contains 315 classes. Instead of using all classes as entity types, our collaborator, a public health expert, identified and validated the essential entity types for this study: TIME, UNIT, DRUG, HEALTH CONDITIONS, INDIVIDUAL RELATED, PYSCHOSTIMULANTS, NETWORK RELATED,

[2] https://mha.ohio.gov/Researchers-and-Media/Workgroups-and-Networks/Ohio-Substance-Abuse-Monitoring-Network

[3] https://spacy.io/models/en

SOCIETY, SOURCES, ROUTE OF ADMINISTRATION, SUBSTANCE RELATED, and UNIT. Our framework allows domain experts to select the entity types deemed as most appropriate for their respective domains.

Modeling

We use our labeled dataset (generated using semantic data annotation with distant supervision) to train state-of-the-art transformer-based language models for the NER task, such as BERT [19] and BOND [15]. We also use these techniques to compare the performance of NER models with and without using domain-specific corpus (e.g., drug trends reports) and DAO. In this section, we describe the baseline approaches and our approaches for the domain-specific NER task. Our framework is outlined in Figure 1.

Baselines: We used the tok2vec model from spaCy, en_core_web_lg, as our very first baseline with results that were reasonably satisfactory (See Section 5.1). In addition, we use transfer learning for BERT [20] by fine-tuning using only the OntoNotes 5.0 dataset to compare the performance for generic entity types. We present the results for tok2vec and the fine-tuned BERT model in the Results Section (see Table 1). Additional experimental results are provided within the supplementary material.

BOND: We trained BOND [15] using the three datasets described above. In addition, we replaced the RoBERTa model [22] with BERT and BioClinicalBERT [21] during Stage I of BOND, while we did not see a significant improvement in performance using Stage II (self-training). Therefore, we did not change the underlying transformer model for State II.

Results

We evaluate the aforementioned models using precision, recall, F1 score, and the macro and weighted averages. Table 1 presents the results from our experiments.

Due to the space restrictions, we only provide three sets of experimental results for: i) OntoNotes (ON) with token2vec, ii) OntoNotes + DrugTrends + DAO (ON + DT + DAO) with BioClinicalBERT, and ii) OntoNotes + DrugTrends + DAO (ON + DT + DAO) with BOND Stage I with RoBERTA.

Findings from our other experiments (not reported): We compare the performance of BERT and token2vec (via spaCy) on the ***ON*** dataset. We find that token2vec provides better performance over BERT. Among token2vec, BERT, and BOND (Stage I and Stage II models) trained over the ***ON + DT*** dataset, BOND with RoBERTA outperforms the others. The results from the fine-tuned BioClinicalBERT models are comparable to that of BOND's Stage I training with BioClinicalBERT for ***ON + DT + DAO*** dataset.

Findings from Table 1: We observe that BioClinicalBERT and BOND (Stage I - RoBERTa) performs better than BERT as expected, since the models were pre-trained on medical corpora and fine-tuned with our ***ON + DT + DAO*** dataset. On the other hand, we observe that BOND - Stage I with RoBERTa outperforms BioClinicalBERT in F1-score. However, we note that the performance gain of BOND with RoBERTa is attributed to generic NER labels, as BioClinicalBERT outperforms for domain-specific entity labels.

Overall, the results indicate that adding Drug Trends reports during the fine-tuning process improves the BERT models' performance (BERT model, BOND with BERT, BOND with RoBERTa) to extract generic NER labels in the OntoNotes dataset. The fine-tuned BioClinicalBert has higher performance for DAO labels when compared to BOND with BioClinicalBert. In our evaluation, we omit results from experiments with poor results, and they are as follows: i) for ***ON*** dataset and BOND (Stage I and Stage II models) ii) for ***ON+DT*** dataset and BOND Stage II model (Self-Training).

Table 1 - Experimental Results

	token2vec (ON)			BioClinical-BERT (ON + DT + DAO)			BOND-RoBERTA (ON + DT + DAO)		
NAMED ENTITY	P	R	F	P	F	R	P	F	R
CARDINAL	0.70	0.89	0.77	0.87	0.92	0.89	0.88	0.93	0.90
DATE	0.78	0.88	0.83	0.87	0.90	0.89	0.88	0.90	0.98
EVENT	0.58	0.37	0.45	0.60	0.51	0.55	0.73	0.57	0.64
FAC	0.43	0.45	0.43	0.42	0.52	0.46	0.50	0.63	0.56
GPE	0.82	0.91	0.87	0.91	0.93	0.92	0.92	0.94	0.93
LANGUAGE	0.76	0.67	0.71	0.83	0.69	0.75	0.73	0.63	0.68
LAW	0.48	0.51	0.50	0.48	0.59	0.53	0.44	0.51	0.47
LOC	0.58	0.60	0.59	0.69	0.69	0.69	0.71	0.75	0.73
MONEY	0.91	0.92	0.91	0.91	0.91	0.91	0.92	0.93	0.93
NORP	0.81	0.93	0.87	0.87	0.90	0.89	0.87	0.89	0.88
ORDINAL	0.66	0.84	0.64	0.75	0.82	0.79	0.76	0.84	0.80
ORG	0.74	0.87	0.80	0.85	0.88	0.87	0.87	0.89	0.88
PERCENT	0.90	0.89	0.89	0.91	0.90	0.91	0.92	0.90	0.91
PERSON	0.67	0.93	0.78	0.89	0.93	0.91	0.91	0.94	0.93
PRODUCT	0.32	0.15	0.2	0.52	0.48	0.50	0.62	0.46	0.53
QUANTITY	0.65	0.64	0.64	0.82	0.84	0.83	0.81	0.86	0.83
TIME	0.62	0.73	0.67	0.73	0.81	0.77	0.75	0.82	0.78
WoA	0.28	0.19	0.22	0.41	0.46	0.43	0.57	0.52	0.55
DRUG	-	-	-	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
ETH_GRP	-	-	-	0.99	1.00	1.00	**1.00**	**1.00**	**1.00**

HLT_COND	-	-	-	**0.97**	**0.99**	**0.98**	**0.97**	**0.99**	**0.98**
INDIV_REL	-	-	-	**0.97**	0.99	**0.98**	**0.97**	**1.00**	**0.98**
NET_REL	-	-	-	0.98	0.99	**0.98**	**0.99**	**1.00**	**0.99**
PSYSTIMU.	-	-	-	**1.00**	**1.00**	**1.00**	1.00	1.00	1.00
RoA	-	-	-	**1.00**	**1.00**	**1.00**	0.99	1.00	1.00
SOCIETY	-	-	-	0.99	0.99	0.99	**1.00**	**1.00**	**1.00**
SOURCES	-	-	-	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
SPAT_REL	-	-	-	**0.99**	0.99	**0.99**	0.98	**1.00**	**0.99**
SUB_REL	-	-	-	**0.97**	**1.00**	**0.98**	0.96	**1.00**	**0.98**
UNIT	-	-	-	**0.99**	**1.00**	**1.00**	0.96	**1.00**	0.98
Macro Avg	0.65	0.68	0.66	0.84	**0.86**	0.85	**0.85**	**0.86**	**0.86**
Weighted Avg	0.75	0.86	0.80	0.90	0.92	0.91	**0.91**	**0.93**	**0.92**

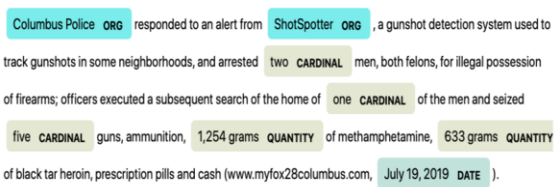

Figure 2 - spaCy (token2vec) Tagged Entities

Discussion

The pipeline in our study indicated that most entities based on the Drug Trends reports and DAO were essential for their relevance to public health practice. Also, the weighted average across the different models indicated that BioClinicalBERT and BOND - Baseline RoBERTa were adequate for applying our pipeline for public health practice in terms of opioid addiction surveillance and planning/policy purposes. Figures 2 and 3 present a sampled sentence tagged with named entities using spaCy (token2vec) and BioClinicalBERT, respectively. From these images, we can see that BioClinicalBERT captures more domain-specific entities. The overall performance of different algorithms on identifying named and domain-specific entities showed the utility of our pipeline for public health practice. Public health departments may use the information generated through this pipeline to quickly identify emerging trends across their region or compare trends across regions given a specific drug type or health conditions. Another potential use for a statewide agency (e.g., Ohio Department of Health or Ohio Department of Medicaid, or Ohio Department of Public Safety) may be to query multiple reports across Ohio based on a given drug type, route of administration, and health conditions. The ability to quickly glean the information from drug trend reports in a standardized manner when the underlying structure of drug trend reports is not available in a standardized format is a significant step forward for public health departments. Our team can automate the pipeline proposed in this study and provide organized data to public health departments in an analysis-ready dataset. Such a dataset may be integrated into existing opioid surveillance systems and complement these existing systems with a rich source of quantitative data extracted from a mainly qualitative data source (i.e., drug trend reports) quickly and efficiently. Future work based on this pipeline will attempt an integration in collaboration with local health departments in Ohio that are building an integrated opioid surveillance system.

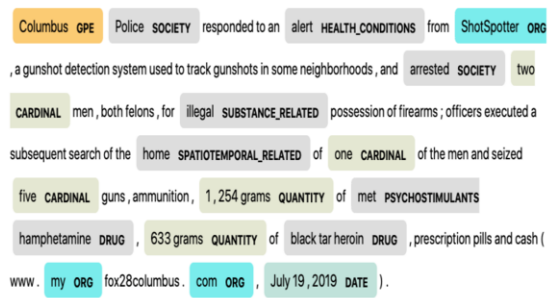

Figure 3 - BioClinicalBERT Tagged Entities

Conclusions

In this work, we utilized deep language models along with Drug Abuse Ontology (DAO) for named entity recognition task in the public health domain. We demonstrate that semantically-annotated data with distant supervision boosts the utility of the state-of-the-art language models, which requires large datasets, for domain-specific tasks. Specifically, we found that this approach allows us to extract both domain-specific and generic named entities with high precision, recall and F1-score without spending extreme resources and time to label training data.

Acknowledgements

We would like to acknowledge NSF Grant #1761969

References

[1] L. Jiang L, M. Yu, M. Zhou, X. Liu, and T. Zhao, Target-dependent twitter sentiment classification. 49th annual meeting of the association for computational linguistics: human language technologies (2011, June), 151-160.

[2] M.A. Khalid, V. Jijkoun, and M. De Rijke, The impact of named entity normalization on information retrieval for question answering. European Conference on Information Retrieval (2008, March), 705-710.

[3] K. Bowden, J. Wu, S. Oraby, A. Misra, and M. Walker, SlugNERDS: A Named Entity Recognition Tool for Open Domain Dialogue Systems. 11th International Conference on Language Resources and Evaluation (2018).

[4] G. Zhou, and J. Su, Named entity recognition using an HMM-based chunk tagger. 40th Annual Meeting of the ACL, (2002), 473-480.
[5] J. Lafferty, A. McCallum, and F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
[6] X. Ma, and E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. 54th Annual Meeting of the ACL (2016). 1064-1074.
[7] Z. Huang, W. Xu, and K. Yu, Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. (2015).
[8] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han, Learning Named Entity Tagger using Domain-Specific Dictionary. 2018 EMNLP (2018), 2054-2064.
[9] J. Fries, S. Wu, A. Ratner, and C. Ré, Swellshark: A generative model for biomedical named entity recognition without labeled data. arXiv preprint arXiv:1704.06360. (2017).
[10] X. Ren, Z. Wu, W. He, M. Qu, ... and J. Han, CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. WWW. (2017).
[11] D. Cameron, G.A. Smith, R. Daniulaityte, A.P. Sheth, D. Dave, L. Chen, and R. Falck, PREDOSE: a semantic web platform for drug abuse epidemiology using social media. Journal of biomedical informatics, **46(6)** (2013), 985-997.
[12] U. Lokala, R. Daniulaityte, F. Lamy, M. Gaur, K. Thirunarayan, U. Kursuncu, and A.P. Sheth, Dao: An ontology for substance use epidemiology on social media and dark web. JMIR Public Health and Surveillance. (2020).
[13] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, …, and M. Franchini, OntoNotes Release 5.0. Linguistic Data Consortium, Philadelphia, PA (2013).
[14] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, Bond: Bert-assisted open-domain named entity recognition with distant supervision. 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (2020), 1054-1064.
[15] Z. Lu, P. Du, and J.Y. Nie, VGCN-BERT: augmenting BERT with graph embedding for text classification. European Conference on Information Retrieval. (2020), (pp. 369-382).
[16] S. Yadav, V. Pallagani, and A. Sheth, Medical Knowledge-enriched Textual Entailment Framework. arXiv preprint arXiv:2011.05257. (2020).
[17] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. (2006).
[18] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.C. Luthi, …, and W.A. Ghali, Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care, (2005). 1130-1139
[19] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. 2019 Conference of the NAACL: Human Language Technologies, **1** (Long and Short Papers). (2019), (pp.4171-4186).
[20] E. Alsentzer, J.R. Murphy, W. Boag, W.H. Weng, D. Jin, T. Naumann, and M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. (2019).
[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).

Address for correspondence

Goonmeet Bajaj, bajaj.32@osu.edu