Training NER Models: Knowledge Graphs in the Loop

Sotirios Karampatakis[1]✉[0000−0001−7436−7620], Alexis Dimitriadis[1], Artem Revenko[1], and Christian Blaschke[1]

Semantic Web Company, Neubaugasse 1, 1070, Vienna, Austria
{sotiris.karampatakis, alexis.dimitriadis, artem.revenko, christian.blaschke}@semantic-web.com
http://www.semantic-web.com

Abstract. Motivated by the need of annotated data for training named entity recognition models, in this work we present our experiments on a distantly supervised approach using domain specific vocabularies and raw texts in the same domain. In the experiments we use MeSH vocabulary and a random sample of PubMed articles to automatically create an annotated corpus and train a named entity recognition model capable to identify diseases in raw text. We evaluate method against the manually curated CoNLL-2003 corpus and the NCBI-disease corpus.

Keywords: Named Entity Recognition · Machine Learning · Linked Data · Vocabulary · Knowledge Graph

1 Introduction

Named Entity Recognition (NER) is a sub-task of information extraction with the objective to identify and classify named entities mentioned in unstructured text. [7,10] NER is commonly approached as a supervised classification problem. This means that annotated training materials are required. Annotated training corpora are obtainable for common NER types like Person, Location and Organization, but training NER models of other named entity types often requires an additional manual effort to annotate texts.

We aim at producing annotated data semi-automatically. As the pre-requisite we require an incomplete vocabulary for a domain. These vocabularies are often produced manually, for example, as a result of terminology extraction. The effort to produce such an incomplete domain specific vocabulary is significantly smaller than manually annotating a corpus. Most of the existing methods either use gazetteers as additional input to the NER model [2,3,4] or introduce new mechanisms to completely automate the task [6,8]. In this paper we investigate a different task of training an NER model with the help of Knowledge Graphs in the form of vocabularies. We reuse existing NER methods and use the vocabulary to create a training set. We investigate how well the modern NER methods can cope with the errors introduced by automatic annotation and if this procedure could be used to also make the domain vocabularies more complete.

2 Methodology

Automatic annotations For getting the automatically annotated training corpus we have conducted the following steps.

First, we obtain or create a fixed vocabulary of a domain (for instance, chemical compound names, animal species, or geographic locations) to extract entity mentions from texts, i.e. to automatically annotate texts. The vocabulary should be sufficiently large, covering the majority of entity mentions in the training set, in order to reduce obscure entities during entity extraction. Instead of using a plain gazetteer, we structure the vocabulary using the SKOS data model [5] as it is easier to maintain and among with other benefits, it can be reused for additional NLP tasks, such as Entity Linking and Entity Disambiguation.

Second, we obtain a sizeable collection of raw texts (without annotations) for the domain of interest. Entity extraction is run on texts to identify occurrences of the vocabulary items. In particular, words in a text undergo morphological analysis and are matched against the contents of the vocabulary, meaning that only terms in the vocabulary can be recognized. As the result we obtain the **automatically annotated** training set. We use PoolParty Semantic Suite[1] as the automatic annotation tool. PoolParty is a tool that is used in many different enterprise use cases, therefore the quality of annotation is assumed to be high. We publish the resulting annotated dataset at https://github.com/semantic-web-company/ner-corpora.

Third, the automatically annotated corpus is then used to train an NER model, which draws on contextual cues to recognize Named Entities that are similar to the training vocabulary. The new classifier is then able to recognize new entities that are not in the training vocabulary.

Human annotations The human annotations are the original annotations manually done by the creators of the dataset. We train the same NER model also on human annotations and then evaluate it on the test set.

3 Experiments

Description of datasets The **CoNLL-2003** shared task corpus [9] is used as a standard benchmark for the NER task. It consists of human annotated text based on the Reuters News. It is annotated by four NE types: Person, Organization, Location and Miscellaneous. The **NCBI-disease** corpus [1] is a collection of 793 PubMed abstracts fully annotated at the mention and concept level to serve as a NER benchmark in the biomedical domain. The public release of the NCBI disease corpus contains 6892 disease mentions, which are mapped to 790 unique disease concepts.

[1] www.poolparty.biz

Table 1. Evaluation results of OpenNLP NER on human annotated test corpora. Annotation method refers to the training corpora in each case. ΔF_1 is the difference in F_1 scores between automatic and human annotations. Vocabulary identifies how the controlled vocabulary for automatic annotations was created: either already provided human annotations were collected and used for automatic re-annotation or *Disease* branch of MeSH-2019.

Dataset	Vocabulary	Entity Type	Annotation Method						ΔF_1
			Human			Automatically			
			PR	RE	F_1	PR	RE	F_1	
CoNLL-2003	Extracted	Person	96.2	86.2	90.9	90.7	72.1	80.3	-10.6
CoNLL-2003	Extracted	Location	94.9	89.1	91.9	81.2	78.3	79.8	-12.2
CoNLL-2003	Extracted	Organization	94.2	65.4	77.2	55.1	70.2	61.7	-15.5
NCBI-disease	Extracted	Disease	82.7	62.1	70.9	75.6	67.1	71.1	0.2
NCBI-disease	MeSH-2019	Disease	82.7	62.1	70.9	55.5	27.7	36.9	-34.0

Set up To set a baseline for our evaluation, we used the human annotated training corpora to train models and then used the evaluation corpora for each dataset to evaluate the models in terms of Precision (PR), Recall (RE) and F_1 score (F_1). For each of the NE types, we created a Concept Scheme based on the labels of the NE found on the training corpora and re-annotated the raw training corpus using the PoolParty Extractor API, configured to use the corresponding Concept Scheme for each of the NE types. Finally, we used this corpus to train NER models for OpenNLP and evaluated the new models using the human annotated evaluation corpus for each. Results are summarized in Table 1.

Results As presented in Table 1 models trained on automatically annotated corpus can achieve comparable results to models trained on human annotated corpus. Regarding the CoNLL-2003 corpus, an average difference of 12.8% indicates that we can actually create high quality training corpus for NER models. The process allowed us to identify common pitfalls in the automated annotation task. For instance homographs, words with the same spelling but different meaning, led to a large number of erroneous annotations on the training corpus. This induced a noisy training corpus for the classifier and as a result a considerable number of misclassified entities reducing both precision and recall. Hiding those entities from the annotator improved the results. In other cases, the annotator missed the correct bounds of the entity. In the case of the NCBI-disease corpus we conducted an additional experiment. For the automatic annotation part, instead of using extracted annotations from the manually curated training corpus, we used the *Disease* branch of the MeSH vocabulary. In this case, the coverage of the vocabulary was not complete to the annotations in the training corpus, leading to unidentified entities and reducing dramatically the performance of the classifier. Additionally, MeSH contains labels of the entities in an inverted form, for instance "Adenoma, Hepatocellular". This form confuses the annotator in an entity reach sentence leading to incorrect bounds of annotation and thus reduced quality of the training corpus. Normalizing the labels improved the quality.

Case Study We wanted to test if we can produce an improved NER model for diseases. The corpus we used for this experiment consists of 10.000 abstracts (100k sentences) harvested from articles published in PubMed[2], filtered in the domain of diseases.

We used the MeSH vocabulary[3] (2019 update) as the target taxonomy to automatically annotate the corpus. We merged the automatically annotated corpus from PubMed with the manually curated NCBI-disease training corpus and then used OpenNLP to train an NER model for diseases. Finally we evaluated the produced model against the NCBI-disease test corpus. The scores of this model are PR 82.2%, RE 74.7% and F_1 78.2%, an improvement of 7.3% in F_1 and 12.5% in recall and 0.5% decline in recall.

4 Conclusion

The performance of the NER models for each specific NE type is dependant on the quality of the training corpus. Manual curated corpus gives the best results, though in this work we presented that it is feasible to produce comparable results using automatically annotated corpus. The quality of the corpus in this case depends on the quality of the vocabulary or Knowledge Graph used. Additionally the case study showed that we can combine a rather small manually curated training corpus with an automatically annotated training corpus to increase the performance of the model.

Acknowledgements This work has been partially funded by the project LYNX which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu.

References

1. Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics **47** (2014), http://www.sciencedirect.com/science/article/pii/S1532046413001974
2. Liu, A., Du, J., Stoyanov, V.: Knowledge-Augmented Language Model and Its Application to Unsupervised Named-Entity Recognition. pp. 1142–1150. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/n19-1117
3. Liu, T., Yao, J.G., Lin, C.Y.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5301–5307 (2019)
4. Magnolini, S., Piccioni, V., Balaraman, V., Guerini, M., Magnini, B.: How to use gazetteers for entity recognition with neural models. In: Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5). pp. 40–49 (2019)

[2] https://www.ncbi.nlm.nih.gov/pubmed/

[3] Provided by National Library of Medicine https://www.nlm.nih.gov/mesh/meshhome.html

5. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. W3C recommendation **18**, W3C (2009), https://www.w3.org/TR/skos-reference/
6. Nadeau, D., Turney, P.D., Matwin, S.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 4013 LNAI (2006), https://doi.org/10.1007/11766247_23
7. Peng, M., Xing, X., Zhang, Q., Fu, J., Huang, X.: Distantly supervised named entity recognition using positive-unlabeled learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2409–2419. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-1231
8. Štravs, M., Zupančič, J.: Named entity recognition using gazetteer of hierarchical entities. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11606 LNAI, pp. 768–776. Springer Verlag (jul 2019), https://doi.org/10.1007/978-3-030-22999-3_65
9. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. p. 142–147. CONLL '03, Association for Computational Linguistics, USA (2003), https://doi.org/10.3115/1119176.1119195
10. Wang, X., Zhang, Y., Li, Q., Ren, X., Shang, J., Han, J.: Distantly supervised biomedical named entity recognition with dictionary expansion. In: Yoo, I., Bi, J., Hu, X. (eds.) 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019. pp. 496–503. IEEE (2019), https://doi.org/10.1109/BIBM47256.2019.8983212