Difficult Relations: Extracting Novel Facts from Text

Ismini Lourentzou[1], Anna Lisa Gentile[1], Daniel Gruhl[1],
Jane Fortner[2], Michele Freemon[2], and Kendra Grande[2]

[1] IBM Research Almaden, CA, US
[2] IBM Watson Health
ismini.lourentzou@ibm.com, annalisa.gentile@ibm.com, dgruhl@us.ibm.com,
jfortner@us.ibm.com, mfreemon@us.ibm.com, kgrande@us.ibm.com

Abstract. Creating, populating, updating and maintaining a knowledge resource requires intense human effort. Automatic Information Extraction techniques play a crucial role for this task, but many ongoing production systems still require a large component of human annotation. In this work we investigate how to better take advantage of human annotations by performing active learning on multiple IE tasks concurrently, specifically Relation Extraction and Named Entity Recognition. Our proposed approach adaptively requests annotations for one task or the other depending on the current overall performance of the combined extraction. We show promising results on a small use case extracting relations expressing Adverse Drug Reactions from unannotated sentences.

1 Introduction

The task of curating Knowledge Bases (KB) has received substantial attention in the recent years. To extract facts from text or other unstructured or semi-structured content sources, many methods have been proposed [11], mostly borrowing ideas from research areas such as Natural Language Processing, Machine Learning, Statistics etc. If we consider the sole task of KB population, the human component appears in all phases at various degrees, from manual efforts from dedicated teams, like WordNet or Cyc, to collaboratively created resources as in the case of Wikipedia, Wikidata, etc. to more automated efforts where facts can be (semi-)automatically extracted from various sources and the human can only be involved at the validation step.

When dealing with highly curated KBs, especially in the medical domain it is of paramount importance that every novel entity, property or relation added to the KB is nearly 100% accurate. For this reason resources such as pharmaceutical KBs, biology KBs (such as genomics data), Information Extraction (IE) systems for clinical trial data etc. have the requirement that every addition is vetted by at least one human. We therefore consider the case of a semi-automatic extraction of novel facts from text, where although the human is ultimately responsible for the addition of new facts to the KB, they can be effectively supported by Information Extraction methods.

We propose an active learning approach that extracts entities and their relation from text in parallel. We develop a system that consists of an active learning pipeline and two neural models for sequence labeling and classification to perform both Named Entity Recognition (NER) and Relation Extraction (RE) tasks. We bootstrap the model by using small amounts of available previously vetted data - in a cold start scenario we collect some annotated examples by "observing" the user in her task. When new text is analyzed, and if a full relation is extracted, we simply pass it for validation. In cases where no entities or relations can be identified we prompt specific annotation tasks to the end user to collect targeted annotations.

The main contribution of this work is a co-training procedure for NER and RE that seamlessly employs the Active Learning paradigm in real-world knowledge curation tasks. We test the approach for the task of extracting Adverse Drug Reactions (ADRs) from medically relevant text, which involves identifying Drugs and Symptoms as well as whether a causal relation among the two is expressed in the text. We show the benefit of performing both NER and RE concurrently by taking full advantage of the continuous interaction with the human (active learning paradigm). The method does not rely on any manually engineered features, nor other Natural Language Processing tools but simply leverages word and character embeddings, therefore can be easily ported to different languages or text styles requiring only a few initial examples.

2 Related Work

The literature on harvesting information from text for the purpose of knowledge creation is extremely vast [11, 13], as well as the literature on specific methods to solve NER and RE as individual tasks [1], where RE systems mostly rely on the assumption that entities have been pre-tagged. Early approaches for joint NER and RE treat these as two separate tasks in a pipeline and exploit their interactions [4], while integrated approaches have also been explored with a diverse set of methods [7, 12]. More recent approaches exploit neural joint models [10, 5] and have also been applied in biomedical literature for extracting ADRs [6]. The main drawback of current joint models are that they are either (i) time-consuming or (ii) rely on complex structures and (iii) on the availability of large annotated datasets, an assumption that *almost never* holds for "difficult" (long tail) entities and relations. While active learning - which has the advantage of producing usable models at early stage of training- has been explored for NER [2] and RE [3], it has not been investigated, to the best of our knowledge, for joint extraction. We aim to fill this gap by designing an active learning experiment for joint NER and RE to extract ADRs from text.

3 Learning entities and relations with limited annotations

We treat NER and RE as separate but interconnected tasks: each of them is solved by a specific neural model and we design a pipeline to ameliorate the

human annotation process. This choice has several benefits, including the ability of concurrently having multiple annotators with different levels of expertise on different tasks. A joint model would limit us to a single annotator per example for both tasks and would require a large pool of training data due to increased complexity. By separating the two tasks we can leverage the interaction between the two modules so as to limit the number of required annotations.

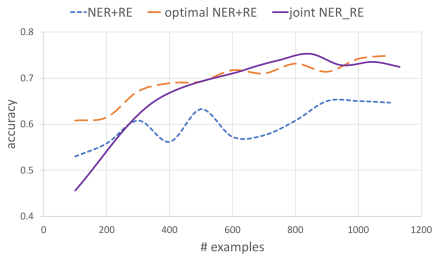
The NER model is an LSTM-CNN-CRF combination similar to [9]. The RE model is mostly based on our previous work [8] which learns one relation at a time. Given a sentence s our goal is to identify whether s: (i) contains entities of interest and (ii) expresses a certain relation r among them. We begin with a pool of unlabeled sentences and we ask our user to label k examples with information about entities[3]. We train our NER on the small batch and use the (imperfect) model to extract entities from the remaining examples and consider sentences containing target entities as positive examples of the target relation. When asked for further annotation for such sentences the human only needs to correct mistakes rather than produce the full annotation for the relation. We collect k of such "approved" annotated relations, train the RE and re-iterate this process. Intuitively, NER does not simply spot all entities of the target class, but only generates entity candidates which are likely to express a relation. The RE module assess if the candidates actually express a relation, which provides feedback to the NER component. This interleaved training paradigm enforces an agreement between the two modules. Minimizing the disagreement of two models is a typical method used in co-training [14]. We test the proposed method to extract ADRs from askapatient.com and involve our medical knowledge curators as humans-in-the-loop. The dataset consists of 1100 (positive and negative) examples of causal relationships between drugs and adverse drug events.

To select examples for NER annotation, we rank unlabeled instances based on their likelihood of being good entity candidates, i.e. likelihood of the generated sequence of tags: $1 - \max_{y_1,\ldots,y_n} P_{\theta_{NER}}(y_1,\ldots,y_n|\mathbf{x})$, which we compute using the Viterbi algorithm. For passing instances to the RE module, we choose between the candidates generated by the NER component or backtrack to uncertainty sampling when the NER cannot find enough instances. In Fig.1 we show the results of our method (**joint NER+RE**) compared to two baselines that apply the NER and RE modules sequentially. **OptimalNER+RE** simulates an optimal NER module by utilizing the gold standard labels of the dataset. Our method achieves comparable performance as starting with an optimal NER.

4 Conclusions

In this work we propose an co-training active learning pipeline for NER and RE to extract entities and their relations from text. Our system continuously trains the models while assisting the knowledge curators in their task of maintaining a

[3] Batch size can adjust based on the task and the annotator, here we set it to 100 examples.



Exp.	A	F1	P	R
optimal NER+RE	74.91	72.16	73.02	71.32
joint NER+RE	72.44	68.03	72.17	64.34
NER+RE	64.66	43.18	80.85	29.46

Fig. 1: Comparison of our proposed method (**joint NER+RE**) with performing NER before RE (**NER+RE**) or having an oracle NER module (**optimal NER+RE**). We report F1, accuracy (A), precision (P) and recall (R).

medical Knowledge Resource up-to-date. We show promising results on a small use case to extract Adverse Drug Reactions from unstructured text.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012)
2. Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C., Xu, H.: A study of active learning methods for named entity recognition in clinical text. Journal of biomedical informatics **58**, 11–18 (2015)
3. Fu, L., Grishman, R.: An efficient active learning framework for new relation types. In: IJCNLP. pp. 692–698 (2013)
4. Ji, H., Grishman, R.: Improving name tagging by reference resolution and relation detection. In: ACL. pp. 411–418. Association for Computational Linguistics (2005)
5. Katiyar, A., Cardie, C.: Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In: ACL. vol. 1, pp. 917–928 (2017)
6. Li, F., Zhang, Y., Zhang, M., Ji, D.: Joint models for extracting adverse drug events from biomedical text. In: IJCAI. pp. 2838–2844 (2016)
7. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: ACL. vol. 1, pp. 402–412 (2014)
8. Lourentzou, I., Alba, A., Coden, A., Gentile, A.L., Gruhl, D., Welch, S.: Mining relations from unstructured content. In: PAKDD. pp. 363–375 (2018), https://doi.org/10.1007/978-3-319-93037-4_29
9. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: ACL. vol. 1, pp. 1064–1074 (2016)
10. Miwa, M., Bansal, M.: End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770 (2016)
11. Paulheim, H.: Automatic Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. SWJ **0**, 1–0 (2015). https://doi.org/10.3233/SW-160218
12. Roth, D., Yih, W.t.: Global inference for entity and relation identification via a linear programming formulation. Introduction to statistical relational learning pp. 553–580 (2007)
13. Weikum, G., Hoffart, J., Suchanek, F.: Ten Years of Knowledge Harvesting: Lessons and Challenges. Data Engineering **5**, 41–50 (2016)
14. Zhou, Z.H., Li, M.: Semi-supervised learning by disagreement. Knowledge and Information Systems **24**(3), 415–439 (2010)