FgER: Fine-Grained Entity Recognition

Abhishek
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Assam, India 781039
abhishek.abhishek@iitg.ernet.in

Introduction

Fine-grained Entity Recognition (FgER) is the task of detecting and classifying entity mentions into more than 100 types. The type set can span various domains including biomedical (e.g., disease, gene), sport (e.g., sports event, sports player), religion and mythology (e.g., religion, god) and entertainment (e.g., movies, music). Most of the existing literature for Entity Recognition (ER) focuses on coarse-grained entity recognition (CgER), i.e., recognition of entities belonging to few types such as person, location and organization. In the past two decades, several manually annotated datasets spanning different genre of texts were created to facilitate the development and evaluation of CgER systems (Nadeau and Sekine 2007). The state-of-the-art CgER systems use supervised statistical learning models trained on manually annotated datasets (Ma and Hovy 2016).

In contrast, FgER systems are yet to match the performance level of CgER systems. There are two major challenges associated with failure of FgER systems. First, manually annotating a large-scale multi-genre training data for FgER task is expensive, time-consuming and error-prone. Note that, a human-annotator will have to choose a subset of types from a large set of types and types for the same entity might differ in sentences based on the contextual information. Second, supervised statistical learning models when trained on automatically generated noisy training data fits to noise, impacting the model's performance.

The objective of my thesis is to create a FgER system by exploring an off the beaten path which can eliminate the need for manually annotating large-scale multi-genre training dataset. The path includes: (1) automatically generating a large-scale single-genre training dataset, (2) noise-aware learning models that learn better in noisy datasets, and (3) use of knowledge transfer approaches to adapt FgER system to different genres of text.

Once realized the proposed work will have a twofold impact. First, FgER system will be beneficial for a plethora of NLP applications where entity recognition is a crucial component. Second, the proposed research path can be applied to several research areas which pose similar challenges.

Our motivation for the proposed direction is from two sources. First, a recently introduced data programming (Ratner et al. 2016) paradigm, where the objective is to quickly generate a noisy training dataset and have a learning model that is noise-aware. Second, the work of transfer learning (Pan and Yang 2010), where the objective is to transfer the knowledge learned in one domain into the new domain.

Proposed Approach

Current Research

Automatic dataset generation The objective is to automatically generate a large-scale training dataset with low noise and no direct human intervention. The annotation quality should be adequate, i.e., the noise present in annotations should be within a permissible limit. The underlying hypothesis is that if the annotation noise is not high, a noise-aware model trained on noisy dataset might perform as good as a supervised learning model trained on a gold dataset.

In this work, we built a dataset creation pipeline which combines heuristics and distant supervision process to generate a large-scale training dataset for the FgER task. Our approach uses Wikipedia and a knowledge base (Freebase[1]) as information sources with no direct human intervention. The pipeline tries to link all entities mentioned in the Wikipedia sentences to Freebase, which then provides the types. The generated dataset contains 31 million annotated sentences with entity mentions assigned types from a type set of size 118. By analyzing a random sample that covers the complete type set of the training dataset, we estimated that in the training dataset around 85% of marked entities are correct and among the correct entities, the dataset captured 89% of them. Compared with existing training dataset (Ling and Weld 2012), our dataset has approx 33% more coverage at similar precision. In addition to automatically generated training dataset, we manually annotated 982 sentences with fine-grained entity types for the sole purpose of evaluating FgER systems. Our dataset will provide a better assessment of FgER systems as it covers almost all the types (99%) in contrast to earlier evaluation data (Ling and Weld 2012) which covers only 38% of types present in the training data. We observed that the skewness in existing dataset leads to a biased evaluation of FgER systems. We are currently in the process to conclude this work and make the dataset public.

[1] https://developers.google.com/freebase/

Noise-aware fine-grained entity classification The entity mention's types are obtained from a knowledge base following distant supervision paradigm in the automatic generation of the training dataset. Types obtained in this manner are independent of the sentence context. For example, every mention of entity *Barack Obama* will receive the same types (e.g., *person*, *author*, *actor*, *artist*, *musician*, *politician*, *organization* and *fictional character*) irrespective of the sentence context. For this problem, we consider context-independent types as noise.

In (Abhishek, Anand, and Awekar 2017) we proposed a non-parametric partial label loss function, that learns better in case of type noise. The loss function does not add any new tunable parameter to existing hyper-parameters and is partial, i.e., it considers few positive context independent types to be correct given an entity mention. In this work, we observed that noise-aware models have an average absolute performance improvement of 3.3% in micro average F1 score on three datasets.

The proposed loss function is similar to the partial loss function proposed in (Ren et al. 2016), which was parametric, and the best parameters vary a lot from dataset to dataset. Our non-parametric loss function was to align with our overall objective of constructing a FgER system, more specifically towards our last problem where we will espouse this work to different multi genre datasets using a partial transfer learning framework.

Future Research

Noise-aware fine-grained entity detection This work will focus on the noise related to entity boundaries, i.e., in the training dataset some of the entities might not be marked, or a non-entity might be marked as an entity, or there is only a partial match between tokens of a marked entity and the actual entity. A learning model trained on this noisy dataset might as well learn the noise. Our objective for this problem is to design a noise-aware sequence labeling loss function that can learn better in this noise.

In (Dredze, Talukdar, and Crammer 2009) the authors showed that for sequence labeling problem, a noise-aware model on noisy dataset could perform on par with supervised learning model on a gold dataset. However, in their experiments, the tokens that have noise (or in their work: multiple labels) were already known. In our problem, the challenge is that we are not aware of the tokens that contain noise.

In this direction we are currently analyzing the following research questions:
1. In the training dataset, can we assign a probabilistic score to annotations? The score will be indicative of the quality of annotation and can be computed based on the heuristic used and a frequent pattern analysis of entities in the corpora.
2. How can we adopt a sequence labeling loss function with a probabilistic annotation score?

The intuition is to have a less stringent penalty in case of incorrectly classifying a label that was annotated with low confidence.

Multi-genre adaptation using partial transfer learning In our initial work of automatically generating training dataset the sentence source was Wikipedia. A learning model trained on one particular genre of text might not generalize well on other genres of text. A FgER system has potential applications in many different genres such as newswire, conversation text, product reviews. Creating training dataset for different genre using heuristics and distant supervision process is a time and resource intensive process and may not yield a dataset with an adequate annotation quality for every genre. However, there are several datasets available for different text genres in which coarse-grained entities are annotated.

In this work, we plan to explore transfer learning approaches (Pan and Yang 2010), however, in our case the challenge is only to transfer the partial knowledge, i.e., a CgER dataset can only provide a subset of types for the FgER task.

Acknowledgments

I thank my thesis advisors Dr. Ashish Anand and Dr. Amit Awekar.

References

Abhishek, A.; Anand, A.; and Awekar, A. 2017. Fine-grained entity type classification by jointly learning representations and label embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 797–807. Association for Computational Linguistics.

Dredze, M.; Talukdar, P. P.; and Crammer, K. 2009. Sequence learning from data with multiple labels. *ECML/PKDD Workshop on Learning from Multi-Label Data (MLD)* 39–48.

Ling, X., and Weld, D. S. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, 94–100. AAAI Press.

Ma, X., and Hovy, E. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1064–1074. Association for Computational Linguistics.

Nadeau, D., and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvistic Investigationes* 30(1):3–26.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Ratner, A. J.; De Sa, C. M.; Wu, S.; Selsam, D.; and Ré, C. 2016. Data programming: Creating large training sets, quickly. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 3567–3575.

Ren, X.; He, W.; Qu, M.; Huang, L.; Ji, H.; and Han, J. 2016. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1369–1378. Association for Computational Linguistics.