International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

A Web Semantic-Based Text Analysis Approach for Enhancing Named Entity Recognition Using PU-Learning and Negative Sampling

Shunqin Zhang, School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, China
Sanguo Zhang, School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, China
Wenduo He, Institute for Network Sciences and Cyberspace (INSC), Tsinghua University, Beijing, China
Xuan Zhang, Tsinghua University, China*

ABSTRACT

The NER task is largely developed based on well-annotated data. However, in many scenarios, the entities may not be fully annotated, leading to serious performance degradation. To address this issue, the authors propose a robust NER approach that combines a novel PU-learning algorithm and negative sampling. Unlike many existing studies, the proposed method adopts a two-step procedure for handling unlabeled entities, thereby enhancing its capability to mitigate the impact of such entities. Moreover, this algorithm demonstrates high versatility and can be integrated into any token-level NER model with ease. The effectiveness of the proposed method is verified on several classic NER models and datasets, demonstrating its strong ability to handle unlabeled entities. Finally, the authors achieve competitive performances on synthetic and real-world datasets.

KEYWORDS

Negative Sampling, NER, PU-Learning, Robustness, Self-Denoising, Token-Level, Two-Step Procedure, Unlabeled Entity Problem

INTRODUCTION

Named-entity recognition (NER) is a well-studied task in natural language processing (NLP) (Tekli et al., 2021; Barbosa et al., 2022; Ehrmann et al., 2023) that has received significant attention (Huang et al., 2015; Ma & Hovy, 2016; Akbik et al., 2018; Li et al., 2020a). In the area of NER, previous methods have had great success (Zhang & Yang, 2018; Gui et al., 2019; Jin et al., 2019; Wang et al., 2023). However, the majority of them rely on well-annotated data and ignore potential unlabeled entities, which are commonly encountered in many cases. Li et al. (2020c) discovered that NER models suffer significantly from the lack of annotations and referred to this as the unlabeled-entity problem.

DOI: 10.4018/IJSWIS.335113 *Corresponding Author

Unlabeled entities often arise from mistakes made by human annotators or the limitations of machine annotators. For instance, distant supervision is a classic method to produce labeled NER data automatically. However, owing to the limited coverage of knowledge resources, datasets generated through distant supervision often retain a significant number of unlabeled entities. Furthermore, enhancing performance with a small set of annotated data could significantly reduce costs. As such, developing an effective and versatile method for NER with unlabeled entities is of great research interest. However, there are several challenges. First, unlabeled entities will misguide the NER training process, causing the model to learn entities as negative instances. It is hard to identify unlabeled entities since they are always confused with negative instances. Second, the reduction of unlabeled entities results in a decrease in learnable data, making it challenging for the model to identify entities correctly. These challenges need to be effectively addressed.

Recently, numerous approaches to alleviate the unlabeled entity problem have been developed. To begin with, Li et al. (2020c) utilized a negative-sampling approach and trained a span-based model to mitigate the misguidance caused by unlabeled entities. They assumed that the unlabeled entities were unknown and thus applied random sampling to cover the unlabeled entities. This line of work was further extended by Li et al. (2022), who used a new weighted sampling distribution to perform a better sampling. Furthermore, Peng et al. (2021) considered reinforcement learning and trained a span selector to enhance the negative-sampling approach.

Another approach makes full use of the labeled data to approximate the true label sequences or detect the potential unlabeled entities. A classic algorithm called positive-unlabeled learning (PU learning) is designed for scenarios where some kinds of samples are easily obtained, but full labeling of all samples is either difficult to obtain or too costly. For instance, Mayhew et al. (2019) proposed the constrained binary learning method, which adaptively trained a binary classifier and assigned weights to each token using the CoDL framework (Chang et al., 2007). Peng et al. (2019) trained a PU-learning (Liu et al., 2002, 2003; Elkan & Noto, 2008; Shunxiang et al., 2023) classifier to perform label prediction; it can unbiasedly and consistently estimate the task loss. Zhang et al. (2022) proposed an adaptive PU-learning technology and then handled the unlabeled-entity problem by integrating it into a machine reading comprehension (MRC) framework. PU learning is widely applied in various fields where obtaining a comprehensive labeled dataset is hard or impractical, offering a solution to effectively utilize limited labeled data along with a larger pool of unlabeled data for better performance.

Another classic algorithm is partial conditional random fields (CRF) (Tsuboi et al., 2008), which is also an effective method for handling unlabeled entities (Yang et al., 2018; Jie et al., 2019; Ding et al., 2023). It functions by generating all potential label sequences for uncertain annotations and subsequently trains on these sequences.

The current methods have achieved great improvement in datasets with unlabeled entities. Despite their success, they also have some limitations. On the one hand, methods that rely on annotated data for self-denoising are heavily dependent on the quality of the available data. These methods often struggle to significantly reduce the impact of unlabeled entities. On the other hand, entirely ignoring annotated information might result in the underuse of valuable data (for example, the negative-sampling techniques). As such, we believe that strategies either entirely dependent on or independent of annotated data are suboptimal. We propose that some of the unlabeled entities can be identified through self-learning methods, yet a certain fraction remains undetected. Thus, combining the advantages of both techniques could offer a more effective solution to the problem of unlabeled entities.

To better solve the unlabeled-entity problem, we propose a robust NER approach that combines a novel PU-learning algorithm and negative sampling. From our empirical studies, we found that the annotated data can be utilized to identify some unlabeled entities in the early stages of model training. However, this capability is constrained, and, notably, a significant portion of unlabeled entities are still confused with negative instances. To address this problem, our approach adopts a two-step strategy for better handling unlabeled entities. Specifically, based on our empirical findings, we propose a

An Overview of and Criteria for the Differentiation and Evaluation of RIA Architectures

Marcel Linnenfelser, Sebastian Weberand Jörg Rech (2010). *Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications (pp. 135-158).*

www.irma-international.org/chapter/overview-criteria-differentiation-evaluation-ria/39168