Abstract

Sequence labelling task traditionally requires large-scale datasets with ground truth labels. As a cheaper and faster alternative approach, crowdsourcing provides sequence labels from multiple noisy annotators at the cost of label quality. In this paper, we propose a novel Teachers-Student framework to aggregate crowd annotations without actual ground truth or prior knowledge about the annotators. Multiple teacher networks reach a consensus on encoding input instances and keep the annotator-specific information in the corresponding crowd components. The student network learns the comprehensive knowledge and aggregates annotator information from multiple teacher networks for better inference.

1 Introduction

Traditional supervised learning systems usually require a large amount of labelled data. However, there are many constraints and difficulties in real world settings. Some tasks are subjective in nature and hard to achieve the consensus on the definition of ground truth (Rodrigues et al., 2014), such as named entity recognition (NER), sentiment analysis and relation extraction. And some tasks require domain-specific knowledge or even professional techniques, such as medical diagnosis.

As an alternative approach, crowdsourcing becomes popular recently and shows its efficiency and benefits in the recent work (Liu et al., 2017; Guan et al., 2017). It collects labelled data with lower cost and higher speed by non-expert contributors with some degradation in quality. Additionally, different annotators may have various levels of reliability and bias, which can lead to disagreement and contradiction. It brings a new challenge to handle multiple annotators and infer the best label for the given instance without any ground truth or prior knowledge about the annotators. This heterogeneous supervision achieves comparable performance in many fields where deep learning shows the advantage, such as multi-class classification (Sheshadri and Lease, 2013), object detection (Su et al., 2012) and information extraction (Liu et al., 2017).

Compared to other domains, much less work focused on sequence labelling task (Finin et al., 2010; Hovy et al., 2014; Nguyen et al., 2017; Rodrigues and Pereira, 2018), which assigns one predefined tag to each token in the input sentence. It needs to consider more context information and label dependency. As an important task in natural language processing (NLP), sequence labelling typically can be applied to many syntactic and semantic problems, such as part-of-speech (POS) tagging and named entity recognition (NER).

To deal with sequential crowd labels, one approach is to aggregate them into a single consensus sequence, then take it as ground truth to train the sequence labelling model. It is straightforward to select the majority label on the token level or selecting the label sequence with more votes as the correct label. It introduces label noise and ignores the difference among the annotators. Another way is to directly train the model on all crowd annotations, using expectation-maximization (EM) algorithm (Dredze et al., 2009; Rodrigues et al., 2014) or deep learning networks (Nguyen et al., 2017; Rodrigues and Pereira, 2018). These methods regard crowd labels as noisy versions of latent ground truth and try to model the annotator-specific biases. However, the training and inference process are not consistent in their works. While only latent representations are used for predicting labels, biased versions are used in the training process.

In this paper, we first investigate the deep learning model with crowd components inspired by Nguyen et al. (2017); Rodrigues and Pereira...
Different from applying crowd components on the hidden states or tag scores, we incorporate crowd matrix inside the structure of linear-chain CRF layer. It is more reasonable to capture consensus knowledge in the encoding layer and annotator-specific biases in the decoding layer.

To address the problem of inconsistency between the training and inference process, we propose Teachers-Student network to better aggregate the crowd representations. The model with multiple crowd components is regarded as teacher networks, which can guide the student network together. The student network directly uses the consensus knowledge in teachers network but also learns how to incorporate teacher-specific information in the crowd component with attention. As the final model, the student network is trained on all annotations and used for inference with the same architecture and parameters.

2 Related Work

As a supervised learning task, sequence labelling typically requires a large amount of precisely labelled data to achieve good performance. Two of the most popular models in the early work are hidden Markov models (HMMs) (Rabiner, 1989) and conditional random fields (CRF) (Lafferty, 2001). Recently, many statistical neural networks show their advantages to learn features, such as convolutional neural networks (CNNs) (Collobert et al., 2011), recurrent neural networks (RNNs) (Goller and Kuchler, 1996) and its extension long-short term models (LSTMs) (Hochreiter and Schmidhuber, 1997; Lample et al., 2016; Ma and Hovy, 2016; Liu et al., 2018). However, obtaining ground truth labels usually costs too much time, human-labour and technical resource.

Dawid and Skene (1979) are the first to address this issue by estimating the hidden ground truth labels from multiple annotations with lower quality. Snow et al. (2008) explored a crowdsourcing platform, Amazon’s Mechanical Turk system, and proved the effectiveness of non-expert annotations for natural language tasks. Later works regarded ground truth labels as a latent variable and focused on EM algorithm (Dempster et al., 1977; Dredze et al., 2009; Raykar et al., 2010) to jointly learn the levels of expertise of each annotator and classification model parameters.

In order to make the marginalization tractable in the sequence labelling task, Rodrigues et al. (2014) took the annotators’ reliabilities as latent variables using the EM algorithm and CRFs. They collected crowdsourced labels of 400 news articles from CoNLL 2003 shared NER task (Sang and De Meulder, 2003) on Amazon’s Mechanical Turk. It consists of 47 valid annotators with enormous levels of label quality and quantity.

To aggregate noisy sequential labels into a high-quality sequence, Nguyen et al. (2017) proposed HMMs with additional crowd component. Although data aggregation can be combined with any traditional supervised algorithm, they are suboptimal compared to jointly learning sequence labelling model on all of the individual annotator. They also introduced additional annotator-specific vector representation on the base of LSTM. It is used to model both label noise and quality for each annotator. The crowd component is integrated into the base model by either added to the tag scores or concatenated to the output of the LSTM layer. The original parameters are supposed to be the latent ground truth representations, which the crowd vectors learn annotator-specific noisy. At test time, the base model predicts the label sequences without any crowd component. It may be inconsistent and problematic because the original representations have never been used for inference during the training process.

Rodrigues and Pereira (2018) also proposed a crowd layer to train the networks end-to-end from multiple noisy annotators for classification, regression and sequence labelling problems. Unlike simply adding an annotator-specific noisy scalar to the tag scores, this work makes a matrix transformation, which is more interpretable and reasonable to model the mislabel probability and adjust the gradients accordingly. However, it is limited to only make the transformation on the tag scores. In this paper, we investigate multiple integration methods to better incorporate annotator-specific knowledge.

To better utilize the learned crowd representations, we treat each annotator as a teacher to train a student network together. Multiple teacher networks are supposed to provide more comprehensive and beneficial guidance (You et al., 2017). Different from the traditional teacher-student framework, the student networks directly take the shared parameters of teachers and learn attention on different teacher-specific knowledge with crowd labels. In this way, the student network
is supposed to be better at sentence encoding and more flexible at fusing teachers’ knowledge in the decoding process.

References


Christoph Goller and Andreas Kuchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In Proceedings of International Conference on Neural Networks (ICNN’96), volume 1, pages 347–352. IEEE.


Liyuan Liu, Xiang Ren, Jingbo Shang, Jian Peng, and Jiawei Han. 2018. Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling. In EMNLP.


Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.