Empirical Co-occurrence Rate Networks For Sequence labeling*

Zhemin Zhu Djoerd Hiemstra Peter Apers Andreas Wombacher
CTIT Dabatase Group, Computer Science, University of Twente
Enschede, The Netherlands
{z.zhu, d.hiemstra, p.m.g.apers, a.wombacher}@utwente.nl

Abstract

Sequence labeling has wide applications in many areas. For example, most of named entity recognition tasks, which extract named entities or events from unstructured data, can be formalized as sequence labeling problems. Sequence labeling has been studied extensively in different communities, such as data mining, natural language processing or machine learning. Many powerful and popular models have been developed, such as hidden Markov models (HMMs) [4], conditional Markov models (CMMs) [3], and conditional random fields (CRFs) [2]. Despite their successes, they suffer from some known problems: (i) HMMs are generative models which suffer from the mismatch problem, and also it is difficult to incorporate overlapping, non-independent features into a HMM explicitly. (ii) CMMs suffer from the label bias problem; (iii) CRFs overcome the problems of HMMs and CMMs, but the global normalization of CRFs can be very expensive. This prevents CRFs from being applied to big datasets (e.g. Tweets).

In this paper, we propose the empirical Co-occurrence Rate Networks (ECRNs) [5] for sequence labeling. CRNs avoid the problems of the existing models mentioned above. To make the training of CRNs as efficient as possible, we simply use the empirical distribution as the parameter estimation. This results in the ECRNs which can be trained orders of magnitude faster and still obtain competitive accuracy to the existing models. ECRN has been applied as a component to the University of Twente system [1] for concept extraction challenge at #MSM2013, which won the best challenge submission awards. ECRNs can be very useful for practitioners on big data.

References

[1] M. B. Habib, M. van Keulen, and Z. Zhu. Concept extraction challenge: University of twente at #msm2013. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', Rio de Janeiro, Brazil*, volume 1019 of *CEUR Workshop Proceedings*, pages 17–20, Aachen, Germany, May 2013. CEUR.

[2] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[3] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML '00*, pages 591–598, 2000.

[4] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[5] Z. Zhu, D. Hiemstra, P. Apers, and A. Wombacher. Empirical co-occurrence rate networks for sequence labeling. In *proceedings of The 12th International Conference on Machine Learning and Applications*, page to appear. IEEE, 2013.

*The software can be requested. This work has been supported by the Dutch national program COMMIT/. Thanks to Maurice van Keulen at University of Twente for his helpful comments.