Empirical Training For Conditional Random Fields

Zhemin Zhu z.zhu@utwente.nl
Djoerd Hiemstra D.Hiemstra@utwente.nl
Peter Apers P.M.G.Apers@utwente.nl
Andreas Wombacher a.wombacher@utwente.nl
CTIT Database Group, Drienerlolaan 5, 7500AE, Enschede, The Netherlands

Conditional Random Fields (CRFs) are undirected graphical models which have been widely applied for sequence labelling, e.g. part-of-speech tagging. Training CRFs (Lafferty et al., 2001) can be very expensive for large-scale applications (Sutton & McCallum, 2009). The standard training **(SD)** of CRFs needs to calculate the partition function $Z_{sd}(X)$ which is a global summation over the whole graph. Piecewise training (**PW**) (Sutton & McCallum, 2009) speeds up the training process by approximating the partition function with an upper bound. But piecewise training is still not scalable to the variable cardinality. Another option for sequence labelling is directed models such as Maximum Entropy Markov Models (MEMMs) (McCallum et al., 2000) which can be trained efficiently. But they suffer from the label bias problem (Lafferty et al., 2001) which may lead to low accuracy.

In this paper (Zhu et al., 2013), we present a practically scalable training method for CRFs called *Empirical Training* **(EP)**. We show that the standard training with unregularized log likelihood can have many maximum likelihood estimations (MLEs). Empirical training has a unique closed form MLE which can be calculated from the empirical distribution very fast. The MLE of the empirical training is also one MLE of the standard training. So empirical training can be competitive in precision to the standard training and piecewise training. And also we show that empirical training is unaffected by the label bias problem even it is a local normalized model. Experiments on two real-world NLP datasets also show that empirical training reduces the training time from weeks to seconds, and obtains competitive results to the standard and piecewise training on linear-chain CRFs, especially when training data are insufficient.

Experiment 1. Brown Corpus is used for the Part-of-Speech (POS) tagging experiment. The size of the tag space is 252. There are 32,623 sentences are used for training and 1,000 sentences are used for testing.

Table 1: Part-of-Speech Tagging Accuracy

Metric	**EP**	SD	PW	PWPL
Accuracy	95.6	95.4	82.9	82.4
Time (s)	3.9	4,571,807	3,791,648	261,021

The method may also suffer from some potential drawbacks. When using large feature vectors the empirical probabilities may become sparse, generalisation from the training data to the test data may be a problem. Also in the experiment we did not try global features. So there is no evidence to show this method works well with global features. Nevertheless, this method is very fast and could be very useful for practitioners who apply CRFs to large scale data sets.

Acknowledgments

This work has been supported by the Dutch national program COMMIT/.

References

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML* (pp. 282–289).

McCallum, A., Freitag, D., & Pereira, F. C. N. (2000). Maximum entropy markov models for information extraction and segmentation. *ICML* (pp. 591–598).

Sutton, C. A., & McCallum, A. (2009). Piecewise training for structured prediction. *Machine Learning*, *77*, 165–194.

Zhu, Z., Hiemstra, D., Apers, P. M. G., & Wombacher, A. (2013). *Closed form maximum likelihood estimator of conditional random fields* (Technical Report TR-CTIT-13-03). CTIT, University of Twente.