Separate Training for Conditional Random Fields Using Co-occurrence Rate Factorization

Zhemin Zhu ✎
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: z.zhu@utwente.nl

Djoerd Hiemstra
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: d.hiemstra@utwente.nl

Peter Apers
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: p.m.g.apers@utwente.nl

Andreas Wombacher
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: a.wombacher@utwente.nl

Conditional Random Fields (CRFs) are undirected graphical models which are well suited to many natural language processing (NLP) tasks, such part-of-speech (POS) tagging and named entity recognition (NER). The standard training method of CRFs can be very slow for large-scale applications. As an alternative to the standard training method, piecewise training divides the full graph into pieces, trains them independently, and combines the learned weights at test time. But piecewise training does not scale well in the variable cardinality. In this paper we present separate training for undirected models based on the novel Co-occurrence Rate factorization (CR-F). Separate training is a local training method without global propagation. In contrast to directed markov models such as MEMMs, separate training is unaffected by the label bias problem even it is a local normalized method. We do experiments on two NLP tasks, i.e., POS tagging and NER. Results show that separate training (i) is unaffected by the label bias problem; (ii) reduces the training time from weeks to seconds; and (iii) obtains competitive results to the standard and piecewise training on linear-chain CRFs. Separate training is a promising technique for scaling undirected models for natural language processing tasks. More details can be found here (http://eprints.eemcs.utwente.nl/22600/).