Extending Hidden Markov (tree) models for word representations

Simon Šuster S.SUSTER@RUG.NL
Department of Language Technology, University of Groningen

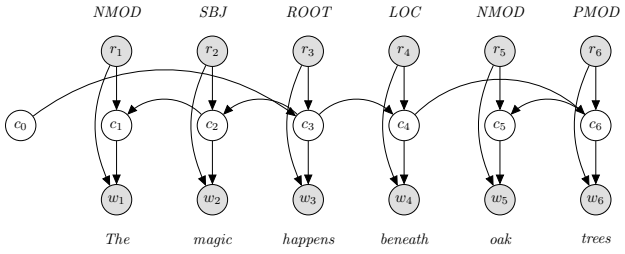
Gertjan van Noord G.J.M.VAN.NOORD@RUG.NL
Department of Language Technology, University of Groningen

Keywords: representation learning, natural language processing, semantic classes, syntactic functions

There is ample research in natural language processing (NLP) on obtaining word representations, including vector space modeling, clustering and techniques derived from language models. Good word representations are vital for overcoming the lexical sparseness inherent to many NLP problems. Much less studied are approaches capturing wider or global context (see e.g. Nepal and Yates (2014)). We are interested in using *syntax* for learning semantic classes, by which a wider and more relevant context can be incorporated (Šuster & Van Noord, 2014). It has been shown that dependency trees can extend *a*) Hidden Markov Models (HMM) in a way that resulting word representations increase performance in NLP classification tasks (Grave et al., 2013), but also *b*) Brown clusters, resulting in higher similarity scores in a wordnet-based experiment (Šuster & Van Noord, 2014). A drawback of the existing approaches is that trees are exploited only partially—dependency links set the *structure* (word context), but the *identity* of dependency links is not part of the model.

To set the intuition for including syntactic functions, consider as example a verb, which typically governs many children. In a bare-bones tree model, all children play an equally important role in constructing the context of the parent, which is a gross oversimplification: for example, subject nouns can be semantically quite unlike object nouns. The notion of the position (left–right) of children with respect to their parents is lost and makes the model in this respect less robust than a simple chain model.

We present our ongoing work towards a probabilistic model capable of including and discriminating between syntactic functions, which we intend to use for obtaining more effective representations. In our initial experiments, we train both chain and tree HMMs and attempt to reproduce the results of Grave et al. (2013) for Dutch. Our learning component consists of combined batch-online expectation-maximization (EM) (Liang & Klein, 2009), and of the forward-backward (for chains) and the belief propagation (for trees) algorithms for inference. We present results from a named-entity recognition task, in which test sequences (or parse trees) are first decoded and then used as features for a structured perceptron classifier. In the second part, we reflect on the extension of HMMs for including syntactic functions. Our proposed architecture adopts the Input-Output HMM (Bengio & Frasconi, 1996). Adapting its left-to-right nature to a tree representation (IOTreeHMM) would allow us to specify two types of observed random variables: *words* (output layer) and *syntactic functions* (input layer). The following example illustrates the model on a concrete sentence as a Bayesian net:



We also consider an alternative view of IOTreeHMM by relaxing the conditioning of the word variable on its observed syntactic function, thus leaving only indirect flow of influence between the two via the hidden layer.

References

Bengio, Y., & Frasconi, P. (1996). Input-output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, *7*.

Grave, E., Obozinski, G., & Bach, F. (2013). Hidden Markov tree models for semantic class induction. *CoNLL*.

Liang, P., & Klein, D. (2009). Online EM for unsupervised models. *HLT-NAACL*.

Nepal, A., & Yates, A. (2014). Factorial Hidden Markov models for learning representations of natural language. *ICLR*.

Šuster, S., & Van Noord, G. (2014). From neighborhood to parenthood: the advantages of dependency representation over bigrams in brown clustering. *COLING*.