Extension of Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms

Stefanie

1 Abstract

In the following drafts, I will share my investigation on many famous papers.

2 Intro

In this paper[1], they describe a new method to estimate parameters for tagging model which relies on Viterbi decoding of training examples. As an example, in order to implement a trigram tagger task, a common approach is to take the parameters to be estimates of conditional probabilities, $\alpha_{x,y,z} = \log P(z|x,y)$ and $\alpha_{t,w} = \log P(w|t)$ where $\alpha_{x,y,z}$ is a trigram, t is a tag and w is a word. Then score will be defined as the log of the joint probability $z_{[1:n]}$ then increase parameter value for features which are missing and downweight parameter values for incorrect features. More general, maximum entropy models represent the tagging task through a feature-vector representation which is denoted by ϕ which could map a history tag pair to a d-dimensional feature vector: $H \times T \rightarrow \mathbb{R}^d$. Therefore, we could define global representation Φ as the sum of the local representation mentioned above[2, 4]:

$$\Phi_s(w[1:n], t[1:n]) = \sum_{i=1}^n \phi_s(h_i, t_i) \quad (1)$$

where h_i is $\langle t_{i-1}, t_{i-2}, w[1:n], t[1:n] \rangle$.

From this equation, we could further define a parameter vector $\bar{\alpha} \in \mathbb{R}^d$ which could be used in probability distribution definition given a history as :

$$p(t|h, \bar{\alpha}) = \frac{e^{\sum_s \alpha_s \phi_s(h,t)}}{Z(h, \bar{\alpha})} \quad (2)$$

where $Z(h, \bar{\alpha})$ is sum of the local representation in the sequence.

The parameter vector $\bar{\alpha}$ is estimated from a training set of sentence/tagged-sequence pairs. Maximum-likelihood parameter values can be estimated using Generalized Iterative Scaling, or gradient descent methods. In some cases it may be preferable to apply a Bayesian approach which includes a prior over parameter values. They describes variants of the perceptron algorithm for tagging problems, an alternative method for estimating parameters of the model[3, 9, 8]. Given a sequence of words $w_{[1:n]}$ and a sequence of part of speech tags $t_{[1:n]}$, we could take the score of a tagged sequence to be

$$\sum_{i=1}^n \sum_{s=1}^d \alpha_s \phi_s(h_i, t_i) = \sum_{s=1}^d \alpha_s \Phi_s(w_{[1:n]}, t_{[1:n]}) \quad (3)$$

This equation does not have normalization term $Z(h, \bar{\alpha})$. Under this method for assigning scores to tagged sequences, the highest scoring sequence of tags for an input sentence can be found using the Viterbi algorithm. If the predicted tag sequence is different from the golden tag sequence, parameters will be updated. Theory justifying the algorithms through a modification of the proof of convergence of the perceptron algorithm for classification problems has been given: The author also mentioned a refinement method for this algorithm which

Initialization: Set parameter vector $\bar{\alpha} = 0$.

Algorithm:

For $t = 1 \dots T, i = 1 \dots n$

- Use the Viterbi algorithm to find the output of the model on the i 'th training sentence with the current parameter settings, i.e.,

$$z_{[1:n_i]} = \arg \max_{u_{[1:n_i]} \in \mathcal{T}^{n_i}} \sum_s \alpha_s \Phi_s(w_{[1:n_i]}^i, u_{[1:n_i]})$$

where $\mathcal{T}^{n_i}$ is the set of all tag sequences of length n_i .

- If $z_{[1:n_i]} \neq t_{[1:n_i]}^i$ then update the parameters

$$\alpha_s = \alpha_s + \Phi_s(w_{[1:n_i]}^i, t_{[1:n_i]}^i) - \Phi_s(w_{[1:n_i]}^i, z_{[1:n_i]})$$

Output: Parameter vector $\bar{\alpha}$.

called averaged parameters. Define $\alpha_s^{t,i}$ to be the value for the s 'th parameter after the i 'th training example has been processed in pass t over the training data. The averaged parameter is defined as: $\gamma_s = \sum_{t=1 \dots T, i=1 \dots n} \alpha_s^{t,i} / nT$ for all $s=1 \dots d$

As for evaluation, part-of-speech tagging and base noun phrase chunking have been implemented by this algorithm and improvements over results for a maximum-entropy tagger has been shown on both of tasks. Also, there are many others paper working on the sequence labelling[5, 6] and speech task[2, 4, 3, 7].

References

- [1] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. Association for Computational Linguistics, July 2002.
- [2] Xiang Kong, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. Analysis of distinctive feature matching with random error generation in a lexical access system. *The Journal of the Acoustical Society of America*, 138(3):1780–1780, 2015.
- [3] Xiang Kong, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5810–5814. IEEE, 2017.
- [4] Xiang Kong, Preethi Jyothi, and Mark Hasegawa-Johnson. Performance improvement of probabilistic transcriptions with language-specific constraints. *Procedia Computer Science*, 81:30–36, 2016.
- [5] Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. Neural machine translation with adequacy-oriented learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6618–6625, 2019.
- [6] Xiang Kong, Qizhe Xie, Zihang Dai, and Eduard Hovy. Fast and simple mixture of softmaxes with bpe and hybrid-lightrnn for language generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6626–6633, 2019.
- [7] Xiang Kong, Xuesong Yang, Mark Hasegawa-Johnson, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. Landmark-based consonant voicing detection on multilingual corpora. *arXiv preprint arXiv:1611.03533*, 2016.
- [8] Kenneth N. Stevens. Modelling affricate consonants. *Speech Communication*, 13(1):33–43, 1993.
- [9] Kenneth N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 114(4):872–1891, 2002.