Exercises

1. Define a second-order MEMM by specifying $P(T_1^n|W_1^n)$, the features used and the Viterbi decoder.

2. Which of the following features are local, allowing the Viterbi algorithm to be used to find $\hat{T}_1^n = \arg\max_{T_1^n} \vec{\theta} \cdot \vec{\phi}(T_1^n, W_1^n)$ given a W_1^n?

   (a) $t_{i-3}t_i$
   (b) $\textsc{Len}(W_1^n)t_i$
   (c) $t_1 t_{i-1}$
   (d) $t_i t_n$
   (e) $t_1 t_n$
   (f) $\textsc{Suffix}(t_i)\textsc{Prefix}(t_{i-2})$

3. Consider the features for POS tagging in Table 8.1.

   (a) Can the feature vector be written as $\phi(t_i, t_{i-1}, w_{i-1}, w_i, w_{i+1})$?

   (b) Why is t_i involved in each feature template?

   (c) If w_1 and w_n are also used for some additional templates, will efficient training or decoding of CRF models be affected?

   (d) If t_1 and t_n are used for additional feature templates, will efficient decoding or training of CRF be affected?

   (e) Are the feature instances sensitive to the value of i? (e.g., if both $w_1 t_1$ and $w_3 t_3$ are "the|DT", will they instantiate to two different feature instances or one feature instance two times?) Consider the case for both MEMMs and CRFs.

   (f) Are Markov assumptions necessary for perceptrons and SVMs for sequence labelling tasks, as they are for CRFs?

   (g) Derive a forward-backward algorithm that calculates $P(t_i, t_{i+1}|w_1^n)$ given a first-order CRF model $\vec{\theta}$.

4. Derive the forward and backward algorithms for calculating $P(t_i = \mathbf{t}|W_1^n)$ given a second-order CRF model.

5. Consider the task of semantic role labelling in Chapter 1. If a predicate is given, cast the problem into sequence labelling. Table 8.1 defines a set of features for POS tagging. Can these features be useful for SRL? Are there other useful features for SRL? If a syntax tree is given, what features can be useful for SRL? Do they complicate the task?

6. Generative sequence labelling models such as HMM turn out not extremely useful for supervised word segmentation, giving significantly lower results compared with CRF. What is the main reason for their lower accuracies?