9.4 Summary

In this chapter we have introduced,

1. Sequence segmentation using Sequence Labeling
2. Discriminative models for directly solving sequence segmentation tasks
3. Semi-Markov Conditional Random Fields
4. A learning guided beam search framework using perceptron training

Exercises

1. Consider a sequence labelling model to solve the NER task. Suppose that the types of entities include person (PER), organization (ORG),

location (LOC) and geo-political entities (GPE). Given a certain input $W_1^n = w_1 w_2 \dots w_n$, how can the marginal probability of $w_j \dots w_k$ being a PER entity be calculated?

2. According to the feature templates in Table 9.5, what is the feature vector $\vec{\phi}(W_1^4)$ for the segmentation "以 (take) 前天(the day before yesterday) 下雨 (rain) 为例 (for example)"? Specify only the non-zero feature instances and their counts.

3. If $w_0 = \langle \text{s} \rangle$ and $w_{|W|+1} = \langle / \text{s} \rangle$ are utilised, which feature templates in Table 9.5 can be applied to instantiating $\vec{\phi}(w_0, w_1)$ and $\vec{\phi}(w_{|W|}, w_{|W|} + 1)$?

4. Discuss possible chunk-level features for (a) syntactic chunking and (b) NER. Compare them with features for word segmentation tasks. What are the salient differences? What are the similarities? Compare the feature templates to Table 9.3 and Table 9.3 for sequence labelling solutions to syntactic chunking and NER, respectively.

5. According to Algorithm 1, how can be highest scored word sequence be recovered from the table *bp*? Given pseudo code here.

6. Where can Algorithm 1 be adjusted if the maximum word size is constrained to M characters? Calculate the asymptotic complexity of the new algorithm.

7. When the feature context is enlarged from word bigram to word trigram for Chinese word segmentation, how could the decoding algorithm be derived?

8. The semi-CRF we discussed in this chapter can be viewed as a first-order semi-CRF. Consider a 0-th order semi-CRF now, for which feature contexts are restricted to only one output segment.

   (a) What is the implication on the decoding process?

   (b) What is the implication to the process of calculating marginal probabilities of segments?

   (c) What is the implication on the training process?

   (d) Now consider labeled cases, where each output segment is assigned a label from L. Specify the decoding and training algorithms for 0th order semi-CRF.

   (e) For labeled sequence segmentation, if the feature context is restricted to $\vec{\phi}(l(s_{j-1}), b(s_j), e(s_j), l(s_j))$, where $l(s)$ represents the label of a segment s, specify the decoding and training algorithms again.

9. Consider labeled sequence segmentation for first order semi-CRF. How can Algorithm 1 for decoding and Algorithm 2, Algorithm 3 and Algorithm 4 for marginal probability calculation be adapted for accommodating label features?