Part-Of-Sentence Task and Basic Generalization of Online Training

Part-of-sentence (POS tagging or POST) is a widely used NLP task. The target is to mark up a word in a corpus (text) as corresponding to a particular part of the sentence based both on its definition (meaning) and context (semantic information). POS tags describe the characteristic structure of lexical terms within a corpus; it can be used for making assumptions about semantics. A basic method is online training.

Online Fixed/Non-fixed Model (Sentence-level): Pseudo Code

The algorithm of online fixed/non-fixed models (figure 1&2) are shown as below.

Online Fixed Learning:
Repeat:
1. Make predictions on the remaining unlabeled data
2. Use mean probability to select Top N sentences
3. Generate new training data
4. Use new training data to train the base_model
5. Remove Top N sentences from unlabeled data
Stop: No remaining data

Online Non-Fixed Learning:
Repeat:
1. Make predictions on the whole unlabeled data
2. Use mean probability to select Top N sentences
3. Generate new training data
4. Use [new training data] to train the base_model
5. Remove Top N sentences from unlabeled data
Stop: The number of changes in Top N sentences is less than 20% of N

Acknowledgments

We are deeply indebted to Barbara Plank for sharing the complete cross-domain datasets. We are extremely grateful to Columbia DSI for providing us with GCP computational power.

References
