

DIL Tutorial 2

Philipp Koehn

February 6, 2006

This tutorial deals with tagging by supervised learning.

Given the following data set:

c/C a/A b/B c/A a/A c/C c/C a/A b/B c/A

We would like to learn models that predict the tag (lowercase character) from the word (uppercase character).

1. **Hidden Markov model:** Give the structure of a simple unigram HMM model for this data.
2. **Transformation-based learning:**
 - (a) Given the unigram HMM, how will the training data be tagged?
 - (b) How many errors does the unigram HMM make?
 - (c) What additional transformation rules would lead to zero error on the training data?
3. **Maximum entropy model:** We define the following features:

$f_1 = \{1, \text{ if } A = 1 \text{ and } a = 1; \text{ otherwise } 0\}$

$f_2 = \{0, \text{ if } A = 1 \text{ and } a = 1; \text{ otherwise } 1\}$

$f_3 = \{1, \text{ if } B = 1 \text{ and } b = 1; \text{ otherwise } 0\}$

$f_4 = \{0, \text{ if } B = 1 \text{ and } b = 1; \text{ otherwise } 1\}$

$f_5 = \{1, \text{ if } C = 1 \text{ and } c = 1; \text{ otherwise } 0\}$

$f_6 = \{0, \text{ if } C = 1 \text{ and } c = 1; \text{ otherwise } 1\}$

- (a) What is the empirical expectation $\tilde{E}(f_j)$ for the six features?

$$\tilde{E}(f_j) = \frac{1}{n} \sum_{i=1}^n f_j(h_i, t_i)$$

- (b) The maximum entropy model is defined by

$$p(h_i, t_i) = \prod_j \lambda_j^{f_j(h_i, t_i)}$$

If we initialize all $\lambda_j = 1$, what is the model expectation $E(f_j)$ for the six features?

$$E(f_j) = \frac{1}{n} \sum_{i=1}^n p(t|h_i) f_j(h_i, t)$$

- (c) We ignored in our presentation of maximum entropy model the normalization factor $\frac{1}{Z}$:

$$p(h_i, t_i) = \frac{1}{Z} \prod_j \lambda_j^{f_j(h_i, t_i)}$$

What purpose does this factor play, and how should it be set in our example? What is the model expectation of the features with this normalization factor?

- (d) Perform one iteration of the Improved Iterative Scaling algorithm on this data.

$$\Delta\lambda_i = \frac{1}{C} \log \frac{\tilde{E}(f_i)}{E(f_i)}$$

- (e) How can the rules from transformation-based learning be used as features in the maximum entropy model?