**NLP Homework 1 – Report**

Snir Lugassy

Yonathan

**Data processing**

Our data processing class “DataProcessing” is implemented under the file “xx”.

This class constructor input is the data file and it is responsible for:

* Reading the data file
* Splitting every sentence into a list of pair (word-tag pair)
* Generating all histories based on the given data (list of history tuples).

**Features**

In both models we used the same set of feature groups (generic definition of features), but the two models generate different features according to given thresholds.

The big model learns from a larger dataset, whereas the small model learns from a much smaller dataset, therefore we will use very low thresholds in the small model and higher thresholds on the big model where we can afford to generalize more.

Notation

Let us denote the set of all sentences in the dataset where each consists of word in the following notation:

In other words, is the i-th word in sentence .

Likewise, the tag of the i-th word in sentence will be notate as .

The set of all histories will be denotate as where each history consists of

Feature definitions

We defined the following features for every history , tag based on every history from the dataset.

Where the learned history is:

And the input history is:

The features groups are defined as:

For :

Where the functions returns the suffix / prefix with size of .

**Feature extraction**

In the feature extraction process implemented in “TaggingFeatureGenerator” we counted the occurrence of every feature mentioned above in the dataset and filter less frequent features according to a specific threshold.

The generated features in the **big model**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature group | Threshold | Number of features |
| 1 |  | 20 | 700 |
| 2 |  | 20 | 599 |
| 3 |  | 10 | 822 |
| 4 |  | 50 | 1061 |
| 5 |  | 50 | 1116 |
| 6 |  | 10 | 15 |
| 7 |  | 10 | 9 |
| 8 |  | 10 | 69 |
| 9 |  | 10 | 24 |
| 10 |  | 10 | 1671 |
| 11 |  | 10 | 534 |
| 12 |  | 10 | 42 |
| 13 |  | 10 | 22 |
| 14 |  | 10 | 21 |
| 15 |  | 10 | 8 |
| 16 |  | 10 | 9 |
| 17 |  | 10 | 1 |
|  |  |  | **Total: 6723** |

The generated features in the **small model**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature group | Threshold | Number of features |
| 1 |  | ? | ? |
| 2 |  | ? | ? |
| 3 |  | ? | ? |
| 4 |  | ? | ? |
| 5 |  | ? | ? |
| 6 |  | ? | ? |
| 7 |  | ? | ? |
| 8 |  | ? | ? |
| 9 |  | ? | ? |
| 10 |  | ? | ? |
| 11 |  | ? | ? |
| 12 |  | ? | ? |
| 13 |  | ? | ? |
| 14 |  | ? | ? |
| 15 |  | ? | ? |
| 16 |  | ? | ? |
| 17 |  | ? | ? |
|  |  |  | **Total: ???** |

**Model accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training set size | Testing set size | Accuracy |
| Big model | ? | ? | 0.92533 (92.5%) |
| Small model | ? | ? |  |

Top 10 incorrect tags confusion matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | NN | JJ | RB | IN | VBD | VBN | NNP | VB | NNPS | VBZ |
| NN | 2986 | 172 | 16 | 8 | 7 | 4 | 38 | 18 | 0 | 1 |
| JJ | 180 | 119 | 19 | 2 | 8 | 28 | 27 | 9 | 0 | 1 |
| RB | 33 | 24 | 651 | 30 | 2 | 5 | 16 | 5 | 0 | 1 |
| IN | 12 | 3 | 47 | 2454 | 0 | 1 | 4 | 0 | 0 | 1 |
| VBD | 15 | 5 | 2 | 3 | 733 | 54 | 0 | 1 | 0 | 0 |
| VBN | 8 | 26 | 0 | 0 | 43 | 416 | 0 | 2 | 0 | 0 |
| NNP | 10 | 14 | 6 | 2 | 1 | 2 | 1921 | 0 | 18 | 0 |
| VB | 26 | 11 | 4 | 0 | 1 | 3 | 2 | 528 | 0 | 0 |
| NNPS | 0 | 4 | 0 | 0 | 0 | 0 | 44 | 0 | 29 | 0 |
| VBZ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 453 |

Inference algorithm - Viterbi