Natural Language Processing 1 (Assignment 2)

Ujjwal Sharma (11392010)

22 November 2016

1 Question 1

1.1

The General equation for sentence probability under the Uni-gram model is

$$\boxed{P(w_i^n) = \prod_{i=1}^{N} P(w_i)}$$

No, this is not a good solution for the *their* vs *there* problem since it does not take the preceding word into context which is essential for deciding which form of this word should be used.

1.2

Under a Bi-gram model, the sentence probability is given by:

$$\boxed{P(w_n|w_{n-1}) = \prod_{n=1}^{N} P(w_{n-1}|w_n)}$$

2 Question 2

Examples of scenario where the independence assumption is violated are:

1. The Food given the cold weather was quite tasty.
2. You are requested to turn over all your IDs.
3. I could have helped you and maybe I would have , but I won't.

3 Question 3

3.1

The transition probabilities for the training data using Maximum-Likelihood Estimation is given

	PER	ORG	OTH	⟨\s ⟩
⟨ s ⟩	0.5	0	0.5	0
PER	0.5	0	0.5	0
ORG	0	0.5625	0.437	0
OTH	0.011	0.077	0.866	0.0444

In the above Transition matrix, the columns are the preceding word of the n-gram and the rows are the succeeding word of the n-gram.

3.2

The transition probabilities using add-one smoothing are:

	PER	ORG	OTH	⟨\s ⟩
⟨ s ⟩	0.428	0.1428	0.428	0
PER	0.4	0.1	0.4	0.1
ORG	0.05	0.5	0.4	0.05
OTH	0.021	0.085	0.8404	0.053

3.3

1. Before Additive Smoothing,

$$P^*(Obama|PER) = \frac{C(W_x)}{N}$$

$$\boxed{P^*(Obama|PER) = \frac{3}{6} = \frac{3}{6} = 0.5}$$

After the Additive-Smoothing, we get

$$P^*(Obama|PER) = \frac{C(W_x) + 1}{N + V}$$

$$P^*(Obama|PER) = \frac{C(Obama|PER) + 1}{N + V}$$

$$\boxed{P^*(Obama|PER) = \frac{3+1}{6+4} = \frac{4}{10} = 0.4}$$

We observe a net decrease in the probability of this Word-Tag pair after additive smoothing.

2.

$$P^*(Obama|ORG) = \frac{C(W_x)}{N}$$

$$\boxed{P^*(Obama|ORG) = \frac{0}{16} = 0}$$

After the Additive-Smoothing, we get

$$P^*(Obama|ORG) = \frac{C(W_x) + 1}{N + V}$$

$$P^*(Obama|ORG) = \frac{C(Obama|ORG) + 1}{N + V}$$

$$\boxed{P^*(Obama|ORG) = \frac{0 + 1}{16 + 4} = \frac{1}{20} = 0.05}$$

We observe a slight increase in the probability of this Word-Tag pair after additive smoothing.

3.4

The LOC tags cannot always disambiguate the information because location often appears as a part of ORG tags and this cannot disambiguate a general text.

”The Daily Bugle has a long history of covering Superman. The Bugle was questioned last week by the authorities for defaming Superman”

The above text refers to Daily Bugle as an ORG tagged entity however a later reference to Bugle implying the organization might be misinterpreted as if the musical instrument called ’Bugle’ was being questioned. This will cause serious conflicts in case acronyms (for example CLAWS) are used which are also proper English words.

3.5

Named-Entity Recognition Systems can improve by classifying all Out-Of-Vocabulary (OOV) words as Non Standard Words (NSW). All NSW’s that do not need to be normalized but are generally correct (examples iPhone) and by adding them to a dictionary that does not discard their probability of occurrence altogether and adds the to a dictionary so that their presence is not so unlikely.

For example iPhone is not a dictionary or In-Vocabulary word but will appear in a multitude of unconnected scenarios so adding it to a list of NSW entries generally helps in a corpus like Twitter where '2mrw' might be used to refer to tomorrow.

A dictionary approach might suggesting converting this phrase to a standard vocabulary entry but this will generally not b useful since corpora like Twitter do not generally incorporate full length standard vocabulary words and contain a large set of NSW's.

4 Question 4

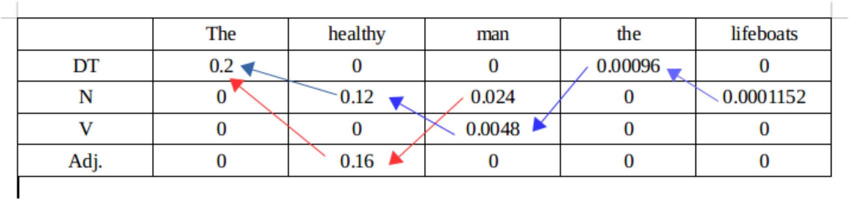
Deciding based on the highest probabilities (represented as number of occurrences in the given data) ,we get :

$$\boxed{The(DT) \rightarrow Healthy(Adj.) \rightarrow man(N) \rightarrow the(DT) \rightarrow lifeboats(N)}$$

The above parse is **grammatically wrong** given the context.

5 Question 5

The Viterbi Algorithm to parse the sentence can be used as follows:



	The	healthy	man	the	lifeboats
DT	0.2	0	0	0.00096	0
N	0	0.12	0.024	0	0.0001152
V	0	0	0.0048	0	0
Adj.	0	0.16	0	0	0

The sentence exhibits the following tag structure when parsed with the Viterbi algorithm

$$\boxed{The(DT) \rightarrow Healthy(N) \rightarrow man(V) \rightarrow the(DT) \rightarrow lifeboats(N)}$$

This represents a more accurate parse of the sentence as compared to a strictly probabilistic bi-gram tagging approach.