Homework Assignment 1

Anand Kumar Singh (11392045)

November 13, 2017

1 Problem 1

1.1 Unigram

Given:

$$P(w_i) = \frac{count(w_i)}{N}$$

where N is the number of words in the corpus. Also we are given that for every word v in the vocabulary $count(v) > 0$

$\therefore$ The sentence probability under the unigram model is given as :

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{N} P(w_i) = \prod_{i=1}^{N} \frac{count(w_i)}{N} \tag{1}$$

No this does not seem like a good solution to *their* vs *there* problem since it doesn't take the context of the word *their* (or *there*) into account. It only focuses on their respective counts and therefore in a corpus which has higher frequency of *their* for example, against *there*, it will allocate a higher probability to *their* and by extension 'p(He saw their was a cat on the street) >p(He saw there was a cat on the street) ', which is incorrect.

1.2 Bigram

The sentence probability under the bigram model is given as :

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{N} P(w_i|w_{i-1}) = \prod_{i=1}^{N} \frac{count(w_i, w_{i-1})}{count(w_i)} \tag{2}$$

Language has long (and short) term dependencies and bigram model would do a better job as compared to a unigram model in capturing a short term dependency atleast, since probability of every word is contingent on the word preceding it. Therefore a bigram model will do a better job at capturing context and by extension a better job at resolving the *their* vs *there* problem for instance. Revisiting the example from previous section, while P(their/saw) and P(there/saw) can have matching frequencies in a corpus, P(was/there) >P(was/their) since the former is more accurate grammatically. Since a bigram model will capture this dependency, it will give 'p(He saw their was a cat on the street) <p(He saw there was a cat on the street) ', which is correct.

2 Problem 2

- In the city *lived* many poor **people**.
- **The man** next to the tall oak tree near the grocery store, *talks* rapidly.
- **The car** with many riders *was speeding* around the curve.

In all the above examples the subject (in bold) is NOT independent of the verb/verb phrase (in italics). However since they are more than 2 words apart, a trigram model will never be able to capture these long term dependencies.

3 Problem 3

3.1 Transition Probability Matrix

Note: $C(t_1 \mid t_2) = \frac{C(t_1, t_2)}{C(t_2)}$

Table 1: Transition Probability Matrix

Tags	**PER**	**ORG**	**OTH**	**</s >**
<s>	0.4	0	0.6	0
PER	0.5	0	0.5	0
ORG	0	0.5625	0.4375	0
OTH	0.011	0.080	0.852	0.057

3.2 Add-1 Smoothing

Note: C(unique POS) = 4 in transition matrix since there is no transition from **</s >**to any other state.

3.2.1 P(***Obama***| PER)

Before smoothing:

$$P(Obama|PER) = \frac{C(Obama|PER)}{C(PER)} = \frac{3}{6} = 0.5 \tag{3}$$

After smoothing:

$$P(Obama|PER) = \frac{C(Obama|PER) + 1}{C(PER) + C(uniquePOS)} = \frac{3+1}{6+4} = \frac{4}{10} = 0.4 \tag{4}$$

3.2.2 P(***Obama***|ORG)

Before smoothing:

$$P(Obama|ORG) = \frac{C(Obama|ORG)}{C(ORG)} = \frac{0}{16} = 0 \tag{5}$$

After smoothing:

$$P(Obama|ORG) = \frac{C(Obama|ORG) + 1}{C(ORG) + C(uniquePOS)} = \frac{0+1}{16+4} = \frac{1}{20} = 0.05 \tag{6}$$

3.3 LOC tag

The context will not necessarily be able to disambiguate between the LOC and ORG tags after being trained on the given training data. For ex: in the general statement written below, Pentagon appears twice, once as an organization and secondly as a location. The context might not be able to resolve this ambiguity.

'A statement regarding Trump's North Korea policy was released by The Pentagon/ORG today. All the offices housed within The Pentagon/LOC are in disagreement with Trump.'

3.4 Improve NER system

A NER system can be improved by incorporating corpora consisting of non standardized words e.g. twitter or tumblr. Such corpora contain many standard and non standard abbreviations for common words. For e.g.

today can be written as 2day/tdy. In the given text Harvard Law School can be abbreviated as HLS and University of Chicago as UChicago. A mapping on such abbreviations to their original versions can make an automatic NER system perform better.