Maico Timmerman
10542590
maico.timmerman@gmail.com

Homework Assignment 1
Natural Language Processing 1, 17/18

2017-11-12

Questions 1

(a) The general equation for the unigram-model can be written as:

$$P(w_1, w_2, \ldots, w_3) = \prod_{i=1}^{N} P(w_i) \tag{1}$$

$$= \prod_{i=1}^{N} \frac{\text{count}(w_i)}{N} \tag{2}$$

$$\tag{3}$$

This is not a good solution to the *their* vs. *there* problem. Using the unigram model, the choice between *their* or *there* is based on the number of occurrences of these words in the training data. This approach will never take into account when a combination of words results in an invalid sentence.

(b) The general equation for the bi-gram model can be written as:

$$P(w_1, w_2, \ldots, w_3) = P(w_1) \cdot \prod_{i=1}^{N-1} P(w_{i+1} \mid w_i) \tag{4}$$

$$\tag{5}$$

This model will take into consideration that some occurrences of their/there might not be valid after certain words. Based on the statistics in the training set, the model will decide which of the two words is valid after, the seen word. However, it is still possible that the model will not choose the correct form.

Question 2

1.

$$\text{I have a dog that walks on the grass.} \tag{6}$$

Here the independence assumption is violated, because it could is possible that:

$$P(\text{walks} \mid \text{a}, \text{dog}) > P(\text{that} \mid \text{a}, \text{dog}) \tag{7}$$

However that would result into the sentence: “I have a dog walks…”, resulting in an invalid sentence. Using the same probabilities also a valid sentence can be constructed: “A dog walks on the grass”.

2.

$$\text{Computers produced in a factory make mistakes too.} \tag{8}$$

violates the independence assumption, as “make” assumes that Computers is plural. Inserting a singular noun, will result in an invalid sentence: “A computer produced in a factory make mistakes too.”. In this case “make” should have been “makes”, due to “A computer” being singular.

3.

$$\text{He told him something about himself.} \tag{9}$$

violates the independence assumption, as the pronoun “himself” is depended on “He”. Changing “He” to “She” will render an invalid sentence, as “himself” should then become “herself”.

Question 2

1. The transition probability matrix for the corpus is:

	<s>	OTH	PER	ORG	</s>
<s>	0	$\frac{3}{5}$	$\frac{2}{5}$	0	0
OTH	0	$\frac{75}{88}$	$\frac{1}{88}$	$\frac{7}{88}$	$\frac{5}{88}$
PER	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
ORG	0	$\frac{7}{16}$	0	$\frac{9}{16}$	0
</s>	0	0	0	0	0

2. $$P(\text{Obama} \mid \text{T}) = \frac{C(\text{Obama} \mid \text{T}) + 1}{C(\text{T}) + V_{\text{T}}} \tag{10}$$

   Where we define V_{T} as the subset of the vocabulary that has the T-tag, $P(C(\text{Obama} \mid \text{T}))$ the number of times "Obama" is tagged with the T-tag, and $C(\text{T})$ is the number of words tagged with the T-tag.

   Using this we calculate $P(\text{Obama} \mid \text{PER})$ and $P(\text{Obama} \mid \text{ORG})$:

   $$P(\text{Obama} \mid \text{PER}) = \frac{3+1}{6+4} = \frac{2}{5} \tag{11}$$

   $$P(\text{Obama} \mid \text{ORG}) = \frac{1}{16+9} = \frac{1}{25} \tag{12}$$

3. Context will not be able to disambiguate between the LOC- and the ORG-tag. Especially since we are only using limited set of four tags, we can not disambiguate the named entity using the preceding words, as they all have the OTH-tag.

   For example "at the University of Chicago Law School" can be interpreted as being handled by the organisation or be interpreted as located at the university that is in Chicago.

4. Other sources of information that might help an automatic NER system could be:

   Capitilization Words containing capital letter at the start or in the middle are more likely to be named entities.

   Hyphenation Words that are hyphenated are more likely to be a named entity.

   Digits words containing digits can often be labeled as named entities.

   Word length Lengthy words could be an indication for named entities.

   Punctuation Named entities like acronyms often contain dots.