Homework 1

Haukur Páll Jónsson
NLP 2017

November 12, 2017

Question 1. *Their* vs *There*

Answer a): Unigram model

The probability of a sentence, $w_1, w_2, ..., w_n$ is given by:

$$p(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} p(w_i|w_1^{i-1})$$

In the unigram model we assume $p(w_i|w_1^{i-1}) = p(w_i)$, i.e. we assume that words are independent of each other. Therefore,

$$p(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} p(w_i)$$

The approach to attempt to solve the *their* vs *there* in terms of a unigram model is a bad idea for the following reason. When evaluating a probability of a sentence; $p(w_1, w_2, ..., w_n)$ which contains either *their* or *there* and we want to see which sentence is more likely, we compare their probabilities and choose the one with higher probability. Namely:

$$argmax_{their,there}\{p(w_1, w_2, ..., w_{n-1})*p(their), p(w_1, w_2, ..., w_{n-1})*p(there)\} = argmax_{their,there}\{p(their), p(t$$

That is, we always pick the word (out of there/their) which has higher probability. That is, the more occurring word will always be considered the "correct" word.

Answer b): Bigram model

When using a bigram we assume; $p(w_i|w_1^{i-1}) = p(w_i|w_{i-1})$, that is, the probability of a word depends on the preceding word. Thus, the formula is:

$$p(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} p(w_i|w_{i-1})$$

The bigram model would do a lot better as it can account for the fact that a noun is more frequently preceded by *their*, rather than *there*. Similarly, a verb is more frequently preceded by *there*, rather than *their*. The language model does not know what a "noun" or a "verb" is, but it knows the probability over all words preceding *their* and *there*, which will have this structure.

Question 2. Independence assumption

Answer a): Consider these sentences as examples of when the independence assumption is broken. "Easier said than done.", "I wish you a merry Christmas." and "Two plus two is four.". In all of these sentences, any deviation from the last word would be highly improbable. We need to consider these phrases as a complete sentences.

Question 3. Hidden Markov Model and Named Entity Recognition

Answer a): Transition matrix

The matrix is represented s.t. we go from column to row, that is, if we sum up the probabilities of a column we get 1.

Table 1: Transition Matrix

	$<s>$	PER	ORG	OTH
$</s>$	0	0	0	0.056
PER	0.4	0.5	0	0.011
ORG	0	0	0.563	0.080
OTH	0.6	0.5	0.437	0.853

Answer b): $p(Obama|PER)$ with add-1 smoothing is 0.0547945. $p(Obama|Org)$ with add-1 smoothing is 0.012048.

Answer c): Consider the sentence "He's going to Columbia next month.". This sentence could mean that a person is going to the University of Columbia, an organization, or that a person is going to the country Columbia, a location. A text-context might not be enough to disambiguate this sentence as the only way to disambiguate might be to know who "he" is and if "he" is more probable to be going to a university or a country.

Answer d): A precompiled list of common domain or application dependent names could be a valuable information and improve the accuracy of the NER system. This would help in cases where the context does not have enough information to disambiguate between LOC and ORG. From the text given, consider "Chicago". In the text this is a LOC but the word could very will be substituted by an ORG and still be a meaningful sentence. Knowing a priori that "Chicago" is most commonly a LOC could improve accuracy.