Department of Computer Science
The University of Melbourne
COMP90042 Web Search and Text Analysis (Semester 1, 2017)

Workshop exercises: Week 8

Discussion

1. What is **Discourse Segmentation**? What do the segments consist of, and what are some methods we can use to find them?

    - In Discourse Segmentation, we try to divide up a text into discrete, cohesive units based on sentences.
    - By interpretting the task as a boundary–finding problem, we can use rule–based or unsupervised methods to find sentences with little lexical overlap (suggesting a discourse boundary). We can also use supervised methods, by training a classifier around paragraph boundaries.

2. What is an **anaphor**?

    - From the lectures: an anaphor is a linguistic expression that refers back to one or more elements in the text (generally preceding the anaphor)
    - These tend to be pronouns (*he*, *she*) but can also be determiners (*which*, *the*, etc.).

    (a) What is **anaphora resolution** and why is it difficult?

    - This is the problem of working out which element (generally a noun or noun phrase, but sometimes a whole clause) a given anaphor is actually referring to.
    - For example:

        > Mary gave John a cat for **his** birthday. (i) **She** is generous. (ii) **He** was surprised. (iii) **He** is fluffy.

        his [birthday] obviously refers to John; (i) (presumably) refers to *Mary*; (ii) (presumably) refers to *John*; and (iii) (presumably) refers to *[the] cat*.

    (b) What are some useful heuristics (or features) to help resolve anaphora?

    - The most obvious (but inherent unreliable) heuristic is the **recency heuristic**: given multiple possible referents (that are consistent in meaning with the anaphor), the mostly intended one is the one most recently used in the text.
    - A better heuristic is that the most likely referent (consistent in meaning with the anaphor) is the focus of the discourse (the "center").
    - We can also build a supervised machine learning model, usually based around the semantic properties of the anaphor/nearby words and the sentence/discourse structure.

3. For the following "corpus" of two documents:

1. `how much wood would a wood chuck chuck if a wood chuck would chuck wood`
2. `a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood`

- I'm going to show the frequencies of the ten different word uni-grams, as it will make life a little easier in a moment:

`a`	`chuck`	`could`	`he`	`how`	`if`	`much`	`the`	`wood`	`would`	Total
4	9	1	1	1	2	1	1	8	4	32

(a) Which of the following sentences: `A`: `a wood could chuck`; `B`: `wood would a chuck`; is more probable, accoding to:

i. An unsmoothed uni-gram language model?

- An unsmoothed uni-gram language model is simply based on the counts of words in the corpus. For example, out of the 32 tokens in the corpus, there were 4 instances of `a`, so $P(\texttt{a}) = \frac{4}{32}$
- To find the probability of a sentence using this model, we simply multiply the probabilities of the individual tokens:

$$\begin{aligned}
P(A) &= P(\texttt{a})P(\texttt{wood})P(\texttt{could})P(\texttt{chuck}) \\
&= \frac{4}{32} \times \frac{8}{32} \times \frac{1}{32} \times \frac{9}{32} \approx 2.75 \times 10^{-4} \\
P(B) &= P(\texttt{wood})P(\texttt{would})P(\texttt{a})P(\texttt{chuck}) \\
&= \frac{8}{32} \times \frac{4}{32} \times \frac{4}{32} \times \frac{9}{32} \approx 1.10 \times 10^{-3}
\end{aligned}$$

- Clearly sentence `B` has the greater likelihood, according to this model.

ii. A uni-gram language model, with Laplacian ("add-one") smoothing?

- Recall that in add-one smoothing, for each probability, we add 1 to the numerator and the size of the vocabulary (in this case, 10) to the denominator. For example, $P_{\mathrm{L}}(\texttt{a}) = \frac{4+1}{32+10} = \frac{5}{42}$.
- Everything else proceeds the same way:

$$\begin{aligned}
P_{\mathrm{L}}(A) &= P_{\mathrm{L}}(\texttt{a})P_{\mathrm{L}}(\texttt{wood})P_{\mathrm{L}}(\texttt{could})P_{\mathrm{L}}(\texttt{chuck}) \\
&= \frac{5}{42} \times \frac{9}{42} \times \frac{2}{42} \times \frac{10}{42} \approx 2.89 \times 10^{-4} \\
P_{\mathrm{L}}(B) &= P_{\mathrm{L}}(\texttt{wood})P_{\mathrm{L}}(\texttt{would})P_{\mathrm{L}}(\texttt{a})P_{\mathrm{L}}(\texttt{chuck}) \\
&= \frac{9}{42} \times \frac{5}{42} \times \frac{5}{42} \times \frac{10}{42} \approx 7.23 \times 10^{-4}
\end{aligned}$$

- Notice that the probability of sentence `A` is larger using this model, because the probability of the unlikely `could` has increased. (The other probabilities have decreased). Sentence `B` is still more likely, however.

iii. An unsmoothed bi-gram language model?

- This time, we're interested in the counts of pairs of word tokens. For example, the probability of the bi-gram `wood would` is based on the count of that sequence of tokens, divided by the count of `wood`: $\frac{1}{8}$ (because only a single `wood` is followed by `would`).

- We are also going to include sentence terminals, so that the first probability in sentence A is $P(\texttt{a})|\texttt{<s>}) = \frac{1}{2}$ — because one of the two sentences in the corpus starts with a. We also need to predict $P(\texttt{</s>}|\texttt{chuck}) = \frac{0}{9}$ — because none of the 9 chucks are followed by the end of the sentence.
- Now, we can substitute:

$$\begin{aligned}P(A) &= P(\texttt{a}|\texttt{<s>})P(\texttt{wood}|\texttt{a})P(\texttt{could}|\texttt{wood})P(\texttt{chuck}|\texttt{could})P(\texttt{</s>}|\texttt{chuck})\\ &= \frac{1}{2}\times\frac{4}{4}\times\frac{0}{8}\times\frac{1}{1}\times\frac{0}{9} = 0\\ P(B) &= P(\texttt{wood}|\texttt{<s>})P(\texttt{would}|\texttt{wood})P(\texttt{a}|\texttt{would})P(\texttt{chuck}|\texttt{a})P(\texttt{</s>}|\texttt{chuck})\\ &= \frac{0}{2}\times\frac{1}{8}\times\frac{1}{4}\times\frac{0}{4}\times\frac{0}{9} = 0\end{aligned}$$

- Because there is a zero–probability element in both of these calculations, they can't be nicely compared, leading us to instead consider:

iv. A bi-gram language model, with Laplacian smoothing?

- We do the same idea as uni-gram add–one smoothing, but now the vocabulary size increases by one (because we're also predicting </s>; we need to do this to ensure that the probabilities of all the events that can following a given token still sum to 1).

$$\begin{aligned}P_{\mathrm{L}}(A) &= P_{\mathrm{L}}(\texttt{a}|\texttt{<s>})P_{\mathrm{L}}(\texttt{wood}|\texttt{a})P_{\mathrm{L}}(\texttt{could}|\texttt{wood})P_{\mathrm{L}}(\texttt{chuck}|\texttt{could})P_{\mathrm{L}}(\texttt{</s>}|\texttt{chuc}\\ &= \frac{2}{13}\times\frac{5}{15}\times\frac{1}{19}\times\frac{2}{12}\times\frac{1}{20} \approx 2.25\times10^{-5}\\ P_{\mathrm{L}}(B) &= P_{\mathrm{L}}(\texttt{wood}|\texttt{<s>})P_{\mathrm{L}}(\texttt{would}|\texttt{wood})P_{\mathrm{L}}(\texttt{a}|\texttt{would})P_{\mathrm{L}}(\texttt{chuck}|\texttt{a})P_{\mathrm{L}}(\texttt{</s>}|\texttt{chuc}\\ &= \frac{1}{13}\times\frac{2}{19}\times\frac{2}{15}\times\frac{1}{15}\times\frac{1}{20} \approx 3.60\times10^{-6}\end{aligned}$$

- This time, sentence A has the greater likelihood, mostly because of the common bi-gram a wood.

v. An unsmoothed tri-gram language model?

- Same idea, longer contexts. Note that we now need two sentence terminals.

$$\begin{aligned}P(A) &= P(\texttt{a}|\texttt{<s2> <s1>})P(\texttt{wood}|\texttt{<s1> a})\cdots P(\texttt{</s2>}|\texttt{chuck </s1>})\\ &= \frac{1}{2}\times\frac{1}{1}\times\frac{0}{4}\times\frac{0}{0}\times\frac{0}{1}\times\frac{0}{0} = ?\\ P(B) &= P(\texttt{wood}|\texttt{<s2> <s1>})P(\texttt{would}|\texttt{<s1> wood})\cdots P(\texttt{</s2>}|\texttt{chuck </s1}\\ &= \frac{0}{2}\times\frac{0}{0}\times\frac{1}{1}\times\frac{0}{1}\times\frac{0}{0}\times\frac{0}{0} = ?\end{aligned}$$

- Given that the unsmoothed bi-gram probabilities were zero, that also means that the unsmoothed tri-gram probabilities will be zero. (Exercise for the reader: why?)
- In this case, they aren't even well–defined, because of the $\frac{0}{0}$ terms, but we wouldn't be able to meaningfully compare these numbers in any case.

vi. A tri-gram language model, with Laplacian smoothing?

- The vocabulary size is now 12 (due to the two sentence terminals); everything else proceeds the same way:

$$\begin{aligned}
P_{\text{L}}(A) &= P_{\text{L}}(\texttt{a|<s2> <s1>})P_{\text{L}}(\texttt{wood|<s1> a})\cdots P_{\text{L}}(\texttt{</s2>|chuck </s1>}) \\
&= \frac{2}{14}\times\frac{2}{13}\times\frac{1}{16}\times\frac{1}{12}\times\frac{1}{13}\times\frac{1}{12}\approx 7.34\times 10^{-7} \\
P_{\text{L}}(B) &= P_{\text{L}}(\texttt{wood|<s2> <s1>})P_{\text{L}}(\texttt{would|<s1> wood})\cdots P_{\text{L}}(\texttt{</s2>|chuck </s} \\
&= \frac{1}{14}\times\frac{1}{12}\times\frac{2}{13}\times\frac{1}{13}\times\frac{1}{12}\times\frac{1}{12}= 4.89\times 10^{-7}
\end{aligned}$$

- Notice that the problem of unseen contexts is now solved (they are just $\frac{1}{12}$).
- Sentence A has a slightly greater likelihood here, mostly because of the a at the start of one of the sentences (note that this will continue to be "seen" even at higher orders of n). You can also see that the numbers are getting very small, which is a good motivation for summing log probabilities (assuming no zeroes) rather than multiplying.

4. What does **back–off** mean, in the context of smoothing a language model? What does **interpolation** refer to?

   - Back–off is a different smoothing strategy, where we incorporate lower–order n-gram models (in particular, for unseen contexts). For example, if we have never seen some tri-gram from our sentence, we can instead consider the bi-gram probability (at some penalty, to maintain the probability of all of the events, given some context, summing to 1). If we haven't seen the bi-gram, we consider the uni-gram probability. If we've never seen the uni-gram (this token doesn't appear in the corpus at all), then we need a so–called "0-gram" probability, which is a default for unseen tokens.
   - Interpolation is a similar idea, but instead of only "falling back" to lower–order n-gram models for unseen events, we can instead consider every probability as a linear combination of all of the relevant n-gram models, where the weights are once more chosen to ensure that the probabilities of all events, given some context, sum to 1.