IN1140 – Uke 4

Sannsynlighetsregning

n-gram

Løsningsforslag 1a

Oblighjelp

Hvordan lese J&M?

absolute discounting

0.75 from the count in the training set! **Absolute discounting** formalizes this intuition by subtracting a fixed (absolute) discount d from each count. The intuition is that since we have good estimates already for the very high counts, a small discount d won't affect them much. It will mainly modify the smaller counts, for which we don't necessarily trust the estimate anyway, and Fig. 3.9 suggests that in practice this discount is actually a good one for bigrams with counts 2 through 9. The equation for interpolated absolute discounting applied to bigrams:

$$P_{\text{AbsoluteDiscounting}}(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i) - d}{\sum_v C(w_{i-1}\,v)} + \lambda(w_{i-1})P(w_i) \tag{3.32}$$

The first term is the discounted bigram, with $0 \leq d \leq 1$, and the second term is the unigram with an interpolation weight λ. By inspection of Fig. 3.9, it looks like just setting all the d values to .75 would work very well, or perhaps keeping a separate second discount value of 0.5 for the bigrams with counts of 1. There are principled methods for setting d; for example, Ney et al. (1994) set d as a function of n_1 and n_2, the number of unigrams that have a count of 1 and a count of 2, respectively:

$$d = \frac{n_1}{n_1 + 2n_2} \tag{3.33}$$

3.7.2 Kneser-Ney Discounting

Kneser-Ney discounting (Kneser and Ney, 1995) augments absolute discounting

To viktige notasjoner

Vi bruker Σ og Π til rekker av henholdsvis summeringer og multiplikasjoner

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15$$

$$\sum_{i=1}^{5} i \cdot 2 = 2 + 4 + 6 + 8 + 10 = 30$$

$$\sum_{i=1}^{5} \frac{i}{2} = \frac{1}{2} + \frac{2}{2} + \frac{3}{2} + \frac{4}{2} + \frac{5}{2} = \frac{15}{2} = 7{,}5$$

$$\prod_{i=1}^{5} i = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$$

$$\prod_{i=1}^{5} \frac{i}{2} = \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{3}{2} \cdot \frac{4}{2} \cdot \frac{5}{2} = \frac{120}{32} = 3{,}75$$

Typiske tegn

- $P(x)$ står for sannsynligheten av x
    - $P(y \;\; x)$ står for sannsynligheten av y gitt x
    - $P(regn \;\; lavtrykk)$ står for sannsynligheten av *regn* gitt *lavtrykk*

- w_n står for ord n i setninga
    - w_1 står for ord 1
    - w_2 står for ord 2

- Mye bra på foilene fra forelesning 5!

n-gram-modeller

- En statistisk modell
  - Laget basert på statistikk
- Kan forutsi sannsynligheten for et ord gitt et foregående ord (eller flere)
  - n står for antall ord (1 = unigram, 2 = bigram, 3 = trigram)
- Brukes typisk i talegjenkjenning, maskinoversettelse, tekstgenerering, ++

<s> Det var en gang en hund </s>

2-gram:
- <<s>, Det>
- <Det, var>
- <var, en>
- <en, gang>
- <gang, en>
- <en, hund>
- <hund, </s>>

3-gram:
- <<s>, Det, var>
- <Det, var, en>
- <var, en, gang>
- <en, gang, en>
- <gang, en, hund>
- <en, hund, </s>>

4-gram:
- <<s>, Det, var, en>
- <Det, var, en, gang>
- <var, en, gang, en>
- <en, gang, en, hund>
- <gang, en, hund, </s>>

Betinget sannsynlighet

Sannsynligheten for ord w_n gitt ordet w_{n-1}:

$$P(w_n \mid w_{n-1}) = \frac{count(w_{n-1}\ w_n)}{count(w_{n-1})}$$ (count forkortes til C i J&M)

Hvis $count(w_n) = 0$ bruker vi smoothing

Et eksempel

- Vi har korpuset:
    1. *<s> Jeg liker ost </s>*
    2. *<s> Du liker musikk </s>*
    3. *<s> Jeg jobber sent </s>*

Når kan vi beregne sannsynligheten for setninga *<s> Jeg liker musikk </s>:*

$$P(Jeg \mid <s>) = \frac{2}{3} \qquad P(musikk \mid liker) = \frac{1}{2}$$

$$P(liker \mid Jeg) = \frac{1}{2} \qquad P(</s> \mid musikk) = \frac{1}{1}$$

Vi summerer: $\frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{1} = \frac{2}{12} = \frac{1}{6} \approx 0{,}167$

1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
	–Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
	–What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done.
	–This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in;
	–It cannot be but so.

Figure 3.4 Eight sentences randomly generated from four n-grams computed from Shakespeare’s works. All characters were mapped to lower-case and punctuation marks were treated as words. Output is hand-corrected for capitalization to improve readability.

Neste uke

- Ingen gruppetime!
- Neste uke der igjen:
    - Ordklasser og ordklassetagging
    - Oblighjelp 2a