IN1140: Introduksjon til språkteknologi

Forelesning #5: Språkmodeller

Lilja Øvrelid

Universitetet i Oslo

19 september 2019

Agenda

Forrige uke

- Regulære uttrykk
- Oblig 1b ute nå (frist 2/10)

I dag

- Grunnleggende sannsynlighetsregning
- Statistiske språkmodeller
- n-gram modeller

Mål for i dag

- Skal se på noen veldig enkle men veldig nyttige modeller.
- n-gram modeller.
- En metode som, gitt eksempler på språkbruk i form av et korpus,
- gir oss en statistisk språkmodell (*language model*),
- som lar oss beregne sannsynligheten for en gitt setning
- eller predikere hva som er neste ord i en sekvens.
- En probabilistisk model; vi trenger først å vite litt om grunnleggende sannsynlighetsregning.

Bare en forsmak

- Trenger litt formell notasjon.
- Men bare akkurat nok til å gjøre det vi vil.
- Jobbe med statistiske språkmodeller (i dag)
- og ordklassetaggere (neste uke).
- Dere kommer til å få grundigere innføring i sannsynlighetsregning og statistiske metoder generelt i senere språkteknologikurs.

Sannsynlighet

- Lar oss kvantifisere usikkerhet.
- Uttrykk for sjanse eller odds.
- $P(\text{sol i morgen})$
- $P(\text{seks på terningen})$
- Betinget sannsynlighet:
  $P(\text{sol i morgen} \mid \text{regn i dag})$

Sannsynlighet

- Klassisk eksempel: terningkast.
  - Hva er sannsynligheten for å få 1?
  - Hva er sannsynligheten for å få 6?
  - Hva er sannsynligheten for å få 1, 2 eller 3?
  - Hva er sannsynligheten for å få 3, gitt av vi vet at resultatet ble noe mellom 1–3?

Litt terminologi og notasjon

Litt mengdelære

- En mengde består av elementer
  - $T = \{1, 2, 3, 4, 5, 6\}$
- Angi om et element tilhører en mengde eller ikke
  - $2 \in T$, $8 \notin T$
- En mengde A er delmengde av B: alle elementer i A fins i B
  - $A = \{1, 3\}$, $B = \{1, 3, 5, 7, 9\}$
  - $A \subseteq B$
- Unionen av to mengder: ny mengde med alle elementer som forekommer i én eller begge mengdene
  - $O = \{1, 3, 5\}$, $P = \{2, 4, 6\}$
  - $O \cup P = \{1, 2, 3, 4, 5, 6\}$
- Snittet av to mengder: ny mengde med alle elementer som forekommer i begge mengdene
  - $O = \{1, 3, 5\}$, $O_2 = \{5, 7, 9\}$
  - $O \cap O_2 = \{5\}$

Litt terminologi og notasjon

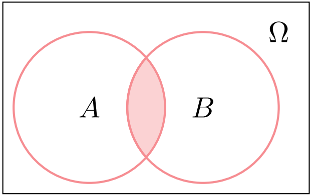
- Sum og produkt
  - $\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n$
  - $\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot \ldots \cdot a_n$
- Eksempler
  - $\sum_{i=1}^{7} 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$
  - $\prod_{i=1}^{7} 1 \cdot 2 \cdot 3 \cdot \ldots \cdot 7 = 5040$

Litt terminologi og notasjon

- Utfallsrommet (*sample space*) er mengden Ω av mulige utfall.
  - For eksemplet med terningkast: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Hendelse (*event*): en delmengde av utfallsmengden, $A \subseteq \Omega$. F.eks:
  - $A = \{1, 2, 3\}$
  - $B = \{5\}$
- Sannsynlighet for en hendelse: en verdi mellom 0 og 1, gitt ved P:
  - $P(A) = 0.5$
  - $P(B) = 1/6$
- Dersom alle utfall er like sannsynlige har vi en uniform distribusjon:
  - $P(A) = \frac{|A|}{|\Omega|}$
- Sannsynlighetene for alle mulige utfall summerer til 1:
  - $\sum_{A \in \Omega} P(A) = 1$

Felles sannsynlighet (*joint probability*)

- $P(A, B)$: sannsynligheten for at både A *og* B skjer.
- Skrives også: $P(A \cap B)$.
- Vi kaster to terninger.
- $|\Omega| = 6 \times 6 = 36$
- $\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,4), (6,5), (6,6)\}$

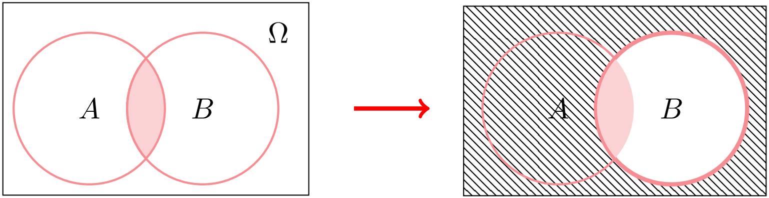


Noen hendelser:

- Summen er 6: $P(A) = \frac{5}{36}$
- Minst en terning viser 1: $P(B) = \frac{11}{36}$
- Summen er 6 *og* minst en terning viser 1: $P(A \cap B) = \frac{2}{36}$

Betinget sannsynlighet (*conditional probability*)

- Vi har ofte *delvis kunnskap* om en hendelse.
- Når vi kaster to terninger, hva er sjansen $P(A|B)$ for at
  - A, summen er 6, *gitt at*
  - B, minst en terning viser 1?



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- F.eks: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{11/36} = \frac{2}{11}$

Produktsetningen (*chain rule*)

- Nyttig omskrivning: $P(A, B) = P(A)\, P(B|A)$
- Symmetrisk: $P(A, B) = P(B)\, P(A|B)$
- Kalles produktsetningen (*chain rule*).
- Mer generelt:

$$P(A_1, \dots, A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\dots P(A_k|A_1^{k-1})$$

- Gir oss ulike måter å bryte opp et komplisert problem til enklere deler.

Uavhengighet

- La A være hendelsen å få 2 på første terningkast, $P(A) = \frac{1}{6}$.
- La B være hendelsen å få 4 på andre terningkast, $P(B) = \frac{1}{6}$.
- $P(B|A) = \frac{1}{6}$.
- Hvis kjennskap til A ikke har noen effekt på sjansen for B, så sier vi at A og B er uavhengige hendelser.

Hvis A og B er *uavhengige*:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A, B) = P(A)P(B|A) = P(A)\,P(B)$

Trenger du mer oppfriskning?

- Bla deg gjerne gjennom sidene her:

`https://www.matematikk.org/trinn8-10/side.html?tid=68810`

- En kortfattet og lettfattelig innføring i sannsynlighetsregning.

Men nå:

- Først litt bakgrunn om statistiske metoder i NLP generelt.
- Så skal vi definere en statistisk språkmodell som beregner sannsynligheten for setninger.

Statistikk og språk

- Statistiske modeller har fått en stadig viktigere rolle i NLP.
- Ofte kalt empiriske eller datadrevne metoder.
- Typisk basert på maskinlæring.
- Settes gjerne i kontrast med regelbaserte eller 'manuelle' metoder.
- Forskjellen (noe overforenklet):

> Håndkoding av kunnskap av en menneskelig ekspert.
> vs.
> Automatisk tilegning av kunnskap fra data av en algoritme.

- 'Kunnskap' om språk i form av statistiske mønstre i data.
- Fordeler i form av skalerbarhet og robusthet.

Forutsigbarhet og føringer

Fyll inn den blanke

- *Det var en ____*
- *Det var en gang*
- *Det var en bil*
- *Det var en hus* *
- *Det var en spiser* **

- Ofte stor grad av forutsigbarhet i språket.
- Tidligere kontekst kan legge føringer på det neste ordet på flere måter:
  - semantisk,
  - syntaktisk,
  - og konvensjonelt.

Meme alert

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Fornuft $\approx$ frekvenser?

- De ulike føringene og konvensjonene manifesterer seg i observerbare frekvenser i språkbruk.
- Selv om vår 'mentale språkmodell' inkorporerer en stor porsjon sunn fornuft, så kan vi komme langt ved å telle ting.
- En statistisk språkmodell bruker korpusfrekvenser for beregne sannsynligheten for sekvenser av ord.
- Akustisk signal: '*jæispiserjærne*'

$$P(\text{jeg spiser gjerne})$$

vs.

$$P(\text{jeg spiser hjerne})$$

Anvendelser av språkmodeller

- Talegjenkjenning
  - P(*recognize speech*) > P(*wreck a nice beach*)
- Maskinoversettelse
  - Mer sannsynlige setninger er ofte bedre oversettelser.
  - P(*She walked home*) > P(*She walked house*)
- Håndskriftgjenkjenning og OCR
- Prediksjon / automatisk fullføring:
  - *Det er ikke lov å nyte medbrakt ???*
- Ordklassetagging, stavekontroll, tekstklassifikasjon, generering, og masse annet...

Språkmodeller

- En probabilistisk språkmodell tilskriver sannsynligheter $P(x)$ til alle ordsekvenser x i et språk L.
- $\Omega = L$ er utfallsrommet
- Gitt en sekvens $w_1, w_2 \ldots w_k = w_1^k$ så ønsker vi å estimere den felles sannsynligheten for ordene $P(w_1^k)$.
- F.eks: $P(\text{jeg}, \text{vil}, \text{drikke}, \text{kaffe}, \text{nå})$
- Skal se på noen ulike muligheter…

Sannsynligheten for en setning (1. forsøk)

- Én mulighet er å la $P(w_1^k) = P(w_1)P(w_2)\dots P(w_k)$,
- $P(\text{jeg}, \text{vil}, \text{drikke}, \text{kaffe}, \text{nå}) \;=\; P(\text{jeg})\,P(\text{vil})\,P(\text{drikke})\,P(\text{kaffe})\,P(\text{nå})$
- Ord er ikke uniformt distribuert. Kan bruke statistisk sannsynlighet basert på relativ frekvens i et korpus:

- $P(w_i) = \frac{Count(w_i)}{Count(*)}$
- Antar at forekomster av ord uavhengige. Stemmer jo ikke!

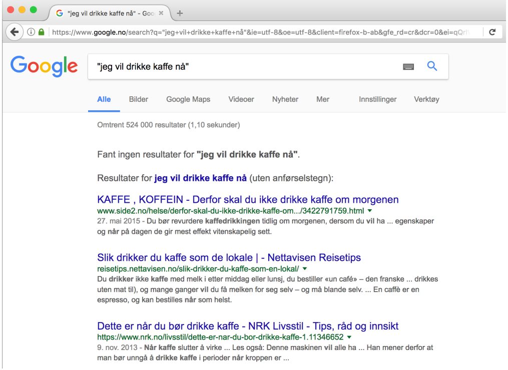
Bør kunne forvente at:

$$P(\text{jeg, vil, drikke, kaffe, nå}) \quad > \quad P(\text{kaffe, vil, nå, drikke, jeg})$$

$$P(\text{kaffe} \,|\, \text{drikke}) \quad > \quad P(\text{kaffe})$$

Sannsynligheten for en setning (2. forsøk)

- Én annen mulighet, estimere felles sannynlighet direkte:
- $P(w_1^k) = \frac{C(w_1, w_2, \ldots, w_k)}{C(*_1, *_2, \ldots, *_k)}$
- Ikke mulig å få et pålitelig estimat.



Sannsynligheten for en setning (3. forsøk)

- Vi kan bruke produktsetningen:

  $$P(w_1 \ldots w_k) = P(w_1)\, P(w_2|w_1)\, P(w_3|w_1, w_2) \ldots P(w_k|w_1^{k-1})$$

- Eksempel:

  $P(\text{jeg, vil, drikke, kaffe, nå}) =$

  $P(\text{jeg}) \times$

  $P(\text{vil} \,|\, \text{jeg}) \times$

  $P(\text{drikke} \,|\, \text{jeg, vil}) \times$

  $P(\text{kaffe} \,|\, \text{jeg, vil, drikke}) \times$

  $P(\text{nå} \,|\, \text{jeg, vil, drikke, kaffe})$

- Har ikke kommet så veldig mye lengre. . . ?

Sannsynligheten for en setning (4. forsøk)

- Vi kan forenkle med Markovantagelsen:
- De $n-1$ siste elementene lar oss tilnærme effekten av å betrakte hele sekvensen.
- Begrenset historikk.
- Eksempel for $n=2$:
  $P\left(w_1^k\right) \approx \prod_{i=1}^{k} P\left(w_i|w_{i-1}\right)$
- For eksempelsetningen vår:
  $P(\text{jeg, vil, drikke, kaffe, nå}) =$
  $P(\text{jeg})\ P(\text{vil}\,|\,\text{jeg})\ P(\text{drikke}\,|\,\text{vil})\ P(\text{kaffe}\,|\,\text{drikke})\ P(\text{nå}\,|\,\text{kaffe})$
- En n-grammodell.

n-grammer

- Et n-gram er en delsekvens på n ord:
- n-grammer i *jeg vil drikke kaffe nå*:

- unigrammer ($n = 1$): ⟨*jeg*⟩, ⟨*vil*⟩, ⟨*drikke*⟩, ⟨*kaffe*⟩, ⟨*nå*⟩
- bigrammer ($n = 2$): ⟨*jeg, vil*⟩, ⟨*vil, drikke*⟩, ⟨*drikke, kaffe*⟩ ⟨*kaffe, nå*⟩
- trigrammer ($n = 3$): ⟨*jeg, vil, drikke*⟩, ⟨*vil, drikke, kaffe*⟩
- 4-grammer ($n = 4$): ⟨*jeg, vil, drikke, kaffe*⟩, ⟨*vil, drikke, kaffe, nå*⟩

Å trene en n-grammodell: MLE

- Vi estimerer P ved å telle n-grammer og se på relativ frekvens:

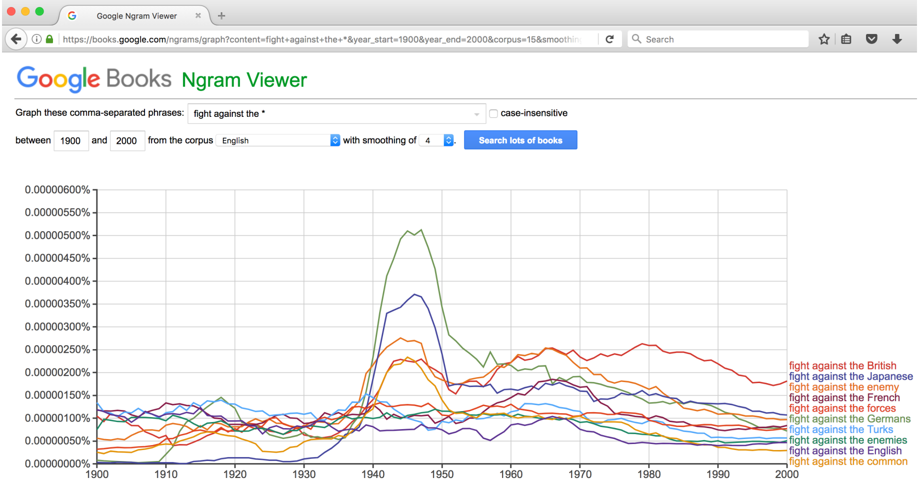
$$P\left(w_i | w_{i-n+1}, \ldots, w_{i-1}\right) = \frac{Count\left(w_{i-n+1}, \ldots, w_i\right)}{Count(w_{i-n+1}, \ldots, w_{i-1})}$$

- F.eks, for bigrammer:

$$P\left(\text{kaffe}|\text{drikke}\right) = \frac{Count\left(\text{drikke, kaffe}\right)}{Count\left(\text{drikke}\right)}$$

- Kalles Maximum Likelihood Estimation (MLE).
- I praksis legger vi gjerne også til markører for start og slutt for sekvensen: <s> og </s>.

Google Ngrams



Basert på Google Books Ngram Corpus:

`http://books.google.com/ngrams`

Eksempel på trening av en bigramsmodell (J&M)

Minikorpus

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

MLE bigramsannsynligheter

$P(\text{I} \,|\, \text{<s>}) = 2/3 = .67$

$P(\text{Sam} \,|\, \text{<s>}) = 1/3 = .33$

$P(\text{am} \,|\, \text{I}) = 2/3 = .67$

$P(\text{</s>} \,|\, \text{Sam}) = 1/2 = 0.5$

$P(\text{Sam} \,|\, \text{am}) = 1/2 = 0.5$

$P(\text{do} \,|\, \text{I}) = 1/3 = 0.33$

Sannsynligheten for en setning:

$P(\text{<s> I am Sam </s>})$

$= P(\text{I} \,|\, \text{<s>})\, P(\text{am} \,|\, \text{I})\, P(\text{Sam} \,|\, \text{am})\, P(\text{</s>} \,|\, \text{Sam})$

$= .67 \times .67 \times .5 \times .5$

$= 0.112225$

Problemer med MLE

- $P(\text{<s> I am Pete </s>}) = 0$
- Ord og n-grammer med frekvens 0 gir oss problemer.
- Ser ut til at vi trenger et 5. forsøk. . .

'Data sparseness'

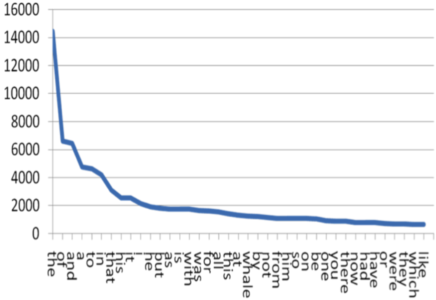
- Uavhengig av korpusets størrelse, vil det alltid finnes både ord og sekvenser vi ikke har sett.
- 'Språkets produktivitet' / kreativitet.
- Vanskelig å få pålitelige estimater selv for enkeltord; jf. Zipfs lov.

En anekdote: Zipfs lov og hapax legomena

Zipfs lov

- Rangér ordene i et stort korpus etter frekvens.
- #1 brukes dobbelt så ofte som #2, tre ganger så ofte som #3, osv.

- Noen få ord forekommer ofte; de fleste forekommer sjelden.
- Typisk er halvparten av ordene hapax legomena: ord som forekommer kun én gang.
- Distribusjon med 'lang hale'.



- Anbefales: `https://www.youtube.com/watch?v=fCn8zs912OE`

Glatting (*smoothing*)

- Omfordeling av sannsynlighetsmassen for å unngå noen av problemene med MLE: Sørg for at alle n-grammer får frekvens > 0.

- Ta fra de rike og gi til de fattige.
- Enkleste metoden; Add-one smoothing (Laplace):

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}$$

der V er antall ordtyper (vokabulæret).

Frampeik

- Neste uke:
  - Ordklasser
  - Ordklassetagging
- Oblig 2a: språkmodeller