Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Modellering av sekvenser

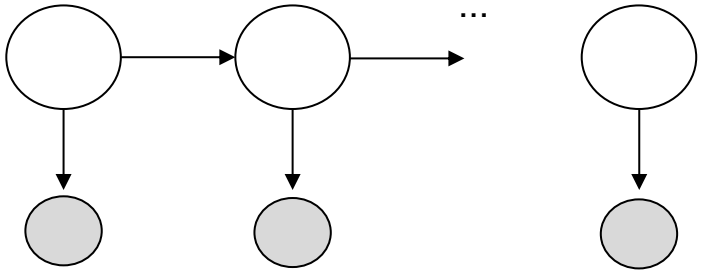
Pierre Lison
plison@nr.no

IN2110: Språkteknologiske metoder

16. mars 2022

Dagens temaer:

Midtveisevaluering



Modellering av sekvenser (del 1): Markovkjeder og Hidden Markov Models

Midtveisevaluering

- 19 svar – **mange takk for tilbakemeldinger**!
- Hva dere likte så langt:
  - Spennende innhold
  - Forelesninger og gruppetimer ser ut til å fungere bra
  - Konkrete oblig med programmeringsoppgaver
- Hva er utfordrende:
  - Matematiske formler
  - Teoretisk innhold (spesielt i boken fra J&M)
  - Noe repetisjon

Midtveisevaluering

Pleier du å delta på forelesningene? *

Svar	Antall	Prosent
Alltid (eller nesten alltid)	15	78,9 %
Nå og da	3	15,8 %
Aldri (eller nesten aldri)	1	5,3 %

Pleier du å delta på gruppetimene? *

Svar	Antall	Prosent
Alltid (eller nesten alltid)	14	73,7 %
Nå og da	3	15,8 %
Aldri (eller nesten aldri)	2	10,5 %

Midtveisevaluering

Hva synes du om vanskelighetsgraden på forelesningene? *

Svar	Antall	Prosent
Veldig vanskelig å henge med	2	**10,5 %**
Litt vanskelig	7	**36,8 %**
Passe	7	**36,8 %**
Litt lett	3	**15,8 %**
Altfor lett	0	**0 %**
Vet ikke	0	**0 %**

Midtveisevaluering

Hva synes du om vanskelighetsgraden på innlevering 1a? *

Svar	Antall	Prosent
Veldig vanskelig	1	5,3 %
Litt vanskelig	5	26,3 %
Passe	10	52,6 %
Litt lett	3	15,8 %
Altfor lett	0	0 %
Vet ikke	0	0 %

Midtveisevaluering

Hva synes du om vanskelighetsgraden på oblig 1b (så langt)? *

Svar	Antall	Prosent
Veldig vanskelig	0	**0 %**
Litt vanskelig	6	**31,6 %**
Passe	7	**36,8 %**
Litt lett	0	**0 %**
Altfor lett	0	**0 %**
Vet ikke	6	**31,6 %**

Midtveisevaluering

Hva synes du om vanskelighetsgraden på lærebøkene? *

Svar	Antall	Prosent
Veldig vanskelig å henge med	5	**26,3 %**
Litt vanskelig	8	**42,1 %**
Passe	6	**31,6 %**
Litt lett	0	**0 %**
Altfor lett	0	**0 %**

Midtveisevaluering

Hva synes du om tempoet / progresjonen i kurset? *

Svar	Antall	Prosent
Det går altfor raskt frem	0	0 %
Litt raskt	7	36,8 %
Passe	9	47,4 %
Litt sakte	3	15,8 %
Altfor sakte	0	0 %

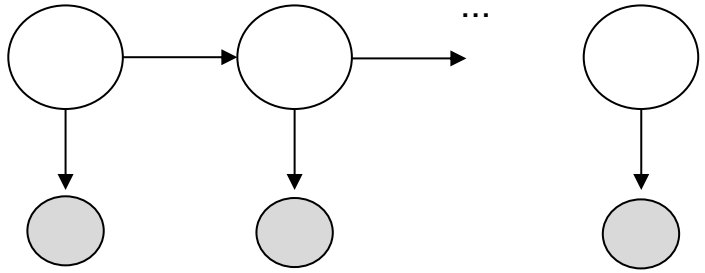
Midtveisevaluering

Hvor mange timer per uke (totalt) jobber du med kurset? *

Svar	Antall	Prosent
<5	0	**0 %**
<10	14	**73,7 %**
<15	5	**26,3 %**
>=15	0	**0 %**

I dag skal vi snakke om 2 temaer:

Midtveisevaluering

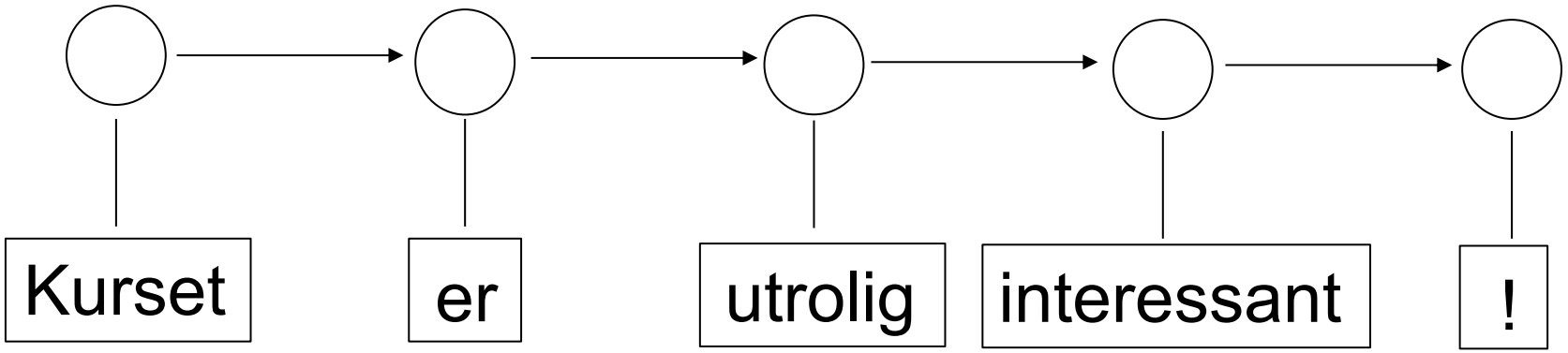


Modellering av sekvenser (del 1): Markovkjeder og Hidden Markov Models

Modellering av sekvenser

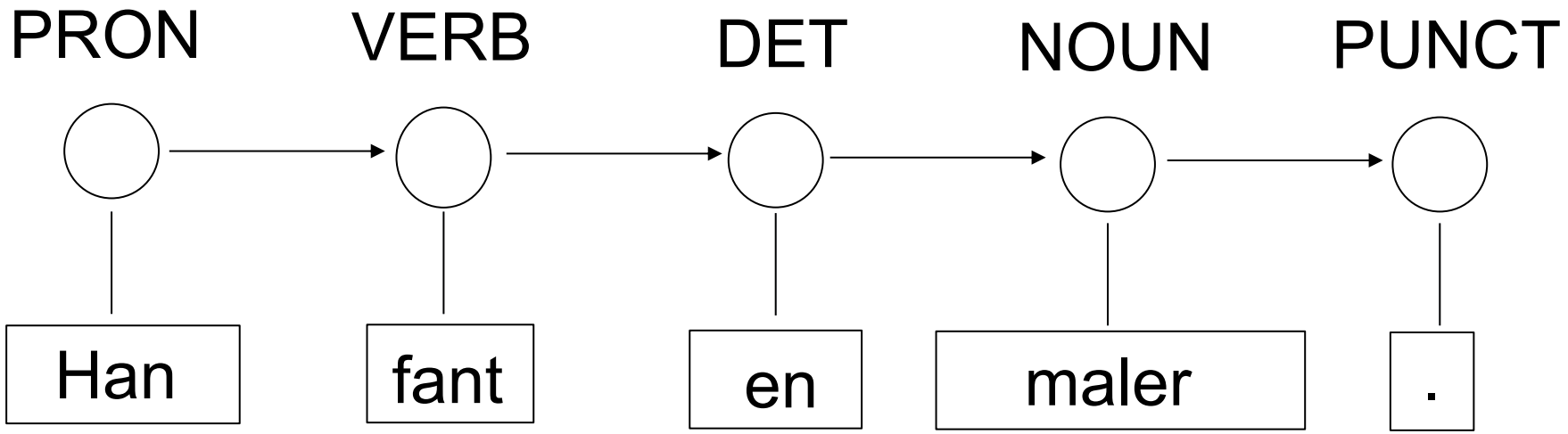
En viktig oppgave i språkteknologi er **sekvensmerking**, hvor vi assosierer hvert element i en sekvens til en *merkelapp*

- Merking må ta hensyn til "konteksten" (dvs. naboene til elementet i seksensen)
- Hva er den mest *sansynnlige sekvensen av merkelapper?*



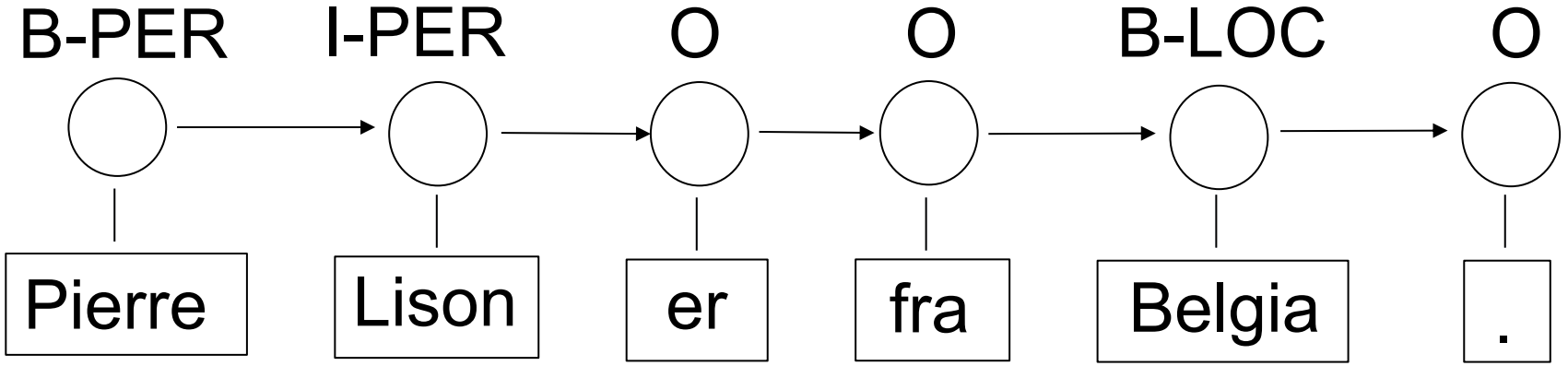
Eksempel: Tagging av ordklasser

- Mål: finne ut ordklassen til hvert ord
- Ordklassen til hvert ord er avhengig av konteksten!
  - "maler": substantiv eller verb, men hvis den forekommer etter en determinativ er det trolig et substantiv



Eksempel: Gjenkjenning av navngitte entiteter (*named entity recognition*)

- Mål: ekstrahere enheter som personer, steder, organisasjoner, beløp, datoer, produkter, osv.
- **BIO**-system: alle ordene er klassifisert som **B**(eginning)-*X*, **I**(nside)-*X* eller **O**(out), hvor *X* er en merkelapp som PER(son), LOC(ation), ORG(anisation), osv.



Spørsmål

Bestak Equinor Petrobras sin tidligere direktør,

ORG ORG

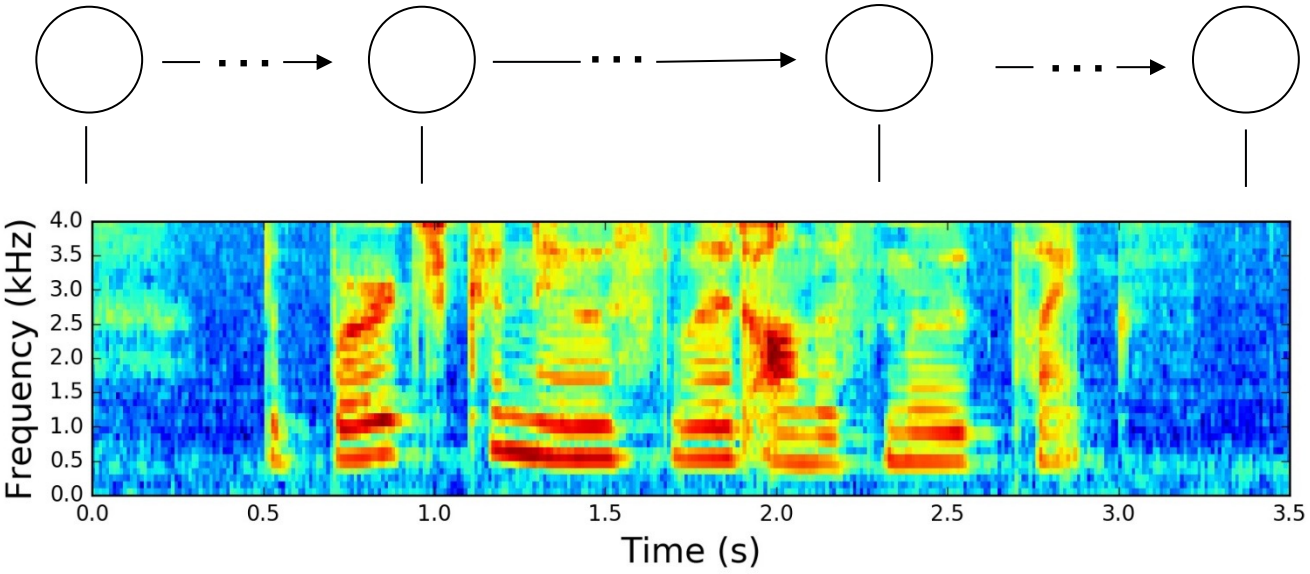
Paulo Roberto Costa i Rio de Janeiro i 2004?

PERSON GPE DATE

- **Spørsmål 1**: hva er sekvensen av BIO-labeller for denne markeringen av navngitte enheter?
- **Spørsmål 2**: hvorfor trenger man å ha BIO-prefikser, egentlig? Hva hadde skjedd om vi skulle bare markere ordene med merkelapp som ORG, PERSON, osv.?

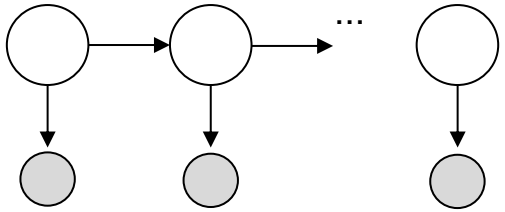
Eksempel: talegjenkjenning

- Mål: gjennkjenne de fonetiske (under)enhetene gitt akustiske observasjoner fra talesignalen
- Lange kjeder (samplingfreksens ofte 8 eller 16 kHz)
- Vi skal snakke (litt) om taleteknologi ved slutten av kurset



Hovedtema for i dag:

Sekvensmodeller, og mer spesielt *Hidden Markov Models*



Modellering: hvordan behandler vi problemer som tar sekvensdata som input og output?

Dekoding: hvordan beregner vi den meste sansynnlige sekvensen av labeller (f.eks. POS tags) gitt en sekvens av observasjoner (f.eks. ord)?

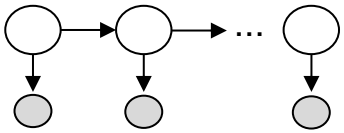
Læring: Hvordan estimerer vi parametrene i HMMs fra (markerte) data?

Sekvensmodeller

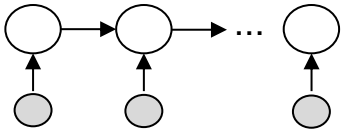
Fokus her på HMMs, som er enklere å forstå
(men som inkluderer mange konsepter som også er til stede i de mer "komplekse" modellene)

Mange ulike modeller for sekvensklassifisering:

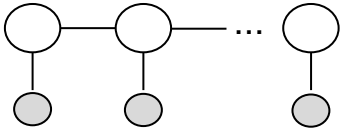
- Hidden Markov Models (HMMs)



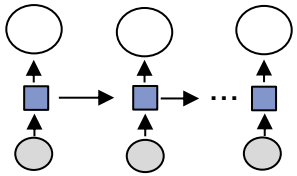
- Maximum Entropy Markov Models (MEMMs)



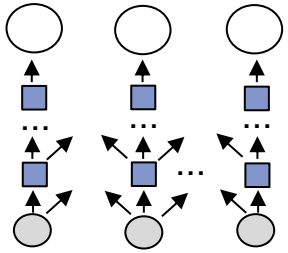
- Conditional Random Fields (CRFs)



- Recurrent Neural Networks (LSTMs, GRUs)

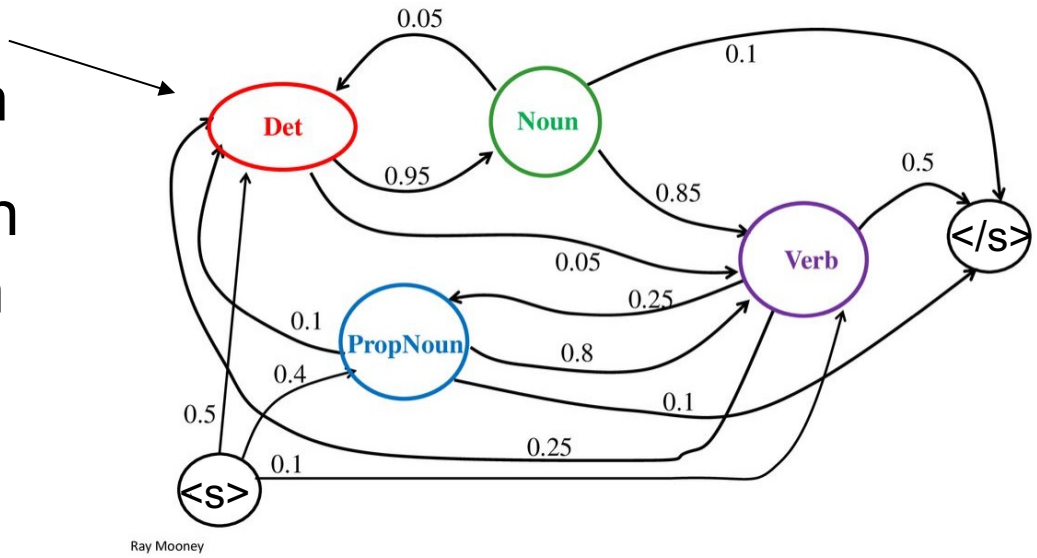


- Andre dype nevrale nettverk (e.g. Google's BERT)



Markovkjeder

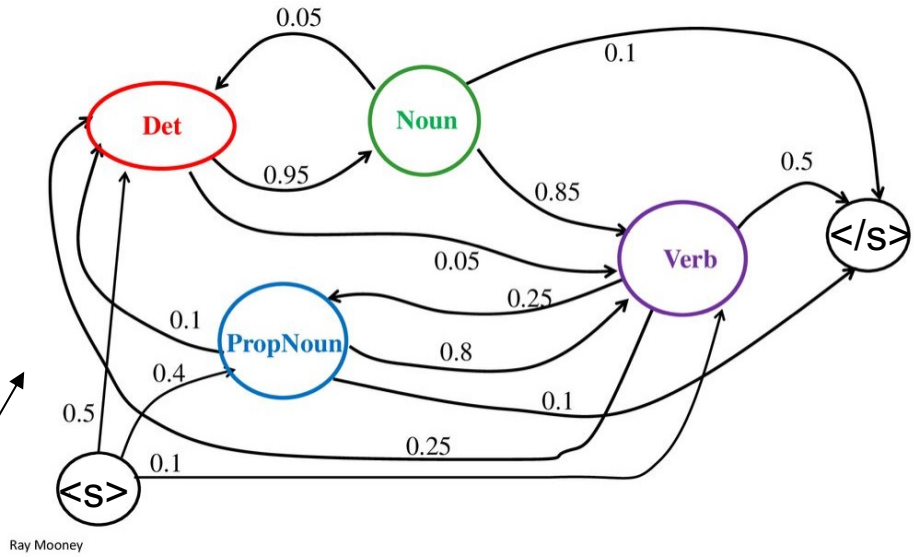
- La oss starte med Markovkjeder, som er en enkel statistisk modell over sekvenser av “tilstander” (=merkelapper)
- Vi definerer først et sett med mulige tilstander, som f.eks. {<s>, Det, Noun, Verb, PropNoun, </s>}
- Vi kan bevege oss fra en tilstand til en annen
- Vi må også indikere en “start” tilstand (eller en sannsynlighet over mulige starttilstander)



Ray Mooney

Transisjonsmodell

= sannsynligheter $P(s_{t+1}|s_t)$
for alle mulige s_t og s_{t+1}



Kan representeres som en **graf** (probabilistisk endelig tilstandsmaskin) …

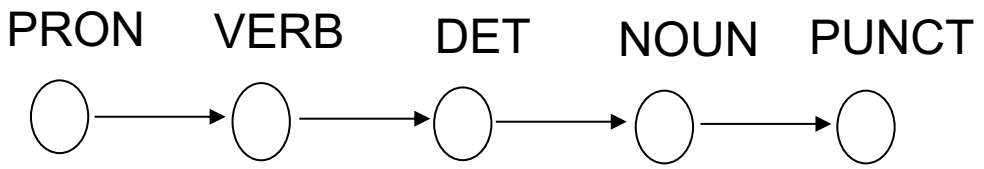
eller som en **matrise**

Fremtidig tilstand s_{t+1}

Nåværende tilstand s_t

	Det	Noun	PropNoun	Verb	</s>
<s>	0.5	0.0	0.4	0.1	0.0
Det	0.0	0.95	0.0	0.05	0.0
Noun	0.05	0.0	0.0	0.85	0.1
PropNoun	0.1	0.0	0.0	0.8	0.1
Verb	0.25	0.0	0.25	0.0	0.5

Markovkjeder



En Markovkjedemodell defineres av 3 komponenter:

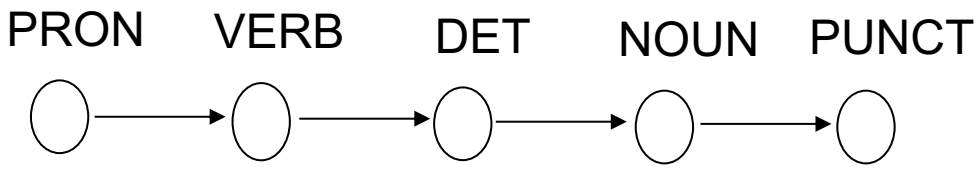
1. Et sett av mulige tilstander $\{s_1, s_2, \ldots, s_N\}$
2. En transisjonsmodell med sannsynlighetene $P(s_{t+1}|s_t)$ av å bevege oss fra s_t til s_{t+1} for alle tilstandene
3. En sannsynlighetsfordeling $P(s_0)$ over starttilstander

	Det	Noun	Prop Noun	Verb	</s>
<s>	.5	.0	.4	.1	.0
Det	.0	.95	.0	.05	.0
Noun	.05	.0	.0	.85	.1
Prop Noun	.1	.0	.0	.8	.1
Verb	.25	.0	.25	.0	.5

Markov-antagelsen: for å predikere s_{t+1} trenger vi kun å vite den nåværende tilstanden s_t

➔ *Eksempel:* Bigram språkmodeller

Markovskjeder



Legg merke på at tallene på hver rekke må være = 1

	Det	Noun	Prop Noun	Verb	</s>
<s>	.5	.0	.4	.1	.0
Det	.0	.95	.0	.05	.0
Noun	.05	.0	.0	.85	.1
Prop Noun	.1	.0	.0	.8	.1
Verb	.25	.0	.25	.0	.5

Vi bruker ofte som konvensjon at hver sekvens starter med <s> og slutter med </s>

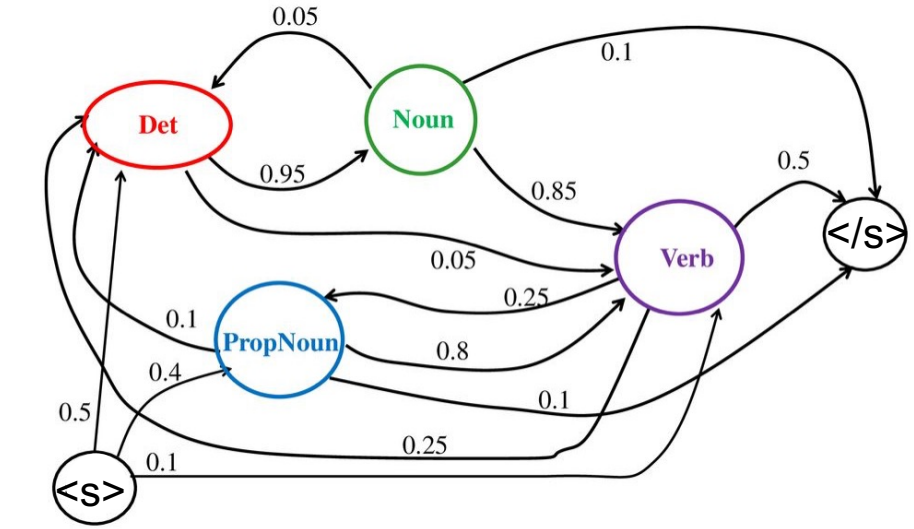
Men ikke kolonnene!

Hidden Markov Models

HMMs er Markovkjeder hvor sekvensen over tilstandene ikke observeres direkte (sekvensen er "skjult")

- Vi har derimot tilgang til en sekvens av **observasjoner** (som gir informasjon om de underliggende tilstandene)
- Kan vi *avlede* tilstandene ut fra observasjonene?

Vi observerer *ordene* i setning, men ikke *ordklassene*!

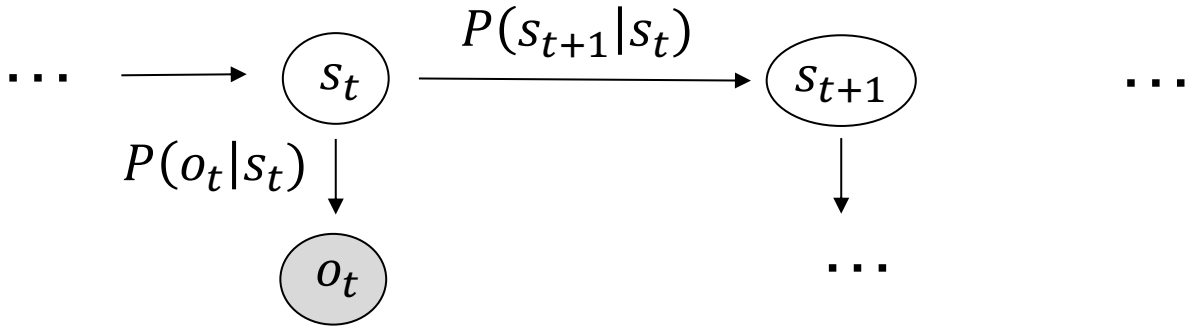


Ray Mooney

Hidden Markov Models

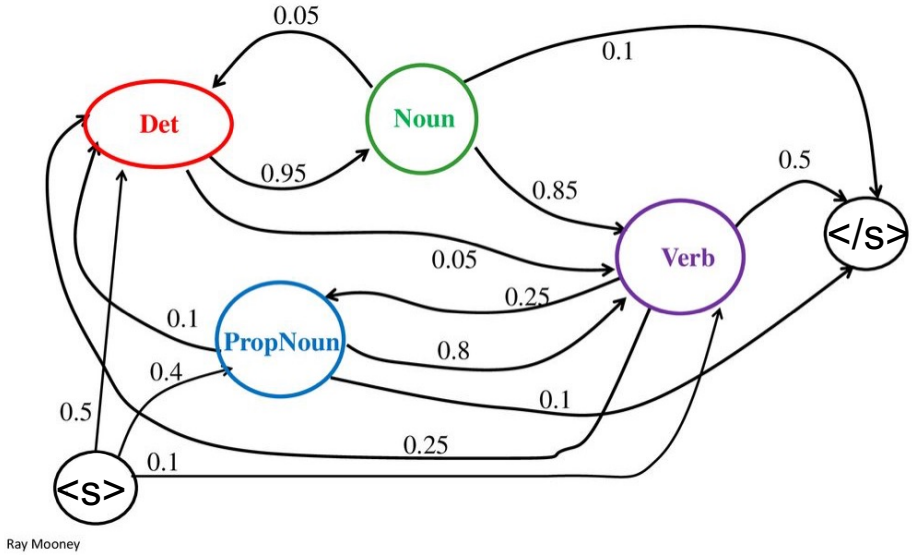
En HMM-modell defineres av 5 komponenter:

1. Et sett av mulige tilstander $\{s_1, s_2, \ldots, s_N\}$
2. Et sett av mulige observasjoner $\{o_1, o_2, \ldots, o_O\}$
3. En transisjonsmodell med sannsynlighetene $P(s_{t+1}|s_t)$ av å bevege oss fra s_t til s_{t+1} for alle tilstandene
4. En sannsynlighetsfordeling $P(s_0)$ over starttilstander
5. En emisjonsmodell med sannsynlighetene $P(o_t|s_t)$ av å observere o_t hvis vi er i tilstanden s_t



Eksempel

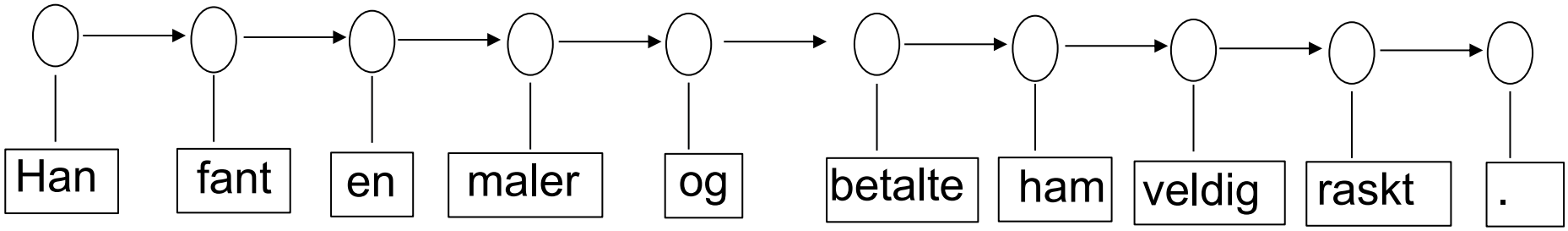
	Det	Noun	PropNoun	Verb	</s>
<s>	0.5	0.0	0.4	0.1	0.0
Det	0.0	0.95	0.0	0.05	0.0
Noun	0.05	0.0	0.0	0.85	0.1
PropNoun	0.1	0.0	0.0	0.8	0.1
Verb	0.25	0.0	0.25	0.0	0.5



	<s>	a	the	book	books	woman	trip	covers	cover	it	...	</s>
<s>	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0
Det	0.0	0.1	0.12	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0
Noun	0.0	1E-19	0.0	3.2E-5	1E-4	2.3E-5	6E-4	2.3E-7	4E-06	3E-8		0.0
PropNoun	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.11		0.0
Verb	0.0	0.0	0.0	4E-4	5E-5	0.0	1E-9	3E-6	5E-5	0.0		0.0
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		1.0

Dekoding

La oss si at vi får har utviklet en HMM for å tagge ordklasser, og vi får en setning med 10 ord



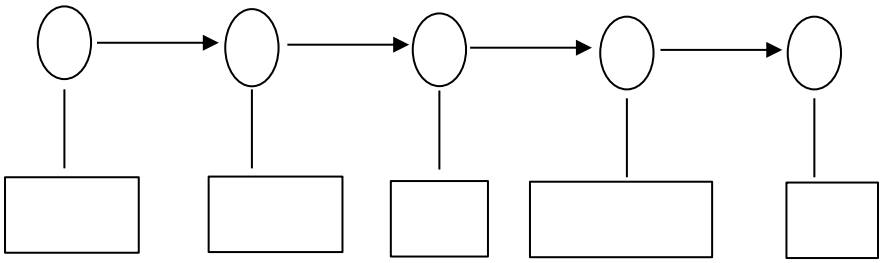
Hvordan kan vi finne den meste sansynnlige sekvensen over ordklassene?

➔ Vi kan beregne sannsynligheten for hver mulig sekvens en etter en … men det er k^{10} kombinasjoner, hvor k er antall ordklassene!

Hvis k=19 (som i Universal Dependencies) er det altså 6 131 milliarder mulige sekvenser!

Viterbi-dekoding

- Heldigvis finnes det en bedre løsning som har en *lineær* algoritmisk kompleksitet: Viterbi-algoritmen!
- Viterbi fungerer ved å behandle sekvensen *ord etter ord*
- For hver posisjon *t* beregner vi sannsynligheter over hver merkelapp tag_t=x, gitt resultatene beregnet for *t-1*
- Vi gjentar beregningen til alle ordene er behandlet



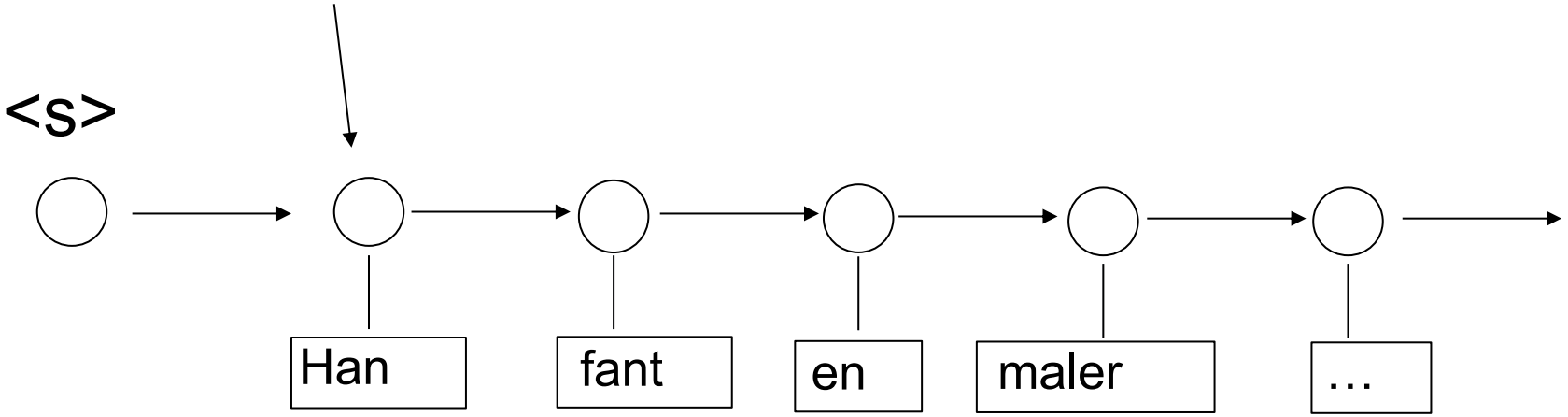
Viterbi-dekoding

$P(tag_1=\underline{x}, tag_0=<s>, word_1=\text{“han”})$
$= P(tag_0=<s>)$
$\quad * P(tag_1=\underline{x} \mid tag_0=<s>)$
$\quad * P(word_1=\text{“han”} \mid tag_1=\underline{x})$

Vi beregner først

$v_1(\underline{x}) = P(tag_1=\underline{x}, tag_0=<s>, word_1=\text{“han”})$

for alle mulige ordklasser *x*



Viterbi-dekoding

Vi går deretter til neste ord:

$v_2(\underline{x}) = \mathbf{max}_y\ P(tag_2=\underline{x}, tag_1=y, word_2=\text{“fant”},, word_1=\text{“han”})$

$= \mathbf{max}_y\ [v_1(y) * P(tag_2=\underline{x} \mid tag_1=y) * P(word_2=\text{"fant"} \mid tag_2=\underline{x})$

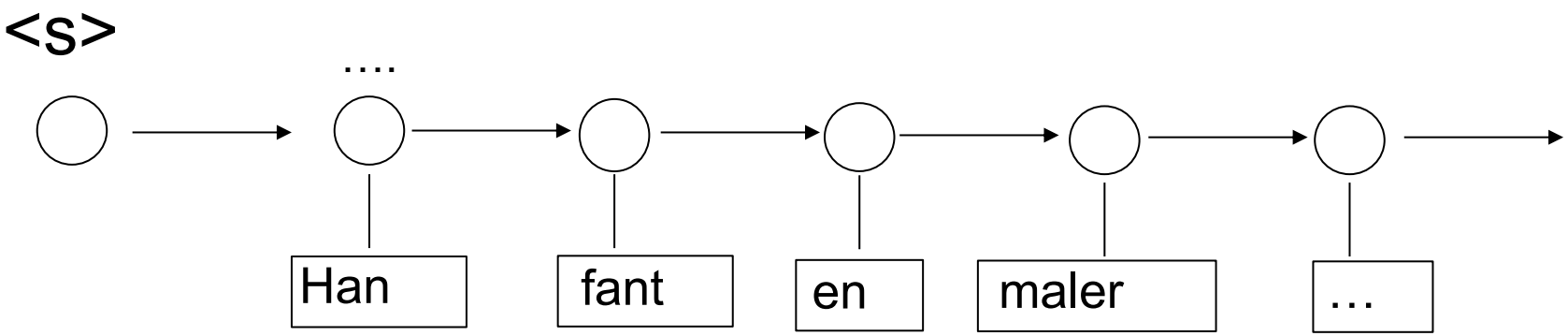
v_1

PRON=.99

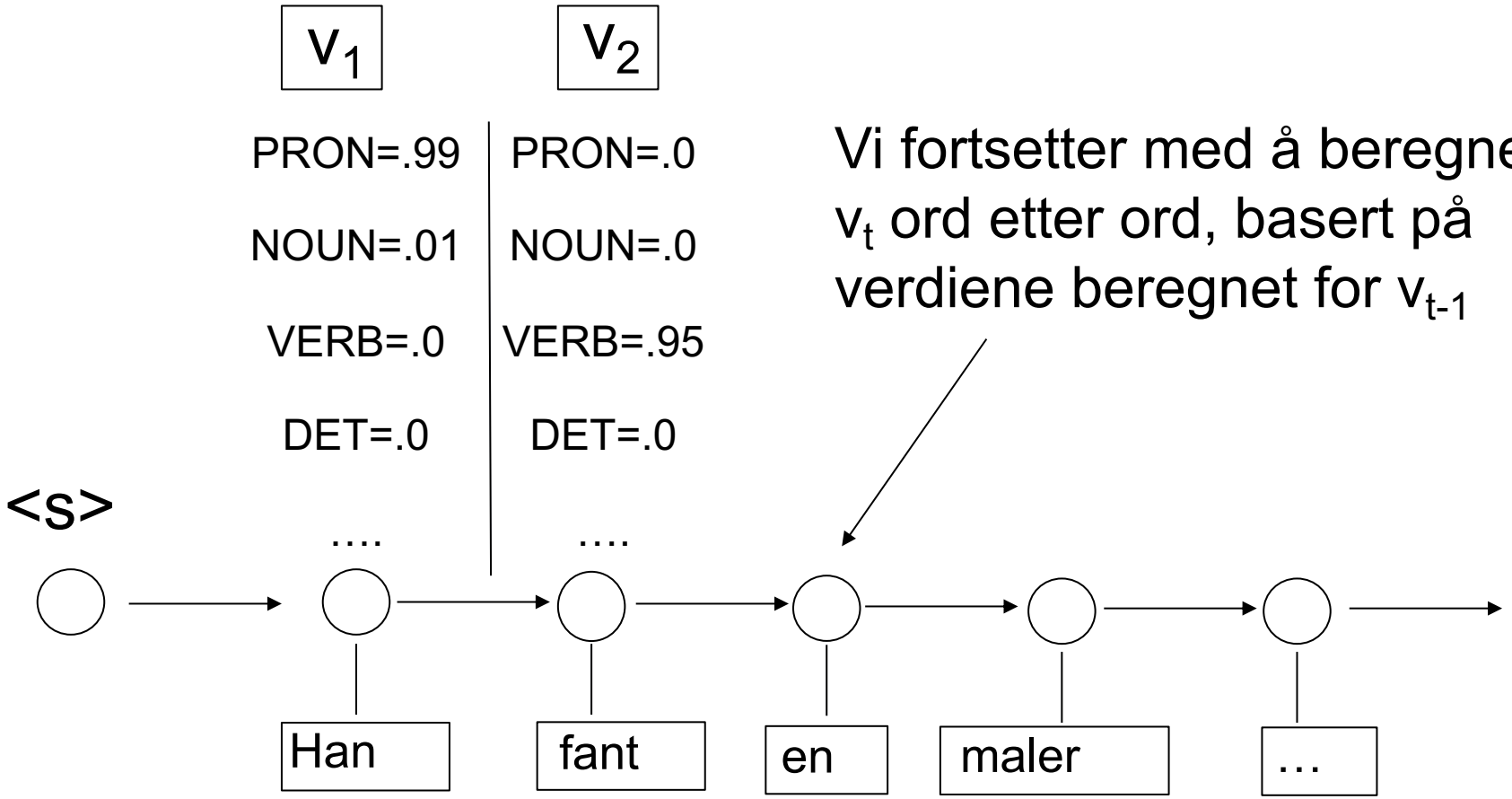
NOUN=.01

VERB=.0

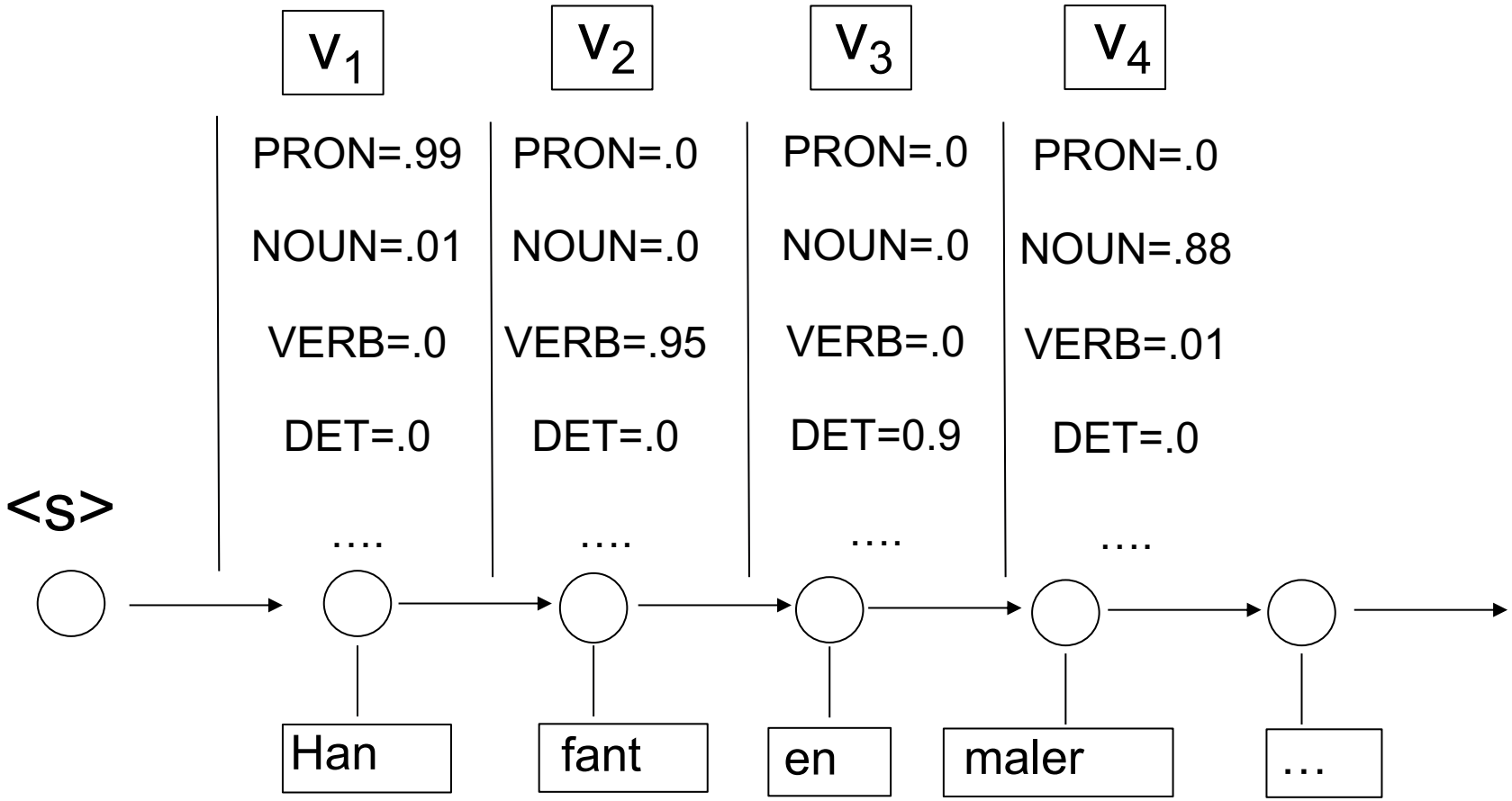
DET=.0



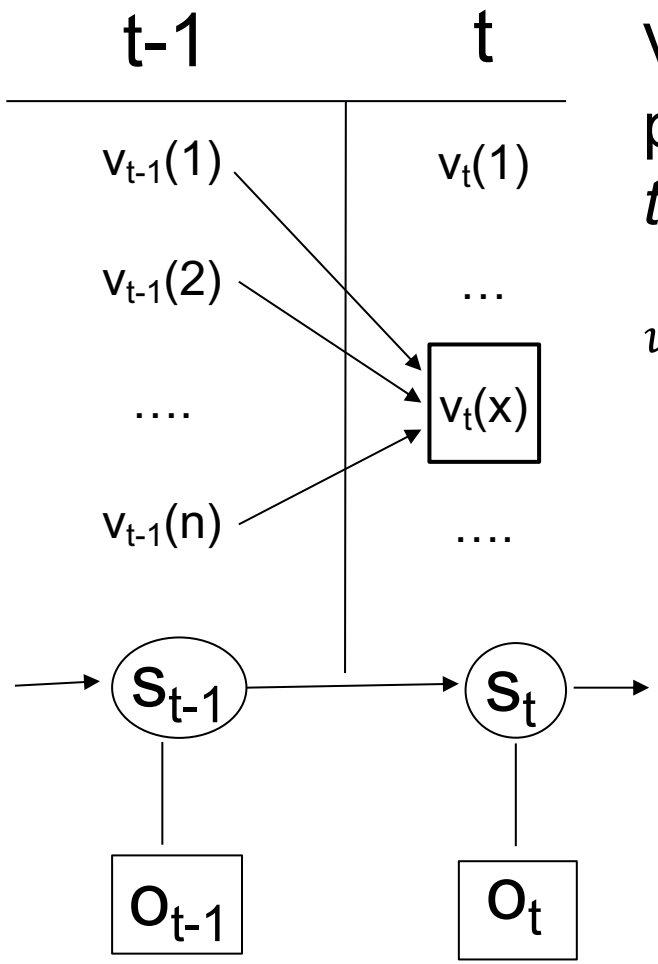
Viterbi-dekoding



Viterbi-dekoding



Viterbi-dekoding



Verdien $v_t(x)$ er sannsynligheten for tag *x* på ordet i posisjon *t og den beste tagsekvensen* fram til *t-1:*

$$v_t(x) = \max_{1 \le y \le N} v_{t-1}(y)\, P(s_t = x | s_{t-1} = y) P(o_t | s_t = x)$$

transisjonsmodell

emisjonsmodell

Vi tar kun hensyn til det *mest sannsynlige tag* y for t-1 (hvorfor?)

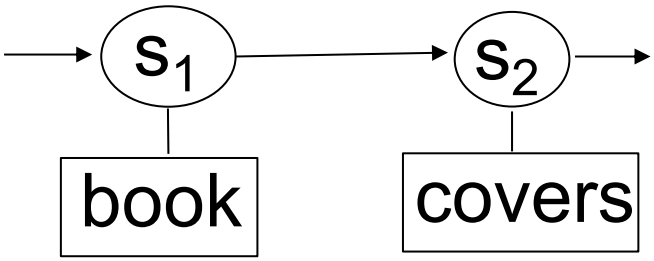
Verdien for tag *y* for det forrige ordet

$$v_t(x) = \max_{1 \le y \le N} v_{t-1}(y)\, P(s_t = x | s_{t-1} = y) P(o_t | s_t = x)$$

Øvelse

Beregn verdiene for v_2, gitt v_1 og HMM-modellen under:

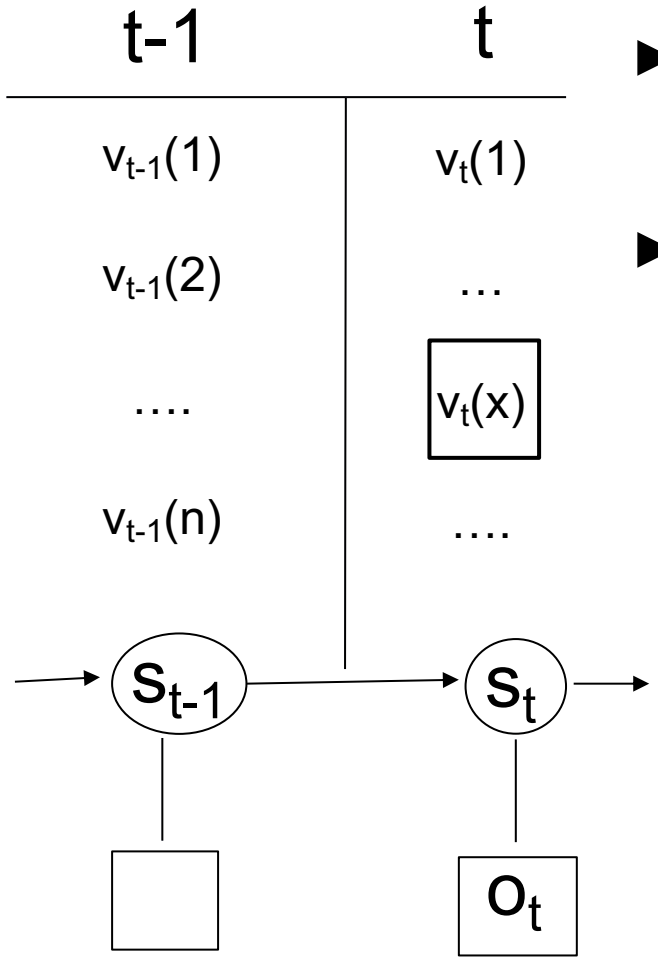
	v_1	v_2?
Det	0	0
Noun	1.6E-6	?
PropNoun	0	0
Verb	4E-5	?



	Det	Noun	PropNoun	Verb	</s>
<s>	0.5	0.05	0.35	0.1	0.0
Det	0.0	0.95	0.0	0.05	0.0
Noun	0.05	0.1	0.0	0.75	0.1
PropNoun	0.1	0.0	0.0	0.8	0.1
Verb	0.25	0.05	0.20	0.0	0.5

	<s>	a	the	book	books	woman	trip	covers	cover	it	...	</s>
<s>	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0
Det	0.0	0.1	0.12	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0
Noun	0.0	1E-19	0.0	3.2E-5	1E-4	2.3E-5	6E-4	2.3E-7	4E-06	3E-8		0.0
PropNoun	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.11		0.0
Verb	0.0	0.0	0.0	4E-4	5E-5	0.0	1E-9	3E-6	5E-5	0.0		0.0
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		1.0

Viterbi-dekoding



- Siste detalj: Viterbi-algorithme må også lagre såkalte **backpointers**
- Backpointer for $v_t(x)$ er tilstanden *y* i formelen vi nettopp har sett, dvs.

$$\max_{1 \le y \le N} v_{t-1}(y) P(s_t = x | s_{t-1} = y) P(o_t | s_t = x)$$

➔ Med disse backpointers kan vi lett ekstrahere de meste sannsynlige merkelappene når hele sekvensen er ferdig behandlet

Eksempel (fra boken)

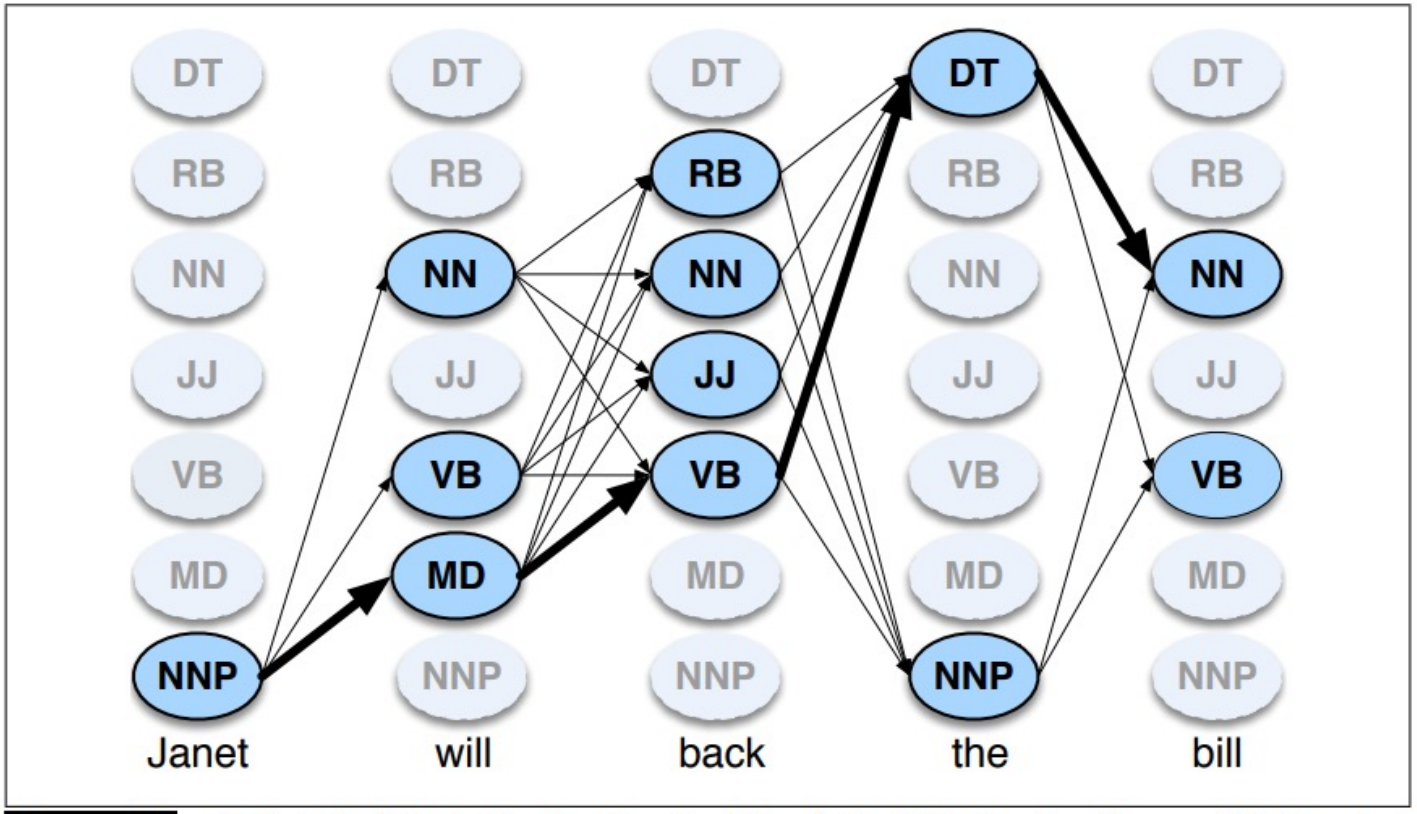


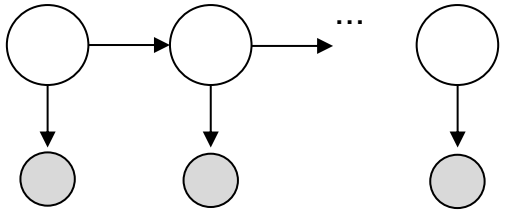
Figure 8.6 A sketch of the lattice for *Janet will back the bill*, showing the possible tags (q_i) for each word and highlighting the path corresponding to the correct tag sequence through the hidden states. States (parts of speech) which have a zero probability of generating a particular word according to the *B* matrix (such as the probability that a determiner DT will be realized as *Janet*) are greyed out.

Kommentarer

- Viterbi basert på "dynamic programming"
  - = vi bryter ned en komplisert beregning i mindre små beregninger som "gjenbruker" hverandres resultater (i vår tilfelle gjenbruker vi v_{t-1} for å beregne v_t)
  - *Edit distance* er et annet eksempel
- Viterbi likevel vanskelig å anvende på mer kompliserte NLP-oppgaver (slik som maskinoversettelse)
  - Long-range dependencies, stor antall mulig outputs
- Det finnes mange andre dekoding-algoritmer, blant annet **beam search** (som fokuserer på en beam av *k* hypoteser)

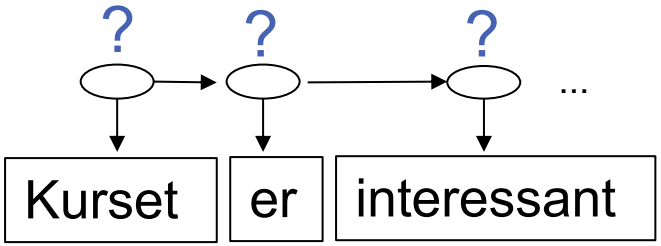
Oppsummering

Sekvensmodeller, og mer spesielt *Hidden Markov Models*



Modellering: hvordan behandler vi problemer som tar sekvensdata som input og output?

Dekoding: hvordan beregner vi den meste sannsynlige sekvensen av labeller (f.eks. POS tags) gitt en sekvens av observasjoner (f.eks. ord)?



NESTE UKE:

Læring: Hvordan estimerer vi parametrene i HMMs fra (markerte) data?