IN1140: Introduksjon til språkteknologi

Forelesning #13

Lilja Øvrelid

Universitetet i Oslo

16 november 2020

Tema for i dag

I dag

- Gjennomgang av digital prøveeksamen
- Obligkonkurranse

Eksamen

- 25 november kl 09:00
- 4 timers hjemmeeksamen i Inspera
- Trøsterunde på Zoom
- *Alle hjelpemidler er tillatt (lærebok, nettressurser, notater osv.) Det er ikke tillatt å samarbeide eller kommunisere med andre under eksamen om oppgavene.*

Prøveeksamen

Oppgave 1

Hvilket av alternativene kan ikke gjenkjennes med det følgende regulære uttrykket:

```
^..j..t.{1,3}$
```

- Majestic
- Adjective
- Adjusting
- Adjunction

Prøveeksamen

Oppgave 1

Hvilket av alternativene kan ikke gjenkjennes med det følgende regulære uttrykket:

`^..j..t.{1,3}$`

- Majestic
- Adjective
- Adjusting
- Adjunction ✓

Prøveeksamen

Oppgave 2

Skriv et regulært uttrykk som gjenkjenner følgende klasse av strenger :

- En streng som starter med "= " etterfulgt av et aritmetisk uttrykk som bruker heltall, addisjon, subtraksjon, og/eller multiplikasjon.

Prøveeksamen

Oppgave 2

Skriv et regulært uttrykk som gjenkjenner følgende klasse av strenger :

- En streng som starter med "= " etterfulgt av et aritmetisk uttrykk som bruker heltall, addisjon, subtraksjon, og/eller multiplikasjon.
- Anta her at heltallet inneholder maksimum fire siffer, og at strenger inneholder maksimum tre aritmetiske operasjoner.

Prøveeksamen

Oppgave 2

Skriv et regulært uttrykk som gjenkjenner følgende klasse av strenger :

- En streng som starter med "= " etterfulgt av et aritmetisk uttrykk som bruker heltall, addisjon, subtraksjon, og/eller multiplikasjon.
- Anta her at heltallet inneholder maksimum fire siffer, og at strenger inneholder maksimum tre aritmetiske operasjoner.
- For eksempel skal strengene "= 2*3+8-1" og "= 2+1000+120+9" gjenkjennes av ditt regulære utrykk.

Prøveeksamen

Oppgave 2

Skriv et regulært uttrykk som gjenkjenner følgende klasse av strenger :

- En streng som starter med "= " etterfulgt av et aritmetisk uttrykk som bruker heltall, addisjon, subtraksjon, og/eller multiplikasjon.
- Anta her at heltallet inneholder maksimum fire siffer, og at strenger inneholder maksimum tre aritmetiske operasjoner.
- For eksempel skal strengene "= 2*3+8-1" og "= 2+1000+120+9" gjenkjennes av ditt regulære utrykk.

Eksempelsvar

- "starter med" `^`
- "addisjon, subtraksjon, multiplikasjon" `(\+-|)|`
- "heltall med maks fire siffer" `\d{1,4}` (alt. `[0-9]{1,4}`)

Prøveeksamen

Oppgave 2

Skriv et regulært uttrykk som gjenkjenner følgende klasse av strenger :

- En streng som starter med "= " etterfulgt av et aritmetisk uttrykk som bruker heltall, addisjon, subtraksjon, og/eller multiplikasjon.
- Anta her at heltallet inneholder maksimum fire siffer, og at strenger inneholder maksimum tre aritmetiske operasjoner.
- For eksempel skal strengene "= 2*3+8-1" og "= 2+1000+120+9" gjenkjennes av ditt regulære utrykk.

Eksempelsvar

- "starter med" `^`
- "addisjon, subtraksjon, multiplikasjon" `(\+-|)|`
- "heltall med maks fire siffer" `\d{1,4}` (alt. `[0-9]{1,4}`)

Det finnes flere måter å løse denne oppgaven på. Følgende er et eksempel:

```
^=\s\d{1,4}((\*|\+|\-)\d{1,4}){1,3}
```

Prøveeksamen

Oppgave 3

Gitt følgende setning (dikt fra André Bjerke):

```
I landet Miramarmora var Farao på ferie hos farmora og
mormora .  En morgen klatret mormora til Farao i furua , og
så begynte moroa .
```

1. Forklar forskjellen mellom begrepene tokens og types.
2. Gi antall tokens og types i setningen.

Eksempelsvar

1. Typer er antall unike ord i teksten, mens tokens er antall løpende ord (der like forekomster telles flere ganger).
2. 26 tokens og 22 typer. (mormora, Farao, og, . forekommer to ganger.)

Prøveeksamen

Oppgave 4

Hvor mange trigram forekommer i teksten under?

`<s> Bjelleklang bjelleklang over skog og hei </s>`
`<s> Hør på bjellens muntre klang når Blakken drar i vei </s>`

Velg ett alternativ

- 10
- 12
- 14
- 16

`<s> Bjelleklang bjelleklang`, `Bjelleklang bjelleklang over`, `bjelleklang over skog` etc.

Prøveeksamen

Oppgave 4

Hvor mange trigram forekommer i teksten under?

`<s> Bjelleklang bjelleklang over skog og hei </s>`
`<s> Hør på bjellens muntre klang når Blakken drar i vei </s>`

Velg ett alternativ

- 10
- 12
- 14
- 16 ✓

`<s> Bjelleklang bjelleklang`, `Bjelleklang bjelleklang over`, `bjelleklang over skog` etc.

Prøveeksamen

Oppgave 5

Ta for deg ett av trigrammene fra forrige oppgave og vis hvordan det kan brukes til å beregne sannsynligheten for et ord gitt de to foregående ordene i en trigrammodell.

Eksempelsvar

Vi benytter tellinger av trigram og bigram for å beregne sannsynligheten for ett ord gitt de to foregående ordene, P($w_i|w_{i-2}w_{i-1}$)
Eksempel:
Trigram: "bjelleklang over skog"
P(skog|bjelleklang over) = count(bjelleklang over skog)/count(bjelleklang over) = 1/1

Oppgave 6

Her skal vi jobbe med følgende setning:

`Underholdende og sofistikert . Herman tok Norge med storm .`

Gitt ordklassene i Tabell 1 under, tildel ordklasser til alle ordene i setningen. Du må velge ett alternativ for hvert ord.

NOUN	Substantiv
ADJ	Adjektiv
VERB	Verb
PROPN	Egennavn
PREP	Preposisjon
CONJ	Konjunksjon
SUBJ	Subjunksjon
ADV	Adverb
DET	Determinativ
PUNCT	Tegnsetting

Prøveeksamen

Oppgave 6

Her skal vi jobbe med følgende setning:

```
Underholdende og sofistikert .  Herman tok Norge med storm
.
```

Svar

```
Underholdende|ADJ og|CONJ sofistikert|ADJ .|PUNKT
Herman|PROPN tok|VERB Norge|PROPN med|PREP storm|NOUN
.|PUNKT
```

Oppgave 7

Gitt at vi har trent en ordklassetagger. Det finnes to strategier for å evaluere en slik modell: Ekstrinsisk og intrinsisk evaluering.
Forklar forskjellen mellom disse to strategiene.

Prøveeksamen

Oppgave 7

Gitt at vi har trent en ordklassetagger. Det finnes to strategier for å evaluere en slik modell: Ekstrinsisk og intrinsisk evaluering.
Forklar forskjellen mellom disse to strategiene.

Eksempelsvar

- Ekstrinsisk evaluering = Vi evaluerer modellen 'indirekte' utfra hvordan den påvirker resultatene for en annen oppgave.
- Intrinsisk evaluering = Bruker et mer direkte mål for hvor bra modellen er på oppgaven den ble trent for.

Oppgave 8

Anta følgende kontekstfrie grammatikk:

S $\rightarrow$ NP VP

NP $\rightarrow$ NP PP

VP $\rightarrow$ V2 NP

VP $\rightarrow$ V1

NP $\rightarrow$ Rudolf, Nissen, nissen, gavene, sleden, grøten

V1 $\rightarrow$ sov, danset

V2 $\rightarrow$ trakk, spiste

PP $\rightarrow$ P NP

P $\rightarrow$ med

Oppgave 8

Grammatikken over er ikke en komplett grammatikk for norsk. For hver av setningene under, ta stilling til om de gis en analyse av grammatikken.

	JA	NEI
Rudolf danset	X	
Nissen danset med Rudolf		X
Rudolf spiste med nissen		X
Nissen spiste gavene	X	
Nissen med gavene sov	X	

Table:

Prøveeksamen

Oppgave 9

Er grammatikken i forrige oppgave rekursiv? Hvis ja, forklar hvorfor. Hvis nei, gi et eksempel på en rekursiv regel.

Prøveeksamen

Oppgave 9

Er grammatikken i forrige oppgave rekursiv? Hvis ja, forklar hvorfor. Hvis nei, gi et eksempel på en rekursiv regel.

Eksempelsvar

Ja, grammatikken er rekursiv. Regelen NP $->$ NP PP er rekursiv fordi den inneholder samme kategori på venstre og høyre side.

Oppgave 10

Vi ønsker å avgjøre hvorvidt *sleden* er en konstituent i setningen *Rudolf trakk sleden*. Beskriv minst to konstituenttester og vis hvordan de kan brukes for å hjelpe oss med å avgjøre om sleden er en konstituent.

Oppgave 10

Vi ønsker å avgjøre hvorvidt *sleden* er en konstituent i setningen *Rudolf trakk sleden*. Beskriv minst to konstituenttester og vis hvordan de kan brukes for å hjelpe oss med å avgjøre om sleden er en konstituent.

Eksempelsvar (forkortet)

- Stå alene: Hva trakk Rudolf? Sleden.
- Erstattes med pronomen: Rudolf trakk den.
- Flyttes som enhet: Det var sleden som Rudolf trakk. Sleden ble trukket av Rudolf.

Oppgave 11

Du skal nå utvide grammatikken slik at den kan håndtere setninger som *Rudolf tror at nissen sov* og *Rudolf tror at nissen spiste grøten*. Hvilke regler må du legge til den opprinnelige grammatikken for å kunne analysere disse setningene?

Prøveeksamen

Oppgave 11

Du skal nå utvide grammatikken slik at den kan håndtere setninger som *Rudolf tror at nissen sov* og *Rudolf tror at nissen spiste grøten*. Hvilke regler må du legge til den opprinnelige grammatikken for å kunne analysere disse setningene?

Eksempelsvar

```
VP -> V3 CP
CP -> C S
C -> at
V3 -> tror
```

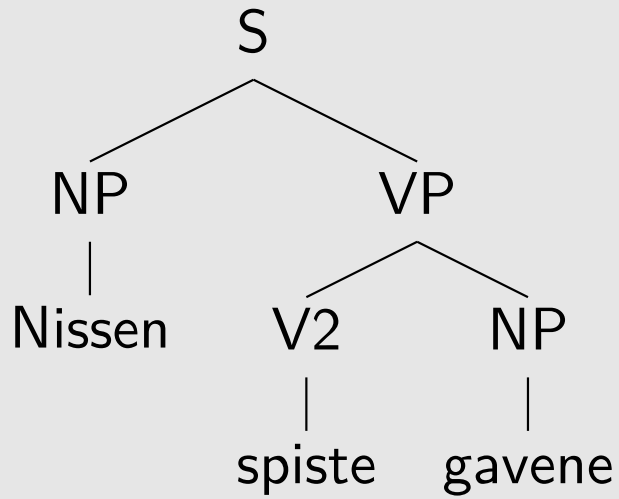
Prøveeksamen

Oppgave 12

Tegn det syntaktiske treet for to setninger som gis en analyse av grammatikken.

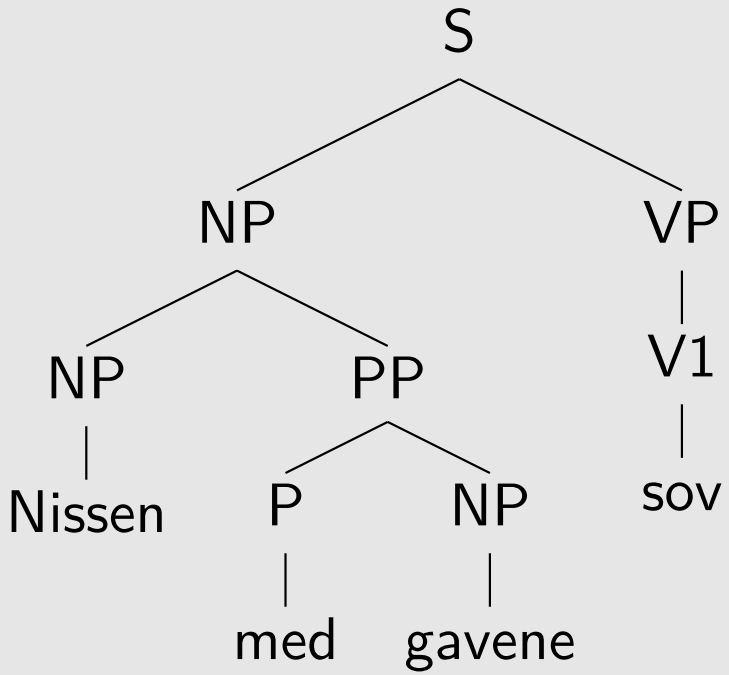
Oppgave 12

Tegn det syntaktiske treet for to setninger som gis en analyse av grammatikken.



Oppgave 12

Tegn det syntaktiske treet for to setninger som gis en analyse av grammatikken.



Oppgave 13

I denne oppgaven har vi et lite utvalg ord fra film-anmeldelser som hører til klassene positiv eller negativ.

1	god, fantastisk, morsom (POS)
2	teit, morsom, gøy (POS)
3	dårlig, kjedelig, morsom (NEG)
4	dårlig, kjedelig (NEG)
5	dårlig, teit, kjedelig (NEG)

Gitt et nytt test-dokumentet D som inneholder følgende ord: *god, teit, fantastisk*, bruk Naive Bayes-formelen til å klassifisere test-dokumentet D. Her skal du:

1. Regne ut sannsynlighetene for de forskjellige ordene. Du trenger bare å regne ut for ordene i test-dokumentet. **Ikke** bruk glatting.
2. Regne ut hvilken verdi som er størst. Blir dokumentet klassifisert som positiv eller negativ?

Oppgave 13 forts.

$$\hat{b} = argmax_{b \in B} \; P(b) \prod_{j=1}^{n} P(v_j|b)$$

Eksempelsvar

Prior sannsynlighetene:
$P(POS) = \frac{2}{5}$, $P(NEG) = \frac{3}{5}$
Sannsynligheten for hvert ord i testdokumentet gitt den positive eller den negative klassen:
$P(god|POS) = \frac{1}{6}$, $P(teit|POS) = \frac{1}{6}$, $P(fantastisk|POS) = \frac{1}{6}$
$P(god|NEG) = \frac{0}{8}$, $P(teit|NEG) = \frac{1}{8}$, $P(fantastisk|NEG) = \frac{0}{8}$
Da får vi:
$P(POS)P(S|POS) = \frac{2}{5} \times \frac{(1 \times 1 \times 1)}{6^3} = \frac{2}{5 \times 6^3}$
$P(NEG)P(S|NEG) = \frac{3}{5} \times \frac{(0 \times 1 \times 0)}{8^3} = 0$
Setningen blir klassifisert som **positiv**.

Oppgave 13 forts.

$$\hat{b} = argmax_{b \in B} \; P(b) \prod_{j=1}^{n} P(v_j|b)$$

Eksempelsvar

Prior sannsynlighetene:
$P(POS) = \frac{2}{5}$, $P(NEG) = \frac{3}{5}$
Sannsynligheten for hvert ord i testdokumentet gitt den positive eller den negative klassen:
$P(god|POS) = \frac{1}{6}$, $P(teit|POS) = \frac{1}{6}$, $P(fantastisk|POS) = \frac{1}{6}$
$P(god|NEG) = \frac{0}{8}$, $P(teit|NEG) = \frac{1}{8}$, $P(fantastisk|NEG) = \frac{0}{8}$
MERK: Vi godtar at studenten teller antall distinkte ord i hver klasse.
Altså at resultatene blir:
$P(POS)P(S|POS) = \frac{2}{5} \times \frac{(1 \times 1 \times 1)}{5^3} = \frac{2}{5^4}$
$P(NEG)P(S|NEG) = \frac{3}{5} \times \frac{(0 \times 1 \times 0)}{4^3} = 0$

Oppgave 14

Hvilken semantisk relasjon holder mellom følgende ord-par? Finn de som passer sammen:

	synonymi	hyponymi	meronymi	antonymi	hypernymi
død - levende				X	
klementin - skall			X		
slede - transportmiddel		X			
snill - slem				X	
dame - kvinne	X				
dyr - reinsdyr					X

Oppgave 15

Gi en kort beskrivelse av hva som kjennetegner de følgende semantiske rollene og illustrer svaret ditt med minst ett eksempel for hver rolle.

1. EXPERIENCER
2. AGENT
3. PATIENT
4. INSTRUMENT

Prøveeksamen

Eksempelsvar

1. EXPERIENCER: deltageren som opplever handlingen beskrevet av verbet uten å kontrollere den, ofte via sansene. F.eks. **Peter** ser toget, **Peter** hører en lyd.
2. AGENT: deltageren som utfører en handling med viten og vilje. F.eks. **Peter** slår Jenny.
3. PATIENT: deltageren som blir påvirket av handlingen som beskrives, endrer ofte tilstand. F.eks. Peter knuser **ruten**.
4. INSTRUMENT: redskapet som brukes til å utføre en handling. F.eks. Peter knuser ruten med **hammeren**.

Oppgave 16

Den vanligste måten å løse Named Entity Recognition på er ved ord-for-ord klassifisering, så kalt BIO klassifisering.
Anta følgende setning:
`Thorbjørn Egner ble født 12.12.1912 i Norge på Kampen i Oslo`
Angi riktig BIO klassifisering med riktig kategori for hvert ord i setningen.
Her opererer vi med følgende katergorier: PER (person), ORG (organisajon), LOC (location), DT (dato), GPE (geopolitical entity) .

Prøveeksamen

Svar

	O	B_P	I_P	B_O	I_O	B_L	I_L	B_D	I_D	B_G	I_G
Thorbjørn		X									
Egner			X								
ble	X										
født	X										
12.12.1912								X			
i	X										
Norge						X				X	
på	X										
Kampen						X				X	
i	X										
Oslo						X				X	

Prøveeksamen

Oppgave 17

Vi har sett på to hovedeksempler av dialogsystemer: oppgaveorienterte dialogagenter og chatbots.
For å kunne utvikle slike systemer må vi først forstå hvilke egenskaper av menneskelige samtaler kan være utfordrende å automatisere.
Gi eksempler på to slike egenskaper, og forklar hvorfor de er utfordrende.

Prøveeksamen

Eksempelsvar

Et lite utvalg av egenskaper av menneskelig språk som kan være utfordrende:

- Hver sin tur til å snakke. Problemet her, er hvordan kan vi vite at den andre er ferdig med å snakke?
- Forankring: systemet bør kunne vise brukeren at den har "forstått" hva brukeren snakker om. Men å identifisere når vi skal gjøre det, eller hvordan er utfordrende.
- Struktur: rekkefølgen av spørsmål/svar er ikke alltid konsekvent når to mennesker har en samtale, og av og til ombestemmer vi oss. Da må maskinen kunne forstå at det vi snakket om tidligere har endret seg, og at den bør oppdatere sin forståelse av samtalen.
- Inferens og implikasjon: Vi mennesker klarer fint å forstå både implikasjon og inferens, men for at en maskin skal kunne forstå det må den ha en bred forståelse av verden rundt oss.
- Flertydighet i språket.

Obligkonkurransen

Obligkonkurransen

Tre studenter har fått (nesten) full pott på alle obliger og er vinnere av obligkonkurransen i IN1140 høsten 2020:

Obligkonkurransen

Tre studenter har fått (nesten) full pott på alle obliger og er vinnere av obligkonkurransen i IN1140 høsten 2020:

- Helle Bollingmo
- Irma Hofseth Fearnley
- Victor Alexander Steen-Olsen

Obligkonkurransen

Tre studenter har fått (nesten) full pott på alle obliger og er vinnere av obligkonkurransen i IN1140 høsten 2020:

- Helle Bollingmo
- Irma Hofseth Fearnley
- Victor Alexander Steen-Olsen

Gratulerer!!!!!!

(Send meg addressen din så kommer en overraskelse i posten...)