i Introduksjon

Eksamen i IN2110

Fredag, 14. juni 2019, kl. 09:00 (4 timer).

Ingen hjelpemidler.

I flervalgsoppgavene får du positive poeng for riktige svar, 0 poeng for ubesvart og negative poeng for gale svar, men du aldri mindre enn 0 poeng på hver oppgave.

Vi anbefaler å lese gjennom hele oppgaveteksten før du begynner. Hvis du føler du mangler informasjon for å løse en oppgave, gjør dine egne antakelser og redegjør for dem.

1.1 TF-IDF

Definer formelen til vektingsmålet TF-IDF ("Term Frequency – Inverse Document Frequency") og forklar notasjonen du bruker. Diskuter kort hva som er hensikten med å anvende TD-IDF.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 7

1.2 Lengde-normalisering

Når vi jobber med vektorrom-representasjoner av dokumenter benytter vi oss ofte av lengde-normalisering. Forklar hva dette innebærer og hvilken praktisk nytte det kan ha.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 6

2.1 Evaluering

Flere av evalueringsmålene vi har sett på i kurset har vært definert på basis av fire mer grunnleggende kategorier av hvordan prediksjonene til en klassifikator kan være riktige eller gale, sammenliknet med gullstandarden; *false positives* (FP), *true negatives* (TN), osv. Matrisen under viser hvordan disse kategoriene er definert:

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Vis hvordan de tre målene *Accuracy*, *Recall* og *Precision* kan defineres på bakgrunn av dette.

Skriv ditt svar her...

Maks poeng: 6

2.2 Accuracy

Tenk deg at vi jobber med binær klassifikasjon og at vi har mange flere eksempler i den negative klassen enn den positive (la oss anta et forhold på 9:10). Diskuter hvorvidt *Accuracy* er et egnet eller uegnet evalueringsmål for dette problemet.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 4

2.3 kNN

Beskriv kort klassifikasjonsmetoden kNN. Diskuter kort dens styrker og svakheter.

Skriv ditt svar her...

Format | B I U x_2 x^2 | I_x | | | | | | | | | | | | | | | ABC | |

Words: 0

Maks poeng: 7

3.1 Leksikalske relasjoner

Ord kan være relatert til hverandre på ulike måter, og i kurset snakket vi blant annet om synonymi, antonymi, hypernymi (overordnet) og hyponymi (underordnet).

Velg riktig relasjon som kan erstatte streken _:

	hyponym	antonym	synonym	hypernym
snill er _ til vennlig	○	○	○	○
jobb er _ til arbeid	○	○	○	○
hund er _ til kjæledyr	○	○	○	○
kvinne er _ til dame	○	○	○	○
frukt er _ til eple	○	○	○	○
bil er _ til kjøretøy	○	○	○	○
ond er _ til slem	○	○	○	○
liten er _ til stor	○	○	○	○

Maks poeng: 2

3.2 Ordvektorer og likhet

Når vi skal regne ut likheten mellom to ordvektorer, kan vi bruke kosinus-likhet (cosine similarity) som et likhetsmål. Kosinus-likhet mellom to vektorer A og B er definert som følger:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Telleren i formelen er prikkproduktet (dot product) til vektorene A og B:

$$\sum_{i=1}^{n} A_i B_i$$

Nevneren her er størrelsen eller lengden til vektorene A og B:

$$\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}$$

Under finner du en matrise med termer og kontekster, som inneholder f.eks. TF-IDF eller PPMI verdier for tre **termer** og 8 *kontekst-ord*. Verdiene her er funnet på, for å gjøre utregningene lettere.

	eple	**pære**	**frukt**
ananas	3	1	0
eple	0	0	2
pære	0	0	1
skjære	2	0	0
dessert	2	1	4
lunsj	3	1	0
grill	1	1	0
Eva	3	0	2

La oss kalle ordvektorene for termene **eple**, **pære** og **frukt** for V_{eple}, $V_{pære}$ and V_{frukt}.

Hva er $\|V_{eple}\|$, dvs. lengden til V_{eple}? Gi svaret som et tall:

Videre skal vi bruke kosinus-liket som likhetsmål.

Hva er likheten mellom de to *likeste* ordene?

Hva er likheten mellom de to *mest ulike* ordene?

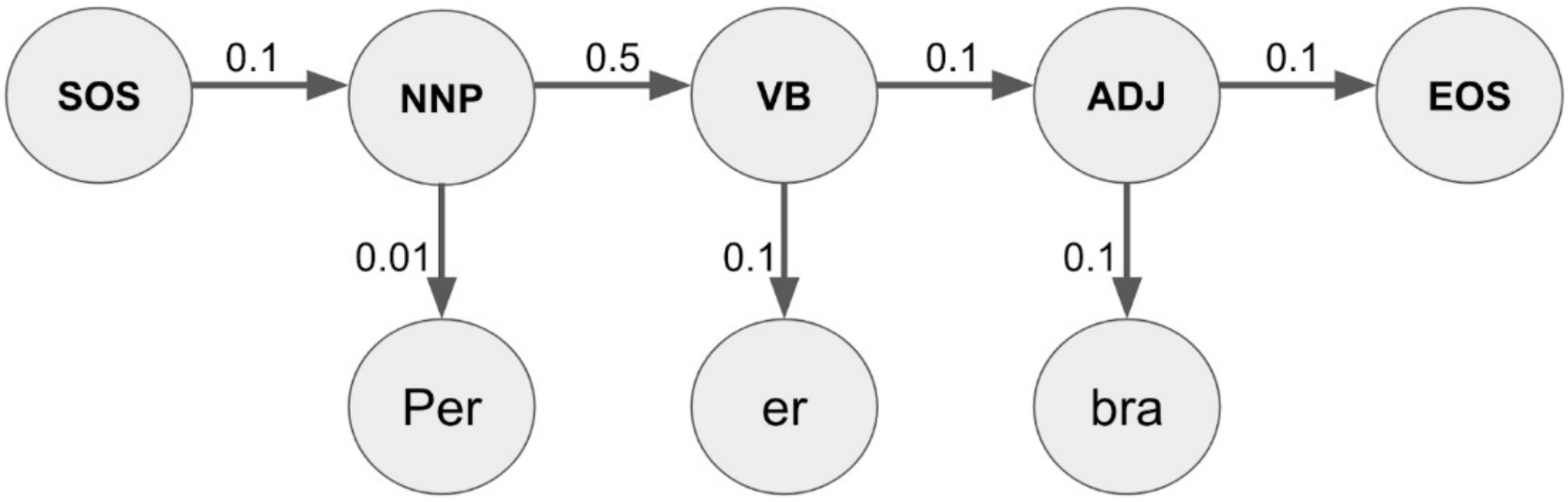
Maks poeng: 8

4

Hidden Markov Models

Du har følgende grafiske framstilling av observerte og skjulte tilstander, med sannsynligheter, gitt en HMM-modell.

Dette er det eneste du vet om modellen.



1 Komponentene i HMM

I pensum i kurset leste du om Hidden Markov Models. Slik Jurafsky og Martin definerte HMM, består den av følgende fem komponenter:

1. $Q = q_1 q_2 \dots q_n$
2. $A = a_{11} a_{12} \dots a_{n1} \dots a_{nn}$
3. $O = o_1 o_2 \dots o_T$
4. $B = b_i(o_t)$
5. q_0, q_F

I denne oppgaven skal du forklare hva hver av de fem komponentene er. (Du vil få en del begreper og hint i de neste oppgavene, så hvis du har problemer med å huske, kan det være lurt å lese gjennom disse først).

I tillegg snakket vi om ordklassetagging som en anvendelse av HMM'er. Der det er relevant, gi eksempler fra ordklassetagging. For eksempel, kan du i den første delen, gi eksempler på hva q kan være i kontekst av en ordklassetagger.

Forklar hver av de fem komponenente i en HMM her...

Format

Words: 0

Maks poeng: 8

2 Vokabular

Hva er vokabularet i denne HMM-en?
Velg et eller flere alternativer

- ADJ
- Per
- NNP
- EOS
- VB
- SOS
- er
- bra

Maks poeng: 1.5

3 Tilstander

Hva er de skjulte tilstandene i denne HMM-en? Ta ikke med de spesielle start- og slutt-tilstandene.
Velg et eller flere alternativer

- NNP
- bra
- EOS
- er
- ADJ
- VB
- SOS
- Per

Maks poeng: 1.5

4 Emmisjonssannsynligheter

Under følger en rekke påstander om emmisjonssannsynligheter, generelt og gitt grafen for "Per er bra".

Kryss av for alle påstander som er korrekt
Velg ett eller flere alternativer

- Emmisjonssannsynligheten for observasjonen "Per" gitt den skjulte tilstanden "NNP" er 0.01
- Emmisjonssannsynligheten for observasjonen "EOS" gitt den skjulte tilstanden "bra" er 0.1
- Emmisjonssannsynligheten for observasjonen "NNP" gitt den skjulte tilstanden "VB" er 0.5
- Den skjulte tilstanden "SOS" har aldri noen emmisjonssannsynligheter

Maks poeng: 1

5 Transisjonssannsynligheter

Gitt at grafen beskrevet er alt du vet om modellen, fyll inn transisjons-sannsynlighetene i denne matrisen, der radene er forrige tilstand *t-1*, og kolonnene er nåværende tilstand *t.*

Skriv inn 0 hvis:

- sannsynligheten for transisjonen faktisk er 0
- du ikke har informasjon om sannsynligheten til denne transisjonen

Du skal (og kan) ikke fylle inn feltene med **-** (strek), bare de tomme boksene.

Transisjonsprobabilitetsmatrise

	SOS_t	NNP_t	VB_t	ADJ_t	EOS_t
SOS_{t-1}	-		-	-	-
NNP_{t-1}		-		-	-
VB_{t-1}	-		-		-
ADJ_{t-1}	-	-		-	
EOS_{t-1}	-	-	-		-

Maks poeng: 2

6 Egennavn og verb

Etter egennavn ("NNP") er det i norsk ganske vanlig å finne et verb ("VB"), spesielt etter egennavn som er subjekter.

Klarer HMM-modellen å fange opp relasjonen mellom et egennavn ("NNP") og påfølgende verb ("VB")?

Forklar hvorfor/hvorfor ikke.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 2

7

Navnet Per

"Per" er et ganske vanlig navn i norsk. Klarer en HMM å uttrykke noe om hvor vanlig dette navnet er?

Forklar hvorfor/hvorfor ikke.

Skriv ditt svar her...

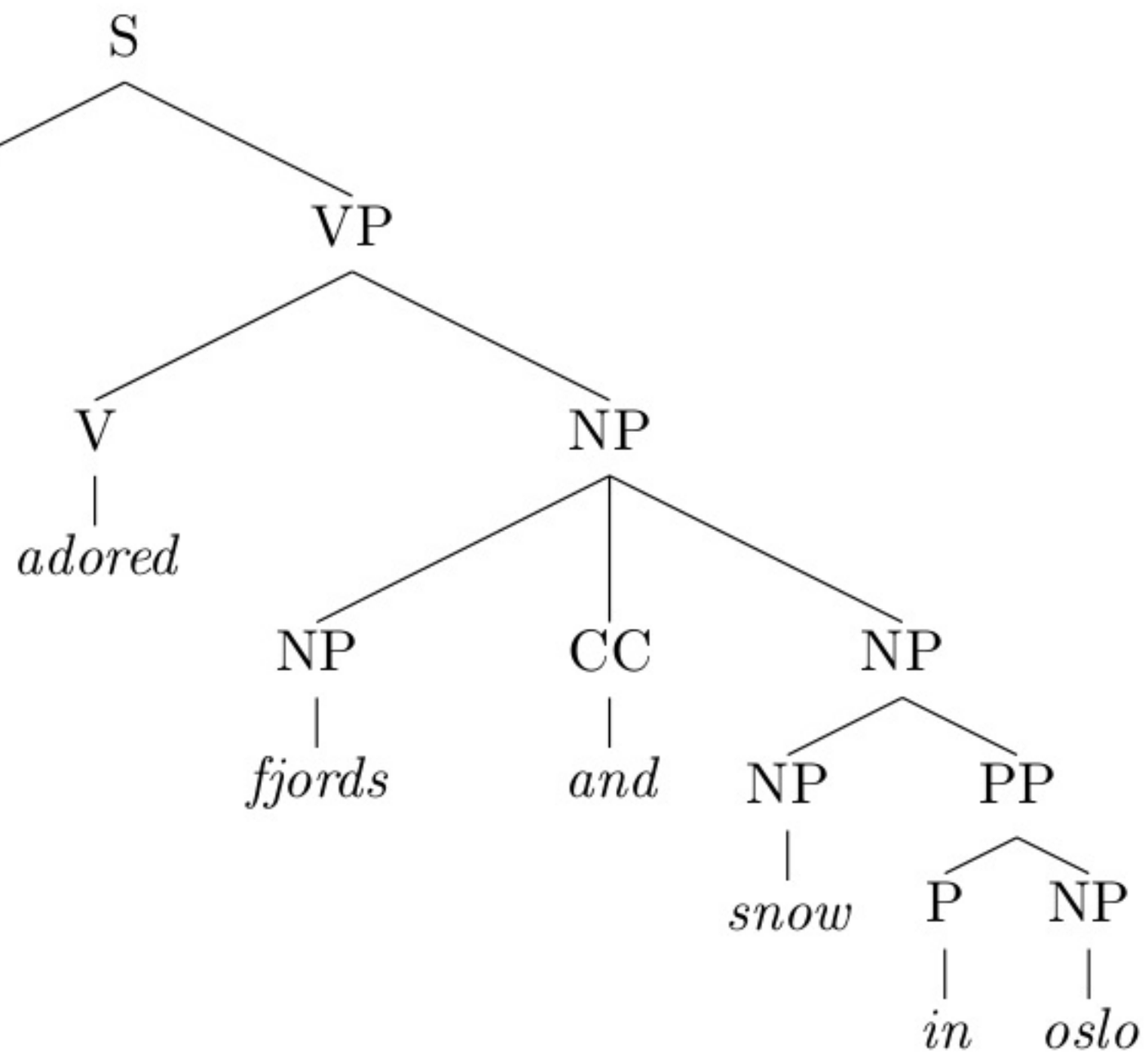
Format

Words: 0

Maks poeng: 2

5

Eksempeltre



1 Kontekstfri grammatikk

Vi antar en kontekstfri grammatikk som vi også brukte i forelesningene:

$\mathrm{S} \rightarrow \mathrm{NP\ VP}$	$\mathrm{NP} \rightarrow \mathit{kim}$
$\mathrm{VP} \rightarrow \mathrm{VP\ PP}$	$\mathrm{NP} \rightarrow \mathit{oslo}$
$\mathrm{NP} \rightarrow \mathrm{NP\ PP}$	$\mathrm{NP} \rightarrow \mathit{snow}$
$\mathrm{VP} \rightarrow \mathrm{V\ NP}$	$\mathrm{NP} \rightarrow \mathit{fjords}$
$\mathrm{VP} \rightarrow \mathrm{V}$	$\mathrm{V} \rightarrow \mathit{adores}$
$\mathrm{PP} \rightarrow \mathrm{P\ NP}$	$\mathrm{P} \rightarrow \mathit{in}$

Denne grammatikken er ikke tilstrekkelig for å parse vårt eksempeltre for setningen *kim adored fjords and snow in oslo*. Det mangler to produksjonsregler; les ut fra treet hvilke regler som mangler og skriv de inn i svarfeltet under.

Vi har definert en kontekstfri grammatikk som en fire-tuppel $\langle C, \Sigma, P, S \rangle$. For grammatikken med de ekstra reglene som du har lagt til, hvilke verdier har C og Σ?

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 6

2 Tvetydighet

Vi har sett mange ganger at det kan være flere mulige tolkninger for en enkelt setning, og at forskjellig syntaktisk struktur vil gi setningen forskjellig betydning. Hvilken tolkning viser eksempeltreet? Gitt vår kontekstfrie grammatikk (med de ekstra reglene), hvor mange flere tolkninger har setningen? Skisser i noen få setninger hvordan deres syntaktiske struktur vil være annerledes enn i eksempeltreet.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 6

3 Dependenssyntaks

I denne oppgaven skal vi gjøre om eksempeltreet til et *dependenstre*. I korte trekk, hva er grunnleggende forskjeller mellom *frasestruktur*- versus *dependenssyntaks*? Bruk gjerne begrep som *konstituent*, *hode*, *dependent*, *grammatikalsk funksjon*, m.m. Forklar i et par setninger hvordan disse begrepene anvendes i henholdsvis et frasestruktur- eller dependenstre.

Heller enn å tegne et dependenstre, kan det skrives ned som en mengde tripler, der hvert element tilsvarer en kant, f.eks. **(2, 1, nsubj)** for en relasjon av type *nsubj* som holder mellom ord #2 og ord #1. Skriv ned alle kanter i et dependenstre som tilsvarer vårt eksempeltre for setningen *kim adored fjords and snow in oslo*.

Setningen inneholder såkalt koordinasjon, der *and* (med ordklasse koordinerende konjunksjon: CC) binder sammen to ledd av samme type. Her står du fritt til å gjøre egne antakelser om hva som skal være den interne strukturen til koordinasjon; Universal Dependencies f.eks. bruker to dependenstyper, *conj* og *cc*, for koordinerte strukturer. Forklar dine valg i én setning.

Skriv ditt svar her...

Format

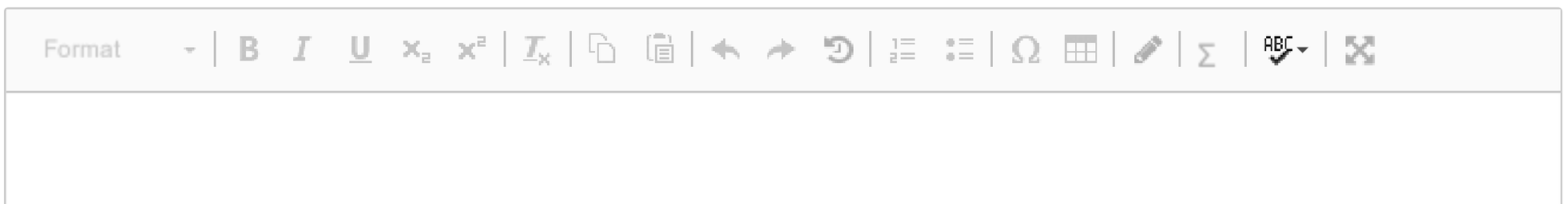
Words: 0

Maks poeng: 6

6.1 CKY

Hvilken egenskap gjør at syntaktisk parsing av naturlige språk er vanskeligere enn parsing av f.eks. programmeringsspråket Python? Hvordan øker antall forskjellige mulige syntaktiske strukturer når man legger til flere og flere preposisjonalfraser, f.eks. *kim adored snow in oslo on monday at noon*? Forklar kort hvordan dette problemet håndteres i CKY-parseren.

Skriv ditt svar her...



Maks poeng: 6

6.2 CKY-tabell

CKY-tabellen under viser parsingen av en setning som f.eks. *kim adored snow in oslo*. Vi vet at setningen har to forskjellige syntaktiske strukter, der forskjellen ligger i hvor PP-en *i oslo* kobles på. I hvilken tabellcelle oppstår tvetydigheten, og hvilke par av celler kombineres med hvilke produksjonsregler for å oppnå de to forskjellige strukturene?

	1	2	3	4	5
0	NP		S		S
1		V	VP		VP
2			NP		NP
3				P	PP
4					NP

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 5

6.3 Overgangsbasert parsing

Forklar kort hvilke datastrukturer og hvilke operasjoner som kjennetegner overgangsbasert parsing.

Skriv ditt svar her...

Format

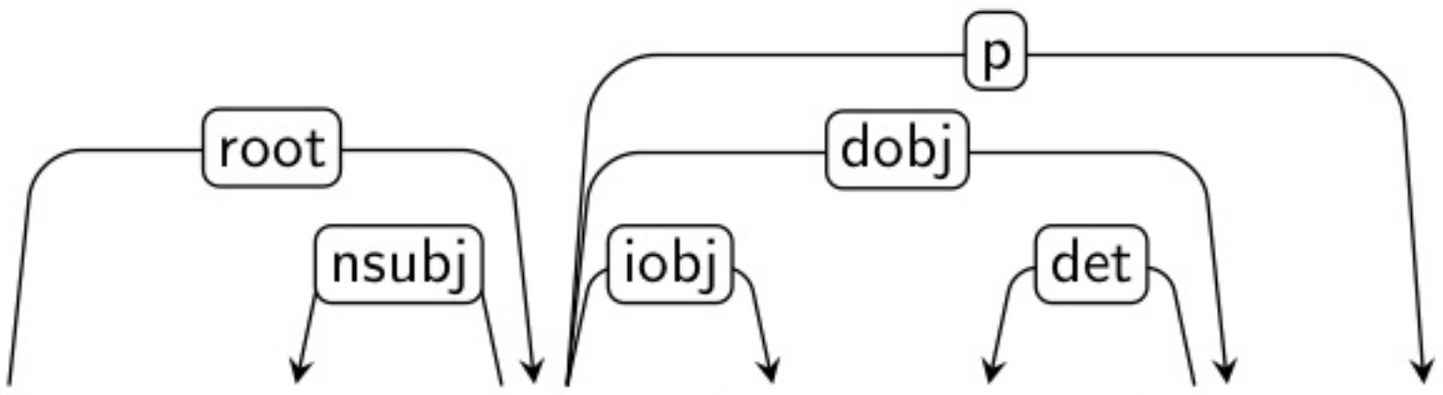
Words: 0

Maks poeng: 5

6.4 Overgangssekvenser

Vi har sett på to forskjellige typer overgangssystemer (*transition systems*), dvs. *arc-eager*– og *arc-standard*–systemene. Forklar i et par setninger hva som er forskjellen mellom de to.

Gitt dependenstreet under, skriv opp overgangssekvensen som gir opphav til dette treet, både for *arc-eager*– og for *arc-standard*–varianten.



Skriv ditt svar her...

Format

Words: 0

Maks poeng: 6