IN1140: Introduksjon til språkteknologi

Løsningsforslag eksamen H2017

Universitetet i Oslo

1. Regulære uttrykk (2 poeng)

Hvilken av de følgende epost-adressene kan ikke gjenkjennes med det følgende regulære uttrykket:

```
[aA-zZ0-9\._%+-]+@[aA-zZ0-9\.-]+\.[aA-zZ]{1,2}
```

Velg ett alternativ

1. Ola.Norman@epost.com
2. olanorman@meg.epost.no
3. OlaNorman@meg.co
4. Ola.Norman@epost.no

1. Regulære uttrykk (2 poeng)

Svar 1: Ola.Norman@epost.com

2. Regulære uttrykk for verb (6 poeng)

Skriv ett regulært uttrykk som gjenkjenner formene imperativ, infinitiv, presens, og preteritum av følgende norske verb: spise, kjøpe, tenke, løpe.

2. Regulære uttrykk for verb (6 poeng)

alternativ 1:

```
((å\s)?(spis|tenk|løp|kjøp)e)|((spis|tenk|løp|kjøp)(er|te)?)
```

alternativ 2:

```
^((å\s)?(spis|tenk|løp|kjøp)e)$|^((spis|tenk|løp|kjøp)(er|
te)?)$
```

3. Ordklasser (5 poeng)

Her skal vi jobbe med følgende setning:

`Jeg ser mannen med kikkerten.`

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

CC	konjunksjon
DET	determinativ
JJ	adjektiv
NN	substantiv
PR	preposisjon
PO	pronomen
RB	adverb
SB	subjunksjon
VB	verb

Table: 1

3. Ordklasser (5 poeng)

```
Jeg ser mannen med kikkerten
=>
Jeg/PO ser/VB mannen/NN med/PR kikkerten/NN
```

4. Grammatikk og strukturell flertydighet (10 poeng)

Setningen i vårt forrige eksempel er strukturelt flertydig. Definer en kontekstfri grammatikk med regler som kan vise ulike analyser av denne setningen. Altså:

`Jeg ser mannen med kikkerten.`

Ta også stilling til om grammatikken din er rekursiv. Begrunn svaret ditt.

4. Grammatikk og strukturell flertydighet (10 poeng)

```
S -> NP VP
NP -> PO | NN | NP PP
PO -> Jeg
NN -> mannen | kikkerten
VP -> VB NP | VB NP PP
VB -> ser
PP -> PR NP
PR -> med
```

5. Oppgave 5 Leksikalsk semantikk/relasjoner (7 poeng

1. Forklar forskjellen mellom homonymi og polysemi. Gi eksempler på begge.
2. Forklar og gi eksempler på relasjonen hyponnymi.
3. Forklar og gi eksempler på relasjonen meronymi.

5. Oppgave 5 Leksikalsk semantikk/relasjoner (7 poeng

1. **Homonymi (1.5 poeng)**: urelaterte betydninger som skrives eller uttales likt, f.eks. “not” og “knot”, “lap” (circuit of course) og “lap” (body part). Bonus: tre undertyper: homografer (lik skrivemåte), homofoner (lik ordlyd, men ulik skrivemåte) og fullstendige homonymer (lik skrivemåte og ordlyd).
2. **Polysemi (1.5 poeng)**: flere betydninger, men betydningene er relatert (konseptuelt eller historisk), f.eks. alle betydninger av et ord som listet i en ordbok – “hook noun. 1. a piece of material, usually metal, curved or bent and used to suspend, catch, hold or pull something. 2. short for fish-hook. 3. a trap or snare 4. a sharply curved spit of land, 5. Boxing a short swinging blow delivered from the side with the elbow bent. . . . etc”
3. **Hyponnymi (2 poeng)**: inkluderingsrelasjon. Et hyponym inkluderer betydningen til et mer generelt ord. F.eks. hund, katt er hyponymer av dyr, spurv er et hyponym av fugl.
4. **Meronymi (2 poeng)**: relasjon mellom del og helhet. X er del av Y / Y har X. F.eks. et omslag er del av en bok, en bok har et omslag.

6. Komposisjonalitet (5 poeng)

Forklar med noen få setninger hva vi mener med at setningssemantikk er komposisjonell (compositional)?

6. Komposisjonalitet (5 poeng)

Vi forstår en frase eller setning på grunnlag av hvordan mindre deler (ord, fraser) er satt sammen. Betydningen er gitt ufra de enkelte delene og reglene som styrer hvordan de settes sammen.

7. Semantiske roller (5 poeng)

Gitt følgende setninger:
"**Markus** ryddet lekene. Plutselig, kastet Nora **brannbilen** på ham og skadet ham. **Edna** så hva som skjedde. Hun ropte på mor som måtte komme og rense såret med et **antibakterielt middel**."
Angi de semantiske rollene for ordene i tabellen under. Du må velge ett alternativ for hvert tilfelle.

Finn de som passer sammen

	INSTRUMENT	BENEFICIARY	PATIENT	EXPERIENCER	THEME	AGENT
Edna	○	○	○	○	○	○
Markus	○	○	○	○	○	○
lekene	○	○	○	○	○	○
et antibakterielt middel	○	○	○	○	○	○
brannbilen	○	○	○	○	○	○

7. Semantiske roller (5 poeng)

- agent: den som setter i gang en handling, i stand til å handle med viten og vilje
- patient: entiteten som påvirkes av en handling, gjennomgår ofte en forandring
- theme: entiteten som blir beveget av en handling eller hvis beliggenhet beskrives
- experiencer: er bevisst på handlingen eller tilstanden, men er ikke i kontroll over den
- beneficiary: entiteten som en handling utføres for
- instrument: middelet som gjør at en handling kan utføres eller finner sted
- goal: entiteten som noe beveger seg mot
- source: entiteten som noe beveger seg fra

7. Semantiske roller (5 poeng)

Finn de som passer sammen

	INSTRUMENT	BENEFICIARY	PATIENT	EXPERIENCER	THEME	AGENT
Edna	○	○	○	✔	○	○
Markus	○	○	○	○	○	✔
lekene	○	○	✔	○	○	○
et antibakterielt middel	✔	○	○	○	○	○
brannbilen	○	○	○	○	✔	○

8. Maximum likelihood estimering (MLE) (5 poeng)

For å estimere sannsynligheter i n-gram språkmodeller har vi benyttet oss av såkalt maximum likelihood estimering (MLE). Vis formlene for hvordan vi fra et korpus estimerer sannsynligheten for et ord wi gitt det foregående ordet w_{i-1} i en sekvens, altså $P(w_i|w_{i-1})$.

8. Maximum likelihood estimering (MLE) (5 poeng)

$$P(w_i|w_{i-1}) = Count(w_{i-1}, w_i)/Count(w_{i-1})$$

9. Glatting (smoothing) (5 poeng)

Å bruke MLE (maximum likelihood estimation) alene kan by på problemer. For språkmodeller utvider vi gjerne MLE med såkalt glatting (smoothing). Forklar kort hvorfor MLE alene byr på problemer. Forklar også generelt hva glatting er, og hvorfor vi bruker det.

9. Glatting (smoothing) (5 poeng)

Opptil 2.5 poeng for hver del:

Største utfordring med MLE er at vi får upålitelige estimater for ord og n-grammer med lav frekvens, og spesielt for hendelser med frekvens 0 (som fører til at hele produktet blir 0).

Uavhengig av korpusets størrelse, vil det alltid finnes både ord og sekvenser vi ikke har sett før (jf. språkets kreativitet / produktivitet).

Glatting er metoder for omfordeling av sannsynlighetsmassen for å unngå dette og sørge for at alle n-grammer får frekvens eller sannsynlighet > 0 (ta fra de rike og gi til de fattige).

Et eksempel på dette er "Add-one smoothing":
$P(w_n|w_{n-1}) = C(w_{n-1}, w_n) + 1/C(w_{n-1}) + V$, der V er antall ordtyper (vokabulæret).

10.Metodologiske paradigmer (5 poeng)

I språkteknologi kan vi grovt skille mellom regelbaserte metoder og empiriske metoder basert på maskinlæring. Forklar med noen få setninger hva som karakteriserer forskjellene.

10.Metodologiske paradigmer (5 poeng)

Regelbaserte metoder: bruker manuelt definerte regler og kunnskap for å f.eks. tildele ord riktig tagg i en gitt kontekst. Håndkoding av kunnskap av en mennesklig ekspert.

Statistiske metoder: beregner en statistisk modell fra data, ofte på grunnlag av et manuelt tagget korpus ("treningskorpus") og ved bruk av maskinlæring. Automatisk tilegning av kunnskap fra data av en algoritme.

11. Lingvistiske nivåer (5 poeng)

Hvilke nivåer av analyse opererer vi med i lingvistikk? Forklar også helt kort (stikkord er nok) hva som er fokuset for analyse på hvert nivå.

11. Lingvistiske nivåer (5 poeng)

1. Fonetikk/fonologi: lyder => ord
2. Morfologi: morfemer => ord, hvordan ord er bygd opp, hvordan ord bøyes, hvordan ord dannes, og hvordan ord deles i ordklasser.
3. Syntaks: studiet av prinsipper og regler for setningsdannelse.
4. Semantikk: Studiet av betydningen til språklige uttrykk
5. Pragmatikk: Studiet av vår språklige forståelse i kontekst.

12. Named Entity Recognition (5 poeng)

Named Entity Recognition (NER) er automatisk identifisering og kategorisering av egennavn. Forklar kort hvorfor oppslag i en navneliste alene er ikke den beste måten å gjøre NER på. Du kan bruke følgende setning for inspirasjon:

```
Flyet lander på John F. Kennedy i kveld.
```

12. Named Entity Recognition (5 poeng)

Får problemer ved flertydighet. Tar ikke hensyn til kontekst. Eksempel: Samme navn kan referere til entiteter av forskjellig type (JFK – person eller flyplass).

13. Språkmodeller (5 poeng)

Hva er en språkmodell? Forklar med noen få setninger. Nevn også noen eksempler på anvendelser av slike modeller.

13. Språkmodeller (5 poeng)

En statistisk språkmodell bruker korpusfrekvenser for å beregne sannsynligheten for sekvenser av ord. En slik modell tilskriver sannsynligheter P(x) til alle ordsekvenser x i et språk L.

Eksempler på anvendelser av slike modeller:

- Talegjenkjenning
- Maskinoversettelse
- Håndskriftgjenkjenning og OCR
- Prediksjon / automatisk fullføring
- Ordklassetagging, stavekontroll, tekstklassifikasjon, generering, og masse annet