

INF1820 V2014 — Oppgave 3b

CFGer og semantikk

Innleveringsfrist fredag 9. mai

Lever inn svarene dine i en fil som angir brukernavnet ditt, slik: `oblig3b_brukernavn.py`. En perfekt besvarelse av denne oppgaven er verdt 100 poeng.

1 Et grammatikkfragment for engelsk (40 poeng)

Engelsk har samsvarsbøyning mellom subjekt og verb: subjekter i tredje person entall krever en spesiell form av verbet (*she likes, it likes* versus *I like, they like*). Engelsk har også, som norsk, noe kasusmarkering for pronomen. For eksempel, *I, he, they* er i nominativ kasus, mens *me, him, them* er i akkusativt kasus. I denne oppgaven skal du skrive en kontekstfri grammatikk som aksepterer setningene i (a), men ikke setningene i (b):

1. Setninger grammatikken skal akseptere:
 - (a) she likes him
 - (b) they like him
 - (c) they like the actress
 - (d) they like the actresses
 - (e) I know her
 - (f) I know the actor whom she likes
 - (g) I know the actress who likes him
 - (h) the actress who likes him sings
 - (i) the actresses who like him sing
2. Setninger grammatikken *ikke* skal akseptere:

- (a) *she likes he
- (b) *they likes him
- (c) *me know her
- (d) *I know the actress whom likes him
- (e) *the actress who likes him sing
- (f) *the actress who like him sings
- (g) *the actresses who likes him sing
- (h) *the actresses who like him sings

NLTK inneholder flere forskjellige *parsere* som tildeler syntaktisk struktur til en setning automatisk, i henhold til en grammatikk. Her skal du bruke `RecursiveDescent`-parseren som står beskrevet i seksjon 8.3. Du kan formulere grammatikken din direkte som en streng, slik:

```
grammar = nltk.parse_cfg("""
S -> NP VP
VP -> V NP | V NP PP
PP -> P NP
V -> "saw" | "ate" | "walked"
NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
Det -> "a" | "an" | "the" | "my"
N -> "man" | "dog" | "cat" | "telescope" | "park"
P -> "in" | "on" | "by" | "with"
""")
```

Merk deg at flere alternativer for samme kategori skal formuleres med disjunksjon `VP -> V NP | V NP PP`, dvs. at VP enten kan bestå av en `V NP` eller `V NP PP`. Merk deg videre at `RecursiveDescent`-parseren ikke håndterer venstre-rekursjon av typen `VP -> VP PP`, så du må formulere grammatikken din uten denne formen for rekursjon.

Du kan teste grammatikken din på en setning slik:

```
sent = "Mary saw Bob".split()
rd_parser = nltk.RecursiveDescentParser(grammar)
for tree in rd_parser.nbest_parse(sent):
    print tree
```

Parseren skriver da ut et tre i klammenotasjon:

```
(S (NP Mary) (VP (V saw) (NP Bob)))
```

Implementer grammatikkfragmentet for engelsk som beskrevet over, test grammatikken din på setningene i (a) og (b) ovenfor og skriv ut resultatet. Pass på at grammatikken din faktisk generer et tre for alle setningene i (a) og ikke genererer noe tre for setningene i (b).

2 Manuell annotering av ordbetydning (25 poeng)

I denne oppgaven skal du gjøre en manuell annotering av ordbetydning og kommentere observasjonene dine. Skriv svarene dine som utkommentert tekst i Python-filen din.

Setningene i (a)-(j) under er hentet fra SemCor-korpuset, et korpus som er annotert med ordbetydning, og alle inneholder verbet *leave*.

1. But questions with which committee members taunted bankers appearing as witnesses **left** little doubt that they will recommend passage of it .
2. The departure of the Giants and the Dodgers to California **left** New York with only the Yankees .
3. After the coach listed all the boy 's faults , Hartweger said , “ Coach before I **leave** here , you 'll get to like me ” .
4. R. H. S. Crossman , M.P. , writing in The Manchester Guardian , states that departures from West Berlin are now running at the rate not of 700 , but of 1700 a week , and applications to **leave** have risen to 1900 a week .
5. The house has been swept so clean that contemporary man has been **left** with no means , or at best with wholly inadequate means , for dealing with his experience of spirit .
6. A second and also good practice is to shear off the tops , **leaving** an inch high stub with just a leaf or two on each branch .
7. No doubt some experiences vanish so completely as to **leave** no trace on the sleeper 's mind .
8. He is a widower , his three children are dead , he has no one **left** on earth ; also he is a drunk , and has lost his job on that account
9. Piepsam tries to stop him by force , receives a push in the chest from “ Life ” , and is **left** standing in impotent and growing rage , while a crowd begins to gather .
10. The audience **leaves** the play under a spell , It is the kind of spell which the exposure to spirit in its living active manifestation always evokes .

Slå opp verbet *leave* i WordNet (bruk “Use WordNet online”: <http://wordnetweb.princeton.edu/perl/webwn>). Du skal ikke ta hensyn til betydningene for substantivet *leave*.

For hver av setningene i (a)-(j) skal du velge *en* betydning (“sense”) fra WordNet for verbet *leave* i setningen og notere valget ditt. På “Use WordNet online”-siden kan du klikke på *Display Options* og velge *Show Sense Numbers* for å få en nummerert oversikt over de forskjellige betydningene. Bruk disse nummerene i svaret ditt.

Videre skal du reflektere rundt arbeidet ditt og besvare følgende spørsmål:

- Hvilke setninger var det vanskelig å annotere og hvorfor?
- Hvilke par (eller grupperinger) av WordNet-betydninger var det vanskelig å skille fra hverandre og hvilke kriterier brukte du for å skille mellom dem?

3 Betydningsdisambiguering (WSD) med en Naive Bayes-klassifiserer (35 poeng)

Filen `wsd_tren.txt` inneholder (fiktive) data annotert med ordbetydning for lemmaet *skim*. Hver linje inneholder en liste med trekk og en kategori. Elementene i hver linje er adskilt med mellomrom. Første element i hver linje er kategorien og de andre elementene er trekk. Første linje ser slik ut:

```
Reading book day novel
```

Dette betyr at betydningskategorien for denne instansen er `READING` og inneholder trekkene *book*, *day* og *novel*.

1. Bruk treningsdataene i `wsd_tren.txt` til å beregne sannsynligheten for betydningen `REMOVING`: beregn sannsynligheten $P(\text{Removing})$ ved Maximum Likelihood Estimation (MLE) fra dataene. Bruk Python til å utføre beregningene dine.
2. Ett av trekkene som forekommer i treningsfilen er *day*. Beregn sannsynligheten for dette trekket, gitt `READING`-betydningen, dvs $P(\text{day}|\text{Reading})$. Bruk MLE over treningsdataene og bruk Python til å utføre beregningene dine.

3. Filen `wsd_test.txt` inneholder en testinstans på samme format som treningsdataene, bortsett fra at kategorien er ukjent:

? paper surface towards

Bruk Naive Bayes-formelen for å beregne den mest sannsynlige betydningen for denne testinstansen. Bruk Python til å utføre beregningene dine.

Husk at i Naive Bayes er den mest sannsynlige betydningen \hat{s} gitt ved:

$$\hat{s} = \operatorname{argmax}_{s \in S} P(s) \prod_{j=1}^n P(f_j | s)$$

Du kan lese mer om bruk av Naive Bayes for WSD-oppgaven i Jurafsky & Martin 20.2.