IN1140: Introduksjon til språkteknologi

Forelesning #6

Sondre Wold

Universitetet i Oslo

25 september 2023



Agenda

Forrige uke

- Språkmodeller
- n-gram

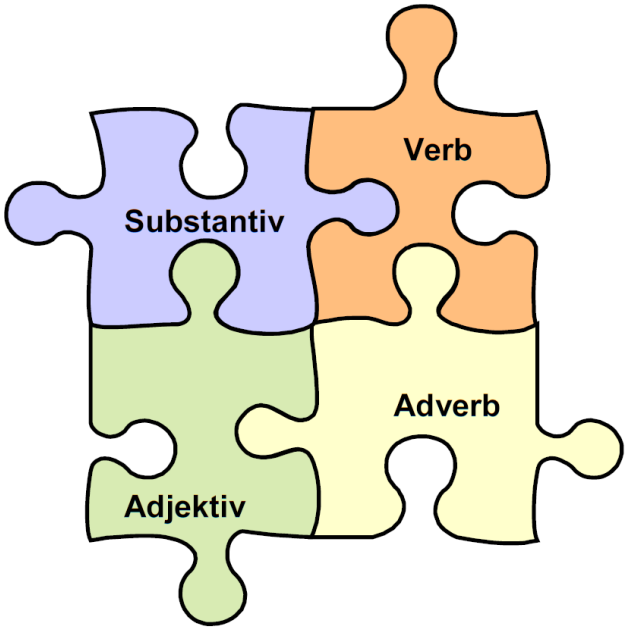
I dag

- Ordklasser
- Ordklassetaggede korpora
- Ordklassetagging

Ordklasser?

- Parts-of-speech (PoS)
- Bindeledd mellom ordet og setningen (syntaks):
  - Sier noe om hva slags kontekster et ord forekommer i
  - Sier noe om uttale (*content*)
- Viktige i en rekke språkteknologiske applikasjoner:
  - Talesyntese
  - Morfologisk analyse
  - "Chunking", syntaktisk parsing
  - Word Sense Disambiguation
  - Informasjonsekstraksjon

Ordklasser



- **Taksonomi** - et system som har kategorier som er uttømmende, gjensidig utelukkende, styrt av et prinsipp
- Alle ord havner i en klasse og ingen ord havner i mer enn én klasse
- Vi trenger **kriterier** for ordklasseinndeling

Ordklassekriterier

3 slags kriterier:

1. **Formelle eller morfologiske kriterier**
   - Hvilke bøyningsformer har ordet?
     - *hare* - *haren* og *redd* - *reddere*
     - **harare* og **redden*

2. **Funksjonelle eller syntaktiske kriterier**
   - Hvordan kan ordet kombineres med andre ord?
     - *en hare*, *en redd hare* og *redd for ilden*
     - **en redd* og **hare for ilden*

3. **Betydningsmessige eller semantiske kriterier**
   - Hva er typiske betydninger hos ord i ordklassen?
     - *hare* - dyr, levende vesen
     - *redd* - egenskap

Ordklassekriterier

Ordet RØD:

- form?
- funksjon?
- betydning?

Ordklassekriterier

Adjektivet RØD:

- form: *rød, rødt (bøyning etter kjønn), røde (bestemt), rødere (komparativ form), rødest (superlativ)*
- funksjon: *et rødt eple* (attributiv funksjon – attributter et egenskap), *Håret hennes er rødt* (predikativ funksjon – en identitesmarkør)
- betydning: betegner en egenskap, typisk for adjektiv
- MEN: *De røde tapte borgerkrigen*??
  - Adjektivet er i en substantiv kontekst: unntak.

- Fokusere på:
  - Vanligste bruken
  - Vekting av kriteriene

Ordklasser: substantiv

Substantiv – *olje, bord, jente, sorg*

1. Bøyes i bestemthet og tall
   - Bestemthet: kan knytte til seg bestemt artikkel som suffiks: *bilen, greina, huset, tanken*, *bordet*
   - Tall: (de fleste har) forskjellige endelser for entall og flertall: *bil-biler, grein-greiner, tanke-tanker*, bord-~~*border*~~

2. Kjerne i substantivfraser, med modifikatorer: *en alldeles fantastisk vakker stol*

3. Betegner "ting" - mennesker, objekter, vesen, steder, fenomener og abstrakte enheter

Unntak – egennavn: bøyes ikke

Ordklasser: verb

Verb (hovedverb) – *sparke, sove, håpe, arbeide, bygge, leve*

1. Bøyes i tid (presens-preteritum)
   inndeles i finitte vs. infinitte former
   - Finitte (kan stå alene i en setning): imperativ, presens, preteritum: *spark, sparker, sparket*
   - Infinitte (kan ikke stå alene, de trenger en fuknsjonell markør/hjelpeverb): infinitiv, perfektum partisipp *(å) sparke, (ha) sparket*

   Transitivitet: transitiv (krever to argumenter: subjekt og objekt) - intransitiv (tar ikke objekt, krever ikke et argument)

2. Kan stå alene som predikat

3. Betegner handlinger, aktiviteter og tilstander

Unntak – hjelpeverb (forekommer typisk med et innholdsverb): *må, skal, bli*

Ordklasser: verb

Hjelpeverb (fra Store norske leksikon):

Hjelpeverb er verb som vanligvis står sammen med infinitiv eller partisipp av et annet verb, kalt hovedverbet, for å uttrykke grammatiske kategorier (perfektum, passiv, futurum og så videre).

I setningen *Hun har gått* er *har* hjelpeverb og *gått* hovedverb.
I setningen *Han må spise* er *må* hjelpeverb og *spise* hovedverb.

Noen norske hjelpeverb er *ha*, *bli*, *være*, *skulle*, *ville*, *kunne*, *måtte* og *burde*.

Ordklasser: adjektiv

Adjektiv – *rød, snill, vanskelig, levende*

1. Samsvarsbøyes i bestemthet, kjønn og tall, *gradbøyes* (*rød, rødere, rødest, interessant, mer interessant, mest interessant*)

2. Modifikator (adledd) til substantiv

3. Betegner egenskaper

- Gradbøyes ved bøyningsendelse eller mer–mest. Betydningen angir et punkt på en skala (feks. ung–gammel / men ikke levende–død).
- Men mange adjektiv har en mer presis betydning som er vanskelig å gradere, f.eks. *død, gift, gratis, nybakt, lovlig*
- Noen av de mest sentrale adjektivene opptrer i par med motsatt betydning **antonymer**: *høy – lav*, *stor – liten*

Ordklasser: adverb

Adverb – *her, ofte, derfor, trolig, ikke, kanskje, nå, vanligvis*

1. Kan være avledet av adjektiv (*raskt*) og kan gradbøyes (*spise fortere*) men mange er også ubøyelige (*ikke, da*).

2. Står som modifikatorer til verb, adjektiv, adverb og setninger

3. Betegner forskjellige omstendigheter - rom, tid, måte m.m.

- **Tidsadverb** uttrykker relativ tid, dvs. et tidspunkt i forhold til et annet
  - *Han kom etterpå* (etter et tidspunkt i fortiden)
  - *Du skal komme da* (på et omtalt tidspunkt i framtiden)
- **Måtesadverb** uttrykker måten noe blir gjort på
  - *Hun gjennomgikk pensum* **stykkevis**
- **Gradsadverb** uttrykker mengde, intensitet eller grad ved verbhandlingen
  - *Jeg fryser* **litt**
  - *Nå har du tullet* **nok**

Ordklasser: preposisjoner

Preposisjoner: Funksjonsord – *ved, på, under, i, foran, av*

1. Ubøyelige

2. Kjerne i preposisjonsfraser, tar substantiv

3. Betegner relasjoner, f.eks.:
   - **rom** og **tid**
     - *Hytta ligger* **ved** *sjøen*
     - *Vi drar* **i** *mai*
     - *Taket* **på** *huset ble nettopp reparert*
   - **måte** eller **middel**
     - *Hun satt* **i** *dype tanker*
     - *Hun åpnet døren* **med** *en rusten nøkkel*
   - **verbalpartikkel**
     - *De sovnet* **inn**
     - *Han brøt* **sammen** *etter løpet*

Ordklasser: pronomen

Pronomen – *jeg, hun, dere, seg, hverandre, hvem, man*

1. Av svært ulik form, uregelmessig bøyning

2. Som substantiv, kan fungere som setningsledd alene

3. Lite eget innhold, får betydning fra sammenhengen (*konteksten*)
   - ***Jeg*** *liker grammatikk*
   - ***Man*** *skal respektere hverandre*
   - **Hvem** *tok vesken?*

- **Personlige pronomen**: (*jeg* – *meg*), (*vi* – *oss*)
- **Refleksivt pronomen**: *seg*. Har antesedent i samme setning, oftest subjektet i setningen
- **Resiproke pronomen**: *hverandre*. Uttrykker en gjensidig relasjon.
- **Interrogative pronomen**: spørreord (hvem, hva).

Ordklasser: determinativ

Determinativ (artikler) – *min, din, denne, alle, noen*

1. Bøyning i kjønn og tall (min bil, mitt hus)
2. Bestemmer til substantiv
3. Bestemmer, spesifiserer substantivets referanse

3 hovedtyper:

a) Possessiver: angir eiendom eller tilhørighet, bøyes i person. (*Det er* **min** *bok*)

b) Demonstrativer: viser til eller peker på en bestemt person eller ting som kan iakttas eller er omtalt. (**Den** *hytta ligger fint til*)

c) Kvantorer: uttrykker mengde eller kvantitet, noen med bøyning (*noen, ingen, en*) og noen uten (*to, tre, visse, enkelte, utallige*). (*Hun har spist opp* **all** *maten*, *Ida har kjøpt* **noen** *bøker*)

Ordklasser: Konjunksjoner

Konjunksjoner – *og, eller, men, for, så*

1. Ubøyelige

2. Binder sammen ledd av samme slag, f.eks. ord, fraser og setninger

3. Grammatisk funksjon, betegner relasjoner
   - *Fullstendig ro* **og** *absolutt trygghet* (nominalfrase og nominalfrase)
   - *Konkret* **eller** *abstrakt betydning* (adjektivfrase og adjektivfrase)
   - *Han var på ski* **og** *hun var i kirken* (setning og setning)

Ordklasser: subjunksjoner

Subjunksjoner – *å, at, om, som, før*

1. Ubøyelige

2. Innleder leddsetninger eller infinitivskonstruksjoner - underordner en setning under en annen

3. Grammatisk funksjon, betegner relasjoner
   - *Hun elsker* **å** *danse*
   - *Vi tror* **at** *det verste snart er over*
   - *Der er hunden* **som** *spiste kaken*



Quiz!

Inndeling av ordklasser

- **åpne** vs. **lukkede** ordklasser
  - **Åpne**: substantiv, verb og adjektiv
    inneholder mange tusen ord, kan enkelt fylle på med nye
    Eksempel: nye bilmodeller - nye farger (*brannbilrød*)
  - **Lukkede**: inneholder mange færre ord enn de åpne
    kan ikke fritt skape nye ord gjennom orddannelse (pronomen)
- **Innholdsord** vs. **funksjonsord**
  - **Innholdsord**: substantiv, verb, adjektiv
    rikt betydningsinnhold,
  - **Funksjonsord**: mer allment betydningsinnhold. Finnes fremst i de lukkede ordklassene.
- Ikke helt én-til-én, feks hjelpeverb.

Språklige data: korpusdata

- Modellere språklig kunnskap $\rightarrow$ trenger språklige data
- Språkteknologi: programmer som generaliserer over språklige mønstre
  - Korpusdata helt sentralt
- Et korpus (tekstkorpus) er en strukturert samling av tekster
- Elektronisk lagret

Ordklassetaggede korpuser

- Brown-korpuset for engelsk (1979):
  - 87 ordklassetagger
  - 1 mill. ord, utvalg fra 500 tekster hentet fra forskjellige sjangere
  - Automatisk tagget og manuelt rettet
- Penn Treebank (1993)
  - 45 ordklassetagger
  - Wall Street Journal (1 mill. ord), Brown-korpuset (tagget versjon), Switchboard, ATIS (sample)
  - Ordklassetagger, syntaktisk struktur (trær som representerer frasestruktur)

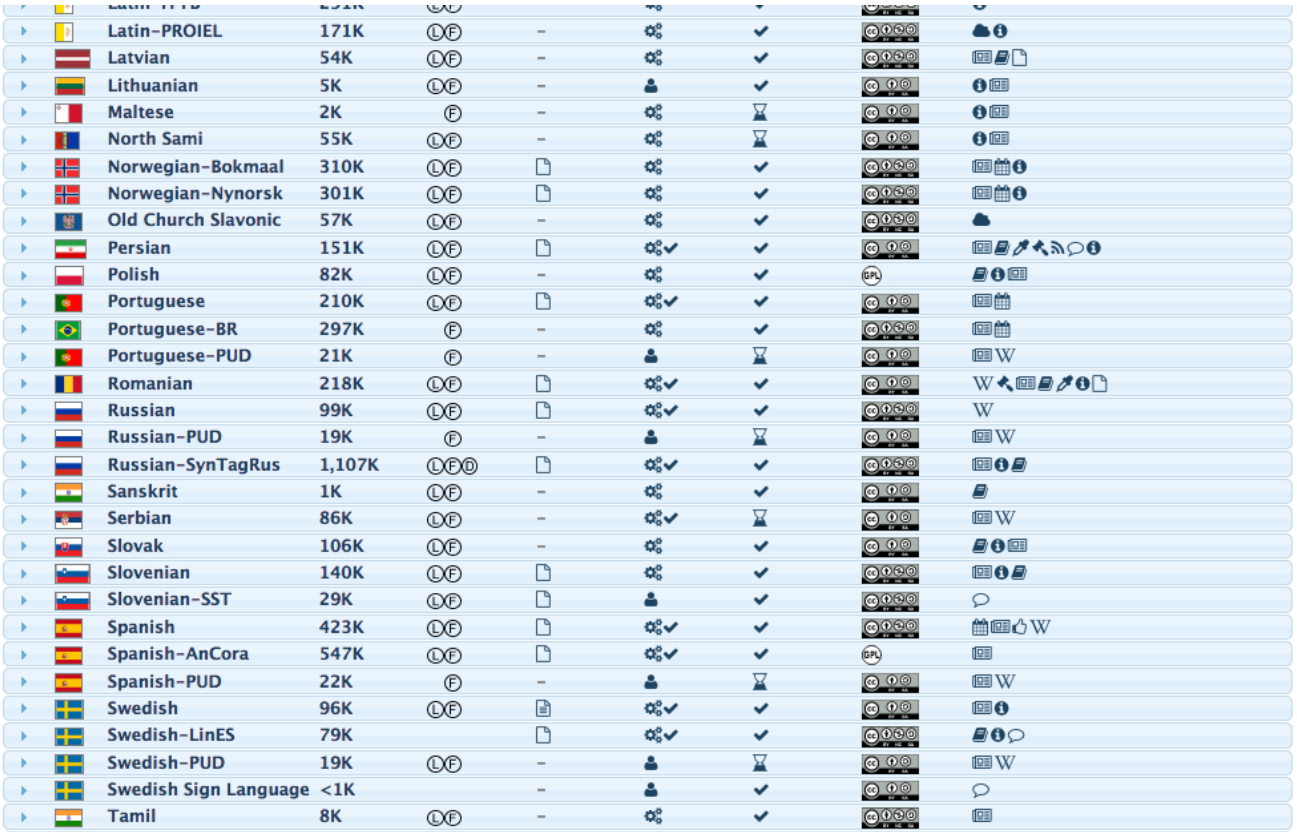
Ordklassetaggede korpuser

- Norsk dependenstrebank (2014)
  - Trebank for norsk
  - Utviklet ved Nasjonalbiblioteket
  - Manuelt tagget (lingvister, 2år)
- Ordklasser samt mye morfologisk informasjon

1	Det	det	pron	nøyt|ent|pers|3
2	er	være	verb	pres
3	hun	hun	pron	fem|ent|pers|hum|3|nom
4	som	som	sbu	_
5	eier	eie	verb	pres
6	og	og	konj	_
7	driver	drive	verb	pres
8	stedet	sted	subst	appell|nøyt|be|ent

Universal Dependencies (UD)

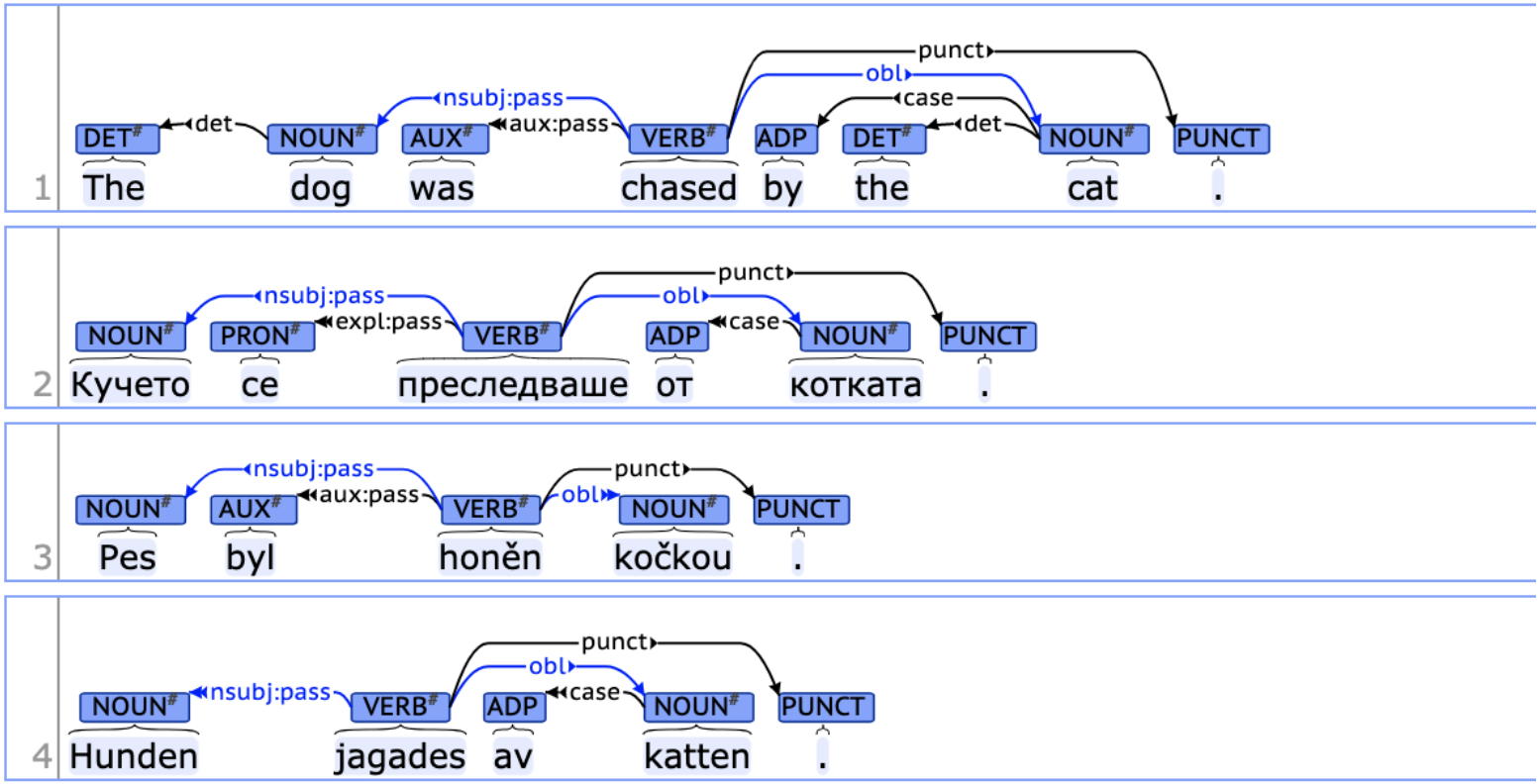
- Annoterte korpuser (trebanker) for mer enn 100 språk (inkludert norsk)
- `http://universaldependencies.org/`



Latin-PROIEL	171K	LF	–	✓	CC BY NC SA	
Latvian	54K	LF	–	✓	CC BY NC SA	
Lithuanian	5K	LF	–	✓	CC BY SA	
Maltese	2K	F	–	⌛	CC BY SA	
North Sami	55K	LF	–	⌛	CC BY SA	
Norwegian-Bokmaal	310K	LF		✓	CC BY NC SA	
Norwegian-Nynorsk	301K	LF		✓	CC BY NC SA	
Old Church Slavonic	57K	LF	–	✓	CC BY NC SA	
Persian	151K	LF		✓	✓	CC BY NC SA	
Polish	82K	LF	–	✓	GPL	
Portuguese	210K	LF		✓	✓	CC BY SA	
Portuguese-BR	297K	F	–	✓	CC BY SA	
Portuguese-PUD	21K	F	–	⌛	CC BY SA	W
Romanian	218K	LF		✓	✓	CC BY NC SA	W
Russian	99K	LF		✓	✓	CC BY NC SA	W
Russian-PUD	19K	F	–	⌛	CC BY SA	W
Russian-SynTagRus	1,107K	LFD		✓	✓	CC BY NC SA	
Sanskrit	1K	LF	–	✓	CC BY NC SA	
Serbian	86K	LF	–	✓	⌛	CC BY NC SA	W
Slovak	106K	LF	–	✓	CC BY NC SA	
Slovenian	140K	LF		✓	✓	CC BY NC SA	
Slovenian-SST	29K	LF		✓	✓	CC BY NC SA	
Spanish	423K	LF		✓	✓	CC BY NC SA	W
Spanish-AnCora	547K	LF		✓	✓	GPL	
Spanish-PUD	22K	F	–	⌛	CC BY SA	W
Swedish	96K	LF		✓	✓	CC BY SA	
Swedish-LinES	79K			✓	✓	CC BY NC SA	
Swedish-PUD	19K	LF	–	⌛	CC BY SA	W
Swedish Sign Language	<1K		–	✓	CC BY NC SA	
Tamil	8K	LF	–	✓	CC BY NC SA	

UD-ordklasser

- Universelle ordklasser: 17 stk



Ordklassetagging

- Tagging følger en manual
- Noen avgjørelser er vanskelige
- Eks: skillet mellom preposisjoner (IN) og adverb (RB)
  - All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB **around**/IN the/DT corner/NN
  - Chateau/NNP Petrus/NNP costs/VBZ **around**/RB 250/CD
- Manualen: preposisjoner er assosiert med en etterfølgende subtstantivfrase. *Around* tagges som adverb i betydningen 'omtrent'

Ordklassetagging

- Oppmerking av ordklasseinformasjon for hvert ord i et korpus
- Språkteknologi: automatiske systemer
- Flertydighet vanskeliggjør dette betydelig
- Ordnivå: Tokenisering

Tokenisering

Tokenisering: dele inn en tekst i ord og setninger. Tidligere har vi gjort det enkelt og bare splittet på mellomrom. Men dette er problematisk:

- Tar ikke hensyn til tegnsetting og gir "ord" som *cents. said, positive."* *Crazy?*.
- Tegnsetting forekommer også innad i ord: *m.p.h. cap'n, AT&T*.
- Tall kan inneholde komma: 555,000

Tokenisering: slik det kan gjøres

Byte-Pair Encoding: En algoritme for tokenisering som er mye brukt de siste årene:

- Ønsker å bryte opp ord i mindre delord (subwords)
- Slår iterativt sammen frekvente sekvenser av ord/tegn til nye ord helt til vi får et vokabular vi er fornøyd med.
- Tekst kan så representeres med disse delordene, slik at frekvente ord har sin egen "token", mens mindre frekvente ord kan kombineres av mindre delord.

Tokenisering: Byte Pair Encoding

Eksempel: Korpus: [“ab”, “bc”, “bcd”, and “cde”].

1. Vokabular = {"a", "b", "c", "d", "e"}
2. Frekvenser = {"a": 1, "b": 2, "c": 3, "d": 2, "e": 1}
3. Hvilket par fra vokabularet vårt har høyest frekvens i korpuset?
   - "bc" forekommer to ganger
4. Slår sammen "b" og "c" til "bc" og legger det til i frekvenstabellen og vokabularet
   - F = {"a": 1, "b": 2, "c": 3, "d": 2, "e": 1, "bc": 2}
   - V = {"a", "b", "c", "d", "e", "bc"}
5. Gjenta fra steg 1

Tekst kan nå representeres med:

- "cde" -> "c" + "de"
- "bcd" -> "bc" + "d"

Ordklassetagging

- Input: streng av ord og en spesifisert mengde tagger
- Output: en tagg per ord

Jeg	*vil*	*drikke*	*kaffe*	*nå*
`pron`	`verb`	`verb`	`subst`	`adv`

- Flertydigheter?

Flertydighet

Jeg	*vil*	*drikke*	*kaffe*	*nå*
`pron`	`verb`	`verb|subst`	`subst`	`verb|adv`

- Tall fra det engelske Brown-korpuset:
  - 12% av ordtypene er flertydige
  - 40% av tokens er flertydige
- De *fleste* engelske ord er *entydige*
- Men mange av de mest *frekvente* ordene er *flertydige*
- Heldigvis er ikke alle lesninger like sannsynlige
  - Både isolert sett og i kontekst.

Ordklassetagging: Noen baselines

- Bør alltid først definere en *baseline*:
  Enklest mulige tilnærming til et problem.
- To ulike majoritets-baserte baselines for PoS-tagging:
  1. Tildel alle ord samme tagg; den mest frekvente (NN)
     - ca. 13% korrekt taggede ord (Brown)
  2. Tildel hvert ord dets mest frekvente ordklassetagg.
     - Vi lagrer de 100 mest frekvente ordene og deres tagger
     - Får ca. 46% korrekt taggede ord (Brown)

Algoritmer for ordklassetagging

- To hovedkategorier:

1. **Regelbaserte taggere**:
   - Manuelt definerte regler for å tildele ord riktig tagg i en gitt kontekst.
   - Eksempel: *drikke* er substantiv, og ikke verb, dersom det følger et adjektiv.

2. **Statistiske taggere:**
   - Bruker et (manuelt) ordklassetagget korpus ('treningskorpus') til å beregne en statistisk model for tagging.

Regelbasert tagging

Typisk to trinn, grovt sett:

1. Morfologisk analyse:
   - Hvert ord tildeles en liste av mulige ordklasser og morfologiske trekk.
   - 'Multitagging'
   - To tilnærminger:
     - Fullformsleksikon: Lister med ord i alle bøyninger (løp, løper, løpt, ...), med tilhørende tagger.
     - To-nivå morfologi: morfologisk analyse som mapper fra overflateform til leksem.
2. Entydiggjøring:
   - Håndskrevne regler (gjerne mange tusen) for å disambiguere ordene.
   - *Constraint Grammar* (CG) – sentral regelformalisme som har resultert i taggere for en rekke språk, deriblant engelsk og norsk.

Regelbasert tagging

- Oslo-Bergen taggeren ('OBT'): PoS-tagger for norsk.
- Constraint Grammar (CG)-regler for entydiggjøring.

```
"<som>"  SELECT:3261 (prep) IF
    (1 pron-akk)
    (NOT 1 pron-nom)
;
#  "Ei jente som (prep) meg"))
```

- Utviklet hos Tekstlaboratoriet.
- Kombinert med statistisk entydiggjøring.

Statistisk tagging

- Bruker et ordklassetagget korpus ('treningskorpus') til å beregne den mest sannsynlige sekvensen av tagger for en gitt setning.
- En mye brukt probabilistisk model: Hidden Markov Model (HMM) – Tagging som klassifiseringsoppgave: Gitt en sekvens med ord, hva er den mest sannsynlige taggsekvensen?
- Ser på ordtaggene som *skjulte variabler* (eller 'tilstander') som vi ønsker å predikere basert på de *observerbare variablene*; ordene.
- Nære bånd til n-grammodeller.
- I økende grad: nevrale modeller som brukes til ordklassetagging

Evaluering

- Gitt at vi har utviklet en ordklassetagger,
- hvordan kan vi evaluere den?
- Generelt to strategier for å evaluere en modell:
- Ekstrinsisk og intrinsisk evaluering.

Ekstrinsisk evaluering

- Vi evaluerer modellen 'indirekte' utfra hvordan den påvirker resultatene for en annen oppgave.
- Oppgavedrevet.
- F.eks se hvordan en ordklassetagger påvirker maskinoversettelse, talegjennkjenning, osv.
- Fordel: kan teste modellen i samme kontekst som vi vil bruke den.
- Ulempe: ofte krevende ift tid/ressurser.

Intrinsisk evaluering

- Bruker et mer direkte mål for hvor bra modellen er på oppgaven den ble trent for.
- PoS-tagging: ønsker en modell som predikerer taggene for et testkorpus med høyest nøyaktighet; *accuracy* $= \frac{\#riktig}{\#tokens}$
- Fordel: ofte rask og billig.
- Ulempe: ikke alltid samsvar mellom ekstrinsiske og intrinsiske mål.

Datasplitter

- Dersom vi tester på treningsdataene får vi urealistisk gode resultater sammenliknet med om vi tester på 'nye' data.
- Kalles *overfitting* dersom en model er for spesifikt tilpasset testdataene til å gi representative målinger for hvordan modellen generaliserer til usette data.
- Trenger minst to datasett: treningsdata og testdata.
- Bruker ofte også en tredje splitt: valideringsdata (*development data*).
- Viktig at datasplittene er balanserte og representative:
- F.eks samme sjanger, domene, osv.

Natural Language ToolKit (NLTK)

- Python-rammeverk for NLP
- Enkel tilgang til en rekke korpuser
- Biblioteker for
  - tokenisering
  - tagging
  - parsing
  - tekstklassifisering
  - ...

Natural Language ToolKit (NLTK)

Fra http://www.nltk.org/book/ch05.html

```
>>> import nltk
>>> from nltk import word_tokenize
>>> text = word_tokenize("And now for something completely
different")
>>> text
['And', 'now', 'for', 'something', 'completely',
'different']
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something',
'NN'), ('completely', 'RB'), ('different', 'JJ')]
```

Natural Language ToolKit (NLTK)

Fra http://www.nltk.org/book/ch05.html

```
>>> text = word_tokenize("They refuse to permit us to
obtain the refugee permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'),
('permit', 'VB'), ('us', 'PRP'), ('to', 'TO'), ('obtain',
'VB'), ('the', 'DT'), ('refugee', 'NN'), ('permit', 'NN')]
```

NLTK

```
>>> nltk.corpus.brown.tagged_words()
[('The', 'AT'), ('Fulton', 'NP-TL'), ...]

>>> nltk.corpus.brown.tagged_words(tagset='universal')
[('The', 'DET'), ('Fulton', 'NOUN'), ...]
```

NLTK

```
>>> from nltk.corpus import brown
>>> brown_news_tagged =
brown.tagged_words(categories='news', tagset='universal')
>>> tag_fd = nltk.FreqDist(tag for (word, tag) in
brown_news_tagged)
>>> tag_fd.most_common()

[('NOUN', 30640), ('VERB', 14399), ('ADP', 12355), ('.',
11928), ('DET', 11389), ('ADJ', 6706), ('ADV', 3349),
('CONJ', 2717), ('PRON', 2535), ('PRT', 2264), ('NUM',
2166), ('X', 106)]
```

Eksempelproblem fra et pågående prosjekt

- Pågående forskningsprosjekt: Hvordan kan man måle hvor vanskelig et ord er, såkalt leksikal kompleksitet?
- Ønsker å produsere et mål som kan brukes til å forenkle tekst, til bruk i psykologisk evaluering.
- Har analysert og evaluert store mengder tekst av ulik vanskelighetsgrad.
- Problem: Kompleksistet er ofte knyttet til eksponering, ergo: vi må basere oss på frekvenser i et korpus.
  - Lukkede ordklasser og stoppord forekommer hele tiden, noe som gjør det vanskelig å beregne hvor ofte et ord forekommer relativt til andre.
  - Løsning: kun se på åpne ordklasser: substantiv, verb, adjektiv, adverb.
  - Hvordan isolere disse?
  - Ordklassetagge!

Neste uke

- Ingen forelesning