IN1140: Introduksjon til språkteknologi

Forelesning #10

Samia Touileb

Universitetet i Oslo

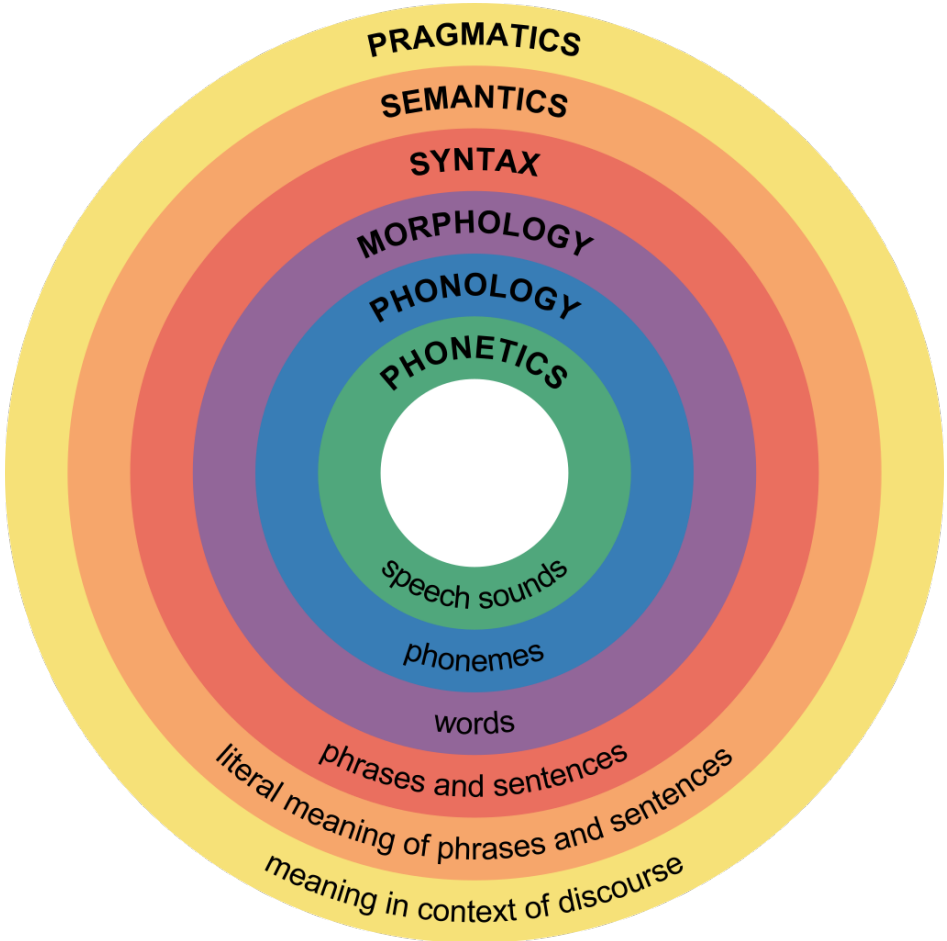
31. oktober 2019

Beregning av poeng og bestått/ikke-bestått for oblig'er

Vi kommer til å operere med en viss margin i beregning av poeng og bestått/ikke bestått per oblig. Som dere vet må dere akkumulere minst 100 poeng for oblig 1 (1a og 1b) for å bestå oppgaven og minst 50 poeng for hver av oblig 2 og oblig 3. Dersom noen mangler poeng på en oblig, vil dere ha mulighet til å hente inn disse på neste oblig. F.eks. dersom dere mangler 10 poeng på oblig 2 (har 40 poeng) må dere få 10 poeng ekstra (dvs minimum 60 poeng) på oblig 3.

November							
Week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
44					1	2	3
45	4	5	6	7	8	9	10
46	11	12	13	14	15	16	17
47	18	19	20	21	22	23	24
48	25	26	27	28	29	30	

Lingvistiske nivåer



Semantikk

Forrige uke:

- Semantikk
- Studiet av betydning slik det uttrykkes gjennom språk
- Betydning til morfemer, ord, fraser og setninger
  - leksikal semantikk
  - setningssemantikk
  - (pragmatikk: hvordan konteksten påvirker betydning)

Semantikk i språkteknologi

Tema for i dag:

- Hva slags oppgaver inngår i semantisk analyse?
- Med hva slags metoder kan disse oppgavene løses?

Ordbetydning ("word sense")

Flertydighet

- They rarely serve red meat, preferring to prepare seafood.
- He served as U.S. ambassador to Norway in 1976 and 1977.
- He might have served his time, come out and led an upstanding life.

Flertydighet

- Which of those flights serve breakfast?
- Does Air France serve Philadelphia?
- ?Does Air France serve breakfast and Philadelphia?

Ordbetydning ("word sense")

Flertydighet

- *The astronomer married the star* – the star
- *You are free to execute your laws, and your citizens, as you see fit* (Star Trek, Next Generation) – execute
- *Oh, flowers are common here, Miss Fairfax, as people are in London* (Oscar Wilde, The Importance of Being Earnest) – common

Semantiske roller

- *The dog bit the mailman* er ikke det samme som *The mailman bit the dog*
- Sammenligne med *The dog was bitten by the mailman*
- Semantiske roller beskriver "hvem som gjør hva mot hvem"
  $[\text{The mailman}]_{\text{AGENT}}$ bit $[\text{the dog}]_{\text{PATIENT}}$

Parafrasering

- *Apple bought Cisco*
- *Apple acquired Cisco*
- *Cisco was taken over by Apple*

Semantiske relasjoner mellom ord

- Synonymer (samme betydning)
  - *ascend/rise*
  - *sweater/pullover*
- Antonymer (motsetninger)
  - *good/evil*
  - *ascend/descend*
- Hypernym/hyponym (mer generell, mer spesifikk)
  - *tree/birch*
  - *animate object / mammal / whale*
- Meronymi (del-helhet)
  - *knob/door*
  - *wheel/car*

Metafor

- Lakoff & Johnson: "Metaphors We Live By". Noen av deres metaforer
  - Argument is War
  - *He shot down all of my arguments*
- Love is a Journey
  - *They hit a bumpy stretch in their relationship*
- time is money
  - *You're wasting my time*
- hvorfor er metaforer problematiske for språkteknologi?

Temporale uttrykk

- *The chairman announced yesterday that they would have the problem solved within three days*
  - Når ble problemet løst dersom setningen forekom i en avis 10 mai, 2007?
- *Book me a flight on the 7:35 tomorrow*

Entailment

- *Apple bought Cisco*
- *Apple acquired Cisco*
- *Cisco was taken over by Apple*

- Spørsmål: *Who bought Cisco?*
- Forventet svar: *Apple bought Cisco*
  - *Cisco's acquisition by Apple* $\rightarrow$ (entails) Apple bought Cisco

Negasjon

- *Apple acquired Cisco*
- *Apple did not acquire Cisco*
- *Apple failed to acquire Cisco*
- *Apple denied not acquiring Cisco*

Negasjon – fra forskningsfronten!

- Automatisk negasjonsanalyse – internasjonal forskningskonkurranse
  - *SEM Shared Task on Negation Resolution
  - system som angir
    - negation **cue**
    - negation scope
    - negated *event*
  - There was **no** *answer*.

Metoder

Metoder

- Klassifisering
  - Gitt et ord i en setning, og en liste av mulige betydninger, velg en betydning (den mest sannsynlige?).
  - Gitt et predikat i en setning, finn dets semantiske roller.
- Automatisk tilegnelse av semantisk informasjon fra rå tekst
  - Hvilke ord og fraser betyr det samme.
  - Distribusjonell semantikk:
    `http://vectors.nlpl.eu/explore/embeddings/en/similar/`
    `https://research.google.com/semantris/`

$\Rightarrow$ **Maskinlæring**

Semantiske ressurser

Semantiske ressurser

- Klassifisering forutsetter treningsdata.
- Leksikalske databaser (WordNet, FrameNet).
- Korpuser annotert med semantisk informasjon (PropBank).

WordNet

- Manuelt konstruert database
- Betydningen til ord karakteriseres gjennom **relasjoner** til andre ord
- Semantiske konsepter karakteriseres gjennom relasjoner til andre konsepter
- Hva slags relasjoner kan det være snakk om?

Relasjoner i WordNet

- Mellom ord:
  - Synonymi (samme betydning).
  - Synonymi-relasjonen grupperer ord i synonymimengder, såkalte **synsets**.
- Mellom konsepter (=synsets)
  - Hypernymi (mer generell, mer spesifikk).
  - Varierer noe, men antonymi og meronymi er også spesifisert for noen synsets.

WordNet

- Elektronisk leksikon
  - Online grensesnitt
  - Lastes ned
  - Tilgjengelig på `http://wordnet.princeton.edu/`
  - Også tilgjengelig via NLTK
- Består av tre separate databaser:
  1. Substantiv (117798 lemmaer)
  2. Verb (11529 lemmaer)
  3. Adjektiv og adverb (22479 adjektiver, 4481 adverb)

WordNet eksempel

- verbet *skim*: synonymer, definisjoner og eksempler

Verb

- S: (v) plane, **skim** (travel on the surface of water)
- S: (v) skim over, **skim** (move or pass swiftly and lightly over the surface of)
- S: (v) scan, **skim**, rake, glance over, run down (examine hastily) *"She scanned the newspaper headlines while waiting for the taxi"*
- S: (v) **skim**, skip, skitter (cause to skip over a surface) *"Skip a stone across the pond"*
- S: (v) **skim** (coat (a liquid) with a layer)
- S: (v) **skim**, skim off, cream off, cream (remove from the surface) *"skim cream from the surface of milk"*
- S: (v) **skim**, skim over (read superficially)

WordNet: substantiver

- Synsets er koblet sammen ved
  - Hyponym/hypernym relasjonen (hovedhierarkiet)
  - Meronymi: del-helhet relasjoner
    - Komponent/del (*leg* – *table*, *finger* – *hand*)
    - Medlem av en gruppe *tree* – *forest*, *student* – *class*
    - Materiale et objekt er laget av (*oxygen* – *water*)
- Ord er koblet sammen ved antonymi

WordNet: substantiv eksempel

- S: (n) **cat**, true cat (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) feline, felid (any of various lithe-bodied roundheaded fissiped mammals, many with retractile claws)
      - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
        - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
          - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
            - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
              - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)

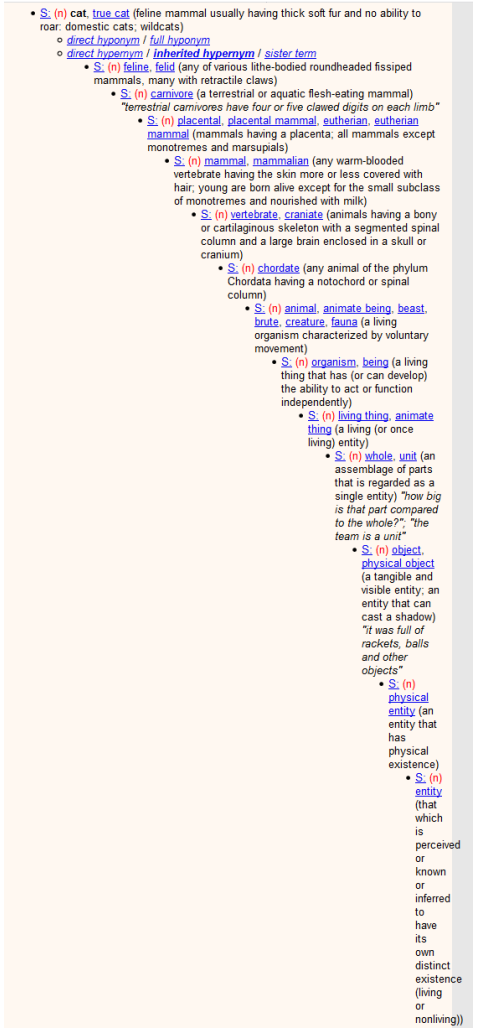
WordNet: substantiv eksempel

- S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
  - S: (n) living thing, animate thing (a living (or once living) entity)
    - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
      - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*

WordNet: substantiv eksempel

- S: (n) physical entity (an entity that has physical existence)
  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

WordNet: substantiv eksempel



WordNet: substantiv – hyponymi, meronymi, holonymi

Noun

- S: (n) **mammal**, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
  - *direct hyponym* / *full hyponym*
    - S: (n) female mammal (animals that nourish their young with milk)
    - S: (n) tusker (any mammal with prominent tusks (especially an elephant or wild boar))
    - S: (n) prototherian (primitive oviparous mammals found only in Australia and Tasmania and New Guinea)
    - S: (n) metatherian (primitive pouched mammals found mainly in Australia and the Americas)
    - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
    - S: (n) fossorial mammal (a burrowing mammal having limbs adapted for digging)
  - *part meronym*
    - S: (n) coat, pelage (growth of hair or wool or fur covering the body of an animal)
    - S: (n) hair, pilus (any of the cylindrical filaments characteristically growing from the epidermis of a mammal) *"there is a hair in my soup"*
  - ***member holonym***
    - S: (n) Mammalia, class Mammalia (warm-blooded vertebrates characterized by mammary glands in the female)

WordNet: verb

- Semantikk for hendelser ('events', verb) er ganske forskjellig fra semantikk for entiteter (substantiver)
- Relasjoner
  - Troponymi:
    - *A verb expressing a specific manner of elaboration of another verb. X is a troponym of Y if to X is to Y in some manner*
    - *wade – walk*
  - Entailment
    - *A verb X entails Y if X cannot be done unless Y is, or has been, done*
    - *walking – stepping*

WordNet: verb eksempel

- S: (v) **jump**, leap, bound, spring (move forward by leaps and bounds) *"The horse bounded across the meadow"; "The child leapt across the puddle"; "Can you jump over the fence?"*
  - *direct troponym* / *full troponym*
    - S: (v) pronk (jump straight up) *"kangaroos pronk"*
    - S: (v) bounce, resile, take a hop, spring, bound, rebound, recoil, reverberate, ricochet (spring back; spring away from an impact) *"The rubber ball bounced"; "These particles do not resile but they unite after they collide"*
    - S: (v) burst (move suddenly, energetically, or violently) *"He burst out of the house into the cool night"*
    - S: (v) bounce (leap suddenly) *"He bounced to his feet"*
    - S: (v) capriole (perform a capriole, of horses in dressage)
    - S: (v) galumph (move around heavily and clumsily) *"the giant tortoises galumphed around in their pen"*
    - S: (v) ski jump (jump on skis)
    - S: (v) saltate (leap or skip, often in dancing) *"These fish swim with a saltating motion"*
    - S: (v) vault (bound vigorously)
    - S: (v) leapfrog (jump across) *"He leapfrogged his classmates"*
    - S: (v) vault, overleap (jump across or leap over (an obstacle))
    - S: (v) curvet (perform a leap where both hind legs come off the ground, of a horse)
    - S: (v) hop, skip, hop-skip (jump lightly)
    - S: (v) caper (jump about playfully)
    - S: (v) hop (make a jump forward or upward)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (v) move (move so as to change position, perform a nontranslational motion) *"He moved his hand slightly to the right"*

WordNet: verb eksempel

Verb

- S: (v) **walk** (use one's feet to advance; advance by steps) *"Walk, don't run!"; "We walked instead of driving"; "She walks with a slight limp"; "The patient cannot walk yet"; "Walk over to the cabinet"*
  - *direct troponym* / *full troponym*
  - *verb group*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - ***entailment***
    - S: (v) step (shift or move by taking a step) *"step back"*
  - *phrasal verb*
  - *antonym*
  - *derivationally related form*
  - *sentence frame*

WordNet: adjektiver

- **Deskriptive** adjektiver
  - Organisert i grupperinger basert på likhet ("similarity"), relatert via antonymi
- **Relasjonelle** adjektiver, linket til sine substantiver
  - *nasal – nose*
- Liten gruppe adjektiver "reference-modifying"
  - *former, alleged*

WordNet: adjektiv eksempel

Adjective

- S: (adj) **cheap**, inexpensive (relatively low in price or charging low prices) *"it would have been cheap at twice the price"; "inexpensive family restaurants"*
  - *similar to*
    - S: (adj) bargain-priced, cut-rate, cut-price (costing less than standard price) *"buying bargain-priced clothes for the children"; "cut-rate goods"*
    - S: (adj) catchpenny (designed to sell quickly without concern for quality) *"catchpenny ornaments"*
    - S: (adj) dirt cheap (very cheap) *"a dirt cheap property"*
    - S: (adj) low-budget (made on or suited to a limited budget) *"a low-budget movie"; "a low-budget menu"*
    - S: (adj) low-cost, low-priced, affordable (that you have the financial means for) *"low-cost housing"*
    - S: (adj) nickel-and-dime (low-paying) *"a nickel-and-dime job"*
    - S: (adj) sixpenny, threepenny, twopenny, tuppeny, two-a-penny, twopenny-halfpenny (of trifling worth)
  - ***antonym***
    - W: (adj) expensive [Opposed to: cheap] (high in price or charging high prices) *"expensive clothes"; "an expensive shop"*

WordNet og polysemi

- Hvor mange betydninger har et ord?
  - Antall synsets ordet forekommer i
- Nærhet i betydning kan utledes fra nærhet i hierarkiet
  - Korteste stien via hyponym/hypernym-linkene mellom synsets

WordNet i språkteknologi

- Utgangspunkt for Word Sense Disambiguation
  - Merke forekomster av et ord med riktig betydning (=synset)
- Generaliserer over ord via hypernymi-relasjonen
  - Fra *cat* til *living being*
- Generalisere over synonymer
- ...

Semantiske roller

- Aspekt ved setningsbetydning: hvilke roller de forskjellige deltagerene inntar
  - *Nina hevet bilen med jekken*
  - Nina – deltageren som er ansvarlig for å utføre handlingen beskrevet av verbet
  - bilen – blir påvirket av handlingen
  - jekken – middelet som Gina bruker til å utføre handlingen
- Semantiske roller beskriver den semantiske relasjonen som argumenter har til handlingen beskrevet av verbet

Semantiske roller

- eksempelet:

Nina	*hevet*	*bilen*	*med*	*jekken*
AGENT		THEME		INSTRUMENT

Problemer med semantiske roller

- Ikke full enighet rundt rolleinventaret
- Vanskelig å formulere formelle definisjoner av roller
- $\Rightarrow$ generaliserte semantiske roller
  - PROTO-AGENT, PROTO-PATIENT
- Verbspesifikke roller
- Semantiske ressurser med informasjon om semantiske roller: **PropBank** og FrameNet

PropBank

- Inneholder alle setningene i Penn Treebank
- Roller er (stort sett) verbspesifikke
  - Arg0, Arg1 = PROTO-AGENT, PROTO-PATIENT
  - Arg2 . . . verbspesifikke

agree.01	
Arg0	Agreer
Arg1	Proposition
Arg2	Other entity agreeing
Ex1	$[_{Arg0}$ *The group*] *agreed* $[_{Arg1}$ *it wouldn't make an offer*]
Ex2	$[_{ArgM-TMP}$ *Usually*] $[_{Arg0}$ *John*] *agrees* $[_{Arg2}$ *with Mary*] $[_{Arg1}$ *on everything*]

PropBank

- Applikasjon: Semantic Role Labeling
- Gitt et predikat i en setning, finn dets semantiske roller
- Gir oss en felles representasjon for:
  - $[_{Arg0}$Big Fruit Co.] increased $[_{Arg1}$the price of bananas]
  - $[_{Arg1}$The price of bananas] was increased again by $[_{Arg0}$Big Fruit Co.]
  - $[_{Arg1}$The price of bananas] increased $[_{Arg2}$5%]

  "Big Fruit Co." er alltid *AGENT* og "the price of bananas" er alltid *PATIENT*

FrameNet

- Roller er “frame-spesifikke”
- Frame air travel = *reservation*, *flight*, *travel*, *buy*, *price*, *cost*, *fare*, *rates*, *meal*, *plane*

Ex1	$[_{Arg1}$ *The price of bananas*$]$ *increased* $[_{Arg2}$ *5%* $]$
Ex2	$[_{Arg1}$ *The price of bananas*$]$ *rose* $[_{Arg2}$ *5%* $]$
Ex3	*There has been a* $[_{Arg2}$ *5%*$]$ *rise* $[_{Arg1}$ *in the price of bananas* $]$

Semantisk klassifisering (WSD)

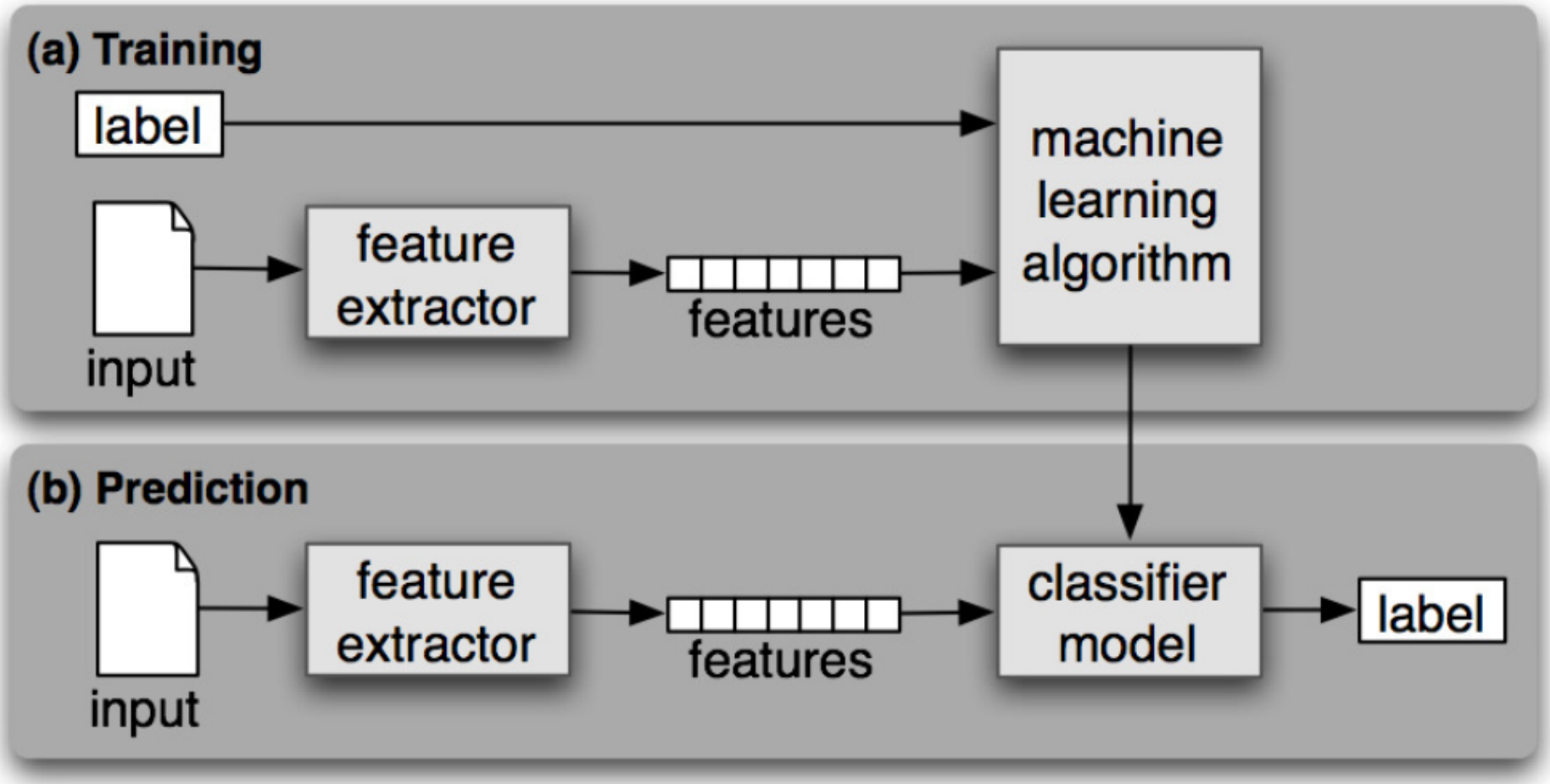
Betydningsdisambiguering

- Word Sense Disambigzation (WSD) – akivt felt innenfor språkteknologi
  - gitt en setning med et spesifikt målord ("target word") og en liste med betydninger (f.eks. fra WordNet)
  - angi korrekt betydning for målordet i den setningen
- Klassifisering basert på et annotert datasett

Statistisk klassifisering

- Sentral metode innenfor maskinlæring
- Automatisk avgjøre hvilken kategori en observasjon tilhører
- Basert på **treningsdata**: observasjoner der kategorien er kjent
  - e-post $\rightarrow$ {spam, ikke-spam}
  - pasient $\rightarrow$ diagnose
- **Supervised** klassifisering: klassifisering som benytter treningsdata

Statistisk klassifisering



Betydningsdisambiguering

- Første skritt består i å hente ut trekk ("features") fra treningsdataene
- Eksempel: setninger merket med betydning
  - SKIM the pages for a clearer insight: Reading
  - She SKIMS through the novel which seems to fascinate them: Reading
  - Remove the vanilla pod, SKIM the jam, and let it cool: Removing
  - We SKIMMED across the surface of that sodding lake whilst all around us gathered the dark hosts of hell: Self_Motion
- Hvilke trekk ("features") kan vi bruke for å skille mellom de forskjellige betydningene?

Trekk: bag-of-words

- SKIM the pages for a clearer insight: Reading
- She SKIMS through the novel which seems to fascinate them: Reading
- Remove the vanilla pod, SKIM the jam, and let it cool: Removing
- We SKIMMED across the surface of that sodding lake whilst all around us gathered the dark hosts of hell: Self_Motion

Henter ut alle ord (**ikke** ordnet):

- a, clearer, for, insight, pages, the: Reading
- fascinate, novel, seems, she, the, them, through, to, which: Reading
- and, cool, it, jam, let, pod, remove, the, the, vanilla: Removing
- across, all, around, dark, gathered,hell, hosts, lake, of, of, sodding, surface, that, the, the, us, we, whilst Self_Motion

Trekk

- Konteksten til målordet kan representeres ved
  - ordformer
  - lemmaer
  - ordklassetagger
  - kombinasjon av disse
  - $[w_{i-2}, \text{POS}_{i-2}, w_{i-1}, \text{POS}_{i-1}, w_{i+1}, \text{POS}_{i+1}, w_{i+2}, \text{POS}_{i+2}]$
- Remove the vanilla pod, SKIM the jam, and let it cool: Removing
  - **trekkvektor**: [vanilla, JJ, pod, NN, the, DT, jam, NN]

Trekk: syntaktiske funksjoner

- SKIM the pages for a clearer insight: Reading
- She SKIMS through the novel which seems to fascinate them: Reading
- Remove the vanilla pod, SKIM the jam, and let it cool: Removing
- We SKIMMED across the surface of that sodding lake whilst all around us gathered the dark hosts of hell: Self_Motion

- Verbets argumenter:
  - direkte_objekt: Reading
  - subjekt, pp_through: Reading
  - direkte_objekt: Removing
  - subjekt, pp_across: Self_Motion
- Kombinasjoner (argumenters hovedord)
  - direkte_objekt / pages: Reading
  - subjekt / she, pp_through / novel: Reading
  - direkte_objekt / jam: Removing
  - subjekt / we, pp_across / surface: Self_Motion

Trekk: n-gram av ord

- N-gram av ord i nærheten av målordet
  - n=1,2,3
  - kan også bruke ordformer, lemmaer, ordklasser
- Eksempel:
  - SKIM the pages for a clearer insight: Reading
  - She SKIMS through the novel which seems to fascinate them: Reading
  - Remove the vanilla pod, SKIM the jam, and let it cool: Removing
  - We SKIMMED across the surface of that sodding lake whilst all around us gathered the dark hosts of hell: Self_Motion
- Trigram:
  - _, _, _, the, pages, for
  - _, _, She, through, the, novel
  - the, vanilla, pod, the, jam, and
  - _, _, We, across, the, surface

Klassifisering

- Gitt treningsdataene og trekkvektorene, kan en rekke forskjellige maskinlæringsalgoritmer brukes til å trene en klassifiserer
- Her skal vi se på **Naive Bayes**-klassifisering
- Bruker informasjon om ord i konteksten for disambiguering av betydning
- Enkel metode, mye brukt i WSD

Naive Bayes

- Naive Bayes klassifiserer

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

- 2 sannsynligheter:
  1. prior-sannsynligheten for betydningen $P(s)$

$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)}$$

  2. sannsynligheten for individuelle trekk $P(f_j|s)$

$$P(f_j|s) = \frac{count(f_j, s)}{count(s)}$$

Eksempel

Trekkvektorer (bag-of-words), der siste element angir betydningsklassen

- [a, clearer, for, insight, pages, the, Reading]
- [fascinate, novel, seems, she, the, them, through, to, which, Reading]
- [and, cool, it, jam, let, pod, remove, the, the, vanilla, Removing]
- [across, all, around, dark, gathered,hell, hosts, lake, of, of, sodding, surface, that, the, the, us, we, whilst, Self_Motion]

1. prior-sannsynligheten for betydningen $P(Reading)$
   - $P(s_i) = \frac{count(s_i, w_j)}{count(w_j)}$
   - $P(Reading) = \frac{count(Reading, skim)}{count(skim)}$
   - $P(Reading) = \frac{2}{4} = 0.5$

Eksempel forts.

- [a, clearer, for, insight, pages, the, Reading]
- [fascinate, novel, seems, she, the, them, through, to, which, Reading]
- [and, cool, it, jam, let, pod, remove, the, the, vanilla, Removing]
- [across, all, around, dark, gathered,hell, hosts, lake, of, of, sodding, surface, that, the, the, us, we, whilst, Self_Motion]

2. sannsynligheten for individuelle trekk $P(f_j|s)$
   - $P(f_j|s) = \frac{count(f_j,s)}{count(s)}$
   - $P(pages|Reading) = \frac{1}{2} = 0.5$
   - $P(pages|Removing) = \frac{0}{1} = 0$
   - $P(pages|Self\_Motion) = \frac{0}{1} = 0$
   - ...
   - $P(jam|Reading) = \frac{0}{2} = 0$
   - $P(jam|Removing) = \frac{1}{1} = 1.0$
   - $P(jam|Self\_Motion) = \frac{0}{1} = 0$
   - ...

Eksempel forts.

- Vi har nå en NB-modell som vi kan bruke til å klassifisere en ny og usett setning:
  - I like to SKIM through the novel
  - [I, like, novel, the, through, to, ??]
- Vi bruker den velkjente formelen:

$$P(s) \prod_{j=1}^{n} P(f_j|s)$$

- Itererer gjennom alle betydningene og trekkene, ganger sammen
- Og velger den betydningen (s) som gir høyest sannsynlighet

Eksempel forts.

$$P(Reading) = \frac{2}{4}$$

$$P(I|Reading) = \frac{0}{2}$$

$$P(like|Reading) = \frac{0}{2}$$

$$P(novel|Reading) = \frac{1}{2}$$

$$P(the|Reading) = \frac{2}{2}$$

$$P(through|Reading) = \frac{0}{2}$$

$$P(to|Reading) = \frac{1}{2}$$

$$P(Removing) = \frac{1}{4}$$

$$P(I|Removing) = \frac{0}{1}$$

$$P(like|Removing) = \frac{0}{1}$$

$$P(novel|Removing) = \frac{0}{1}$$

$$P(the|Removing) = \frac{2}{1}$$

$$P(through|Removing) = \frac{0}{1}$$

$$P(to|Removing) = \frac{0}{1}$$

Eksempel forts.

$$P(Self\_motion) = \frac{1}{4}$$

$$P(I|Self\_motion) = \frac{0}{1}$$

$$P(like|Self\_motion) = \frac{0}{1}$$

$$P(novel|Self\_motion) = \frac{0}{1}$$

$$P(the|Self\_motion) = \frac{2}{1}$$

$$P(through|Self\_motion) = \frac{0}{1}$$

$$P(to|Self\_motion) = \frac{0}{1}$$

- Som dere kan se både her og de to forrige sett av beregninger (forrige slide), vi kommer til å få 0 overalt pga sannsynlighetene som har verdien 0 (siden vi kommer til å multiplisere dem sammen ved å bruke Naive Bayes formelen).
- Vi må derfor bruke glatting (add-one smoothing – se neste slide).

Eksempel forts. add-one smoothing

$$P(Reading) = \frac{2}{4}$$

$$P(I|Reading) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(like|Reading) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(novel|Reading) = \frac{1+1}{2+3} = \frac{2}{5}$$

$$P(the|Reading) = \frac{2+1}{2+3} = \frac{3}{5}$$

$$P(through|Reading) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(to|Reading) = \frac{1+1}{2+3} = \frac{2}{5}$$

Eksempel forts. add-one smoothing

$$P(Reading) \times \frac{\overbrace{1 \times 1 \times 2 \times 3 \times 1 \times 2}^{12}}{5^6} = 0,000384 = 3.84 \times 10^{-4}$$

Eksempel forts. add-one smoothing

$$P(Removing) = \frac{1}{4}$$

$$P(I|Removing) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(like|Removing) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(novel|Removing) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(the|Removing) = \frac{2+1}{1+3} = \frac{3}{4}$$

$$P(through|Removing) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(to|Removing) = \frac{0+1}{1+3} = \frac{1}{4}$$

Eksempel forts. add-one smoothing

$$P(Removing) \times \frac{\overbrace{1 \times 1 \times 1 \times 3 \times 1 \times 1}^{3}}{4^6} = 0,000183105 = 1.83 \times 10^{-4}$$

Eksempel forts. add-one smoothing

$$P(Self\_motion) = \frac{1}{4}$$

$$P(I|Self\_motion) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(like|Self\_motion) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(novel|Self\_motion) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(the|Self\_motion) = \frac{2+1}{1+3} = \frac{3}{4}$$

$$P(through|Self\_motion) = \frac{0+1}{1+3} = \frac{1}{4}$$

$$P(to|Self\_motion) = \frac{0+1}{1+3} = \frac{1}{4}$$

Eksempel forts. add-one smoothing

$$P(Self\_motion) \times \frac{\overbrace{1 \times 1 \times 1 \times 3 \times 1 \times 1}^{3}}{4^6} = 0,000183105 = 1.83 \times 10^{-4}$$

Eksempel forts. add-one smoothing

$$P(Reading) \times \frac{\overbrace{1 \times 1 \times 2 \times 3 \times 1 \times 2}^{12}}{5^6} = 0,000384 = 3.84 \times 10^{-4}$$

$$P(Removing) \times \frac{\overbrace{1 \times 1 \times 1 \times 3 \times 1 \times 1}^{3}}{4^6} = 0,000183105 = 1.83 \times 10^{-4}$$

$$P(Self\_motion) \times \frac{\overbrace{1 \times 1 \times 1 \times 3 \times 1 \times 1}^{3}}{4^6} = 0,000183105 = 1.83 \times 10^{-4}$$

$\Rightarrow$ Den betydningen som gir høyest sannsynlighet for testsetningen vår [I, like, novel, the, through, to] er Reading

Naive Bayes – To måter å klassifisere

Merk at vi her har beregnet sannsynlighetene på en litt annen måte enn det vi gjorde under forelesning 8 om maskinlæring og klassifisering.

- Her har vi brukt:

$$P(f_j|s) = \frac{count(f_j, s)}{count(s)}$$

  Hvor vi teller antall ganger et ord f_j fra trekkvektoren forekommer med betydning s, delt på antall setninger klassifisert som betydning s.

- Når vi brukte Naive bayes for klassifisering brukte vi:

$$P(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)}$$

  Hvor vi teller antall ganger et ord w_i forekommer i klassen c delt på antall ord typer som finnes i klassen c

Naive Bayes – To måter å klassifisere

Merk at det samme gjelder glatting!

- For Word Sense Disambiguation legger vi $+1$ i telleren som vanlig. Men i nevneren legger vi til antall senses (betydninger) som finnes i treningsdataen vår. I eksemplet på slide 60, ble $P(I|Reading) = \frac{0+1}{2+3}$ fordi det finnes 3 betydninger i treningsdataen vår (altså $Reading$, $Removing$, og $Self\_motion$.

- For klassifisering av klasser legger vi $+1$ i telleren, og $+ |V|$ i nevneren som representerer alle ord typer i hele treningskorpuset:

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} (count(w, c) + 1)} = \frac{count(w_i, c) + 1}{(\sum_{w \in V} (count(w, c)) + |V|}$$

Eksempel 2– Naive Bayes

Treningsdata:

```
Reading       clearer insight pages
Reading       fascinate novel seems
Reading       like novel book pages
Reading       book action suspense
Reading       book pages thriller killer
Reading       book novel fascinate like
Removing      jam bread like breakfast
Removing      cool jam let pod remove vanilla
Removing      milk surface cream
Self_Motion   lake surface close novel surfboard
Self_Motion   dark gathered hell hosts lake sodding surface
```

Eksempel – Naive Bayes

Test data:

`? like novel`

##

Compute Naive Bayes

##

$$\textbf{P(Reading)P(S|Reading)} = \frac{6}{11} \times \frac{3 \times 4}{9^2} = \textbf{0,080} = \textbf{8.00} \times 10^{-2}$$

$$P(Removing)P(S|Removing) = \frac{3}{11} \times \frac{2 \times 1}{6^2} = 0,0151 = 1.51 \times 10^{-2}$$

$$P(Self_motion)P(S|Self_motion) = \frac{2}{11} \times \frac{1 \times 2}{5^2} = 0,0145 = 1.45 \times 10^{-2}$$

Oppsummering

- Rekke oppgaver inngår i semantisk analyse
  - ordbetydningsdisambiguering (WSD)
  - semantiske roller
  - parafrasering
  - temporal analyse
  - entailment
  - negasjon
  - ...
- Sentral metode: klassifisering

Oppsummering

- For klassifisering trenger vi treningsdata
- Semantiske ressurser
  - WordNet
    - leksikal database
    - innholdsord: substantiver, verb, adjektiver
    - bygget rundt leksikale relasjoner som synonymi, hyponymi, meronymi, etc.
  - PropBank/FrameNet
    - forskjellige ressurser for semantiske roller
    - korpus vs database
    - verbspesifikk vs ramme ("frame")
  - ...

Oppsummering

- Nærmere kikk på betydningsdisambiguering
- Trekkrepresentasjon av treningsdata
  - ord
  - lemma
  - ordklasse
  - syntaktisk funksjon
  - etc.
- Naive Bayes-klassifisering
  - hvordan vi kan beregne den mest sannsynlige betydningen for et ord:
    $\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s|\overrightarrow{f})$