IN1140: Introduksjon til språkteknologi

Forelesning #10

Egil Rønningstad

Universitetet i Oslo

30. oktober 2023



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 innen semantisk analyse kan løses med språkteknologi?
- Med hva slags metoder kan disse oppgavene løses?

Semantikk i språkteknologi

Tema for i dag:

- Hva slags oppgaver innen semantisk analyse kan løses med språkteknologi?
- Med hva slags metoder kan disse oppgavene løses?
- Tre nivåer av betydning:
  1. Ord: Word Sense Disambiguation (WSD)
  2. Fraser: Named Entity Recognition (NER)
  3. Setninger: Semantic Role Labeling (SRL)

Ordbetydning ("word sense")

stjerne

SUBSTANTIV *hankjønn eller hunkjønn* VIS BØYNING +

OPPHAV norrønt *stjarna*; beslektet med astronom og stella betydning 4 fra engelsk

BETYDNING OG BRUK

1. fjernt himmellegeme som er synlig som en lysende prikk

   Eksempel

   fiksstjerne, morgenstjerne

   en klar himmel med millioner av stjerner

   i vest blinket en stor stjerne

   være født under en lykkelig stjerne – etter den oppfatning at stjernenes stilling i det
   skjebne

   - noe som står lysende for ens tanke

     tenne håpets stjerne

2. lysende, flimrende glimt

   Eksempel

Ordbetydning ("word sense")

3. noe som ligner en stjerne (1)

   Eksempel

   et juletre med stjerne i toppen

   et ord merket med stjerne

   en hest med stjerne i panna

   campingplassen har tre stjerner – brukt for å angi en bestemt standard

4. berømt person

   Eksempel

   en stjerne på vitenskapens himmel

   en ny stjerne i langrennssporet

   filmstjerne, idrettsstjerne, popstjerne

5. i sammensetninger i dyre- og plantenavn:

   Eksempel

   sjøstjerne, gullstjerne, skogstjerne

Ordbetydning ("word sense")

Når er det ulik betydning? **J&M Kapittel 18**

ably three distinct senses of *serve*. One practical technique for determining if two senses are distinct is to conjoin two uses of a word in a single sentence; this kind of conjunction of antagonistic readings is called **zeugma**. Consider the following examples:

(18.4) Which of those flights serve breakfast?

(18.5) Does Air France serve Philadelphia?

(18.6) ?Does Air France serve breakfast and Philadelphia?

We use (?) to mark those examples that are semantically ill-formed. The oddness of the invented third example (a case of zeugma) indicates there is no sensible way to make a single sense of *serve* work for both breakfast and Philadelphia. We can use this as evidence that *serve* has two different senses in this case.

Ordbetydning ("word sense")

sende

VERB VIS BØYNING +

OPPHAV norrønt *senda*

BETYDNING OG BRUK

1. la noe bli tatt og ført (av andre) til det stedet det skal, få av gårde, skikke (2),

Eksempel

sende en delegasjon til Paris

sende tropper til unnsetning

sende bud etter en

sende en til sjøs

sende et brev

sende varer med jernbanen

sende rundt kakefatet

Ordbetydning ("word sense")

Øvelse: Lag en setning som som kombinerer to av eksemplene på sende (1)

sende

VERB VIS BØYNING +

OPPHAV norrønt *senda*

BETYDNING OG BRUK

1. la noe bli tatt og ført (av andre) til det stedet det skal, få av gårde, skikke (2),

   Eksempel

   sende en delegasjon til Paris

   sende tropper til unnsetning

   sende bud etter en

   sende en til sjøs

   sende et brev

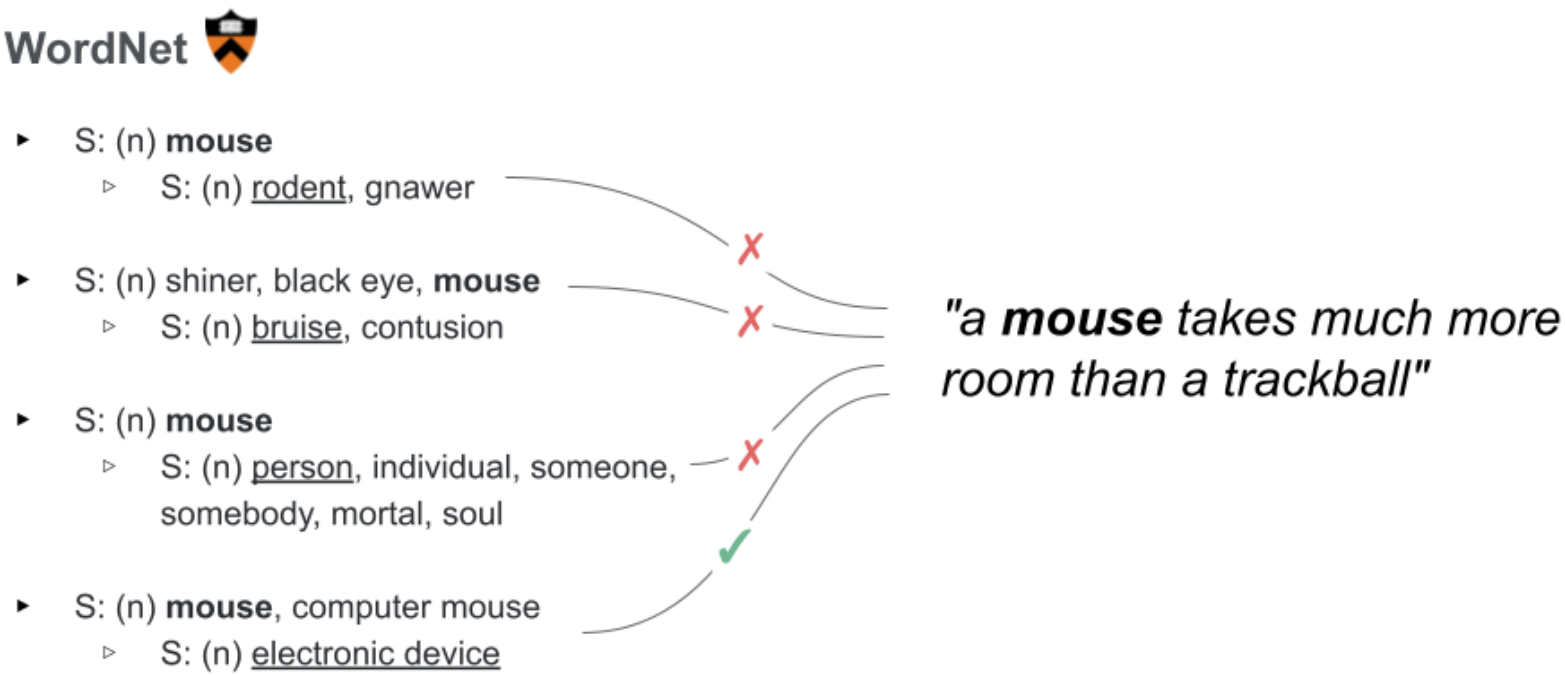
   sende varer med jernbanen

   sende rundt kakefatet

Betydningsdisambiguering

- Word Sense Disambiguation (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



Named Entity Recognition (NER)

- Egennavn inneholder viktig semantisk informasjon
  - enkeltord *Erna, Oslo*
  - NP-fraser *Universitetet i Oslo*
- Automatisk **gjenkjenning** og **kategorisering** av egennavn
- Vanlige kategorier: person, organisasjon, sted (lokasjon), geo-politisk entitet

	ORG						GPE_LOC	
Den	internasjonale	domstolen	har	sete	i	Haag	.	
The	*International*	*Court of Justice*	*has*	*its seat*	*in*	*The Hague*	*.*	

Named Entity Recognition (NER)

- Kategorier

NE Type	**Eksempler**
ORGANIZATION	Omnicom, WHO
PERSON	George Washington, President Obama
LOCATION	Downing St., Mississippi River, Norway
DATE	June, 2011-05-03, 03/05/2011
TIME	two fifty a.m., 1:30 p.m.
MONEY	175 million Canadian Dollars, GBP 10.40
FACILITY	Washington Monument, Stonehenge
GPE	Washington D.C., Norway

Eksempel på å trekke ut tabellarisk informasjon fra ustrukturert tekst.

Named Entity Recognition (NER)

Oppslag i en navneliste (feks. “gazetteer”)?

Named Entity Recognition (NER)

Oppslag i en navneliste (feks. "gazetteer")?
https://www.ssb.no/statbank/table/10501/tableViewLayout1/

Anniken	2 638	2 649	2 666	2 686
Annlaug	557	551	544	539
Anny	2 308	2 235	2 150	2 080
Antonia	.	.	.	.
Aria	.	.	.	228
Ariana	265	293	332	353
Ariel	298	325	352	394
Arna	708	681	654	632
Arnhild	1 436	1 418	1 402	1 376
Asbjørg	1 356	1 322	1 253	1 200
Asha	243	247	254	255
Aslaug	4 021	3 799	3 604	3 374
Asma	219	221	240	258

Named Entity Recognition (NER)

Oppslag i en navneliste (feks. “gazetteer”)?

Lena	Lene	Tone
Irene	Trine	Linda
Siv	Line	Hilde
Gro	Helene	Eva
Lise	Mari	Anna
Linn	Nora	Ida
Mette	Tove	Liv

Named Entity Recognition (NER)

Oppslag i en navneliste (feks. “gazetteer”)?

- Tar ikke hensyn til kontekst
- Dårlig dekningsgrad, er statisk (må oppdateres)
- En entitet kan strekke seg over flere ord “*Stanford University*”
- Navn kan inneholde andre navn “*Cecil H. Green Library*”

Named Entity Recognition (NER)

Flertydighet

- Samme navn kan referere til forskjellige entiteter av samme type
  - JFK – presidenten og hans sønn
- Samme navn kan referere til entiteter av forskjellig type
  - JFK – flyplass
  - **Metonymi**: et systematisk forhold der vi bruker ett aspekt ved et konsept for å referere til et annet aspekt ved konseptet
    f.eks. bygning-for-organisasjon: *The White House claims that . . .*
- Trenger manuelt annotert korpus

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 $[\text{the dog}]_{\text{PATIENT}}$

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*
- Også her har vi behov for et manuelt annotert korpus

Andre semantiske oppgaver: Entailment

- *Apple bought Cisco*
- *Apple acquired Cisco*
- *Cisco was taken over by Apple*
- Spørsmål: *Who bought Cisco?*

Andre semantiske oppgaver: 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

Andre semantiske oppgaver: Negasjon

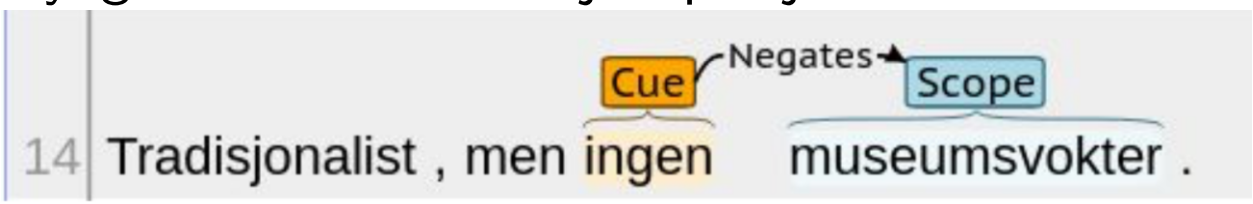
- *Apple acquired Cisco*
- *Apple did not acquire Cisco*
- *Apple failed to acquire Cisco*
- *Apple denied not acquiring Cisco*

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*.

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*.
- Nylig avsluttet annotasjonsprosjekt ved LTG: norsk negasjon

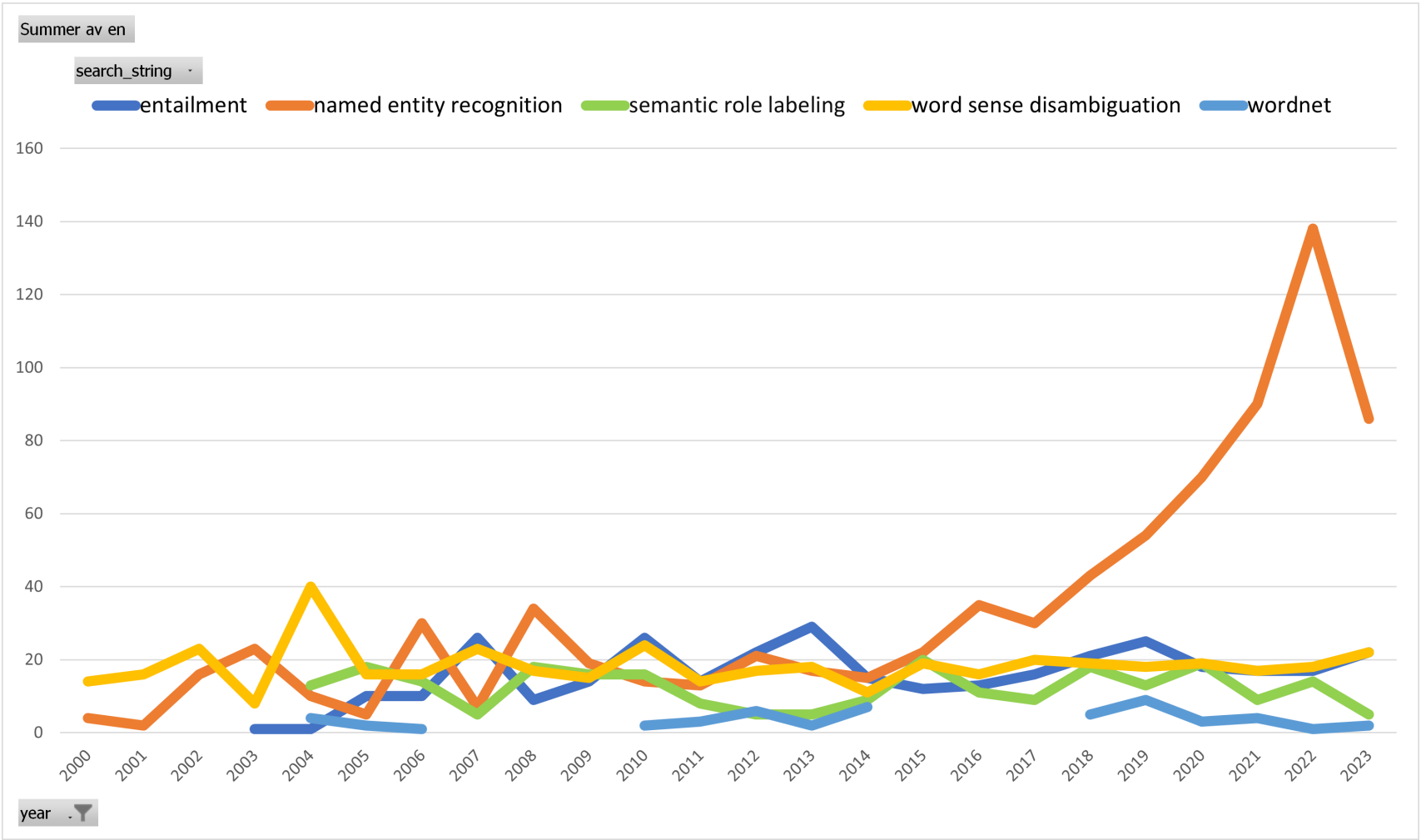


Semantikk i språkteknologi

Tema for i dag:

- Hva slags oppgaver innen semantisk analyse kan løses med språkteknologi?
- Med hva slags metoder kan disse oppgavene løses?
- Tre nivåer av betydning:
  1. Ord: Word Sense Disambiguation (WSD)
  2. Fraser: Named Entity Recognition (NER)
  3. Setninger: Semantic Role Labeling (SRL)

Trender i språkteknologi





Metoder

Metoder

- Veiledet klassifisering på ulike nivåer
  - Gitt et ord i en setning, og en liste av mulige betydninger, velg en betydning (WSD)
  - For hvert ord i en setning, avgjør om det tilhører en entitet av en viss kategori (NER)
  - Gitt et predikat i en setning, finn dets semantiske roller. (SRL)

Metoder

- Veiledet klassifisering på ulike nivåer
  - Gitt et ord i en setning, og en liste av mulige betydninger, velg en betydning (WSD)
  - For hvert ord i en setning, avgjør om det tilhører en entitet av en viss kategori (NER)
  - Gitt et predikat i en setning, finn dets semantiske roller. (SRL)
- Ikke-veiledet tilegnelse av semantisk informasjon fra rå tekst
  - Distribusjonell semantikk:
  - Ord som forekommer i lignende kontekster har lik betydning
    `http://vectors.nlpl.eu/explore/embeddings/en/similar/`

$\Rightarrow$ **Maskinlæring**

Semantiske ressurser

Semantiske ressurser

- Klassifisering forutsetter treningsdata.
- Leksikalske databaser (WordNet, FrameNet).
- Korpuser annotert med semantisk informasjon (SemCor, CoNLL03/NorNE, 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)
  - Hyponymi (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
- hver betydning utgjør ett synset

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)

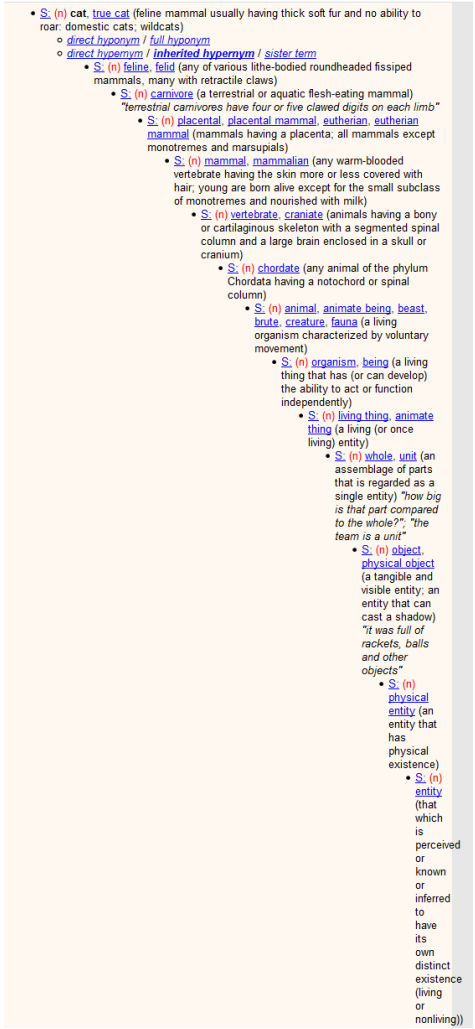
WordNet: substantiv eksempel

- 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 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 opp forekomster av et ord med riktig betydning (=synset)
  - Men trenger også korpus der ord er annotert med betydning (SemCor)
- Informasjon brukes som trekk i mange semantiske NLP-oppgaver
- Generalisere over synonymer

The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms–words that denote the same concept and are interchangeable in many contexts–are grouped into unordered sets (synsets). Each of WordNet's 117 000 synsets is linked to other synsets by means of a small number of "conceptual relations." Additionally, a synset contains a brief definition ("gloss") and, in most cases, one or more short sentences illustrating the use of the synset members.
https://wordnet.princeton.edu/

NE-annoterte korpuser

- Flere korpuser for engelsk: CoNLL03, ACE, etc.
- Her fokusere på et norsk korpus: NorNE
  - Første fritt tilgjengelige NER-datasett for norsk
  - Samarbeid mellom Schibsted, Språkbanken (Nasjonalbiblioteket) og LTG

NorNE

- Norsk Dependenstrebank (NDT) beriket med semantiske kategorier for egennavn, både Bokmål og Nynorsk
- $\sim$300K tokens for hver, hvorav $\sim$20K er del av et egennavn
- Distributert i såkalt CoNLL-U format med BIO-oppmerking. Forenklet versjon:

1	Den	den	DET	name=B-ORG
2	internasjonale	internasjonal	ADJ	name=I-ORG
3	domstolen	domstol	NOUN	name=I-ORG
4	har	ha	VERB	name=O
5	sete	sete	NOUN	name=O
6	i	i	ADP	name=O
7	Haag	Haag	PROPN	name=B-GPE_LOC
8	.	$.	PUNCT	name=O

NorNE entitetkategorier (Bokmål)

Type	Train	Dev	Test	Total
PER	4033	607	560	5200
ORG	2828	400	283	3511
GPE_LOC	2132	258	257	2647
PROD	671	162	71	904
LOC	613	109	103	825
GPE_ORG	388	55	50	493
DRV	519	77	48	644
EVT	131	9	5	145
MISC	8	0	0	0

https://github.com/ltgoslo/norne/

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
- $\Rightarrow$ Verbspesifikke roller
- Semantiske ressurser med informasjon om semantiske roller: **PropBank** og FrameNet

PropBank

- Korpus som inneholder alle setningene i Penn Treebank
- Annotert med informasjon om semantiske roller
- 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*

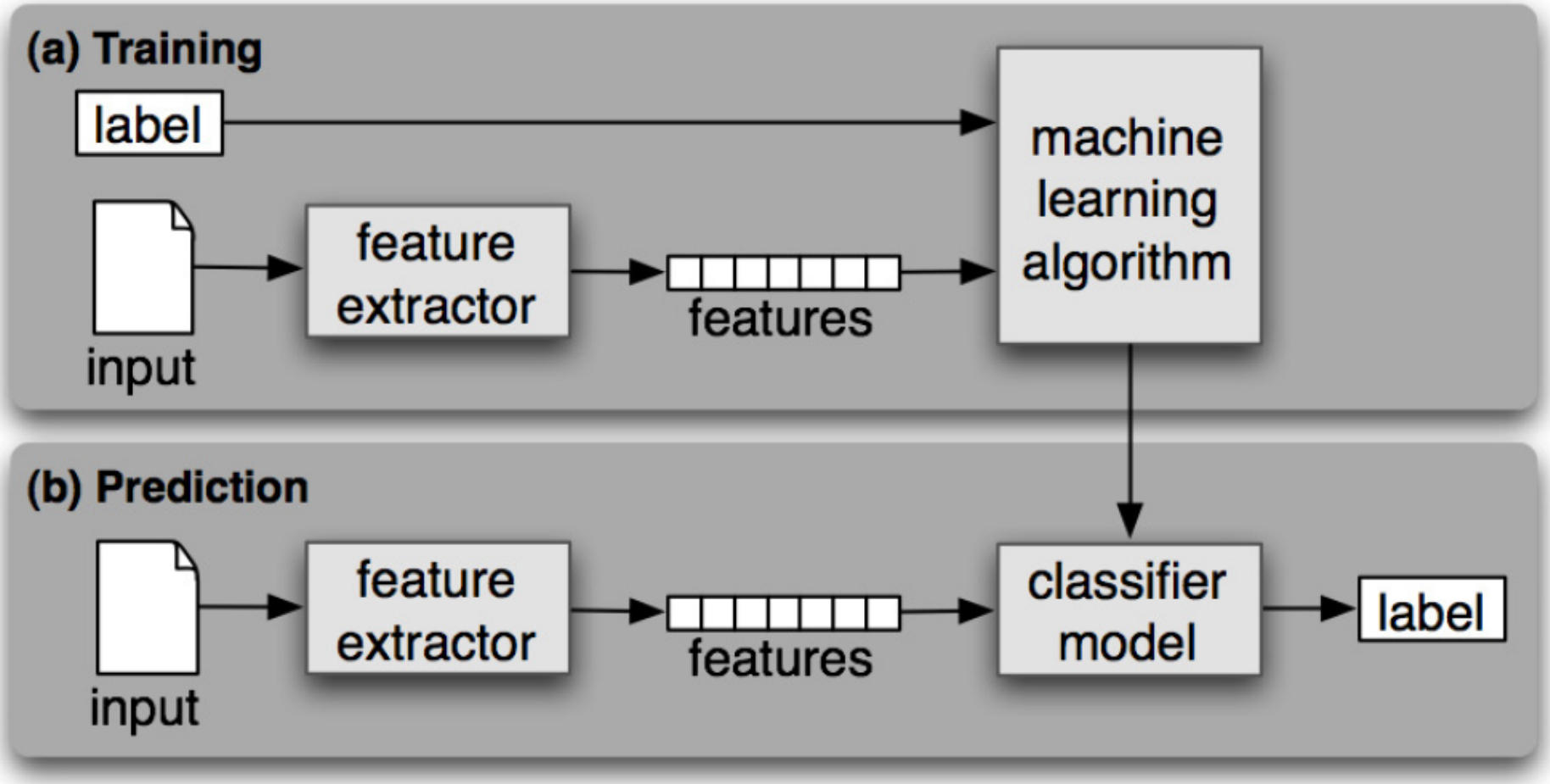
Semantisk klassifisering

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
- **Veiledet** klassifisering: klassifisering som benytter manualt annoterte treningsdata

Klassifisering

Petal: Kronblad. Sepal: Begerblad



Klassifisering Sverdliljeslekta

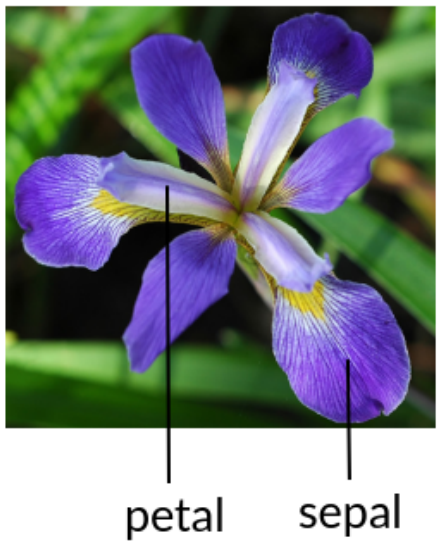
iris setosa



iris versicolor



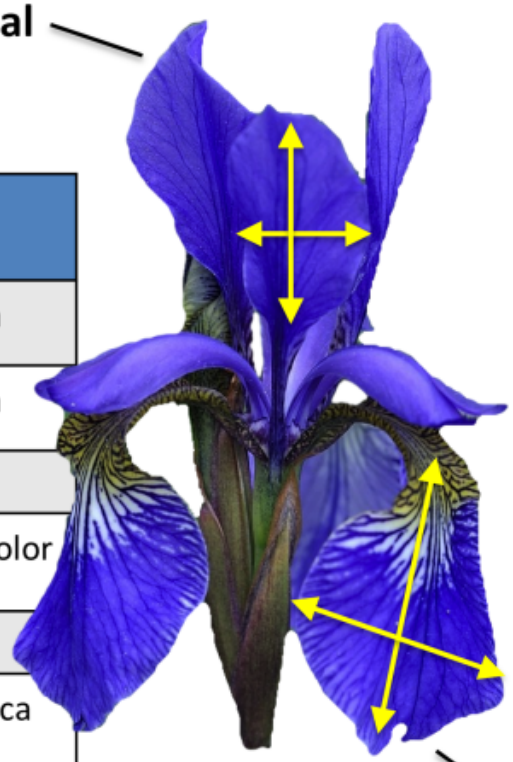
iris virginica



Klassifisering Sverdliljeslekta

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

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

Andre trekk

- Konteksten til målordet kan representeres ved
  - ordformer
  - n-gram
  - 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]

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

WordNet eksempel

- verbet *skim*: synonymer, definisjoner og eksempler
- hver betydning utgjør ett synset

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)

Wikipedia eksempel

- Entity Linking
- Etternavnet *Stoltenberg*: Kan vi ut fra sammenhengen avgjøre hvem av disse teksten snakker om?

Stoltenberg

Fra Wikipedia, den frie encyklopedi

Stoltenberg har flere betydninger.

Norske personer [rediger | rediger kilde]

- Camilla Stoltenberg, lege og forsker
- Jens Stoltenberg, Norges 34. statsminister, NATOs 13. generalsekretær
- Karin Stoltenberg, tidligere statssekretær
- Matthias Stoltenberg, maler
- Robert Stoltenberg, komiker
- Harald Stoltenberg, skuespiller og forfatter
- Thorvald Stoltenberg, politiker og diplomat
- Helge Stoltenberg, lokalpolitiker i Gol

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 $P(f_j|s)$

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

Named Entity Recognition (NER)

Vanligste måten å løse denne oppgaven på er ved ord-for-ord klassifisering

- Metoder for sekvensklassifisering, f.eks. Hidden Markov Models (HMM), CRF, LSTM
- BIO-klassifisering: taggen indikerer om ordet befinner seg i begynnelsen (B), innenfor (I) eller utenfor (O) et egennavn, samt indikerer kategori.

BIO-klassifisering

```
honor          O
of             O
George         B_pers
Washington     I_pers
,              O
who            O
...           ...
```

Named Entity Recognition (NER)

American	
Airlines	
,	
a	
unit	
of	
AMR	
Corp.	
,	
immediately	
matched	
the	
move	
,	
spokesman	
Tim	
Wagner	
said	
.	

Named Entity Recognition (NER)

American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

Named Entity Recognition (NER)

Data kan representeres ved **trekk** ("features")

- ordform (tokenisering): *of, George, Washington, led*
- lemma: *of, George, Washington, lead*
- shape: lower, capital, capital, lower
- affikser: *of, rge, ton, ead*
- ordklasse: IN, NNP, NNP, VBD
- chunk-kategori: PP, NP, NP, _
- navneliste: 0, 1, 1, 0

(Eller nevrale nettverk som lærer beste representasjon av dataene)

Named Entity Recognition (NER)

- NER-systemer ofte kombinasjon av
  - Lister ("gazetteers")
  - Regler (regel-baserte systemer)
  - Veiledet ("supervised") klassifisering
- Beste systemer for engelsk: 93.5% totalt
- Ulike resultater for ulike klasser

NorNE: resultater for norsk

- Jørgensen, Aasmoe, Husevåg, Øvrelid & Velldal (2020). *NorNE: Annotating Named Entities for Norwegian*
- Nevrale modeller for sekvensklassifisering (LSTM-CRF)

	IOB	IOBE	IOBS	IOBES
NorNE-full	90.75	90.45	**90.76**	90.58
NorNE-7	91.86	91.80	91.99	**92.29**
NorNE-6	90.95	90.50	**91.85**	91.38

Table:

NorNE: resultater for norsk

- Jørgensen, Aasmoe, Husevåg, Øvrelid & Velldal (2020). *NorNE: Annotating Named Entities for Norwegian*
- Nevrale modell for sekvensklassifisering (LSTM-CRF)

	Development		Heldout	
Training	BM	NN	BM	NN
BM	89.47	*82.34*	**83.89**	*81.59*
NN	*84.01*	86.53	*76.88*	83.89
BM+NN	**90.92**	**88.03**	83.48	**85.32**

Table:

SRL

- Oppgaven: finne semantiske roller for hvert argument i setninger
- Hvorfor?
  - felles semantisk representasjon for setninger
  - forbedrer en rekke NLP-applikasjoner: spørsmål-svar (QA), maskinoversettelse

SRL

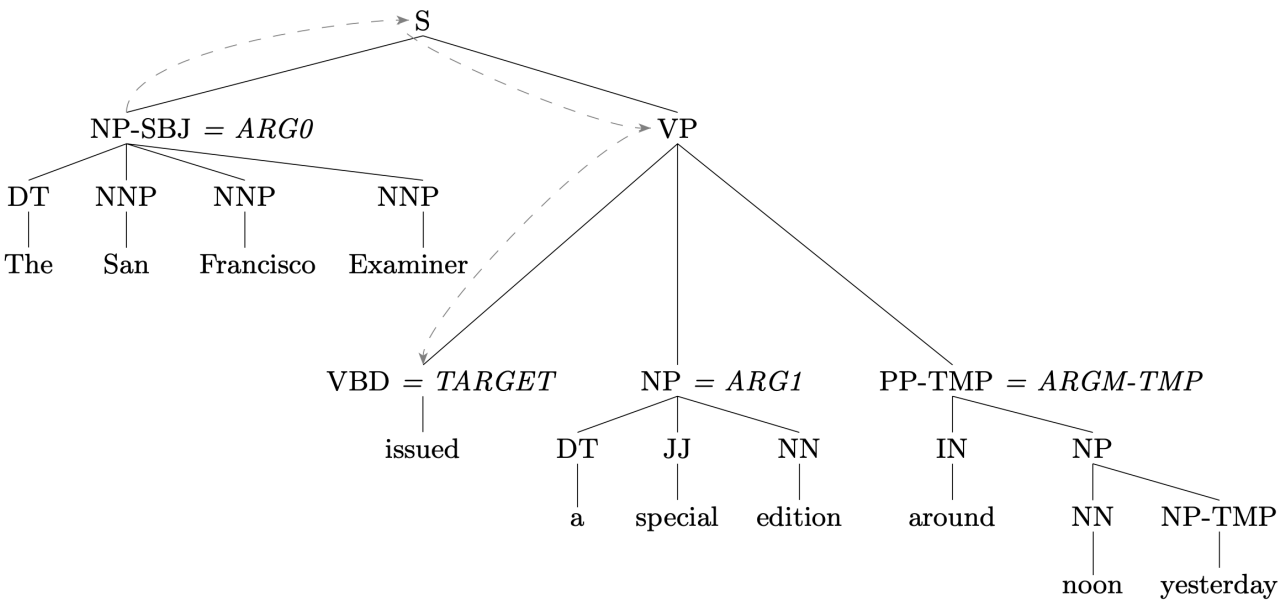
- Oppgaven: finne semantiske roller for hvert argument i setninger
- Hvorfor?
  - felles semantisk representasjon for setninger
  - forbedrer en rekke NLP-applikasjoner: spørsmål-svar (QA), maskinoversettelse
- Vanlig å utgå fra syntaktisk analyse

function SEMANTICROLELABEL(*words*) **returns** labeled tree

```
parse ← PARSE(words)
for each predicate in parse do
    for each node in parse do
        featurevector ← EXTRACTFEATURES(node, predicate, parse)
        CLASSIFYNODE(node, featurevector, parse)
```

SRL

- Klassifiserer noder (konstituenter) i det syntaktiske treet
- Trekk (eksempler):
  - Ordformen til hodet (Examiner)
  - Ordklassen til hodet (NNP)
  - Predikatets struktur (`V->VBP NP PP`)
  - NER-klassen til konstituenten (ORG)
  - ...



Oppsummering

- En rekke oppgaver inngår i semantisk analyse
  - ordbetydningsdisambiguering (WSD)
  - navnegjenkjenning (NER)
  - semantiske roller (SRL)
  - 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.
  - Annoterte korpuser:
    - NorNE
    - PropBank

Oppsummering

- Trekkrepresentasjon av treningsdata
  - ord
  - lemma
  - ordklasse
  - NE-klasse
  - syntaktisk kategori
  - etc.
- Forskjellige typer klassifisering
- Naive Bayes-klassifisering for WSD
  - hvordan vi kan beregne den mest sannsynlige betydningen for et ord:
    $\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s|\overrightarrow{f})$