Semantikk i språkteknologi

(NB: Dette er et sammendrag, flere detaljer i forelesningen)

Tre språkteknologiske felt/oppgaver for hvert nivå innen betydning:

1. Ord: Word Sense Disambiguation (WSD)
2. Fraser: Named Entity Recognition (NER)
3. Setninger: Semantic Role Labeling (SRL)

WSD (betydningsdisambiguering)

Disambiguering: Analyse av flertydighet

Gitt en setning med et spesifikt målord og en liste med betydninger, angi korrekt betydning for målordet i setningen

«Jeg dro til en bank i dag»

✅ Bank: substantiv, forretningsinstitusjon

❌ Bank: substantiv, banking, pryl

❌ Bank: verb, å banke (på)

Klassifisering basert på et annotert datasett

Named Entity Recognition (NER)

Automatisk gjenkjenning og kategorisering av egennavn

PERSON GPE ORG

Jonas Gahr Støre er statsminister i Norge, og sitter på Stortinget

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

Semantic Role Labeling (SRL)

Gitt et predikat i en setning, finn dets semantiske roller

[*arg0* Læreren] leste [*arg1* en bok]

[*arg1* Boka] ble lest av [*arg0* læreren]

Læreren er alltid *agent* og boka er alltid *instrument*

Disse oppgavene har noe til felles: Veiledet klassifisering

- Maskinen får se ferdigannotert data (treningsdata) før den klassifiserer på egenhånd

Bruker trekk for å representere treningsdataen:

-Ord

-Lemma

-ordklasse

-NE-klasse

-Syntaktisk kategori

..osv..

Ikke-veiledet klassifisering: Maskinen finner mønster uten forhåndsannotert data

Naive Bayes-klassifisering

Populær metode for klassifisering innenfor WSD (men også innen mye annet)

Bruker informasjon om ord i konteksten for disambiguering av betydning

1. Prior-sannsynlighet for betydningen P(s)

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

2. Sannsynlighet for individuelle trekk P(fj|s)

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

BIO-klassifisering

Metode for klassifisering innenfor NER. Ord-for-ord klassifisering.

B: begynnelsen av egennavn, I: Innenfor egennavn, O: Utenfor (outside) egennavn

B_PERS	I_PERS	I_PERS	O	O	O	B_LOC	O	O	O	B_ORG
Jonas	Gahr	Støre	er	statsminister	i	Norge	og	sitter	på	Stortinget

Syntaktisk analyse

Metode for klassifisering innenfor SRL

Klassifiserer noder (konstituenter) i det syntaktiske treet