INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Tolvte forelesning

Lilja Øvrelid

8 mai, 2017

Språkteknologiske applikasjoner

- Mange språkteknologiske applikasjoner kombinerer teknikker vi har sett på tidligere i kurset
- Motiverer de teknikkene vi har lært
- Introdusere en ny (og mye brukt) metode: vektorromrepresentasjoner
- Se på noen eksempler
  - teknikker i informasjonsekstraksjon
  - informasjonsgjenfinning, websøk
  - maskinoversettelse

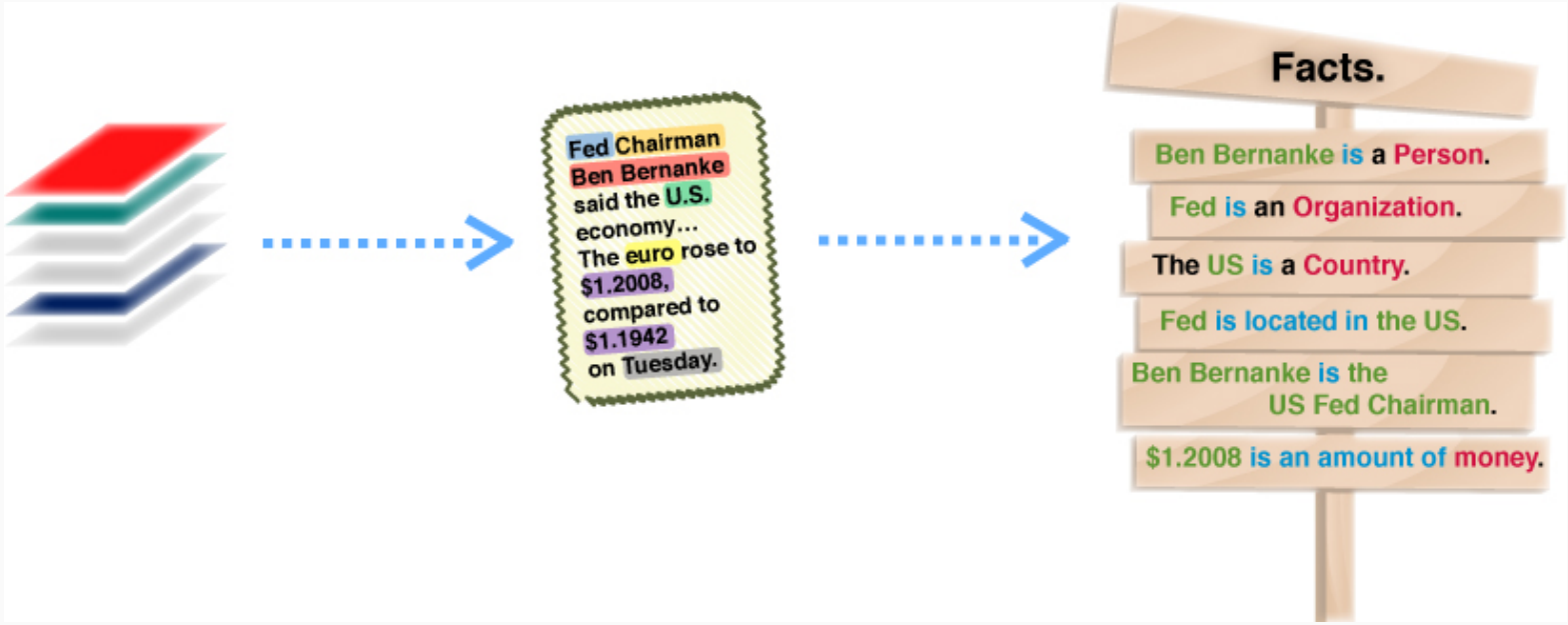
Informasjonsekstraksjon (IE)

Informasjonsekstraksjon (IE)

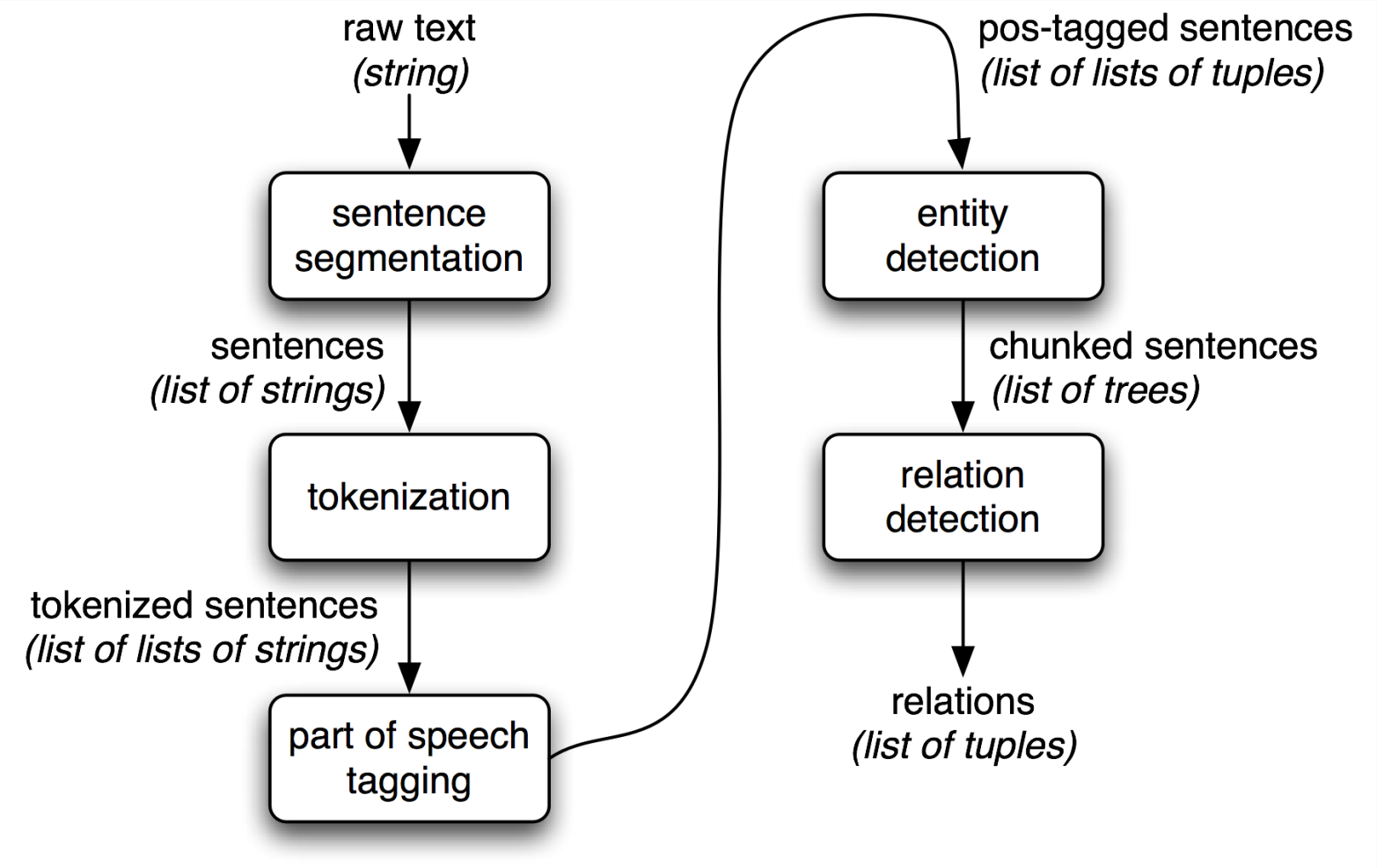
- Samlebegrep rundt teknikker som henter ut (ekstraherer) forskjellige typer semantisk informasjon fra tekst
- Ustruktert informasjon $\Rightarrow$ strukturerte data (relasjonsdatabase)
  - F.eks. relasjoner mellom organisasjoner og steder

Organisasjon	**Steder**
Omnicom	New York
DDB Needham	New York
Kaplan Thaler Group	New York
BBDO South	Atlanta
Georgia Pacific	Atlanta

Informasjonsekstraksjon (IE)



Informasjonsekstraksjon (IE)



Named Entity Recognition (NER)

- Automatisk **gjenkjenning**
- og **kategorisering** av egennavn

Output

[*The Washington Monument*] *is the most prominent structure in* [*Washington, D.C.*] *and one of the city's early attractions. It was built in honor of* [*George Washington*]*, who led the country to independence and then became its first President.*

- FACILITY
- GPE
- PERSON

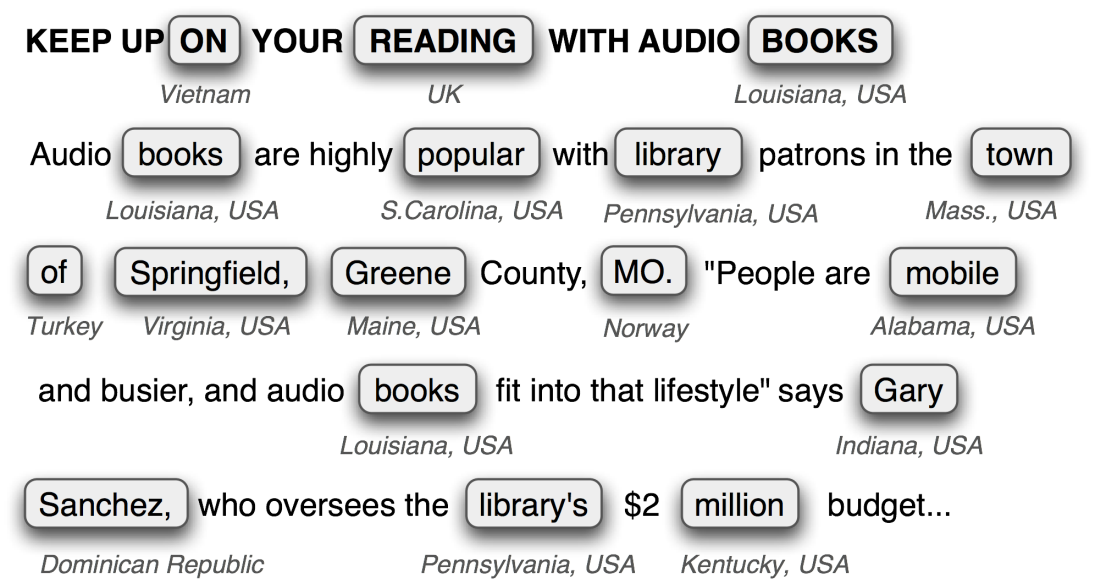
Named Entity Recognition (NER)

- 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

Named Entity Recognition (NER)

Oppslag i en navneliste ("gazetteer")?



Named Entity Recognition (NER)

Oppslag i en navneliste ("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 ...*

Named Entity Recognition (NER)

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

- 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)

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

- ordform: *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

Named Entity Recognition (NER)

- NER-systemer oftest kombinasjon av
  - lister
  - regler
  - veiledet ("supervised") klassifisering
- Beste systemer for engelsk: 92% for PERSON, LOCATION, 84% for ORGANIZATION

Relasjonsekstraksjon

Finne fram til **relasjoner** mellom entitetene i en tekst

Input

Citing high fuel prices, [$_{\text{ORG}}$ *United Airlines*] *said* [$_{\text{TIME}}$ *Friday*] *it has increased fares by* [$_{\text{MONEY}}$ *$6*] *per round trip on flights to some cities also served by lower-cost carriers.* [$_{\text{ORG}}$ *American Airlines*]*, a unit of* [$_{\text{ORG}}$ *AMR Corp.*]*, immediately matched the move, spokesman* [$_{\text{PER}}$ *Tim Wagner*] *said.*

- PER $\rightarrow$ ORG: spokesman_of(Tim Wagner, American Airlines)
- ORG $\rightarrow$ ORG: unit_of(American Airlines, AMR Corp)

Metoder for relasjonsekstraksjon

Regulære uttrykk over NE-tagget tekst

- (X, α, Y), der X, Y er egennavn og α er ordstrengen som forekommer mellom dem
- søke etter spesifikke ord/frase i α, feks *in*
  - [$_{ORG}$ WHYY] 'in' [$_{LOC}$Philadephia]
  - [$_{ORG}$ Brookings Institution] 'the research group in' [$_{LOC}$Washington]
  - [$_{ORG}$ OpenText] 'based in' [$_{LOC}$Waterloo]
  - [$_{ORG}$ Omnicom] 'in' [$_{LOC}$New York]

Metoder for relasjonsekstraksjon

Klassifisering, to oppgaver

1. hvorvidt to entiteter står i en relasjon (binær klassifiserer)
2. tildele kategori til relasjonene

Viktig del av dette er å finne fram til gode **trekk**

- NE-kategoriene (PER, ORG, LOC etc.)
- Hovedordene i NE-argumentene
- Ord hentet fra konteksten (teksten): bag-of-words, lemmaer, distanse, antall entiteter mellom, etc.
- Syntaktisk struktur fra konteksten: chunk-kategorier, funksjoner, frasestrukturkategorier, etc.

IR og vektorrommodellen

IR

- Informasjonsgjenfinning (Information Retrieval – IR): lagring og gjenfinning av alle slags media (fokus her: tekst)
- Stort fagfelt, eget ISK-kurs (INF3800)
- Automatisk finne fram til dokumenter som er relevante for en søkestreng
- Vektorrommodellen brukes i de fleste moderne systemer, inkludert søkemotorer som Google

IR

Litt terminologi:

- **dokument**: tekstenhet som er indeksert i systemet og tilgjengelig for gjenfinning, f.eks. en artikkel, en webside, deler av en webside, etc.
- **samling** ("collection"): mengde dokumenter
- **term**: leksikal enhet som forekommer i samlingen (som regel ord)
- **søkestreng** ("query"): representerer brukerens informasjonsbehov ved en mengde termer

Dokumenter som samlinger av ord

Doc1: *Osama bin Laden, mastermind, attack, American soil, hunted, world, killed, firefight, United States, forces, Pakistan, President Obama, announced, Sunday*
Doc2: *story, love, blood, samurai, movie, "13 assassins", man, courtyard, mask, shirt, blade, hand, director, Takasji Miike, showing, tale, revenge, liberation, stakes*
Doc3: *University, Delaware, year, demoting, men's, track, cross-country, teams, college, sports*

Samlinger av ord og mening?

- Hvilke meningsaspekter kan vi fange inn via en mengde ord?
  - generelle tema, hva handler dokumentet om?
  - entiteter (mennesker, steder, datoer) som omtales i dokumentet
- Hvordan kan vi bruke slikes samlinger innenfor språkteknologi?

  - dokumenter om lignende temaer inneholder like ord
  - bruk i informasjonsgjenfinning (Information Retrieval (IR), dvs søk)

Samlinger av ord og mening?

- Dokumenter kan representeres som frekvenstabeller

> **Doc2:** *story, love, blood, samurai, movie, "13 assassins", man, courtyard, mask, shirt, blade, hand, director, Takasji Miike, showing, tale, revenge, liberation, stakes*

story	love	blood	samurai	movie	"13 assassins"	man	mask
4	6	10	14	13	3	2	1

Samlinger av ord og mening?

- Vi kan sammenligne dokumenter ved å sammenligne frekvenstabeller
- Hva kan vi si om det andre dokumentet?

story	love	blood	samurai	movie	”13 assassins”	man	mask
4	6	10	14	13	3	2	1

story	love	blood	samurai	movie	”13 assassins”	man	mask
4	16	0	0	13	0	2	0

Tabeller og vektorer

story	love	blood	samurai	movie	”13 assassins”	man	mask
4	16	0	0	13	0	2	0

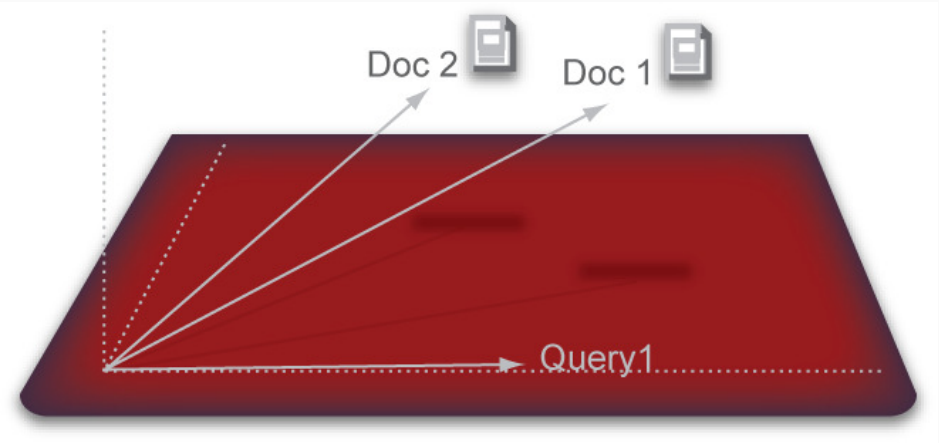
- Tolke tabell som en vektor
- Hvert oppslag er en dimensjon
  - *story* er en dimensjon, koordinat: 4
  - *love* er en dimensjon, koordinat 16
- Dette dokumentet er et punkt i et 10-dimensjonalt rom
- vektoren $\overrightarrow{d} = (4, 16, 0, 0, 13, 0, 2, 0)$

Tabeller og vektorer

- Eksempel: to-dimensjonalt rom *fried* og *chicken*
  - $q = (1, 1)$
  - $d_1 = (2, 8)$ (fried chicken recipe)
  - $d_2 = (0, 6)$ (poached chicken recipe)

Tabeller og vektorer

- Likhet mellom dokumenter blir til distanse i rommet
- Representasjonen kan enkelt (og effektivt) beregnes: trenger bare å telle ord
- Likhet i vektorrommet forutsier likhet i tema
    - Dokumenter som inneholder like ord handler ofte om det samme!



Vektorlikhet

- Hva mener vi med likheten mellom to vektorer?
- Flere standardmetoder fra vektormatematikk som baserer seg på avstand i vektorrommet
  - Cosine similarity (cosinus av vinkelen mellom dokumentene)

$$sim(\overrightarrow{p}, \overrightarrow{q}) = \frac{\sum_{i=1}^{n}(p_i \times q_i)}{\sqrt{\sum_{i=1}^{n} p_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}$$

  - Identiske dokumenter $sim(\overrightarrow{p}, \overrightarrow{q}) = 1$

Vektorlikhet

- Vi kan beregne likhet mellom dokumenter
  - cosinus
- Vi kan også representere en **søkestreng** ("query") som en vektor:
  - Teller ordene i søkestrengen
- Dermed kan vi søke etter dokumenter som ligner på søkestrengen

Termvekting

- Alle ord er ikke like informative
  - *movie* vs *the*
  - en samling om bildeler vil inneholde *car* i hvert eneste dokument
- Hvordan kan vi vekte ord slik at informative ord teller mer og ikke-informative ord teller mindre?
- bruke dokumentfrekvensen til en term df_t – antall dokumenter som inneholder termen

Termvekting

- **Inverse document frequency (IDF)**:
  - N dokumenter, n_{movie} av dem inneholder *movie* (df_{movie})

$$\text{idf}_{\text{movie}} = \log \frac{N}{n_{\text{movie}}}$$

  - idf for en sjelden term er høy, men for frekvente termer vil den være lav
  - logaritmisk for å skalere for høyt antall dokumenter

term	df_t	idf_t
car	18165	1.65
auto	6723	2.08
insurance	19241	1.62
best	25235	1.5

(Reuters collection; 806791 dokumenter)

Termvekting

- Kan bruke IDF til å vekte frekvenser for termer (ord):
  - forekomster av *movie* i Doc2: $\mathrm{tf}_{2,\mathrm{movie}}$
  - **tf-idf vekting:**

$$\mathrm{v}_{1,\mathrm{movie}} = \mathrm{tf}_{2,\mathrm{movie}} \times \mathrm{idf}_{\mathrm{movie}}$$

  - foretrekker ord som er vanlige i Doc2, men sjeldne i samlingen som helhet

Stemming og stopplister

- Vanlig førprosessering
  - **Tokenisering**
  - **Stemming**: slå sammen morfologiske varianter av ord *process*, *processing*, *processed* $\rightarrow$ *process*
  - **Stoppliste**: liste med høyfrekvente ord som utelates fra vektorrepresentasjonen
  - Høyfrekvente ord er ikke så informative

Fra dokumenter til ord

- Det samme gjelder for ord som for dokumenter: kontekstord er gode indikatorer på betydning
  - like ord forekommer i like kontekster
- Hva er en kontekst? Hvordan teller vi her?
  - alle forekomster av målordet i en stor tekst
  - bestemmer et kontekstvindu (f.eks. 10 ord til hver side)
  - tell alle ordene som befinner seg innenfor vinduet

Fra dokumenter til ord

A stirring, unexpectedly moving story of love and blood, the samurai **movie** 13 Assassins opens with a dignified man seated alone in a large courtyard.

stirring 1	unexpectedly 1	moving 1	story 1	of 1	love 1	and 1	blood 1	the 1	samurai 1
”13 assassins” 1	opens 1	with 1	a 2	dignified 1	man 1	seated 1	alone 1	in 1	

- teller for målordet *movie*, vindu på 10 ord
- en enkelt forekomst inneholder ikke så mye informasjon
  - gå gjennom alle forekomster av *movie* i et stort korpus
    - tell alle ord i et 10-ords vindu og pluss sammen

Fra tabeller til vektorer

- Tellinger for ordene *letter* og *surprise* fra "Pride and Prejudice"

	admirer	all	allow	almost	am	and	angry
letter	1	8	1	2	2	56	1
surprise	0	7	0	0	4	22	0

- Tolker tabellen som en vektor
  - hvert kontekstord er en dimensjon
    - *admirer* er en dimensjon
      koordinat for *letter*: 1; koordinat for *surprise*: 0
    - *all* er en dimensjon
      koordinat for *letter*: 8; koordinat for *surprise*: 7
    - ...

Hva er dette godt for?

- Ordrepresentasjoner i vektorrom kan beregnes automatisk: teller bare ord i konteksten
- Likhet i vektorrommet har vist seg å henge sammen med likhet i betydning
  - ord som forekommer i like kontekster har som regel lik betydning
  - synonymer er nær hverandre i vektorrommet

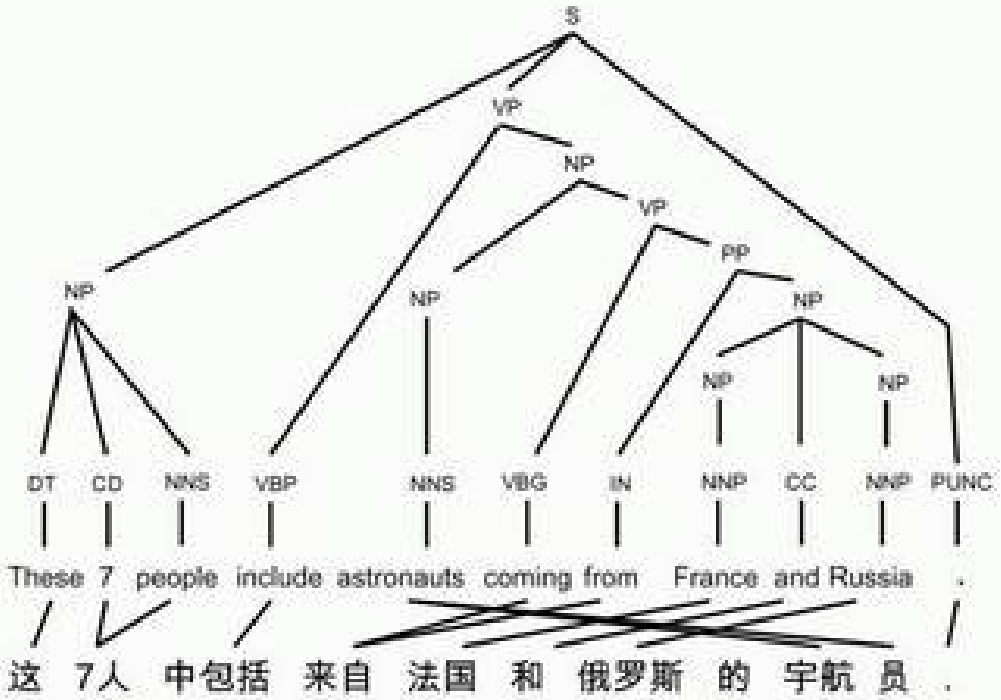
Maskinoversettelse (MT)

Maskinoversettelse (MT)

- Automatisk oversette fra et menneskelig språk til et annet
- For et eksempel, søk på en fremmedspråklig side i Google, klikk på "Translate this page"
- Hvilke erfaringer har dere med maskinoversettelse?

Hvorfor er maskinoversettelse vanskelig?

- Typologiske forskjeller mellom språk:
  - morfologisk
  - syntaktisk



Hvorfor er maskinoversettelse vanskelig?

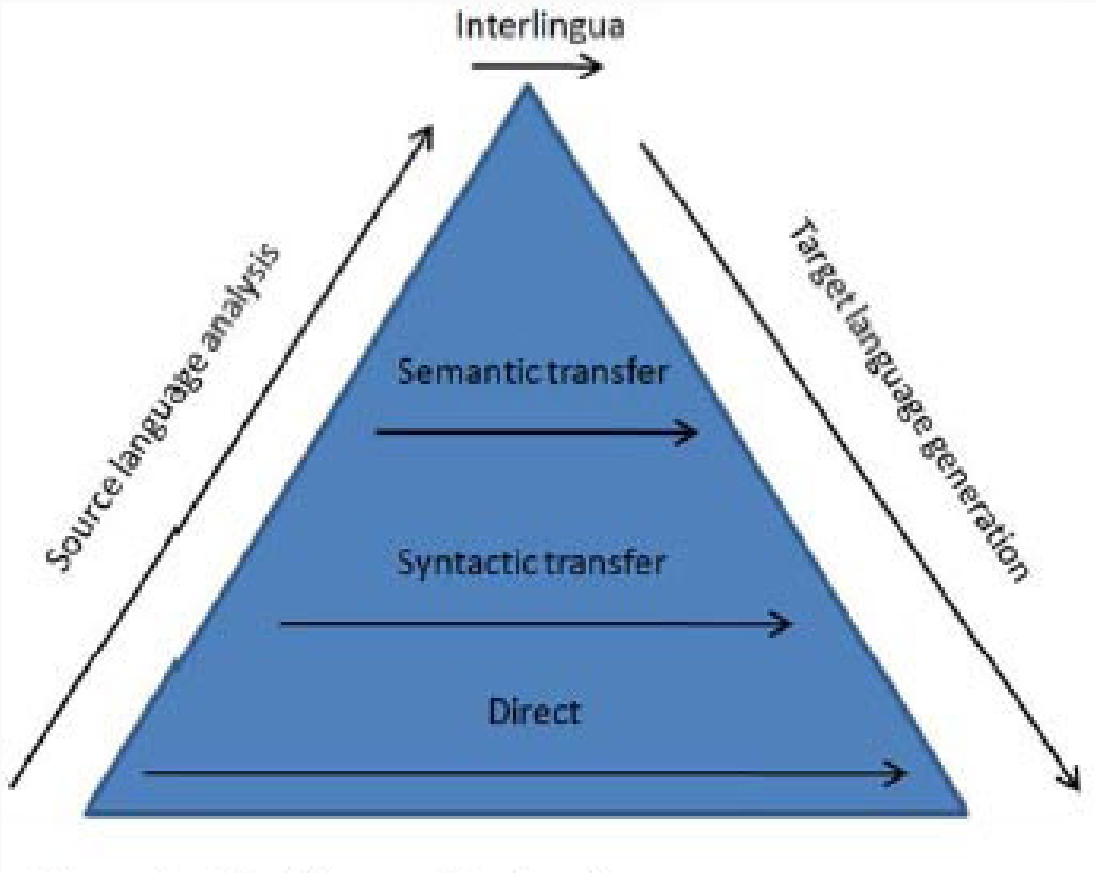
- Leksikale forskjeller
  - homonymi/polysemi
    - The spirit is willing but the flesh is weak
    - The vodka is excellent but the meat is lousy
    - (Fra http://babelfish.altavista.com/, teknologi utviklet av SYSTRAN)
  - språkspesifikke distinksjoner,
    - engelsk *know* $\rightarrow$ fransk *savoir/connaître*
    - engelsk *wall* $\rightarrow$ tysk *Wand/Mauer*
  - "lexical gaps" (Japansk *privacy*)
  - idiomatisk uttrykk: *kick the bucket*, *gå bort* (begge = *å dø*)

Kvalitet

- Hva slags kvalitet kan vi forvente?
  - bra nok for oversettelse av litteratur?
  - ujevn men forståelig?
- Hva kan vi gjøre med en ujevn men forståelig oversettelse?

Metoder

- 3 hovedmetoder:
  - **direkte**: oversetter ord for ord
  - **transfer**: syntaktisk analyse og deretter transformasjonsregler
  - **interlingua**: semantisk analyse til abstrakt meningsrepresentasjon



Direkte MT

- Ord for ord gjennom kildeteksten, oversetter underveis
- Lite prosessering (morfologisk analyse)
- Basert på et stort tospråklig leksikon
- Enkle transformasjonsregler på ordnivå

Direkte MT

Et eksempel: engelsk $\rightarrow$ spansk

- Morfologisk analyse:
  - Mary didn't slap the green witch
  - Mary DO-PAST not slap the green witch
- Leksikalsk transfer:
  - Mary DO-PAST not slap the green witch
  - Maria no dar:PAST una bofetada a la verde bruja
- Leksikalske transformasjonsregler:
  - Maria no dar:PAST una bofetada a la verde bruja
  - Maria no dar:PAST una bofetada a la bruja verde
- Morfologisk generering:
  - Maria no dió una bofetada a la bruja verde

Direkte MT

- Ingen kunnskap om syntaktisk struktur (f.eks. fraser)
- Kan ikke håndtere leddstillingsforskjeller som involverer grupper av ord (fraser)
- engelsk $\rightarrow$ tysk
  - The green witch is at home this week
  - Diese Woche ist die grüne Hexe zu Hause
- engelsk $\rightarrow$ japansk
  - He adores listening to music
  - kare ha ongaku wo kiku no ga daisuki desu
    (he music to listening adores)

Transfer MT

- Forskjeller i leddstilling mellom språk er **systematiske**
- Disse forskjellene kan beskrives i systemet
  (s = source, t = target)
  1. analyse ($\text{input}_s \rightarrow \text{syntaks}_s$)
  2. transfer ($\text{syntaks}_s \rightarrow \text{syntaks}_t$)
     i. syntaktisk transfer
     ii. leksikalsk transfer
  3. generering ($\text{syntaks}_t \rightarrow \text{output}_t$)

Transfer MT

- Syntaktisk transfer: transformasjonsregler
  - engelsk $\rightarrow$ spansk:
    - NP $\rightarrow$ Adj N $\Rightarrow$ NP $\rightarrow$ N Adj
  - engelsk $\rightarrow$ japansk:
    - VP $\rightarrow$ V NP $\Rightarrow$ VP $\rightarrow$ NP V
    - PP $\rightarrow$ P NP $\Rightarrow$ PP $\rightarrow$ NP P
    - NP $\rightarrow$ NP RelClause $\Rightarrow$ NP $\rightarrow$ RelClause NP
- Leksikalsk transfer
  - basert på tospråklig leksikon
  - oversettelse for idiomatiske fraser, f.eks. *at home* $\rightarrow$ *zu Hause*
  - noen systemer bruker også betydningsdisambiguering (WSD)

Kombinasjoner

Systran-systemet:

- **Analyse**
  - morfologisk analyse
  - chunking (NP, PP, etc.)
  - parsing (syntaktiske trær/funksjonell analyse)
- **Transfer**
  - oversettelse av idiomatiske uttrykk
  - WSD
  - tildele preposisjoner basert på verb
- **Syntese**
  - leksikal overettelse (tospråklig leksikon)
  - transformasjoner
  - morfologisk generering

Interlingua MT

- Annet perspektiv:
  - trekke ut **mening** fra kildesetningen og uttrykke denne i målspråket
- **Interlingua**: representerer alle setninger som betyr det samme på samme måte, uavhengig av språket det er uttrykt i
- en språkuavhengig, generell form
- hvordan kan den se ut?
  - $\exists$ x: is_witch(x) $\wedge$ is_green(x) $\wedge$ is_def(x) $\wedge\neg$(slap(m,w))
  - temporal informasjon

Interlingua MT

- semantiske roller

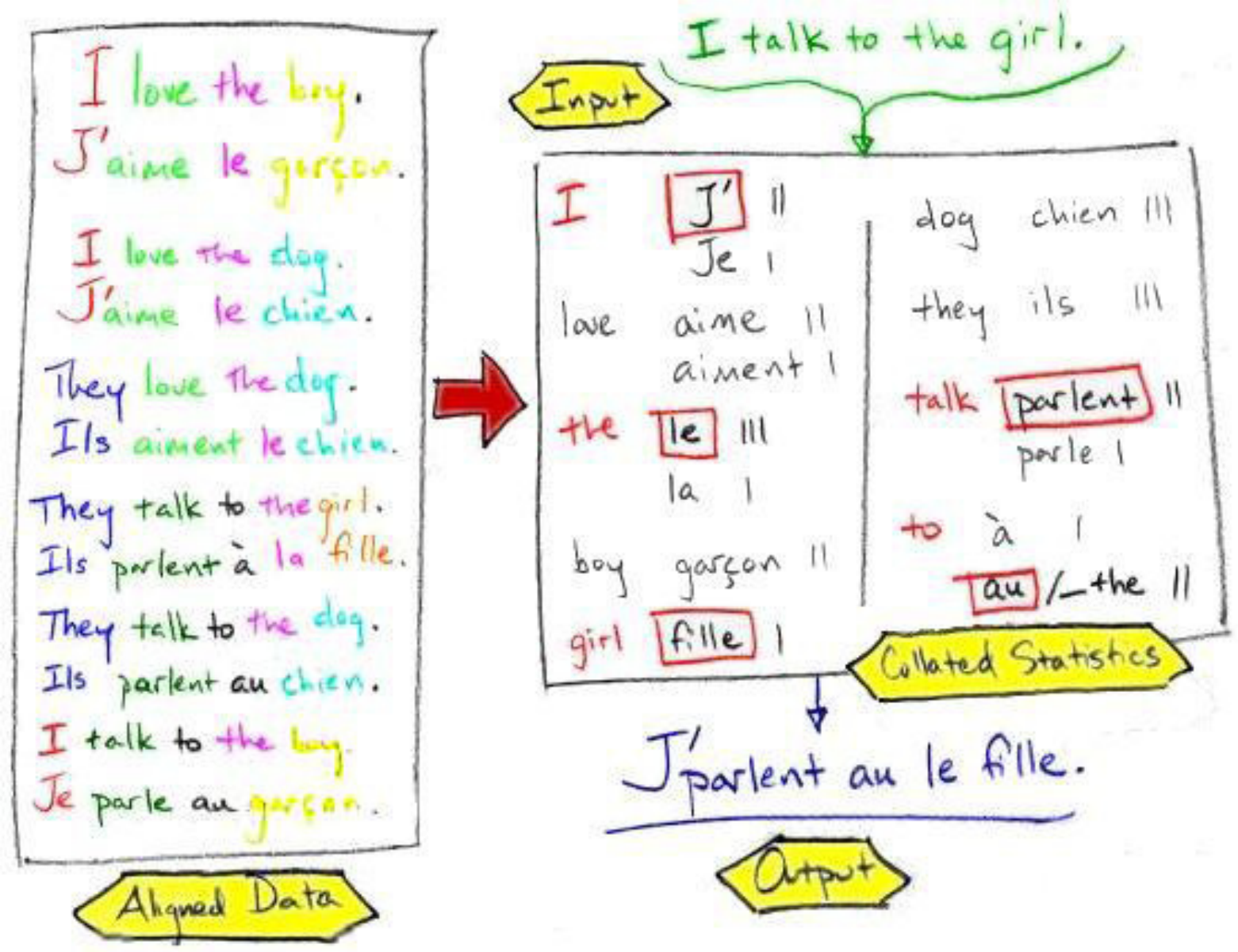
EVENT	SLAPPING	
AGENT	MARY	
TENSE	PAST	
POLARITY	NEG	
PATIENT	WITCH	
	DEFINITENESS	DEF
	ATTRIBUTES	GREEN

- generering direkte fra semantisk form

Metoder

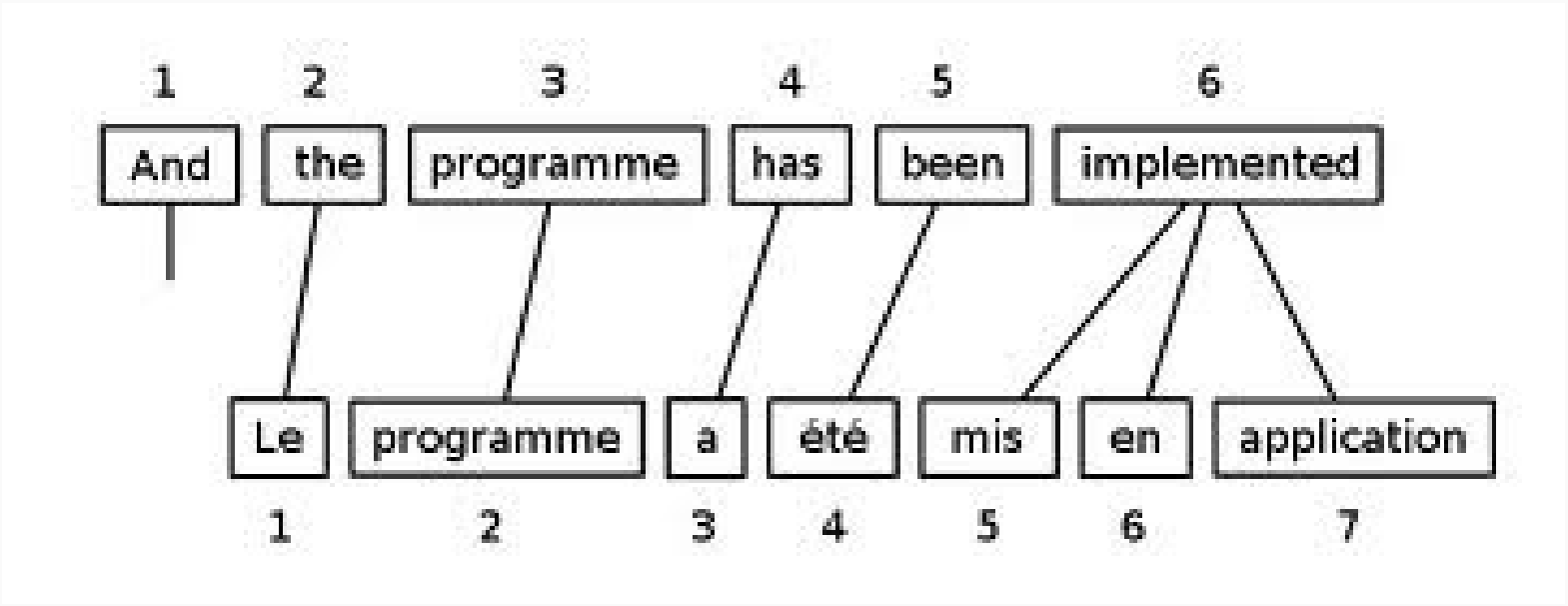
- Tidligere systemer (direkte, transfer, interlingua) stort sett **regelbaserte** systemer
- Fokus på oversettelsesprosessen: hvilke representasjoner, hvilke skritt
- Alternativ: fokus på resultatet
  - representerer et kompromiss
  - output maksimerer en funksjon som tar hensyn til både trofasthet (mot kildesetningen) og flyt (for målsetningen)
- Dominerende metoden de siste 15 årene: statistisk MT

Statistisk MT



Statistisk MT

- Fra engelsk $\rightarrow$ "gebrokken" fransk
  - via **parallelt korpus**: samme setninger på hvert språk
  - lærer hvor ofte hvert engelsk ord forekommer sammen med forskjellige franske ord: P(Fransk|Engelsk)
  - beregner sannsynlighetene ved ordlenkinger (word alignments)



Statistisk MT

- Fra gebrokken fransk $\rightarrow$ fransk
  - bruker en **språkmodell**
  - husker dere hva en **språkmodell** gjør?
  - hvordan kan vi bruke den her?

Statistisk MT

- Fra gebrokken fransk $\rightarrow$ fransk
  - bruker en **språkmodell**
  - husker dere hva en **språkmodell** gjør?
  - hvordan kan vi bruke den her?
- **Språkmodell**: P(F), sannsynligheten for en fransk setning
- **Oversettelsesmodell**: P(E|F)

$$\hat{F} = \underset{F \in \text{French}}{\text{argmax}}\, P(E|F)P(F)$$

Oppsummering

- Sett på en rekke språkteknologiske applikasjoner
- Berører kjente **lingvistiske nivåer**:
  - morfologi (NER, MT)
  - ordklasser (IE)
  - syntaks (IE, NER)
  - semantikk (MT)
- Berører kjente **språkteknologiske oppgaver**:
  - tokenisering/stemming (IE, IR, MT)
  - ordklassetagging (IE, MT)
  - morfologisk analyse (NER, MT)
  - chunking (NER)
  - betydningsdisambiguering (MT)

Oppsummering

- Berører kjente **metoder/modeller**:
  - regulære uttrykk/FSA (IE, NER)
  - språkmodeller (MT)
  - klassifisering (NER)
- Nytt: vektorromrepresentasjoner