IN1140: Introduksjon til språkteknologi

Forelesning #11

Samia Touileb

Universitetet i Oslo

07. november 2019

Tema for i dag

- Språkteknologiske applikasjoner
- Se på noen eksempler
    - Teknikker i informasjonsekstraksjon
    - Leksikon for sentimentanalyse
    - Dialogsystemer og chatbots

Språkteknologiske applikasjoner

- Mange språkteknologiske applikasjoner kombinerer teknikker vi har sett på tidligere i kurset.
- Motiverer de teknikkene vi har lært.
- Introdusere nye metoder.

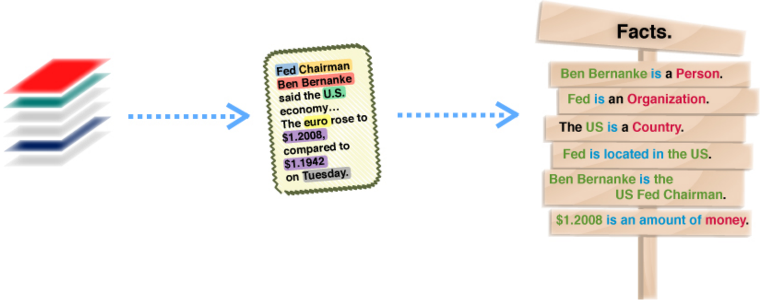
Informasjonsekstraksjon (IE)

Informasjonsekstraksjon (IE)

- Samlebegrep rundt teknikker som henter ut (ekstraherer) forskjellige typer semantisk informasjon fra tekst.
- Ustruktert informasjon ⇒ 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)



Fed Chairman Ben Bernanke said the U.S. economy… The euro rose to $1.2008, compared to $1.1942 on Tuesday.

Facts.

Ben Bernanke is a Person.

Fed is an Organization.

The US is a Country.

Fed is located in the US.

Ben Bernanke is the US Fed Chairman.

$1.2008 is an amount of money.

- Named entity recognition
- Relation extraction
- Coreference resolution
- Event extraction

Named Entity Recognition (NER)

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp. said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Named Entity Recognition (NER)

Citing high fuel prices, [$_{ORG}$ United Airlines] said [$_{TIME}$ Friday] it has increased fares by [$_{MONEY}$ \$6] per round trip on flights to some cities also served by lower-cost carriers. [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PER}$ Tim Wagner] said. [$_{ORG}$ United], a unit of [$_{ORG}$ UAL Corp]. said the increase took effect [$_{TIME}$ Thursday] and applies to most routes where it competes against discount carriers, such as [$_{LOC}$ Chicago] to [$_{LOC}$ Dallas] and [$_{LOC}$ Denver] to [$_{LOC}$ San Francisco].

Relation Extraction

- SpokesmanOf(Tim Wagner, American Airlines)
- UnitOf(United, UAL Corp.)
- UnitOf(American Airlines, AMR Corp.)

Coreference Resolution

But the little prince could not restrain admiration: "Oh! How beautiful you are!"

"Am I not?" the flower responded, sweetly. "And I was born at the same moment as the sun ..."

The little prince could guess easily enough that she was not any too modest–but how moving–and exciting–she was!

"I think it is time for breakfast," she added an instant later. "If you would have the kindness to think of my needs–"

And the little prince, completely abashed, went to look for a sprinkling-can of fresh water. So, he tended the flower.

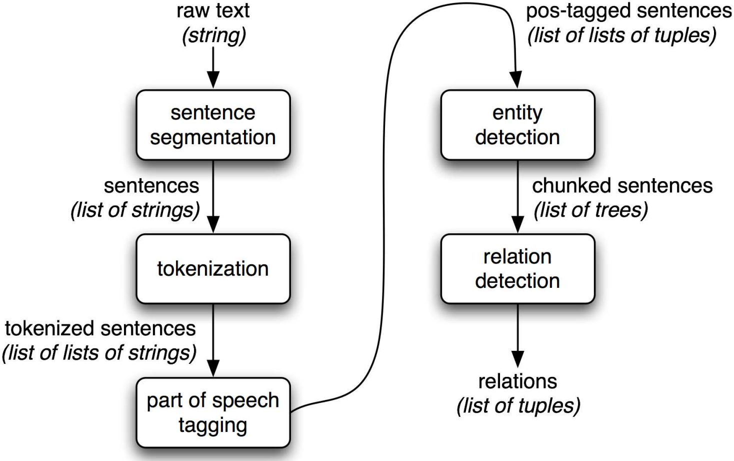
Event Extraction

- Who did what to whom, when, where, through what methods (instruments) and why.
- Innebærer bruk av NER og relation extraction.

"Masked gunmen armed with assault rifles and grenades attacked a wedding party in mainly Kurdish southeast Turkey, killing at least 44 people."

perpetrators	masked gunmen
victims	people
number of killed/injured	at least 44
weapons and means used	rifles and grenades
location	southeast Turkey
event	attacked

Informasjonsekstraksjon (IE)



raw text
(string)

sentence segmentation

sentences
(list of strings)

tokenization

tokenized sentences
(list of lists of strings)

part of speech tagging

pos-tagged sentences
(list of lists of tuples)

entity detection

chunked sentences
(list of trees)

relation detection

relations
(list of tuples)

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 (Geopolitical entity)
- PERSON

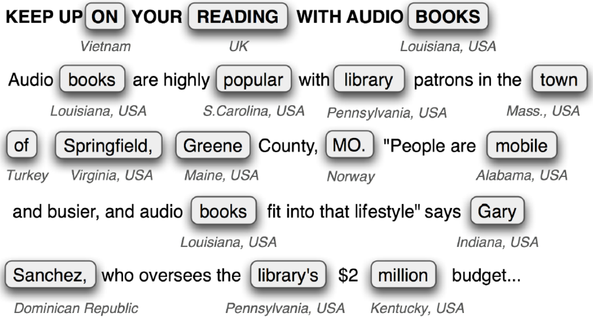
Named Entity Recognition (NER)

- 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 (feks. “gazetteer”)?



KEEP UP ON YOUR READING WITH AUDIO BOOKS

Audio books are highly popular with library patrons in the town

of Springfield, Greene County, MO. "People are mobile

and busier, and audio books fit into that lifestyle" says Gary

Sanchez, who oversees the library's $2 million budget...

Vietnam
UK
Louisiana, USA
Louisiana, USA
S.Carolina, USA
Pennsylvania, USA
Mass., USA
Turkey
Virginia, USA
Maine, USA
Norway
Alabama, USA
Louisiana, USA
Indiana, USA
Dominican Republic
Pennsylvania, USA
Kentucky, USA

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 ...*

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)

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 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

Named Entity Recognition (NER)

- NER-systemer oftest kombinasjon av
  - Lister
  - Regler (regel-baserte systemer)
  - Veiledet ("supervised") klassifisering
- Beste systemer for engelsk: 92% for PERSON, LOCATION, 84% for ORGANIZATION
- Finnes ikke for alle typer for norsk foreløpig (kommer veldig snart), finnes NEC (Named Entity Chunking) og NER (Named Entity Recognition) for noen entity typer (Johansen, 2015 og 2019).
- Annotert korpus for norsk NorNE (2019) (`https://www.nb.no/en/artikler/norwegian-named-entity-recognition/`).

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 → ORG: spokesman_of(Tim Wagner, American Airlines)
- ORG → 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/fraser 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 (NP, VP, ...), etc.

Temporal Expression Extraction

Absolute	**Relative**	**Durations**
April 24, 1916	yesterday	four hours
The summer of ’77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2016	last quarter	the last three quarters

Leksikon for sentimentanalyse

Affective states

- Affective meaning = emotion, sentiment, personality, mood, attitudes.
- Affective meaning is related to subjectivity (the study of a writer/speaker's evaluations, opinions, emotions, and speculations).

Affective states

Emotion: angry, sad, joyful, ashamed, proud, elated, desperate.
Mood: cheerful, gloomy, irritable, listles, depressed.
Interpersonal stance: distant, cold, warm, supportive, friendly.
Attitude: liking, loving, hating, valuing, desiring.
Personality trait: nervous, anxious, reckles, morose, hostile, jealous.

Emotions – 2 teorier

- Teori 1: Plutchik (1980) wheel of emotion.

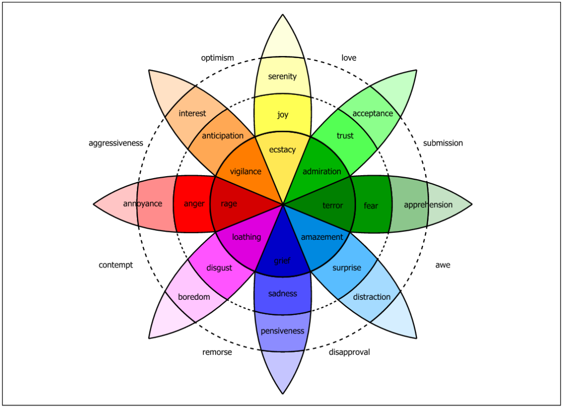


Figure 21.2 Plutchik wheel of emotion.

`https://web.stanford.edu/~jurafsky/slp3/21.pdf`

- Teori 2: Russel (1980):
  - valence: pleasantness of the stimulus.
  - arousal: the intensity of emotion provoked by the stimulus.
  - dominance: the degree of control exerted by the stimulus.

Sentiment Analaysis

`https://www.clickworker.de/wp-content/uploads/2017/02/Sentiment-Analyse.jpg`

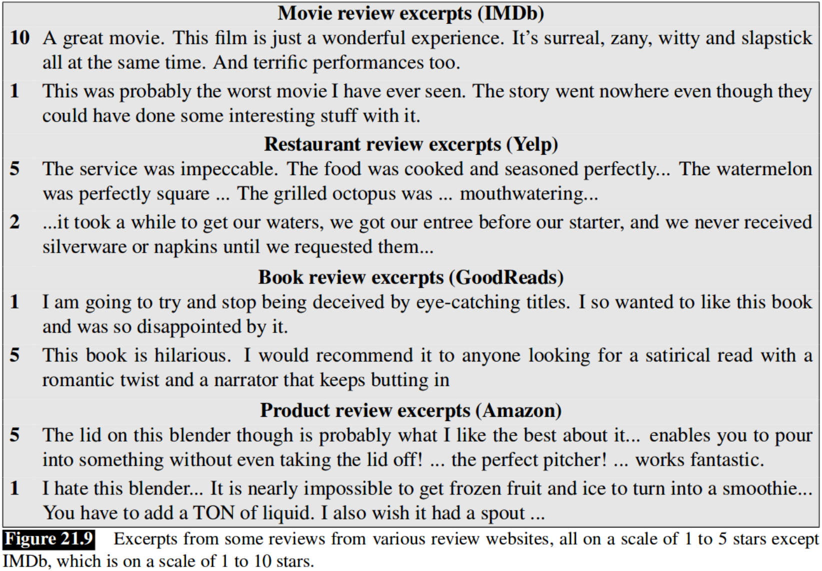
Sentiment leksikon

- General Inquirer: eldste, 1966. Inneholder 1915 positive ord og 2291 negative ord.
- MPQA (2005): 2718 positive og 4912 negative ord.
- Hu and Liu (2004): 2006 positive og 4783 negative ord.
- NRC Valence, Arousal, Dominance (VAD, 2018): 20.000 ord.
- EmoLex (2013): 14.000 ord (8 emotions).
- Norsk sentiment leksikon (2019): 1.601 positive og 3.917 negative ord. Fullform: 6.103 positive og 14.839 negative ord. `https://github.com/ltgoslo/norsentlex`

Å lage sentiment-leksikoner

- Manuelt: eksperter lager lister av positive og negative ord.
- Semi-supervised: starte med “seed words” f.eks. *god* og *dålig*. Regne ut hvorvidt hvert ord w som skal bli klassifisert, er lik ordet *god*, og hvorvidt det er annerledes fra ordet *dårlig*
- Supervised: bruke anmeldelser (f.eks terningkast, stjerner) for å lage leksion av positive og negative ord.

Supervised



	Movie review excerpts (IMDb)
10	A great movie. This film is just a wonderful experience. It's surreal, zany, witty and slapstick all at the same time. And terrific performances too.
1	This was probably the worst movie I have ever seen. The story went nowhere even though they could have done some interesting stuff with it.
	Restaurant review excerpts (Yelp)
5	The service was impeccable. The food was cooked and seasoned perfectly... The watermelon was perfectly square ... The grilled octopus was ... mouthwatering...
2	...it took a while to get our waters, we got our entree before our starter, and we never received silverware or napkins until we requested them...
	Book review excerpts (GoodReads)
1	I am going to try and stop being deceived by eye-catching titles. I so wanted to like this book and was so disappointed by it.
5	This book is hilarious. I would recommend it to anyone looking for a satirical read with a romantic twist and a narrator that keeps butting in
	Product review excerpts (Amazon)
5	The lid on this blender though is probably what I like the best about it... enables you to pour into something without even taking the lid off! ... the perfect pitcher! ... works fantastic.
1	I hate this blender... It is nearly impossible to get frozen fruit and ice to turn into a smoothie... You have to add a TON of liquid. I also wish it had a spout ...

Figure 21.9 Excerpts from some reviews from various review websites, all on a scale of 1 to 5 stars except IMDb, which is on a scale of 1 to 10 stars.

`https://web.stanford.edu/~jurafsky/slp3/21.pdf`

- Kan bruke anmeldelser: positive ord forekommer mest sannsynlig i 5-stjerners anmeldelser, negative ord i 1-stjerners anmeldelser.
- Slike anmeldelser kan brukes for mer enn kun binære klassifisseringer.
- Vi kan identifisere fordeling over stjerner/terningkast for hvert ord.

Potts Score

Potts Score (Potts, 2011):

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

$$PottsScore(w) = \frac{P(w|c)}{\sum_{c} P(w|c)}$$

Potts Score

For eksempel:
Vi kan beregne IMDB sannsynligheten av forekomsten av ordet *disapointing* i en 1-stjerners anmeldelse.

$$P(disapointing|1) = \frac{count(disapointing, 1)}{\sum_{w \in 1} count(w, 1)} = \frac{8557}{25395214} = 0.0003$$

$$PottsScore(disapointing) = \frac{0.0003}{\sum_{c=1} P(w|1)} = 0.10$$

Og ordet *disapointing* er da associert med vektoren *[0.10, 0.12, 0.14, 0.14, 0.13, 0.11, 0.08, 0.06, 0.06, 0.05]* som representerer Potts score for hver klasse fra 1-10.

Potts Score

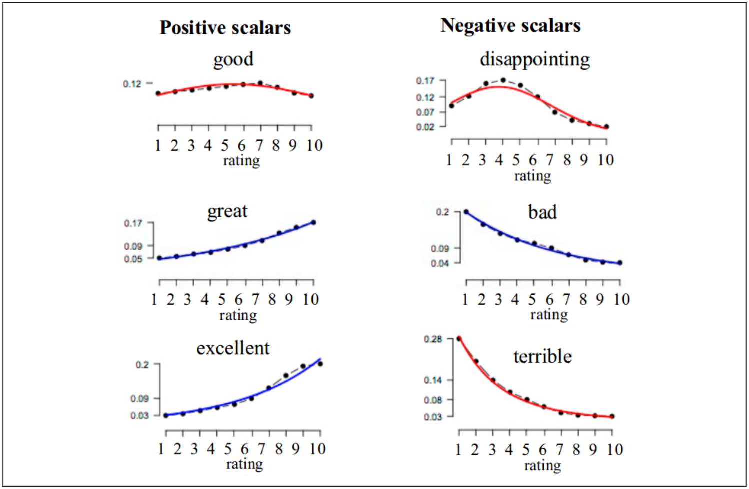


Figure 21.10 Potts diagrams (Potts, 2011) for positive and negative scalar adjectives, showing the J-shape and reverse J-shape for strongly positive and negative adjectives, and the hump-shape for more weakly polarized adjectives.

`https://web.stanford.edu/~jurafsky/slp3/21.pdf`

Bruk av leksikon

Sentiment leksikon kan:

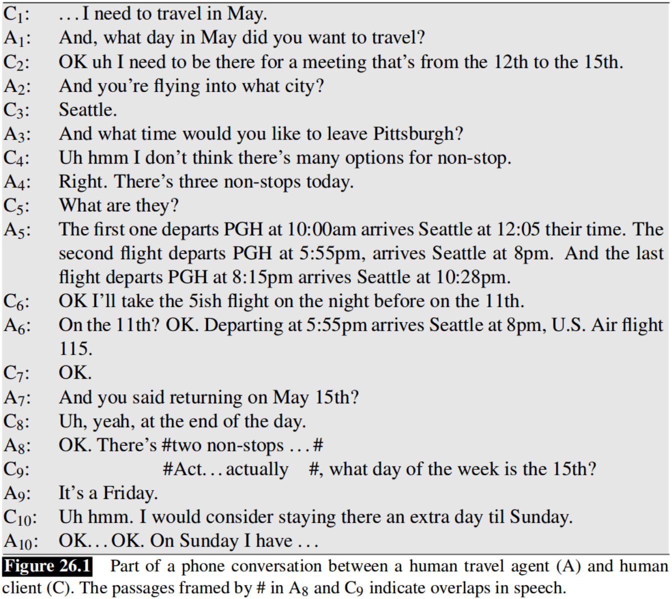
- brukes direkte ved å telle forekomster av ord.
- brukes som trekk i en Naive Bayes klassifissering.
- brukes med andre maskinlærings-algoritmer.

Dialogsystemer og chatbots

Dialogsystemer

- To hovedalgoritmer av dialogsystemer:
  1. Oppgaveorienterte dialogagenter.
  2. Chatbots.

Egenskaper av menneskelig samtaler



C_1: . . . I need to travel in May.
A_1: And, what day in May did you want to travel?
C_2: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2: And you're flying into what city?
C_3: Seattle.
A_3: And what time would you like to leave Pittsburgh?
C_4: Uh hmm I don't think there's many options for non-stop.
A_4: Right. There's three non-stops today.
C_5: What are they?
A_5: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_6: OK I'll take the 5ish flight on the night before on the 11th.
A_6: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7: OK.
A_7: And you said returning on May 15th?
C_8: Uh, yeah, at the end of the day.
A_8: OK. There's #two non-stops . . . #
C_9: #Act. . . actually #, what day of the week is the 15th?
A_9: It's a Friday.
C_{10}: Uh hmm. I would consider staying there an extra day til Sunday.
A_{10}: OK. . . OK. On Sunday I have . . .

Figure 26.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A_8 and C_9 indicate overlaps in speech.

`https://web.stanford.edu/~jurafsky/slp3/26.pdf`

Samtaler

- Hver sin tur til å snakke.
    - Men hvordan vet vi at den andre er ferdig med å snakke?
- Talehandling: be noen å gjøre noe, stille et spørsmål som krever et svar, gi en begrensning, takke. F.eks linje C_2 og C_1 i forrige bilde.
- Forankring: å si ok, eller repetere det den andre sa. F.eks. linje A_8, A_{10}, A_1.
- Struktur: rekkefølgende er ikke alltid konsekvent, av og til ombestemmer vi oss. F.eks. fra C_9 til A_{10}.
- Inferens og implikasjon: f.eks. A_1 og C_2.

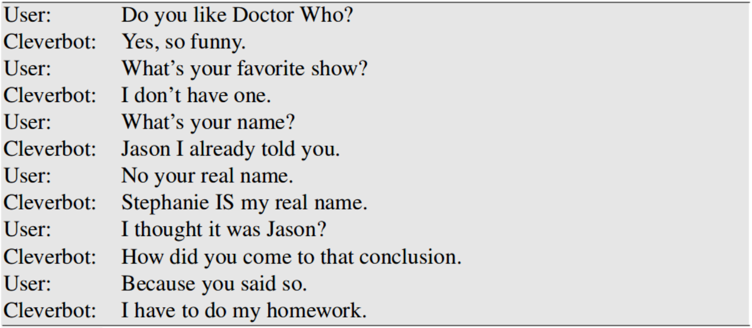
Regel-basert chatbot

- Første: Eliza (Weizenbaum, 1966).
- Regel-basert: bruk av regulære utrykk-lignende regler.

User: *Men are all alike*
ELIZA: IN WHAT WAY?
User: *They're always bugging us about something*
ELIZA: CAN YOU THINK OF AN EXAMPLE?
User: *Well, my boyfriend made me come here*
ELIZA: YOUR BOYFRIEND MADE YOU COME HERE.
User: *He says I am depressed much of the time*
ELIZA: I'M SORRY TO HEAR THAT YOU ARE DEPRESSED.

Chatbots

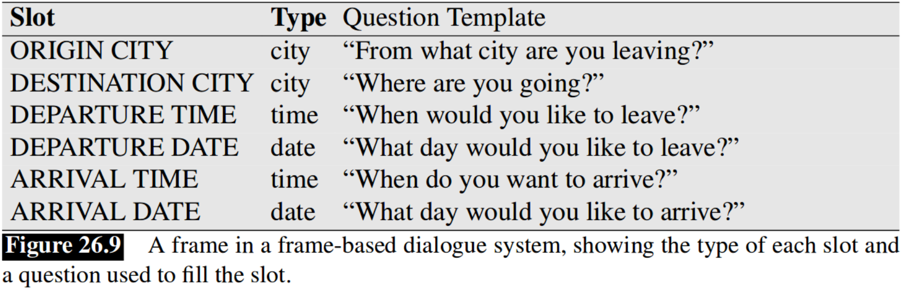
- Finnes også korpus-basert chatbots.
- Krever stor dataset av menneske – menneske samtaler.
- Bruker blant annet maskinlæringsmetoder for å trene modeller.
- Fokuset er å generere setninger, og de fleste chatbots sliter fortsatt med å beholde en sammenhengde samtale.



User:	Do you like Doctor Who?
Cleverbot:	Yes, so funny.
User:	What's your favorite show?
Cleverbot:	I don't have one.
User:	What's your name?
Cleverbot:	Jason I already told you.
User:	No your real name.
Cleverbot:	Stephanie IS my real name.
User:	I thought it was Jason?
Cleverbot:	How did you come to that conclusion.
User:	Because you said so.
Cleverbot:	I have to do my homework.

Figure 26.2 Chatbots: A conversation between a user and Cleverbot, an IR-based chatbot (Carpenter, 2017).

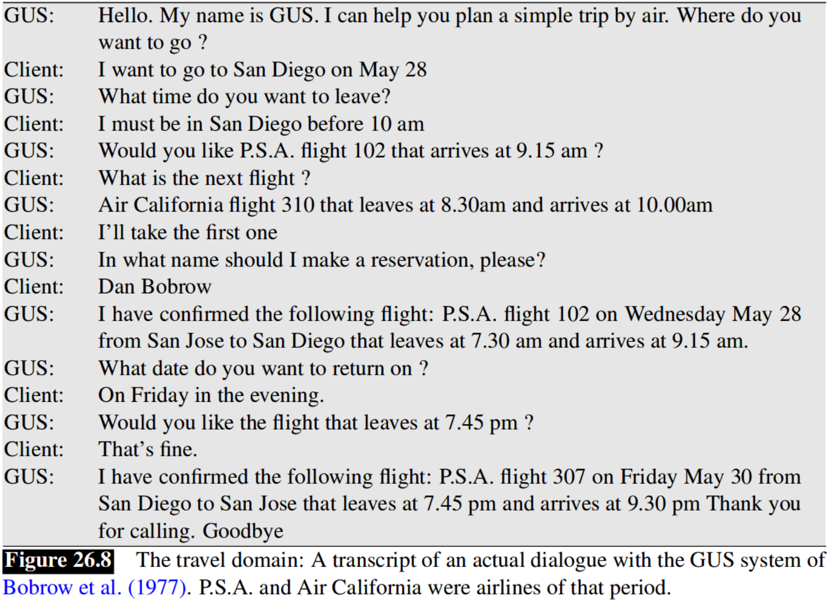
Ramme-basert dialogsystemer



Slot	**Type**	Question Template
ORIGIN CITY	city	"From what city are you leaving?"
DESTINATION CITY	city	"Where are you going?"
DEPARTURE TIME	time	"When would you like to leave?"
DEPARTURE DATE	date	"What day would you like to leave?"
ARRIVAL TIME	time	"When do you want to arrive?"
ARRIVAL DATE	date	"What day would you like to arrive?"

Figure 26.9 A frame in a frame-based dialogue system, showing the type of each slot and a question used to fill the slot.

Ramme-basert dialogsystemer



GUS: Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?
Client: I want to go to San Diego on May 28
GUS: What time do you want to leave?
Client: I must be in San Diego before 10 am
GUS: Would you like P.S.A. flight 102 that arrives at 9.15 am ?
Client: What is the next flight ?
GUS: Air California flight 310 that leaves at 8.30am and arrives at 10.00am
Client: I'll take the first one
GUS: In what name should I make a reservation, please?
Client: Dan Bobrow
GUS: I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.
GUS: What date do you want to return on ?
Client: On Friday in the evening.
GUS: Would you like the flight that leaves at 7.45 pm ?
Client: That's fine.
GUS: I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

Figure 26.8 The travel domain: A transcript of an actual dialogue with the GUS system of Bobrow et al. (1977). P.S.A. and Air California were airlines of that period.

Oppsummering

- Sett på en rekke språkteknologiske applikasjoner
- Berører kjente **lingvistiske nivåer**:
    - Morfologi
    - Ordklasser
    - Syntaks
    - Semantikk
- Berører kjente **språkteknologiske oppgaver**:
    - Tokenisering/stemming
    - Ordklassetagging
    - Morfologisk analyse
    - Chunking
    - Betydningsdisambiguering

Oppsummering

- Berører kjente **metoder/modeller**:
    - Regulære uttrykk
    - Klassifisering

Prøveeksamen IN1140

Prøveeksamen er et frivillig tilbud hvor du kan teste deg på reelle eksamensoppgaver i det systemet som brukes til eksamen. Logg inn på `uio.inspera.no`, så skal du finne link til prøveeksamen.

Oppgavesettet er tilgjengelig fra **fredag 8.11 til onsdag 27.11**.

Løsningsforslag vil bli gjennomgått på forelesningen **torsdag 21.11**. Fint om du kan jobbe med eksamen før forelesningen, da får du mer ut av løsningsforslaget.

Du kan gå inn og ut av oppgavesettet så mange ganger du vil i løpet av perioden, men for treningens skyld anbefaler vi at du setter av 4 sammenhengende timer for å trene mest mulig på selve eksamenssettingen.