IN1140: Introduksjon til språkteknologi

Forelesning #11

Egil Rønningstad

Universitetet i Oslo

November 2024

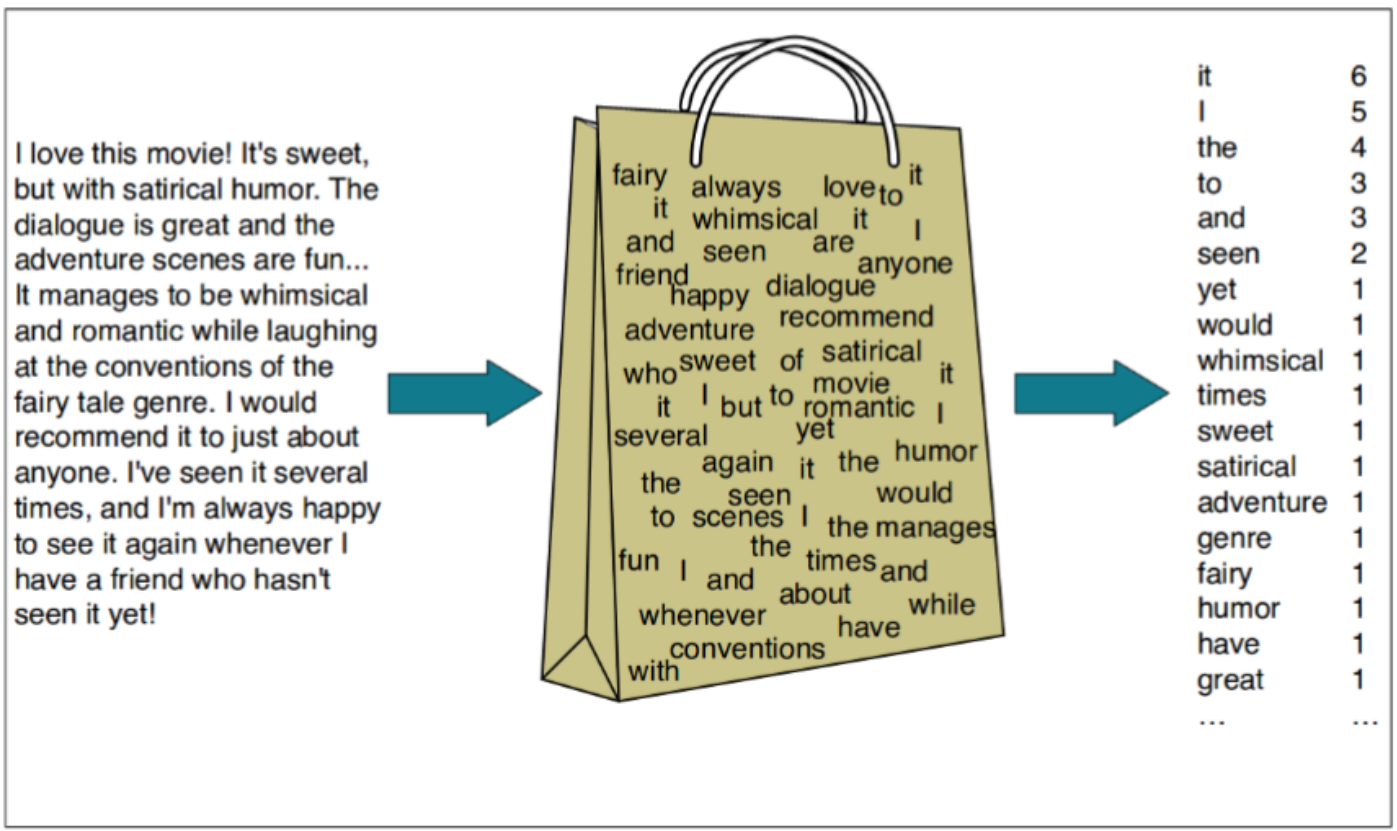


Tema for i dag

- Språkteknologiske applikasjoner
- Se på noen eksempler
  - Informasjonsekstraksjon
  - Sentimentanalyse
  - Question Answering
  - Dialogsystemer og chatbots

Repetere: Representere tekst som tall

- Vanligste trekk-type for tekstklassifikasjon: bag-of-words / BoW
- Hvert ord som forekommer i dokumentet utgjør et trekk, bryr oss ikke om rekkefølge eller posisjon.
- Unionen av alle trekkene, altså alle ordtypene, kaller vi vokabularet (V).



Repetere: BoW Trekkvektor

a)	Månen er en rund ost
b)	Jeg spiser og spiser en rund ost
c)	Jeg spiser hver måned

	en	er	hver	jeg	måned	månen	og	ost	rund	spiser
a)	1	1	0	0	0	1	0	1	1	0
b)	1	0	0	1	0	0	1	1	1	2
c)	0	0	1	1	1	0	0	0	0	1

Informasjonsekstraksjon (IE)

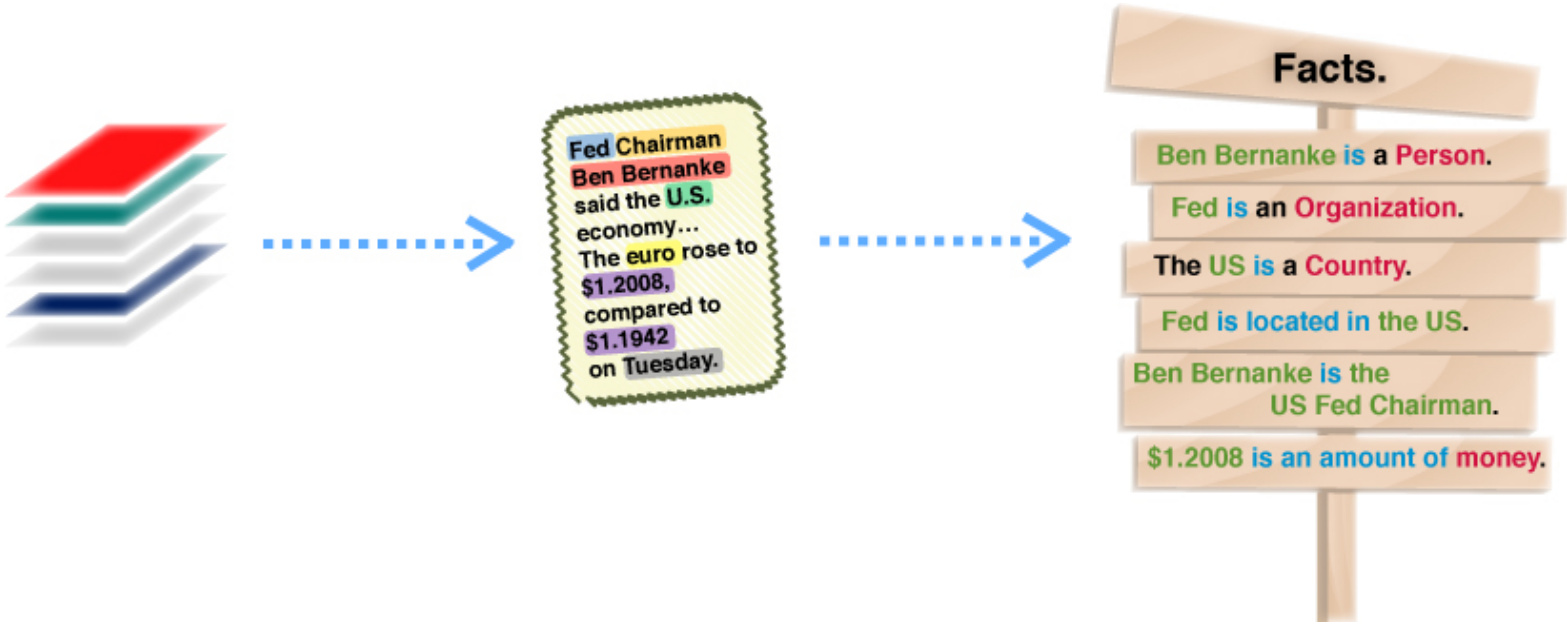
Informasjonsekstraksjon (IE) Forrige forelesning: Navnegjenkjenning. Finne egennavn til personer, steder og organisasjoner. Hvordan kan vi ta dette videre?

Informasjonsekstraksjon (IE)

- Samlebegrep for 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)



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

Relasjonsekstraksjon

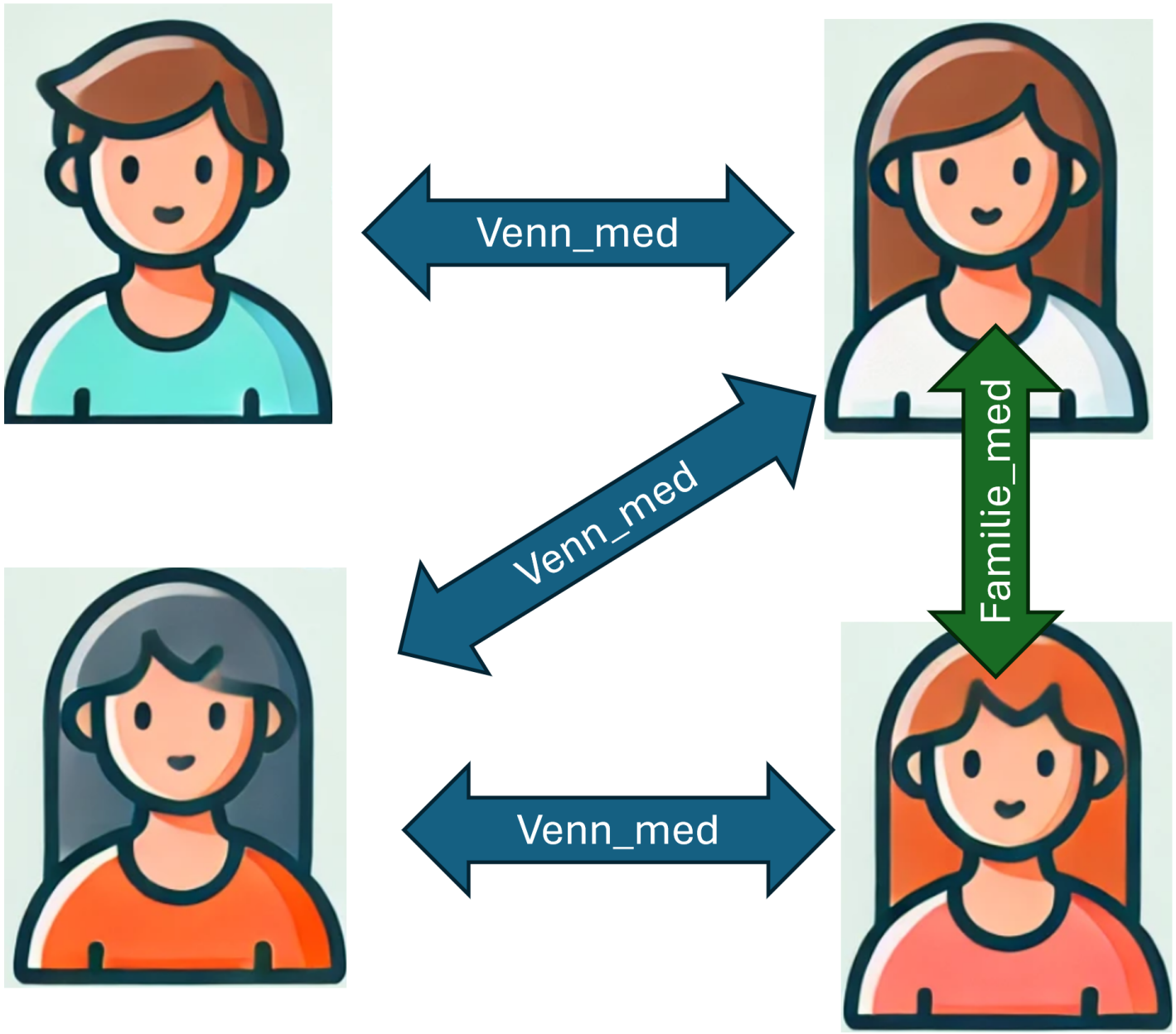
Finne fram til **relasjoner** mellom entitetene i en tekst

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)

Relasjonsekstraksjon



Relasjonsekstraksjon

Noen relasjoner går begge veier, mens andre relasjoner går kun én vei. Når datamengden blir stor, kan vi lagre informasjonen i dedikerte graf-databaser.

Venn_med(Diana, Alice)
Familie_med(Hannah, Charlie)
Venn_med(George, Ethan)
Familie_med(Bob, Fiona)
Venn_med(Alice, George)

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 (supervised learning), to oppgaver:

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

```
function FindRelations(words) returns relations
    relations ← nil
    entities ← FindEntities(words)
    forall entity pairs <e1, e2> in entities do
      if Related?(e1, e2)
        relations ← relations+ClassifyRelation(e1, e2)
```

Metoder for relasjonsekstraksjon

Klassifisering (supervised learning):

- Viktig del er å finne fram til gode **trekk**
  - NE-kategoriene (PER, ORG, LOC etc.)
  - 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.

Metoder for relasjonsekstraksjon

Kan også bruke:

- **Semisupervised learning via Bootstrapping**: bruke lite treningsdata, liten håndfull regulære uttrykk (RE mønstre), **bootstrap** en klassifiserer.
- **Distant supervision**: bruke en stor database for å hente ut stor andel eksempler av relasjoner og lage flere mønstre (med mye støy) og kombinære dem i en supervised klassifiserer.
- **Unsupervised**: også kjent som **open information extraction**. Bruker ikke annotert treningsdata, men bruker masse data fra f.eks. internet. Relasjonene her anses som strenger av ord som vanligvis begynner med et verb.

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

Event Extraction

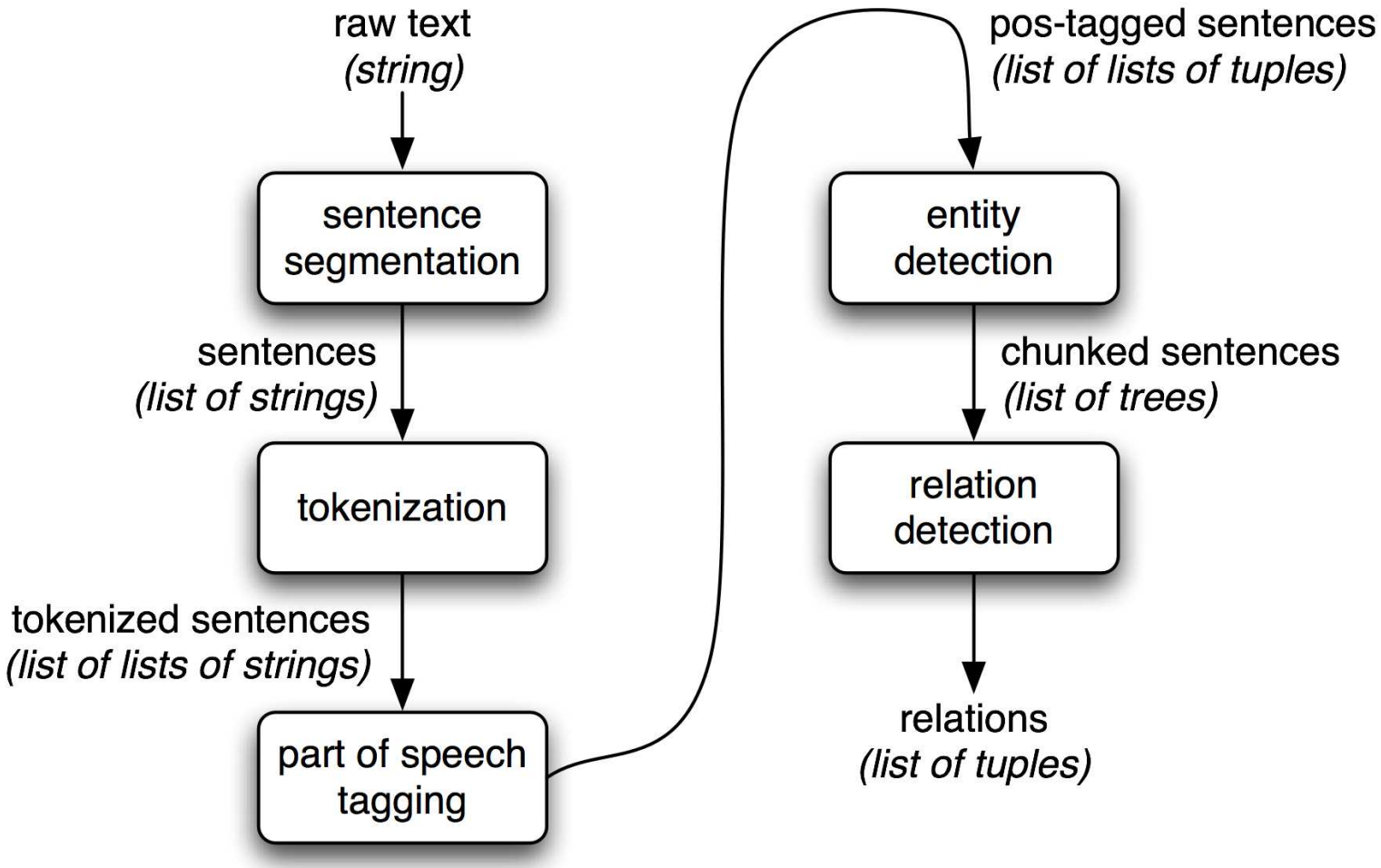
- Who did what to whom, when, where, through what methods (instruments) and why.

- 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)



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.

Sentimentanalyse



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

Sentimentleksikon

- Kan vi analysere sentimentet i en tekst ved å telle positive og negative ord?
- Hva gjør et ord til positivt eller negativt?

Sentimentleksikon

- 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 Sentimentleksikon (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 sentimentleksikon

- **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 leksikon 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 20.9 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/slp3/20.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 klassifiseringer.
- 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.

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

$$PottsScore(disapointing) = \frac{0.0003}{\sum_c P(w|C)} = 0.10$$

Og ordet *disapointing* er da assosiert 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

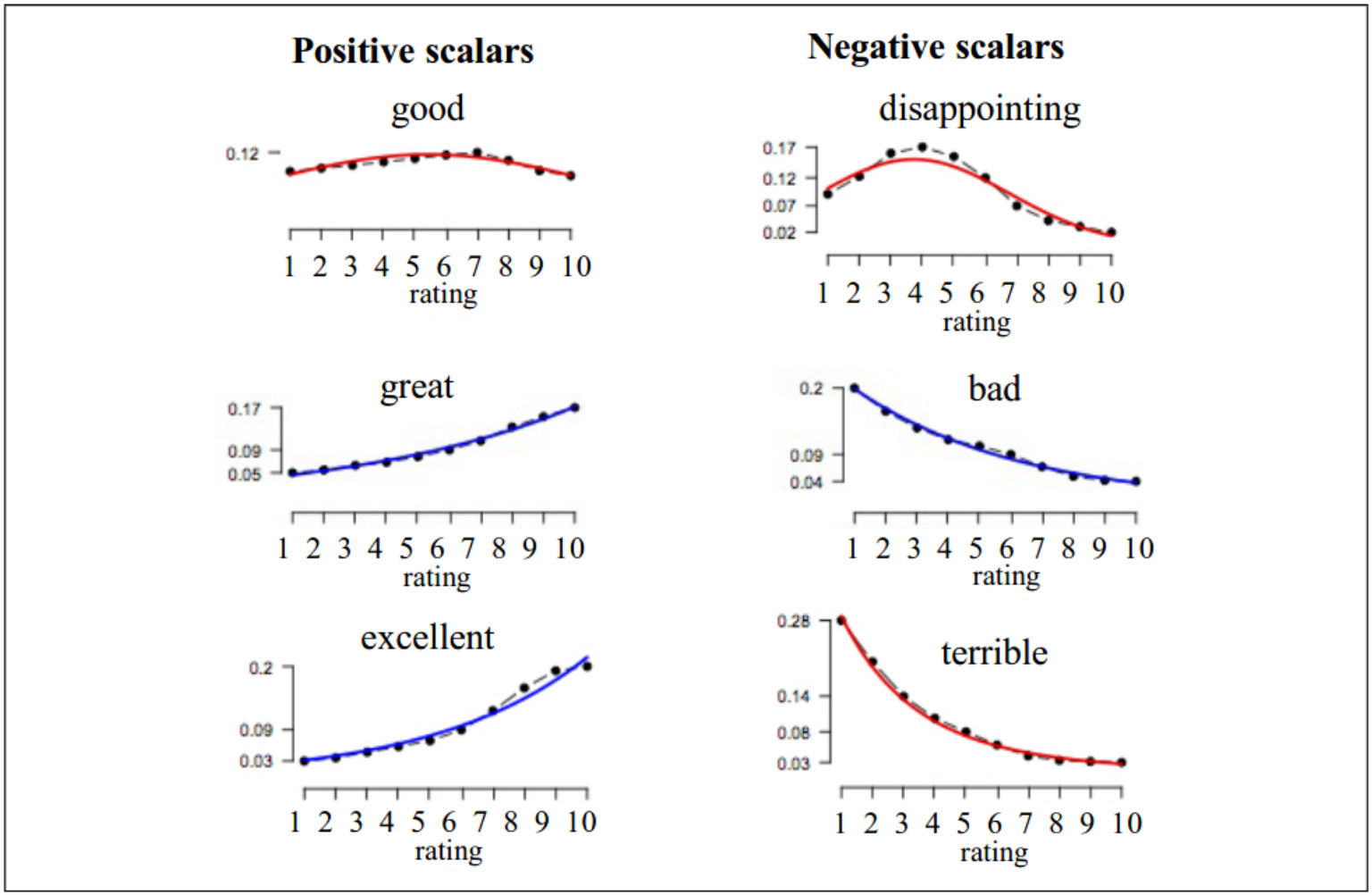


Figure 20.10 Potts diagrams (Potts, 2011) for positive and negative scalar adjectives, showing the J-shape and reverse J-shape for strongly positive and negative adjectives, and the hump-shape for more weakly polarized adjectives.

https://web.stanford.edu/~jurafsky/slp3/20.pdf

Trekkvektor for sentimentanalyse

	1	2	3	4	5	6	7	8	9	10
disappointing	0.10	0.12	0.14	0.14	0.13	0.11	0.08	0.06	0.06	0.05

Trekkvektor for sentimentanalyse

Ord	POS	NEG
jeg	0,09	0,16
liker	0,07	0,06
alltid	0,29	0,06
utenlandske	0,04	0,15
filmer	0,08	0,11

Ulike former for sentimentanalyse

- **Dokument-nivå:** Positivt eller negativt? Hvor positivt eller negativt
- **Setningsnivå:** Positiv eller negativ? Hvor positivt eller negativt
- **Aspekt-orientert:** Brukeromtaler: Vi kjenner hva som omtales, og hvilken rating brukeren ga, men hvilke **aspekter** ved det som omtales, var viktig for brukeren?
- **Entitets-orientert:** Litt tilsvarende men knyttet til personer og organisasjoner: Hvilket sentiment formidler teksten som et hele mot en gitt entitet som er nevnt i teksten?
- *"Structured sentiment analysis"*: For hver setning, hvilke ord representerer noe som blir omtalt positivt eller negativt? Hvilke ord representerer sentimentet som uttrykkes? Hvilke ord representerer hvem som har dette sentimentet?
- Stor forskjell på korte tekster (Brukeromtaler, Twitter) og lange tekster (Avisomtaler og -artikler)

Oppsummering trekkvektor

Trekkvektorer beskriver tekst med tall.
Hvert tall i trekkvektoren beskriver én egenskap ved tallet.
Første steg er å lage et vokabular, som tildeler et nummer til alle ord vi skal forholde oss til.

- Nabo-ord
- POS-tagging
- dependens-parsing
- NER-tagging
- Meta-data (For eksempel sentiment-score for dokumentet)

Trekkvektorer kan utarbeides både for enkeltord, for setninger, og for hele dokumenter. Trekkvektorer er grunnlag for maskinlæring på tekst, slik som Naïve Bayes.

Question Answering

Question Answering

- Begynnelsen av 1960-tallet: systemene brukte de to store paradigmene – informasjonsinnhentingsbasert (Information Retrieval) og kunnskapsbasert – for å svare på spørsmål om baseballstatistikk eller vitenskapelige fakta.
- Uekte maskiner: *Deep Thought*, datamaskinen i "The Hitchhiker's Guide to the Galaxy" klarte å svare på "*The Ultimate Question Of Life, The Universe, and Everything*"[1]
- 2011: IBMs **Watson** question-answering system (brukte en hybridarkitektur) vant TV-showet Jeopardy!

[1] 42!

Question Answering

- De fleste Question-Answering (QA) systemer fokuserer på fakta-basert spørsmål, spørsmål som kan besvares med enkle fakta uttrykt i korte tekster.

- Svarene på spørsmålene nedenfor kan uttrykkes med navn, tidsmessig uttrykk, eller sted:
  1. Hvem grunnla Virgin Airlines?
  2. Hva er gjennomsnittsalderen i Real Madrid?
  3. Hvor er SAS basert?

Question Answering – IR-based

SQL:

```
SELECT AVG(AGE)
FROM PLAYERS
WHERE CLUB = 'Real Madrid';
```

Question Answering – IR-based

- informasjonsinnhentingsbasert eller IR-basert QA er avhengig av store mengder tekster fra f.eks. nettet.
- Gitt et brukerspørsmål:
  - finner relevante dokumenter og avsnitt.
  - systemene deretter bruker (feature-based, nevrale eller mix av begge) "reading comprehension" algoritmer for å lese hentede dokumenter/avsnitter.
  - hente svar direkte fra tekstene (sekvenser av strengene).

Question Answering – IR-based

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What are the names of Odin's ravens?	Huginn and Muninn
What kind of nuts are used in marzipan?	almonds
What instrument did Max Roach play?	drums
What's the official language of Algeria?	Arabic

Figure 23.9 Some factoid questions and their answers.

https://web.stanford.edu/~jurafsky/slp3/23.pdf

Question Answering – IR-based

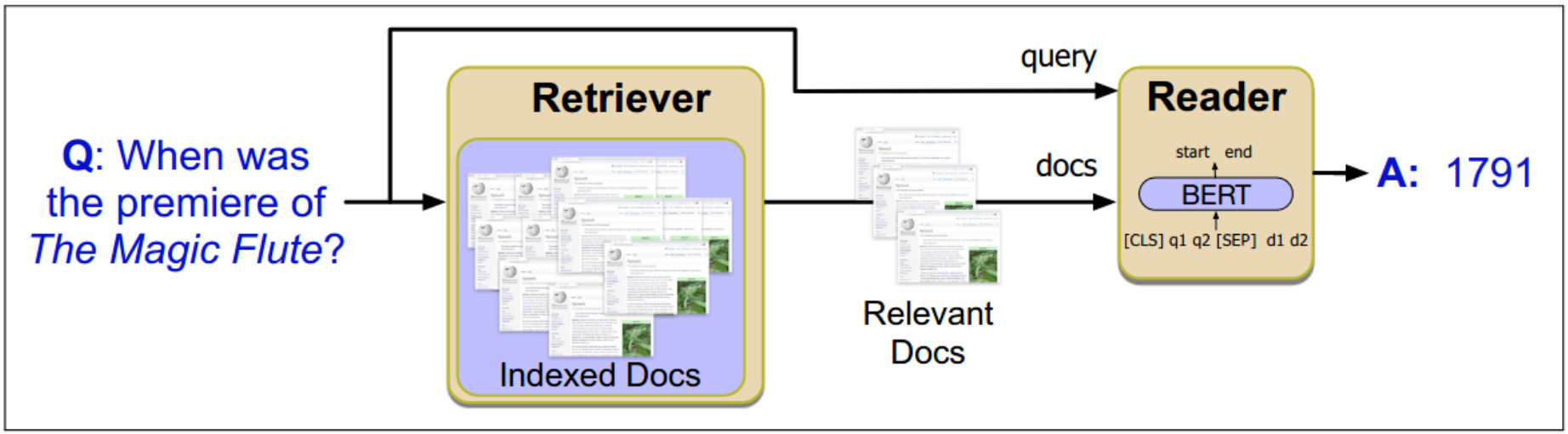


Figure 23.10 IR-based factoid question answering has two stages: **retrieval**, which returns relevant documents from the collection, and **reading**, in which a neural reading comprehension system extracts answer spans.

https://web.stanford.edu/~jurafsky/slp3/23.pdf

NorQuAD

"context": "Molde-gutten kan slå Rosenborg i kveld: – Jeg har et sterkt hat til dem Sander Svendsen er tilbake i norsk fotball og skal spille for Odd fram til sommeren. Foto: Teigen, Trond Reidar / NTB scanpix Sander Svendsen (t.h.) var med da MFK og RBK spilte 1 – 1 på Lerkendal stadion i 2015. Foto: Alley, Ned / NTB scanpix Sander Sander Svendsen, som selv er Molde-gutt og spilte for MFK i sin oppvekst og fram til han ble solgt for litt under to år siden, er klar i sin tale når han snakker om kveldens motstander. – Ingenting er bedre enn å slå Rosenborg – på Lerkendal. Jeg har et sterkt hat til den klubben, og det har jeg hatt siden jeg var en 13 – 14 og spilte elitefotball mot dem, sier Svendsen til avisa Varden. – Jeg har hatt noen tunge opplevelser der oppe. Så det er på tide med en seier. Det hadde gjort godt, sier moldenseren - som foreløpig aldri har kunnet reise fra Lerkendal med tre poeng i lomma. Mandagens kamp har avspark klokka 19. ...",

qas: [{question: Når på mandag starter kampen?",
answer": { text: "klokka 19", answer_start: 1214, answer_end: 1223}
"is_impossible": false }],

Dialogsystemer og chatbots

Dialogsystemer

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

Regelbasert chatbot

- Første: Eliza (Weizenbaum, 1966).
- Regelbasert: 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.

`https://web.njit.edu/~ronkowit/eliza.html`

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 har slitt med å beholde en sammenhengende samtale.
- Maksimene: Quantity, Quality, Relevance, Manner

Evaulering

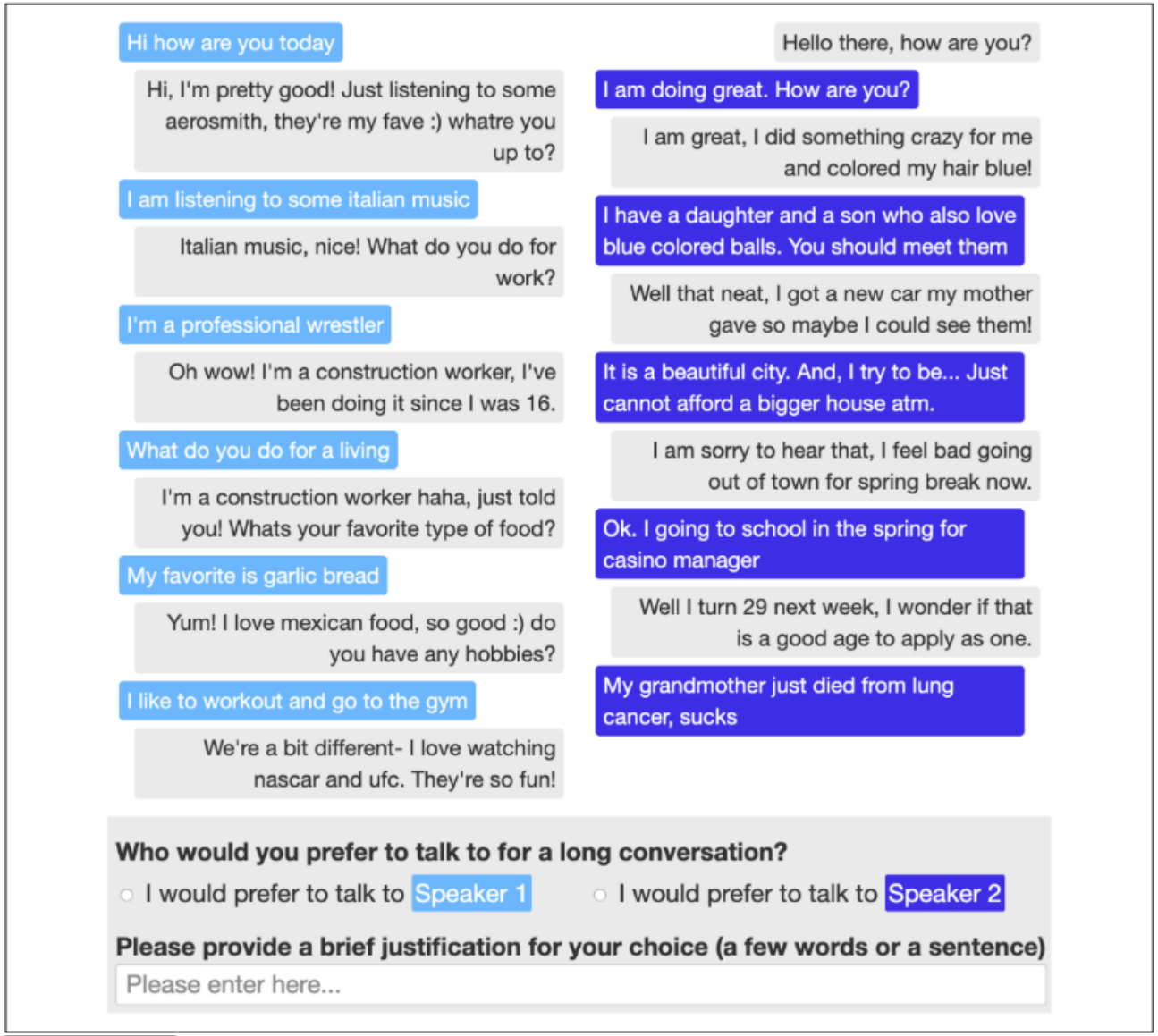


Figure 24.19 The ACUTE-EVAL method asks annotators to compare two dialogues and choose between Speaker 1 (light blue) and Speaker 2 (dark blue), independent of the gray speaker. Figure from Li et al. (2019).

Menti.com

8591 7425



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 24.10 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 24.9 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

Oppsummering

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

Oppsummering

- Ser på en rekke applikasjoner:
  - Teknikker i informasjonsekstraksjon
  - Leksikon for sentimentanalyse
  - Question Answering
  - Dialogsystemer og chatbots