IN1140: Introduksjon til språkteknologi

Forelesning #11

Egil Rønningstad

Universitetet i Oslo

9. nov[illegible]er 2022



Tema for i dag

- Språkteknologiske anvendelser
- Utviklingen av språkteknologiske metoder
- Bygge intuisjon

Språkteknologiske applikasjoner

Takket være forskning innen språkteknologi har vi:

- Bedre stavekontroll
- Maskinoversettelse
- Bedre søkemotorer
- Senteimentanalyse av tale og tekst
- Analyseverktøy for medisinske tekster
- Stemmestyring og dialogsystemer
- Chatboter
- Bedre innformasjonshenting

Språkteknologiske applikasjoner

Forskiningen bygger på

- Forståelse av språkets mekanismer med sine byggeklosser, regler og unntak
- Forståelse av variasjonene mellom språk og domener
- Forståelse av digital representasjon av språk
- Forståelse av algoritmenes muligheter og begrensninger
- Kjennskap til forskningens State-of-the-art

Forskning vs produksjon



Construction of igloos (source: Inuits By Frank E. Kleinschmidt -Library of Congress , Public Domain).

I forskning utvikler vi og analyserer byggeklossene som andre setter sammen til systemer avhengig av bruksområdet.

Utvikling innen NLP

- Oppslag
- Regelstyrt
- Statistiske modeller
- Nevrale modeller

Representasjon av tekst

- Indeksering
- Statiske word embeddings
- Kontekstuelle word embeddings

Representasjon av tekst

Indeksering

Hvert ord har sitt eget nummer. Ord kan være lemmatisert og lowercase for å bruke samme nummer på flere varianter av ordet. Men ulike nummer angir (helt) ulike ord. I utgangspunktet ingen grunn til å tenke at ord nr 500 og 501 "ligner", er synonymer eller i samme ordklasse.

Statiske word embeddings

Hvert ord har sin egen vektor (tallrekke av desimaltall) Ut fra hvordan ord opptrer i treningskorpus, vil "like" ord (som synonymer) få representasjoner som ligner. (cosine similarity). Hver ordform (som *bank* eller *hoppe*) får en (og kun en) vektorrepresentasjon, som formes av alle steder ordet opptrer i treningskorpus.

Kontekstuelle word embeddings

I tillegg til en representasjon ut fra hvilke setninger ordet opptrer i, har ordet også med sen en komprimert representasjon av kontekst. På den måten kan verbet og substantivet *hoppe* få ulik representasjon utfra konteksten til ordet i hvert konkrete tilfelle.

Utvikling innen NLP

Oppslag på indeksert korpus: Databaser.
Søkestreng: *Ost fra Gudbrandsdalen* `[214, 9, 4999]`
Dokumenter:
`[188, 25, 9, 18, 214 ...]`
`[9, 214, 4900, 18, 214 ...]`
`[36, 1, 600, 214, 9, 4999, 3210 ...]`

Regelstyrt på indeksert representasjon:

```
hvis <ord> ender på "isk" -> Klassifiser som adjektiv

hvis <positivt sentimentord> har "ikke" foran ->
Klassifiser som negativt sentiment

hvis <ord> etterfølges av "." og <Stor bokstav> ikke i
<ord> og <ord> ikke i <liste over forkortelser> og "."
etterfølges av <\s> -> tokeniser "." separat.
```

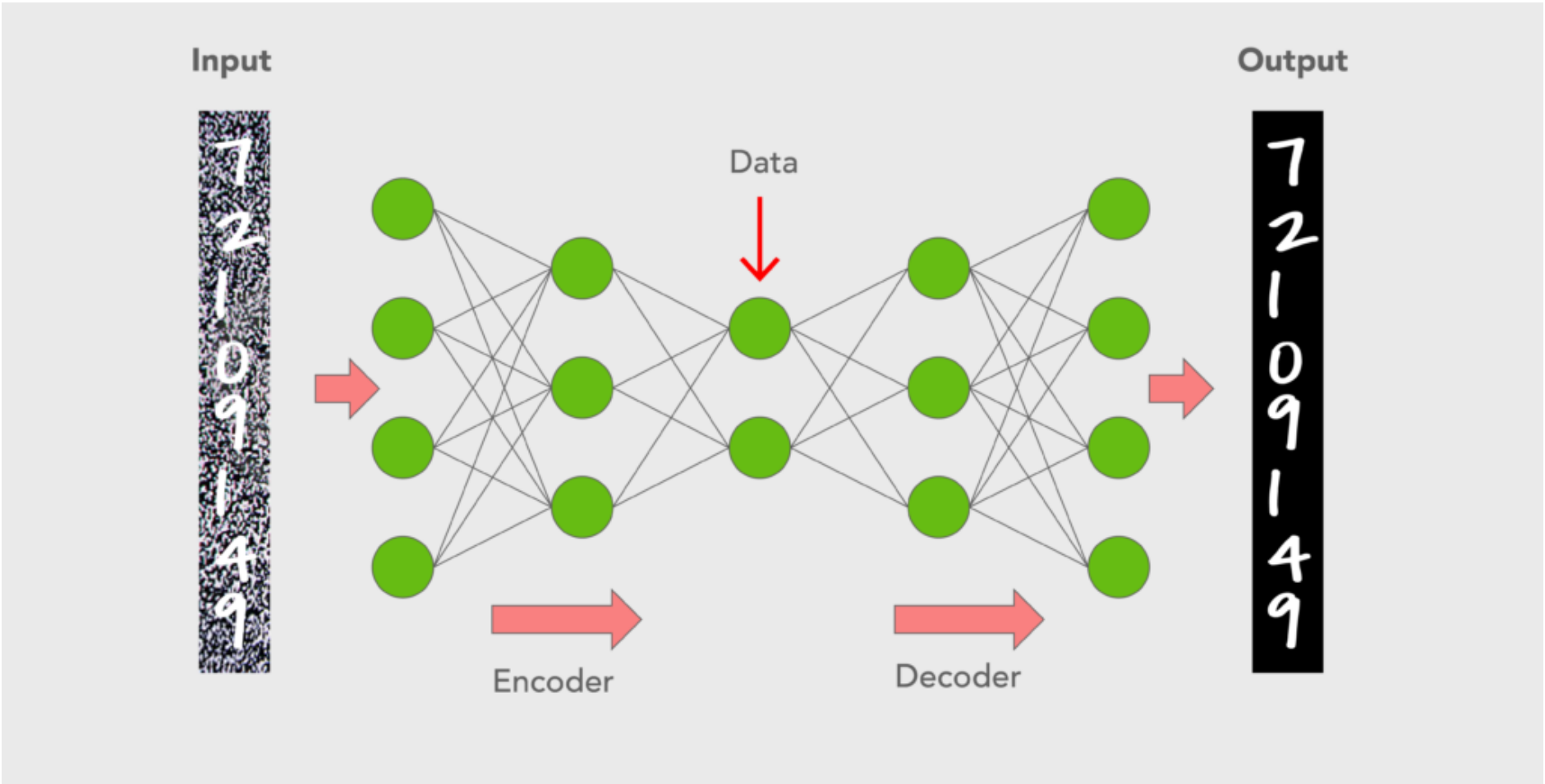
Representasjon	Oppslag	Regelstyrt	Statistiske	Nevrale	Algoritmer / Modeller
Kontekstuelle embeddings			■	■	
Statiske embeddings			■	■	
Indeksering	■	■	■		

Tilbake til start: Retrieval-augmented language models.

Komprimering



Encoder-decoder



https://www.mygreatlearning.com/blog/autoencoder/

Encoder - decoder

Vi trenger ikke dekode tilbake til originale. **Encoder**: Skape representasjon for engelske setninger. **Decoder**: Utvide representasjonen til franske setninger.
Encoder: Skape representasjon av vitenskapelig tekst. **Decoder**: Utvide representasjonen til tekstsammendrag.
Viktig: Vi kan ha ulikt språk, ulikt antall tokens osv, så lenge vi har parvise treningseksempler.
Andre bruksområder:
Alle Transformersmodeller er encoders, eventuelt også decoders.

- Word Sense Disambiguation J&M 18.8
- Question Answering J&M 23.1-2
- Chatbots & dialugue systems J&M 24.1-2
- Talegjenkjenning J&M 26 intro

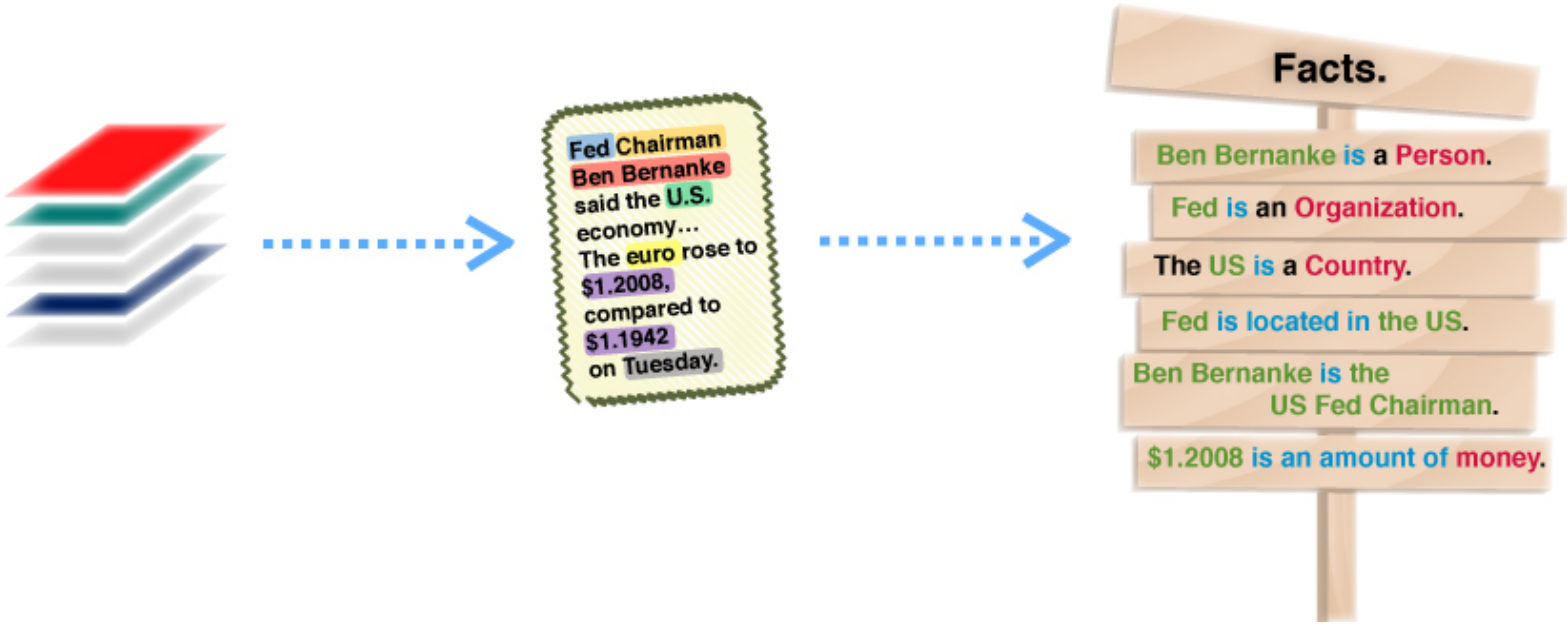
Informasjonsekstraksjon (IE)

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

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].

text = Widerøe vurderte å flytte ledig kapasitet frå Sandane anten til Bringeland eller til Haukåsen under utbygginga.

1	Widerøe	Widerøe	PROPN	_	_	2	nsubj	_	name=B-ORG
2	vurderte	vurdere	VERB	_	Mood=Ind|Tense=Past|VerbForm=Fin	0	root	_	name=O
3	å	å	PART	_	_	4	mark	_	name=O
4	flytte	flytte	VERB	_	VerbForm=Inf	2	xcomp	_	name=O
5	ledig	ledig	ADJ	_	Definite=Ind|Degree=Pos|Number=Sing	6amod	_	name=O	
6	kapasitet	kapasitet	NOUN	_	Definite=Ind|Gender=Masc|Number=Sing	4obj	_	name=O	
7	frå	frå	ADP	_	_	8	case	_	name=O
8	Sandane	Sandane	PROPN	_	_	6	nmod	_	name=B-LOC
9	anten	anten	CCONJ	_	_	11	cc	_	name=O
10	til	til	ADP	_	_	11	case	_	name=O
11	Bringeland	Bringeland	PROPN	_	_	8	obl	_	name=B-LOC
12	eller	eller	CCONJ	_	_	14	cc	_	name=O
13	til	til	ADP	_	_	14	case	_	name=O
14	Haukåsen	Haukåsen	PROPN	_	_	11	conj	_	name=B-LOC
15	under	under	ADP	_	_	16	case	_	name=O
16	utbygginga	utbygging	NOUN	_	Definite=Def|Gender=Fem|Number=Sing	4obl	_	SpaceAfter=No|name=O	

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

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)
```

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

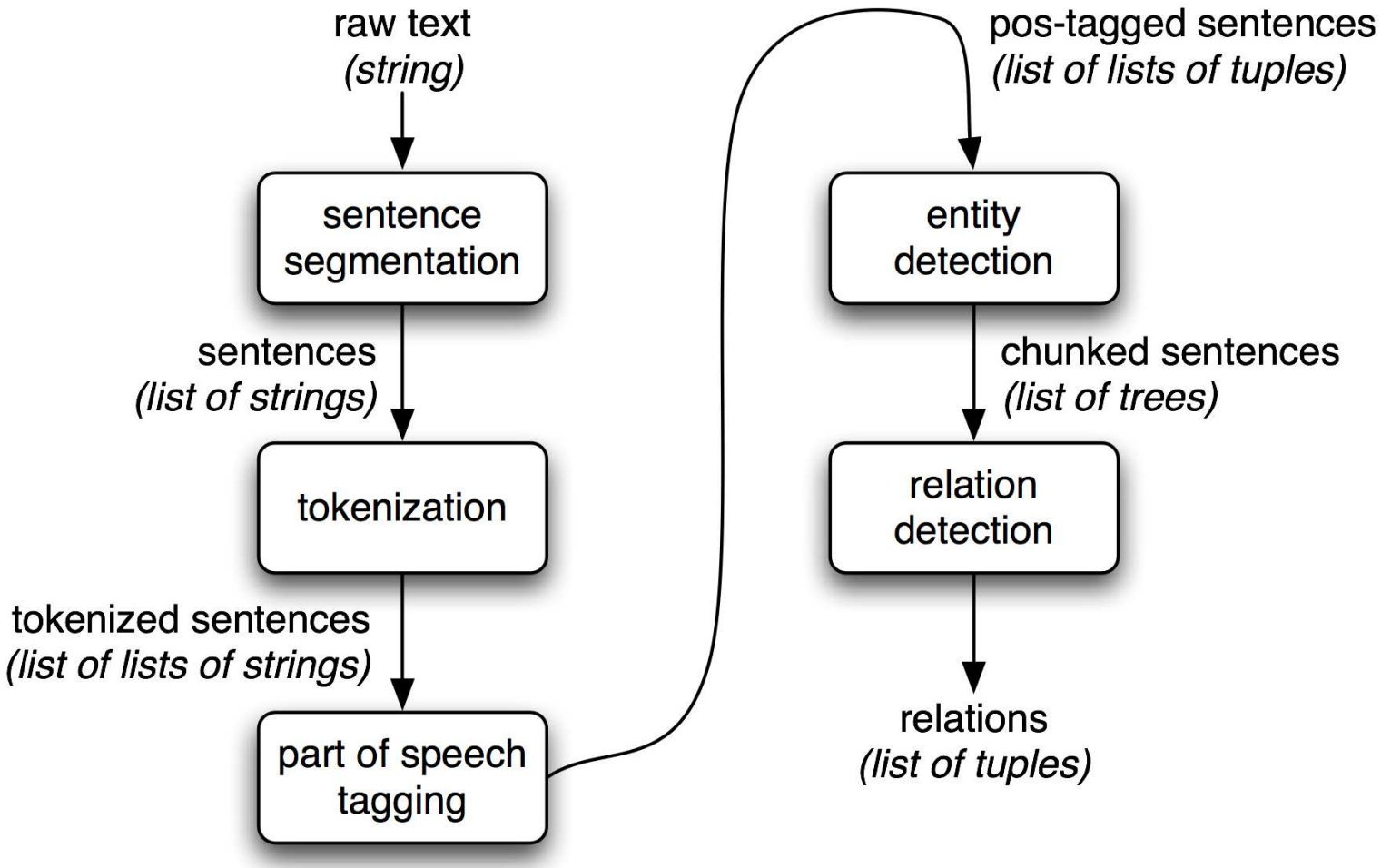
- Who did what to whom, when, where, through what methods (instruments) and why.

- 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)



Supervised learning: Sentimentanalyse

	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`

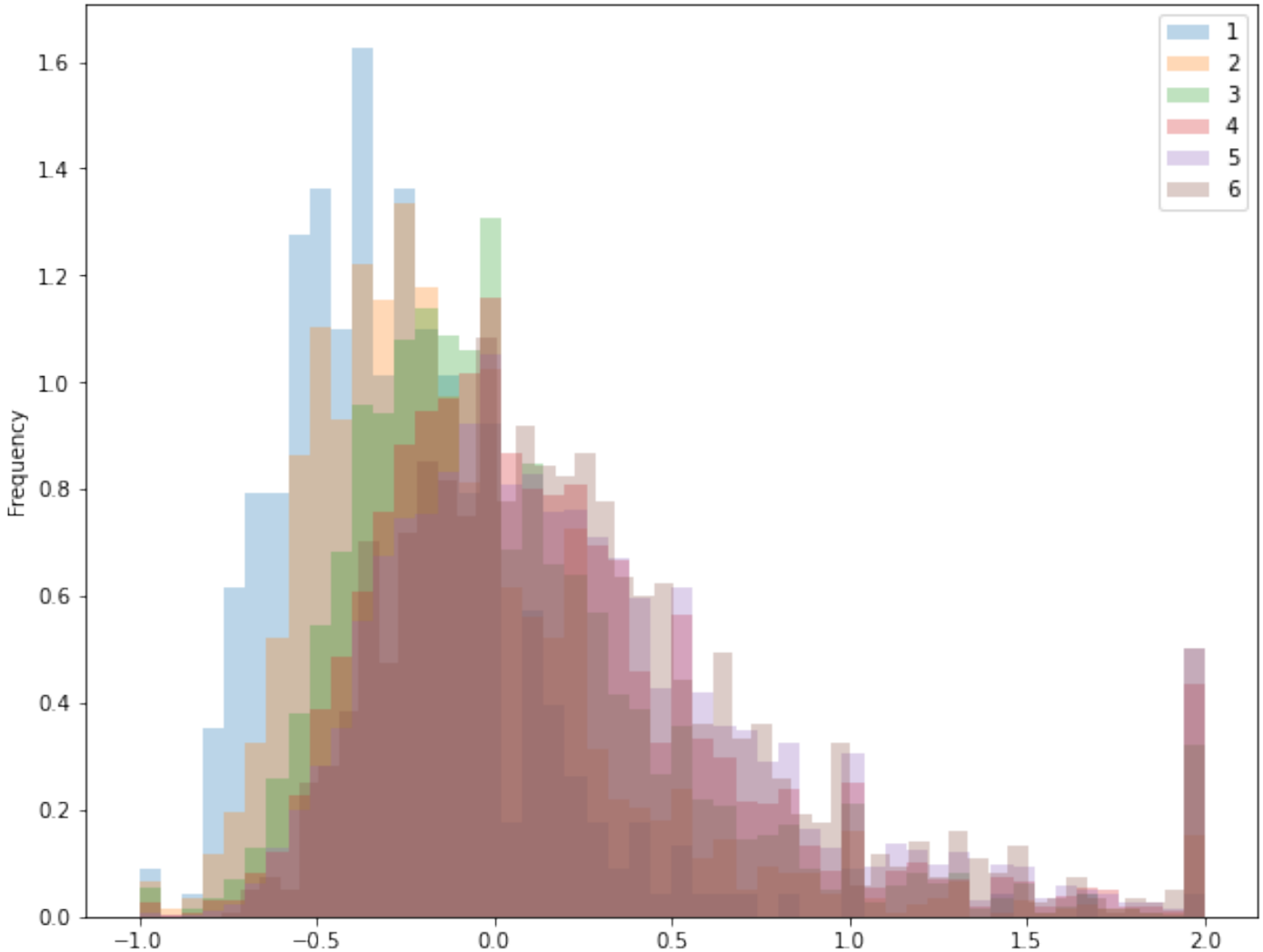
Supervised learning: Sentimentanalyse



- **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)

Sentimentanalyse

Oppslag for norec i sentiment-leksikon norsentlex. (#pos -#neg) / #neg




```
1 -0.27 0.362 2 -0.105 0.459 3 0.057 0.513
4 0.179 0.546 5 0.263 0.568 6 0.28 0.559
```

Norec-fine

{"sent_id": "201911-07-01", "text": "Skjermen er helt klart visuelt utfordrende , noen vil synes den er overharry , andre vil bli begeistret .",
"opinions": [{"Source": [[], []], "Target": [["Skjermen"], ["0:8"]], "Polar_expression": [["visuelt utfordrende"], ["23:42"]], "Polarity": "Positive", "Intensity": "Standard", "NOT": false, "Source_is_author": true, "Target_is_general": false, "Type": "E"},

{"Source": [["noen"], ["45:49"]], "Target": [["Skjermen"], ["0:8"]], "Polar_expression": [["vil synes den er overharry"], ["50:76"]], "Polarity": "Negative", "Intensity": "Standard", "NOT": false, "Source_is_author": false, "Target_is_general": false, "NFP": true, "Type": "E"},

{"Source": [["andre"], ["79:84"]], "Target": [["Skjermen"], ["0:8"]], "Polar_expression": [["vil bli begeistret"], ["85:103"]], "Polarity": "Positive", "Intensity": "Strong", "NOT": false, "Source_is_author": false, "Target_is_general": false, "NFP": true, "Type": "E"}]},

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 (bruker 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

- 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

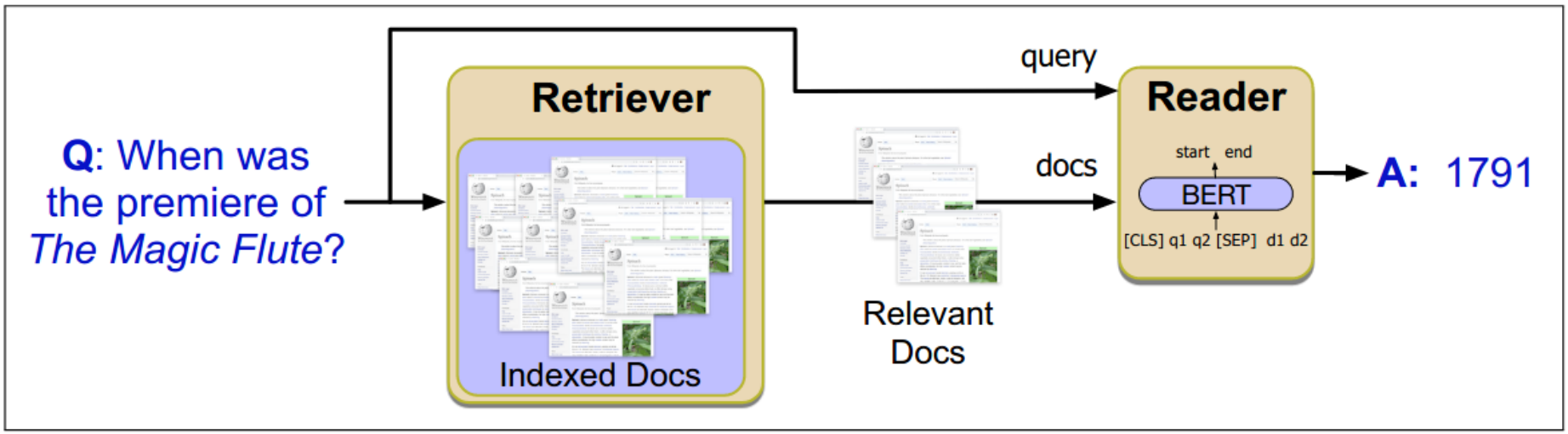


Figure 23.10 IR-based factoid question answering has two stages: **retrieval**, which returns relevant documents from the collection, and **reading**, in which a neural reading comprehension system extracts answer spans.

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

Watson – IBM

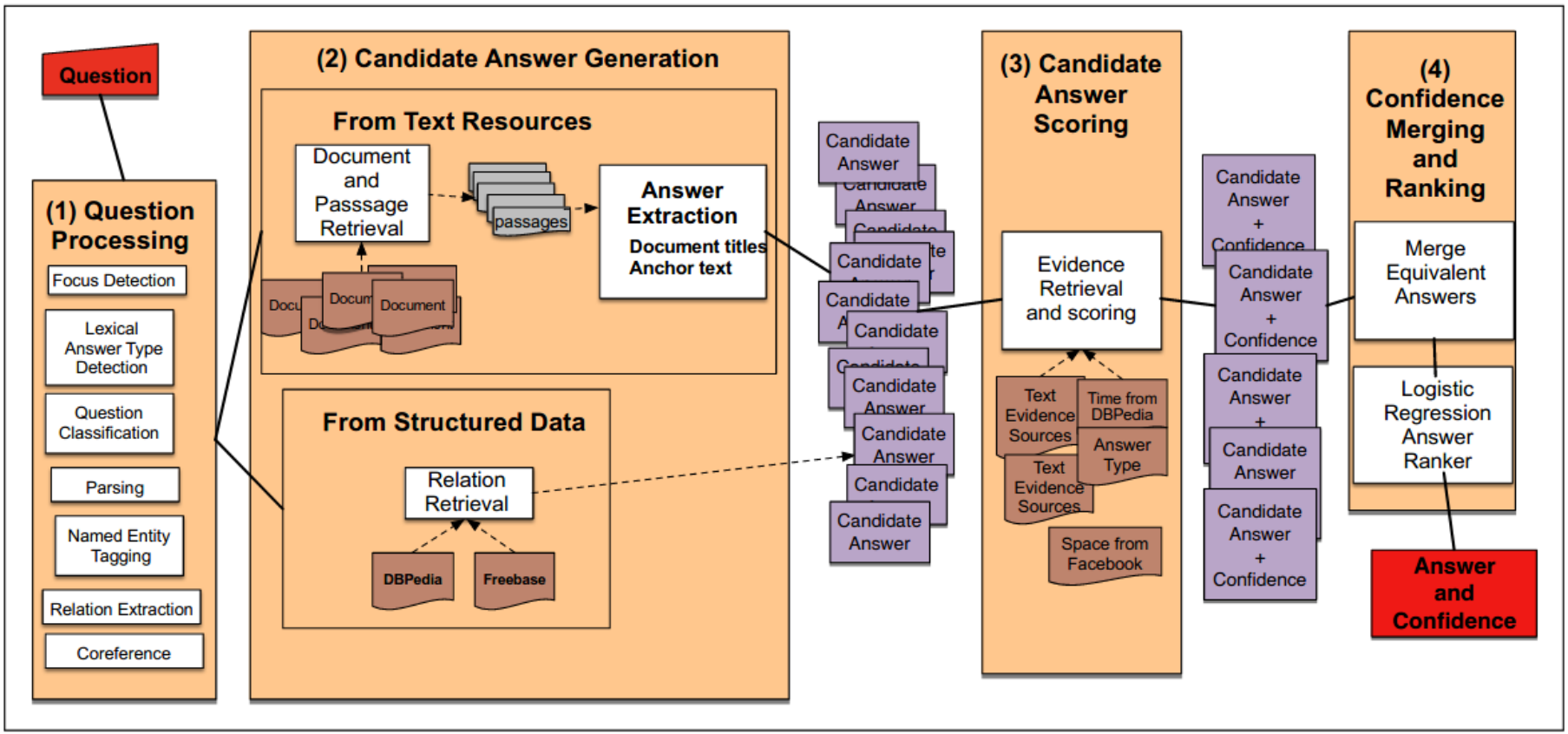
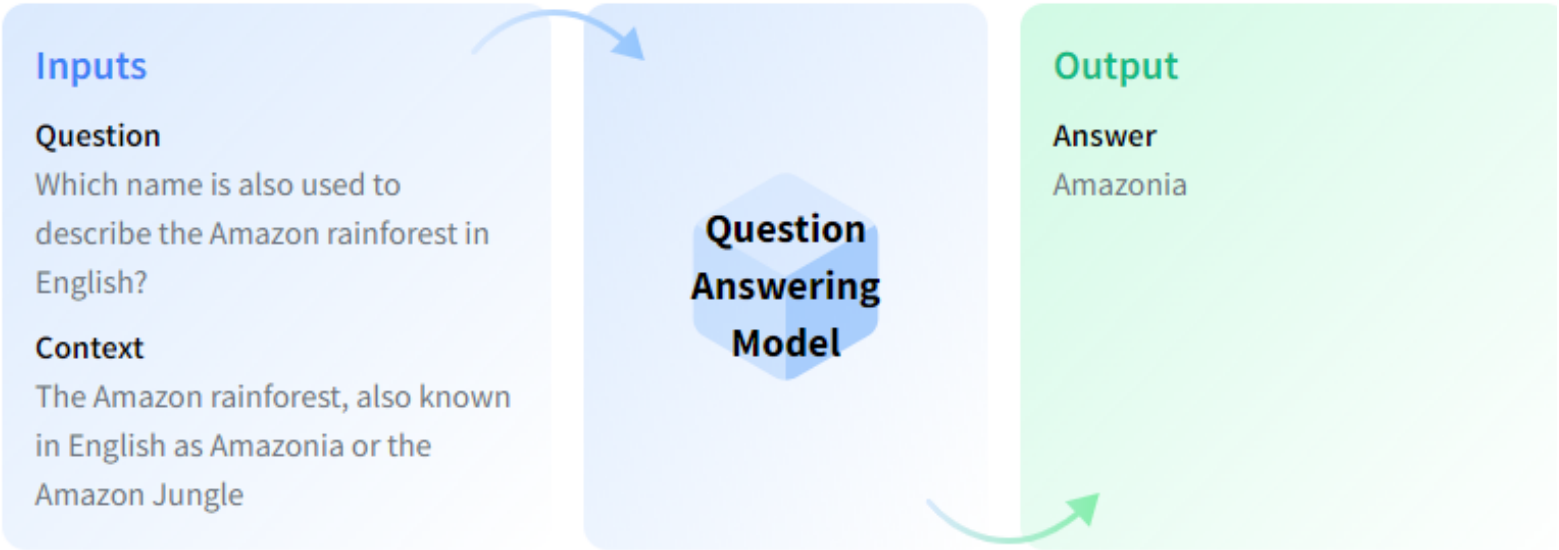


Figure 23.17 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

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

QA



https://huggingface.co/tasks/question-answering

https://www.uio.no/studier/emner/matnat/ifi/IN1140/index-eng.html



Dialogsystemer og chatbots

Teksten herfra og ut ble ikke dekket i forelesningen og er ikke med i grunnlaget for eksamineringen

Dialogsystemer

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

Egenskaper av menneskelig samtaler

C_1:	. . . I need to travel in May.
A_2:	And, what day in May did you want to travel?
C_3:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_4:	And you're flying into what city?
C_5:	Seattle.
A_6:	And what time would you like to leave Pittsburgh?
C_7:	Uh hmm I don't think there's many options for non-stop.
A_8:	Right. There's three non-stops today.
C_9:	What are they?
A_{10}:	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_{11}:	OK I'll take the 5ish flight on the night before on the 11th.
A_{12}:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_{13}:	OK.
A_{14}:	And you said returning on May 15th?
C_{15}:	Uh, yeah, at the end of the day.
A_{16}:	OK. There's #two non-stops . . . #
C_{17}:	#Act. . . actually #, what day of the week is the 15th?
A_{18}:	It's a Friday.
C_{19}:	Uh hmm. I would consider staying there an extra day til Sunday.
A_{20}:	OK. . . OK. On Sunday I have . . .

Figure 24.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A_{16} and C_{17} indicate overlaps in speech.

https://web.stanford.edu/~jurafsky/slp3/24.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_1, A_8, A_{10}.
- 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.

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.

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.

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.

Evaulering

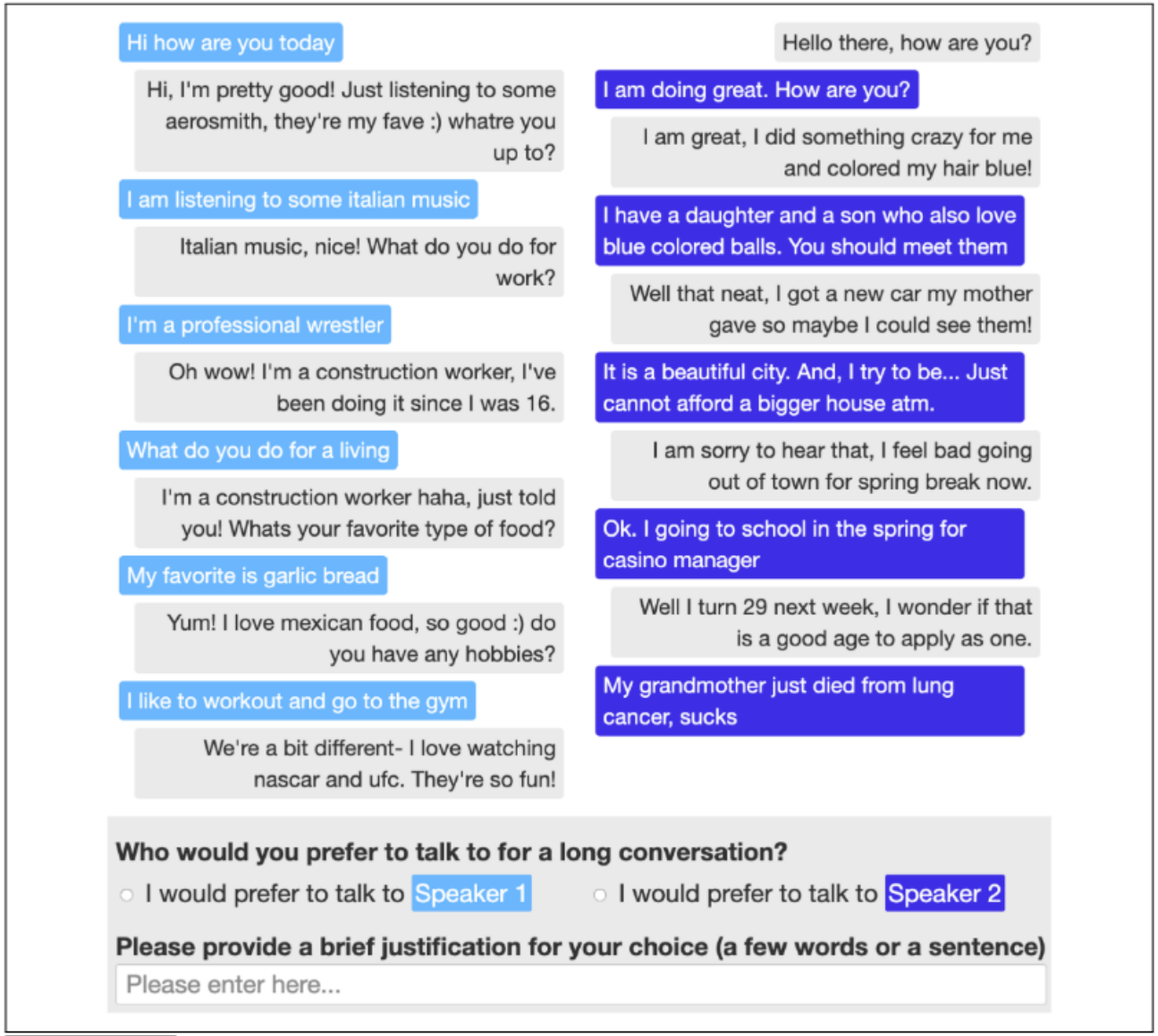


Figure 24.19 The ACUTE-EVAL method asks annotators to compare two dialogues and choose between Speaker 1 (light blue) and Speaker 2 (dark blue), independent of the gray speaker. Figure from Li et al. (2019).

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

- Ser på en rekke applikasjoner:
  - Teknikker i informasjonsekstraksjon
  - Sentimentanalyse
  - Question Answering
  - Dialogsystemer og chatbots