

Interaktive systemer

Pierre Lison
plison@nr.no

IN2110: Språkteknologiske metoder
30. april 2024



Oblig 2b

- ▶ 2 deler:
 - **MT:** Oversettelseseksperimenter med filmtekstinger fra Ringenes Herre (tysk → engelsk)
 - **Interaktive systemer:** utvikling av en enkel praterobot basert på filmtekstinger



- ▶ Oppgaven skal publiseres i de neste dagene
- ▶ Innleveringsfrist: **7. mai**

Plan for i dag

- ▶ Siste del om MT: BLEU
- ▶ Hva er en samtale?
- ▶ Prateroboter
- ▶ Etske betrakninger

Evaluering

- ▶ Alternativ: bestemme oversettelseskvaliteten ut fra dens avstand til en eller flere «fasiter»
 - skrevet av menneskelige eksperter
- ▶ Det er ikke nok å bare se hvor mange ord er korrekt: vi må også ta hensyn til *ordstillingen*!
- ▶ N-gram overlapp med menneskelige oversettelser

Source: À l'évidence, son mari était un gros fumeur

Reference: Tydeligvis var mannen hennes en storrøyker

Output 1: Mannen hennes var åpenbart en storrøyker

Output 2: Hans ektemann var en tung røyker

Unigrams
Bigrams

5 2

2 0

BLEU

- ▶ Det mest kjente evalueringsmetode er **BLEU**, som beregnes ved å se på overlapp mellom N-grams fra fasiten(e) og oversettelsene fra systemet.
- ▶ Vi ekstraherer for hver setning alle N-grams (med N fra 1 til 4) fra både systemet og fasiten, og beregner hva som er "precision" for $i \in 1, 2, 3, 4$:

$$precision_i = \frac{\text{Antall } i\text{-grams som forekommer i både system og fasit}}{\text{Antall } i\text{-grams i setningene fra systemet}}$$

(beregnes setning for setning!)

BLEU

Deretter beregnes BLEU slikt:

$$BLEU = brevity_penalty * \left(\prod_{i=1}^4 precision_i \right)^{\frac{1}{4}}$$

hvor “*brevity penalty*” brukes til å “straffe” modeller som produserer for korte oversettelser:

$$brevity_penalty = \min\left(1, \frac{\text{Antall ord i systemets setninger}}{\text{Antall ord i fasitens setninger}}\right)$$

BLEU

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

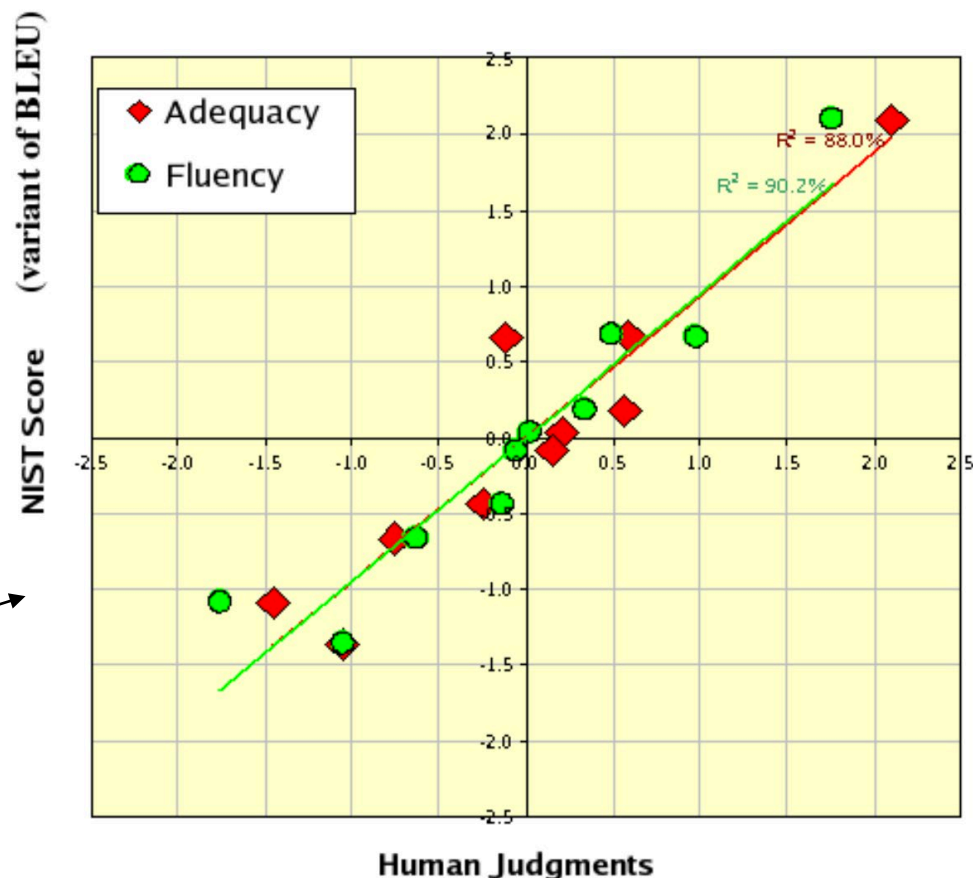
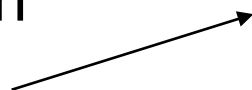
REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)		
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

BLEU

- ▶ Hovedfordelen med BLEU: kan beregnes automatisk!
- ▶ Ikke perfekt, men korrelert med oversettelses kvalitet



- ▶ Men BLEU har også viktige svakheter:
 - Ignorerer semantikken i setningen ("ikke" er bare ett ord, men et viktig ord i en setning!)
 - Ignorerer den globale koherensen / strukturen i setningen

Plan for i dag

- ▶ Siste del om MT: BLEU
- ▶ Hva er en samtale?
- ▶ Prateroboter
- ▶ Ethiske betrakninger

Dialogsystemer?

= dataprogrammer ("bots") utviklet til å samhandle med mennesker ved hjelp av (talte eller tekstbaserte) naturlige språk

Dialogsystem

brukerytring



maskinytring

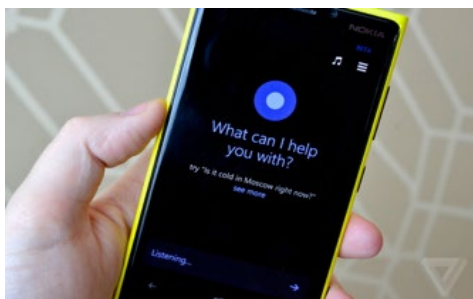
Hvorfor?

- *Brukervennlig*: du trenger bare å kunne snake/skrive!
- Kan uttrykke *komplekse forespørsler*
- For talebaserte systemer: *berøringsfri grensesnitt* (viktig for f.eks. bilkjøring)

Smarthusprodukter

Anvendelser

Virtuelle assistenter
(Siri, Cortana, osv.)



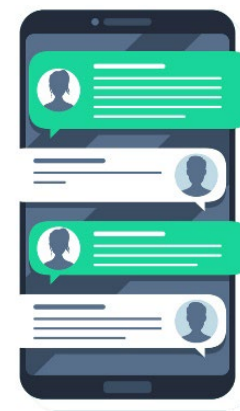
Stemmebasert navigering
og styring i bilen



Tjenesteroboter



Chatbots

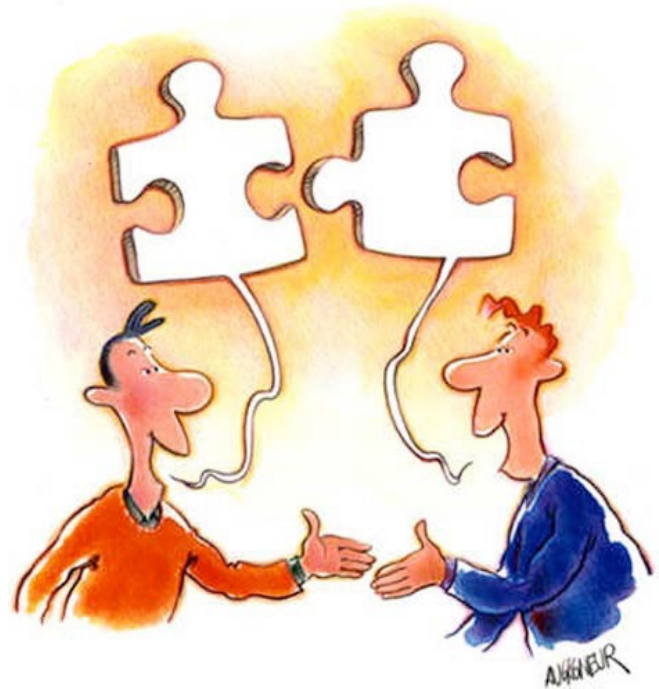


Instruksjonstilpassede språkmodeller



Plan for i dag

- ▶ Siste del om MT: BLEU
- ▶ **Hva er en samtale?**
- ▶ Prateroboter
- ▶ Etiske betrakninger



Turtaking (turn-taking)

- ▶ Deltakerne i dialog tar *turer*:
 - En tur = et kontinuerlig bidrag fra en samtalepartner
- ▶ Turtaking er en beskrivelse av hvordan ordet fordeles mellom samtalepartnere
- ▶ Overraskende flytende i dagligdagssamtaler
 - Forsøk på å minimere både tomrom (ingen som snakker) og overlapp (flere som snakker over hverandre)
- ▶ Intervallet mellom turer er rundt 250 ms

[Duncan (1972): «Some Signals and Rules for Taking Speaking Turns in Conversations», in *Journal of Personality and Social Psychology*]

Turtaking

- ▶ Hvordan er turer tildelt (tatt/frigitt)?
- ▶ Ulike markører:
 - Komplette syntaktisk/semantisk enhet?
 - Dialogstruktur (hilsen → hilsen, spørsmål → svar osv.)
 - Intonasjon (fallende intonasjon signaliserer ofte at turen er ferdig)
 - Ikke-verbale signaler som blikk, gester
 - Markører for stillhet og nøling som "eh" (utfylte pauser ≠ fylte pauser)
 - Sosiale konvensjoner



My Turn



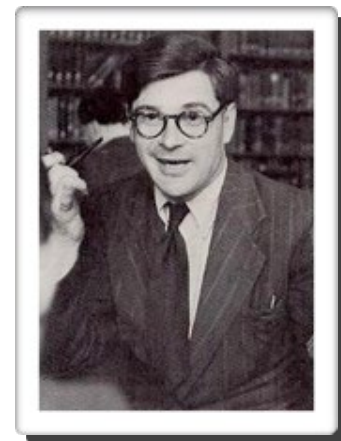
Your Turn

Eksempel av turtaking

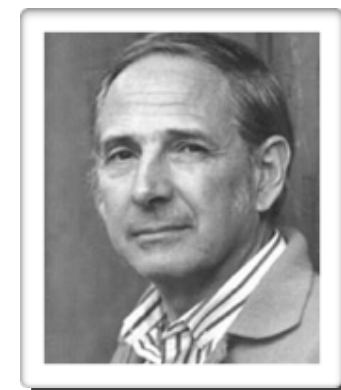
Person 1:	han vil bo i skogen ?
Person 2:	# altså hvis jeg hadde kommet og sagt " skal vi flytte i skogen ? " så hadde han sagt ja
Person 1:	mm
Person 2:	men jeg vil ikke bo i skogen
Person 1:	nei det skjønner jeg
Person 2:	så vi må jo finne et sted som er mellomting og det jeg vil ikke bo utpå landet # i hvilken som helst (uforståelig) ...
Person 1:	* men det kommer jo an på hvor i skogen da

Språkhandlinger

- ▶ Hver ytring er en **handling** utført av taleren
- ▶ Taleren har et spesifikt mål (som kanskje "bare" er å etablere eller opprettholde et sosial kontakt med samtalepartnere)
- ▶ Ytringer leder til bestemte *virkning*er på samtalepartnere (eller på verden forøvrig)
- ▶ «Språk som handling» -perspektiv



J.L. Austin (1911-1960)
språkfilosof



J. Searle (1932, -)
språkfilosof

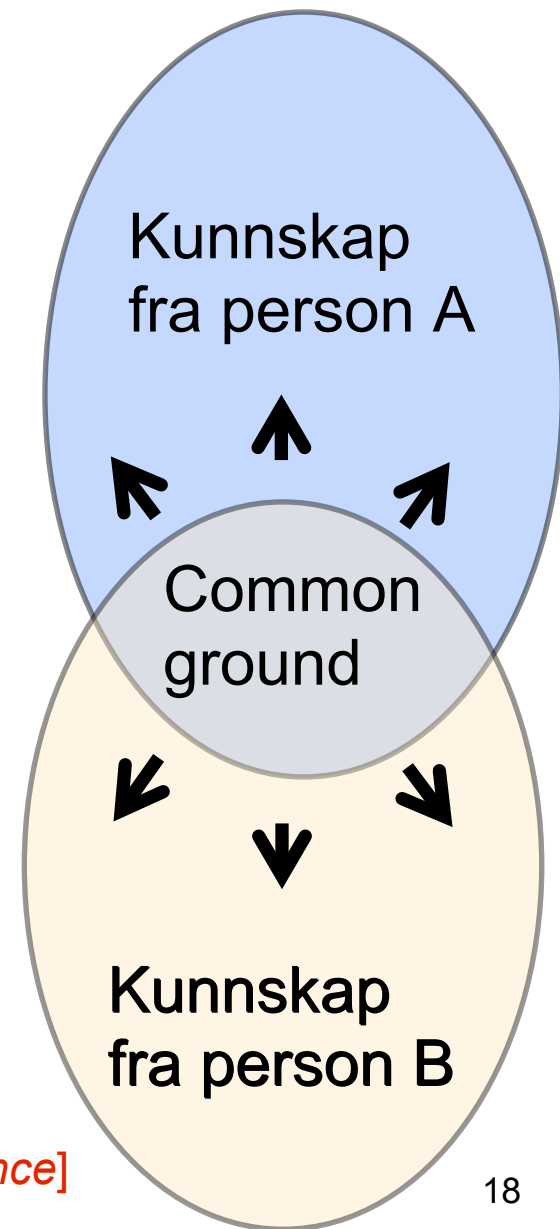
Språkhandlinger



- ▶ Morens reaksjonen har et bestemt *formål*, nemlig å formidle hennes forbauselse/sinne og stoppe Calvin
- ▶ Spørsmålet hennes leder til bestemte virkninger:
 - En psykologisk reaksjon fra Calvin (overraskelse)
 - + en virkning på selve verdenen (Calvin stopper)

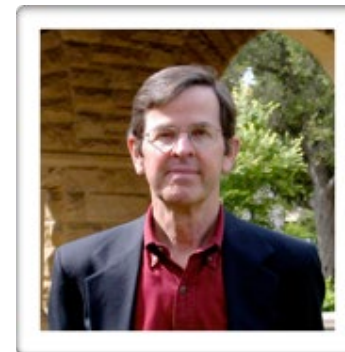
Grounding

- ▶ Dialog er et samarbeid mellom deltakerne i samtalen
 - Trenger å sikre gjensidig forståelse – "vi er på samme linje"
- ▶ Gradvis utvidelse og berikelse av vår "common ground"
- ▶ "common ground" = kunnskap som er felles for samtaledeltakerne



Grounding

- ▶ Grounding betegner prosessen av å gradvis utvide vår "common ground" i løpet av samtalen
 - Bred register av signaler og strategier
- ▶ Flere nivåer
 - Kontakt (oppmerksomhet overfor partneren)
 - Oppfatning av ytringen
 - Forståelse av ytringen
 - Holdninger (enighet, empati osv.)



Herbert H. Clark
psykolingvist



Jens Allwood
lingvist

[Jens Allwood (1992), «On discourse cohesion», in
Gothenburg papers in Theoretical Linguistics.]

Grounding actions

- ▶ Backchannels: «*uh-uh*», «*mm*», «*ja*»
- ▶ Eksplisitte tilbakemeldinger: «*ja det skjønner jeg*»
- ▶ Implisitte tilbakemeldinger: A: «*Jeg ønsker å fly til Roma*» → B: «*På onsdag er det to flyvninger til Roma: ...* »
- ▶ Klarifiseringsstrategier: «*Mente du til Roma eller til Goa?*», "*Kan du bekrefte at ...* "
- ▶ Fiksingsstrategier: «*OK, dere ønsker altså ikke å fly til Goa. Hvor ønsker dere å fly da?*"

Eksempel av grounding

Person 1:	vi vasker den hver dag vi # vi har mopp
Person 2:	mm ## ja det er fort og faren til M27 legger nytt teppe han # det er gjort på to timer ## så det er fort gjort
Person 1:	ja ## da er ikke noe sak
Person 2:	vi har skifta teppe tre ganger allerede han gjør det gratis
Person 1:	hæ ?
Person 2:	vi har skifta teppe tre ganger og # han han ...
Person 1:	* jeg skjønner ikke hvorfor dere har teppe
Person 2:	jeg syns det var rart jeg òg # men e # (sibilant)

Eksempel av grounding (2)

Person 1:	e # nei det er ikke mange
Person 2:	ja * nei
Person 1:	men heldigvis så var ikke Petter Rudi tatt ut denne gangen da
Person 2:	ja # jeg skjønner ikke hva han skal på landslaget å gjøre
Person 1:	* nei han har ingen ting på landslaget
Person 2:	nei # definitivt
Person 1:	å gjøre # han er ubrukelig
Person 2:	* moldensere
Person 1:	hm?
Person 2:	ja disse moldenserne
Person 1:	en gang til?
Person 2:	disse moldenserne
Person 1:	* å ja (fremre klikkelyd) # unnskyld # jeg hørte ikke hva du sa

Implisitt tilbakemelding
(gjentakelse av *landslaget*)

klarifisering



Plan for i dag

- ▶ Siste del om MT: BLEU
- ▶ Hva er en samtale?
- ▶ **Prateroboter**
- ▶ Etiske betrakninger

Grunnarkitektur

Brukerens "intent"

NLU (Natural
Language
Understanding)

Generation /
response selection

inngangssignal
(brukerytring)

utgangssignal
(maskinytring)



Bruker

Regelbaserte systemer

- Pattern-action rule:

(0 YOU 0 ME) [*pattern*]

→

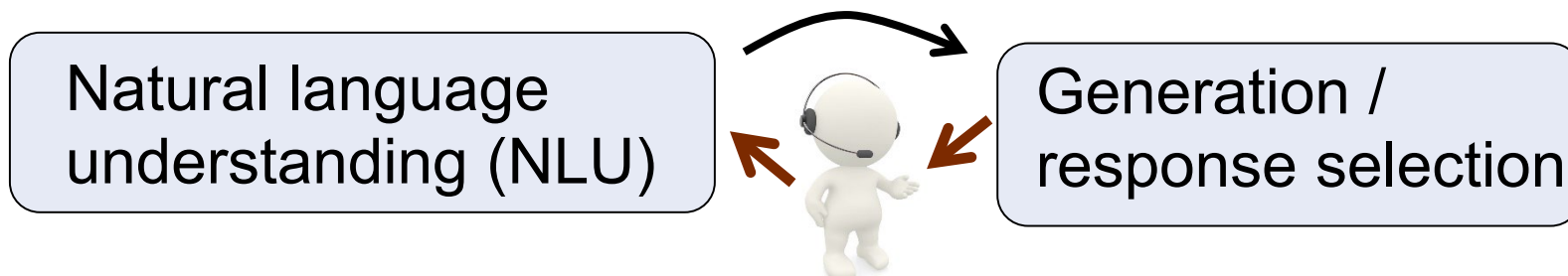
(WHAT MAKES YOU THINK I 3 YOU) [*transform*]

- For eksempel:

You hate me

WHAT MAKES YOU THINK I HATE YOU

Intentgjenkjenning



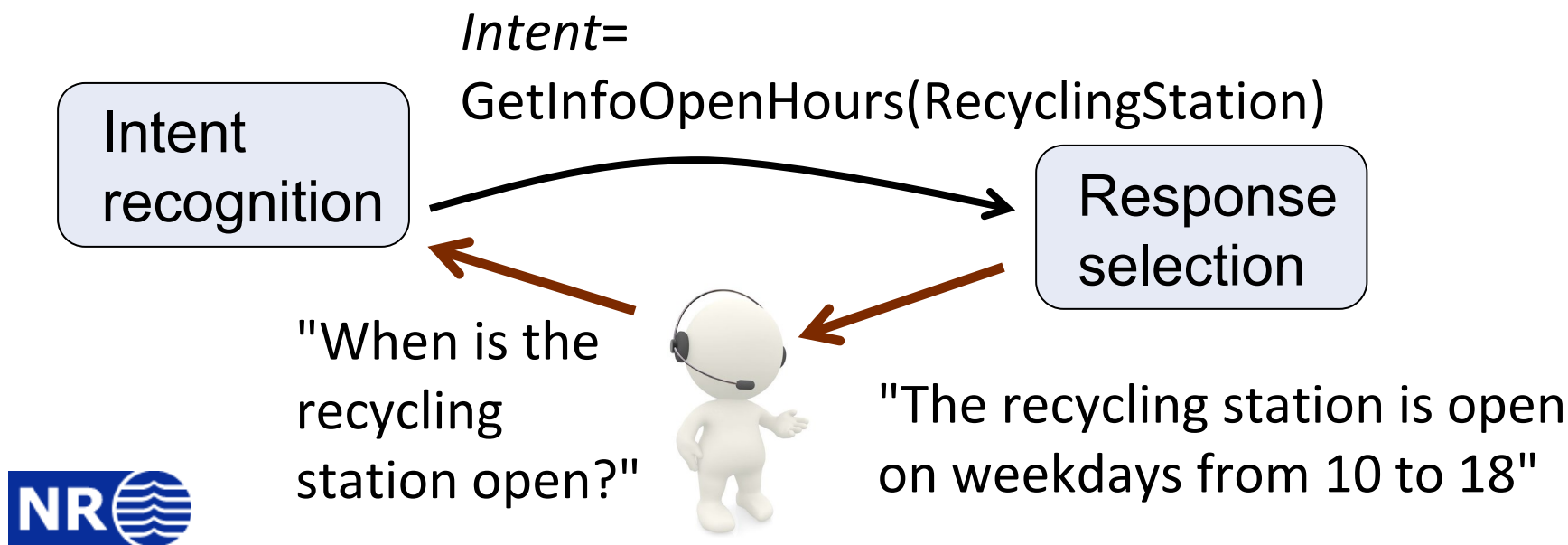
Mange chatbots tar i bruk en *hybrid tilnærming* som kombinerer en maskinlæringsbasert NLU-modul med håndskrevne svar

- ▶ NLU-komponenten er trent fra markerte data (klassifiseringsproblem: koble brukerytringen til dets *intent*)
- ▶ Svarene er forhåndsskrevet av utvikleren

Intentgjenkjenning

Definert fra en liste
forhåndsdefinerte kategorier
(muligens med argumenter,
ofte kalt slots)

- ▶ *Input*: selve brukerytringen (+ kontekst)
- ▶ *Output*: **intent** (= hva brukeren forsøker å oppnå)



IR modeller

- ▶ Alternativ: bruke en datadrevet tilnærming og svare brukeren basert på et *dialogkorpus*
- ▶ Prosedyren:
 - Gitt en brukerinngang **q**, finn ytringen **t** i dialogkorpuset som ligner mest på **q**
 - Så returner som svar enten (a) selve **t** eller (b) ytringen som følger **t** i korpuset

IR modeller

$$r = \text{response} \left(\operatorname{argmax}_{t \in C} \frac{q^T t}{||q|| ||t||} \right)$$

Hvordan kan vi bestemme hvilken ytring **t** fra korpuset «ligner mest» på inputytringen **q**?

- ▶ Cosine similarity mellom vektorer
- ▶ Vektorene kan være TF-IDF vektorer (hvor hver ytring tilsvarer et eget "dokument")
- ▶ Eller embeddings beregnet fra et nevralt nettverk

Eksempel

TF vektorer:

Korpuset:

1. hei ! →
2. hei ! har du det bra ? →
3. ja , hva med deg ? →
4. bare bra →
5. har du spist ? →
6. ja →

bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
							1					1	
	1		1	1		1	1					1	1
		1			1			1	1		1		1
1	1												
				1		1				1			
					1								

Eksempel

$$\log(6) \approx 0.78$$

$$\log\left(\frac{6}{2}\right) \approx 0.48$$

TF-IDF vektorer:

Korpuset:

1. hei ! →
2. hei ! har du det bra ? →
3. ja , hva med deg ? →
4. bare bra →
5. har du spist ? →
6. ja →

	bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
1. hei !								.48					.48	
2. hei ! har du det bra ?		.48		.78	.48		.48	.48					.48	.48
3. ja , hva med deg ?			.78			.48			.78	.78		.78		.48
4. bare bra	.78	.48												
5. har du spist ?					.48		.48				.78			
6. ja						.48								

Ny brukerinput q : "går det bra med deg?"

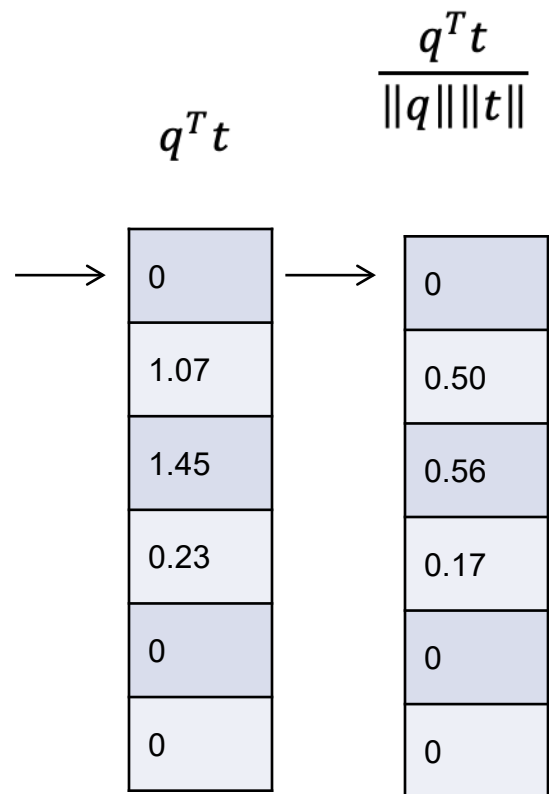
TF-IDF vektor:

	.48	.78	.78							.78				.48
--	-----	-----	-----	--	--	--	--	--	--	-----	--	--	--	-----

Eksempel

	bare	bra	deg	det	du	ja	har	hei	hva	med	spist	,	!	?
1.								.48					.48	
2.		.48		.78	.48		.48	.48					.48	.48
3.			.78			.48			.78	.78		.78		.48
4.	.78	.48												
5.					.48		.48				.78			
6.						.48								

	.48	.78	.78						.78					.48
--	-----	-----	-----	--	--	--	--	--	-----	--	--	--	--	-----



Eksempel

$$\frac{q^T t}{\|q\| \|t\|}$$

Korpuset:

1.	hei !	→	0
2.	hei ! har du det bra ?	→	0.50
3.	ja , hva med deg ?	→	0.56
4.	bare bra	→	0.17
5.	har du spist ?	→	0
6.	ja	→	0

→ Det betyr at ytringen som ligner mest på q er ytringen 3: "ja, hva med deg?"

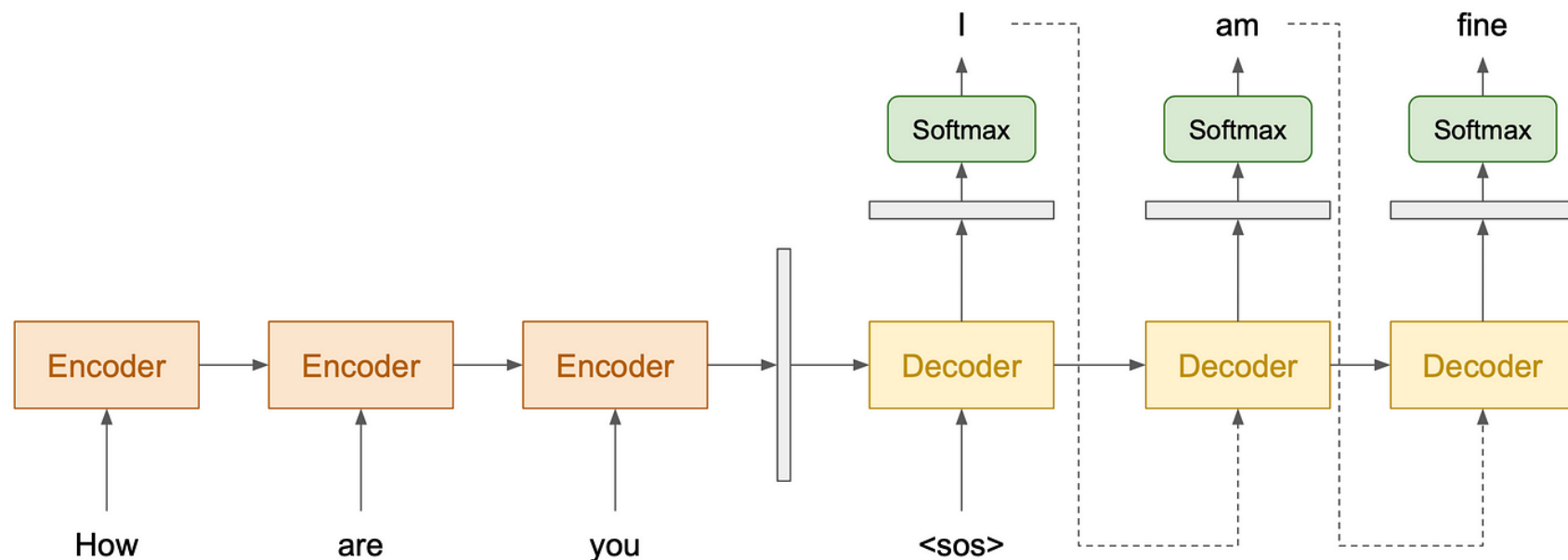
... og at systemet kan velge som respons ytringen som kommer rett etter 3

Ny brukerinput q : "går det bra med deg?"

→ **Respons fra systemet: "bare bra"**

Generative modeller

- ▶ IR modeller kan kun velge ytringer fra korpuset
- ▶ Generative modeller kan lage *nye* ytringer, ord for ord
 - Veldig lik modellene brukt for maskinoversettelse
 - *Decoding* av svaret token etter token



Generative modeller

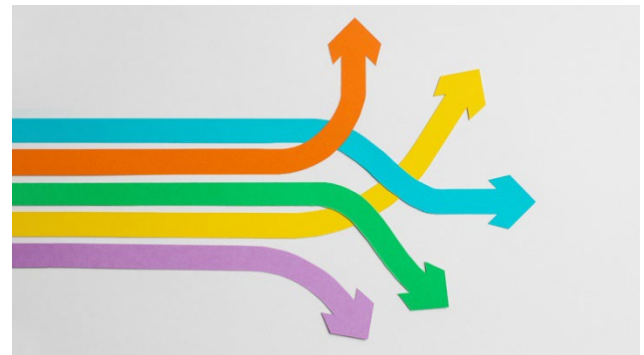
Generative modeller for chatbots er *store språkmodeller* som er *tilpasset* til å fullføre en gitt oppgave



Modelltilpasning: vi ønsker at språkmodellen skal *følge instruksjoner* og svare brukeren på en bestemt måte

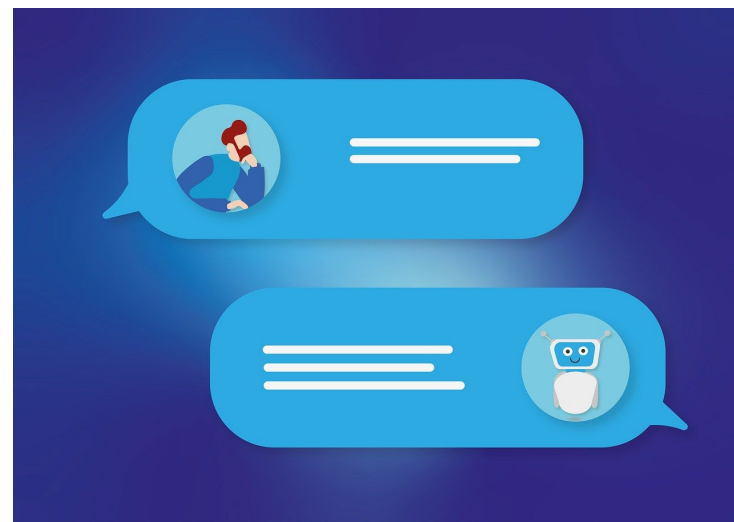
Flere mulige *tilpasningsstrategier*:

- Prompting
- Finjustering
- Preferanseoptimering



Prompting

- ▶ En *prompt* = en instruksjon i tekstform som gis som input til en språkmodell
- ▶ Ulike formuleringer vil gi ulike resultater!
- ▶ Man kan legge til *eksempler* av <input, output> i prompt (såkalt «*in-context learning*»)



- ⊕ Lett å bruke, også med lukkede modeller
- ⊖ Bra for “few-shots”, men ikke optimalt hvis man har tilgang til et større treningssett

Finjustering (fine-tuning)

= “videretrening” av språkmodellen basert på data spesifikke for domenen eller oppgaven

- ▶ *Forutsetning*: ha tilgang til modelvektene (modellen må være open-source) og minst en GPU
- ▶ Vi kan enten optimere alle vektene (kraftkrevende!) eller kun et *utvalg* av disse



Såkalt “**parameter-efficient fine-tuning**”, slik som LoRA



Kan ta i bruk tusenvis av treningseksempler



Litt mer komplisert å få til

Preferanseoptimering



- ▶ Vi kan også justere språkmodellen ved å *belønne* gode svar og *straffe* dårlige svar
- ▶ Kan gjøres gjennom reinforcement learning
 1. Estimering av en belønningsmodell fra menneskelige vurderinger
 2. Optimering av språkmodellen for å maksimere sannsynligheten for å få “tommel opp” for svaret
- ▶ Eller ved å direkte optimere språkmodellen for å fremme svar som fører til høye menneskelige vurderinger (DPO: direct preference optimization)

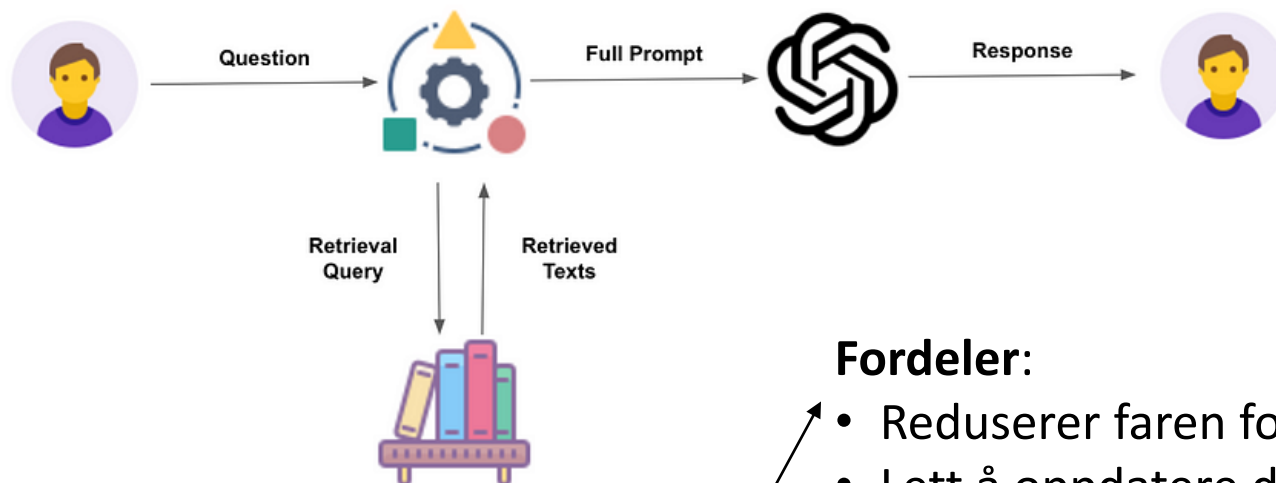


Lettere å samle ratings enn eksempler av “fasittsvar”



Krever en allerede god modell, og nøye retningslinjer

Innhenting fra eksterne kilder



Fordeler:

- Reduserer faren for oppdiktete fakta
- Lett å oppdatere databasen
- Kan bruke «interne» datakilder

Retrieval-augmented Generation (RAG):

- ▶ Modellen leter etter relevante opplysninger i en database...
- ▶ ... og produsere sine svar basert på hva som ble funnet

Utfordringer

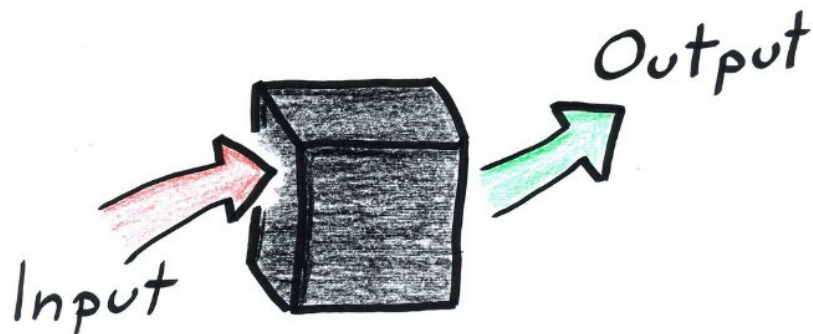
► Manglende saklighet:

språkmodeller kan

hallusinere (dvs. dikte opp fakta)

- De er optimert til å produsere plausible tekster, og har ikke noe forhold til sannhet som sådan!
- Hallusinering kan skje selv med et “perfekt” treningsett

► Manglende kontroll: språkmodeller er “sorte bokser” som er vanskelig å tolke, styre og redigere



Plan for i dag

- ▶ Siste del om MT: BLEU
- ▶ Hva er en samtale?
- ▶ Prateroboter
- ▶ **Etiske betrakninger**

Etiske betrakninger

1. Maskinlærings kan gjenskape (og noen ganger forsterke) **skjevheter** som oppstod i treningsdataene
 - Hatytringer, fordommer, krenkende språk, kjønnskjevhet, osv.
2. Hvordan beskytte brukerens **personvern**?
3. Forsterker virtuelle assistenter **kjønnsstereotyper**?
4. Hvordan unngå at prateroboter brukes til å **bedra** folk?