JM2050 – Natural Language Processing

NLP tasks II

September 26th, 2024

Prof.dr.ir. U. Kaymak

Recap

- Representing text
  - Vector space model
  - Both sparse and dense representations
- Part-of-speech (POS) tagging
  - Identifies the syntactic type of the words
- Text classification
  - Divide documents into pre-determined classes
  - Supervised learning
  - Evaluation of multiple class learners
    - Micro-averaging vs. macro-averaging of one-against-all classifiers

[2]

Outline

- Named Entity Recognition
- Dialogue systems
- Linguistic summaries

[3]

www.jads.nl

Jheronimus Academy of Data Science

Examples of NLP tasks: Named entity recognition

Based on slides from Jurafsky et al. (2020)

Jheronimus Academy of Data Science

Named Entities

- Named entity, in its core usage, means anything that can be referred to with a proper name.
- Most common four tags are:
  - PER (Person): “Marie Curie”
  - LOC (Location): “Amsterdam”
  - ORG (Organization): “Eindhoven University of Technology”
  - GPE (Geo-Political Entity): “Noord-Brabant, The Netherlands"
- Often multi-word phrases
- But the term is also extended to things that aren't entities:
  - dates, times, prices

Named Entity tagging

The task of named entity recognition (NER)

- Find spans of text that constitute proper names
- Tag the type of the entity
- In more complex cases, determine relations between entities

www.jads.nl

Jheronimus Academy of Data Science

NER output

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. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

[7]

www.jads.nl

Jheronimus Academy of Data Science

Why NER?

- Sentiment analysis
  - consumer's sentiment toward a particular company or person
- Question Answering
  - answer questions about an entity
- Information Extraction
  - extracting facts about entities from text
- Anonymization, pseudonymization

[8]

www.jads.nl

Jheronimus Academy of Data Science

Why NER is hard?

- Segmentation
  - In POS tagging, no segmentation problem since each word gets one tag
  - In NER, we have to find and segment the entities!

- Type ambiguity

[$_{PER}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.

BIO Tagging

- How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?
- [PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

www.jads.nl

BIO Tagging

- [PER Jane Villanueva] of [ORG United Airlines Holding] discussed the [LOC Chicago] route.

Words	Label
Jane	R
Villanueva	R
of	
United	RG
Airlines	G
Holding	G
discussed	
the	
Chicago	C
route	
.	

Now we have one tag per token!

www.jads.nl

BIO Tagging

- B: token that *begins* a span
- I: tokens *inside* a span
- O: tokens outside of any span

- # of tags (where n is #entity types):
- 1 O tag,
- *n* B tags,
- *n* I tags
- total of *2n+1*

Words	Label
Jane	R
Villanueva	R
of	
United	RG
Airlines	G
Holding	G
discussed	
the	
Chicago	C
route	
.	

www.jads.nl

Jheronimus Academy of Data Science

BIO Tagging variants: IO and BIOES

- [PER Jane Villanueva] of [ORG United Airlines Holding] discussed the [LOC Chicago] route.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

www.jads.nl

Standard algorithms for NER

Supervised Machine Learning given a human-labeled training set of text annotated with tags

- Hidden Markov Models
- Conditional Random Fields (CRF)
- Maximum Entropy Markov Models (MEMM)
- Neural sequence models (RNNs or Transformers)
- Large Language Models (like BERT), finetuned

www.jads.nl

Jheronimus Academy of Data Science

Typical features

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

[15]

Example

Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

[16]

www.jads.nl

Jheronimus Academy of Data Science

Rule-based methods

Still in use, especially in production (commercial) systems

Pragmatic combination of lists and rules

- Often specified in a formal (propriety) query language

Common approach: multiple rule-based passes over the text

- Use high-precision rules to tag unambiguous entity mentions
- Search for substring matches of the previously detected names
- Use application-specific name lists to find likely domain-specific mentions
- Apply supervised sequence labeling techniques that use tags from previous stages as additional features

[17]

www.jads.nl

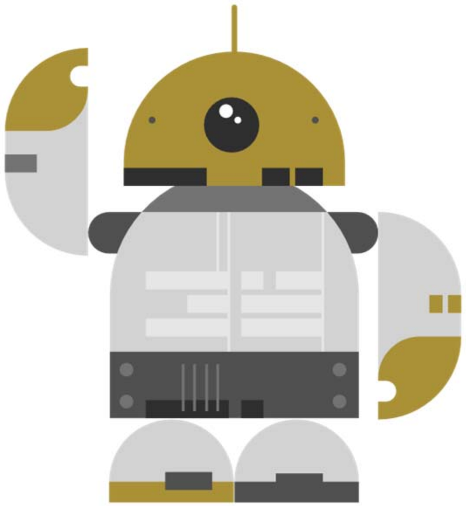
Examples of NLP tasks: Dialogue systems (conversational agents)

Based on slides from Jurafsky et al. (2020)

18

Jheronimus Academy of Data Science

Conversational Agents – Dialogue Systems – Dialogue Agents – Chatbots



Personal Assistants on phones or other devices

- SIRI, Alexa, Cortana, Google Assistant
- Playing music, setting timers and clocks

Chatting for fun

Booking travel reservations

Clinical uses for mental health

[19]

www.jads.nl

Jheronimus Academy of Data Science

Two kind of conversational agents

Chatbots

- mimic informal human chatting
- for fun, or even for therapy

(Task-based) Dialogue Agents

- interfaces to personal assistants
- cars, robots, appliances
- booking flights or restaurants

[20]

www.jads.nl

Jheronimus Academy of Data Science

Chatbot Architectures

Rule-based

- Pattern-action rules (ELIZA)
- + A mental model (PARRY)
  The first system to pass the Turing Test!

Corpus-based

- Information Retrieval (XiaoIce)
- Neural encoder-decoder (BlenderBot)

[21]

www.jads.nl

Jheronimus Academy of Data Science

BlenderBot (Roller et al. 2020)

Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of “Sing Along with Mitch”?

Yes, I would love that.

OK, I’ll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I’m a teacher, so I don’t have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: “Baking is my favorite thing to do in the world.”

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It’s so catchy. I can’t get it out of my head. [22]

www.jads.nl

Task-based dialogue agents

"Task-based" or "goal-based" dialogue agents

- Systems that have the goal of helping a user solve a task
  - Setting a timer
  - Making a travel reservation
  - Playing a song
  - Buying a product

Architecture:

- Frames with slots and values
- A knowledge structure representing user intentions

[23]

www.jads.nl

Jheronimus Academy of Data Science

The Frame

- A set of **slots**, to be filled with information of a given **type**
- Each associated with a **question** to the user

Slot	Type	Question
ORIGIN	city	"What city are you leaving from?
DEST	city	"Where are you going?
DEP DATE	date	"What day would you like to leave?
DEP TIME	time	"What time would you like to leave?
AIRLINE	line	"What is your preferred airline?

A telephone conversation between a human travel agent (A) and a human client (C)

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 . . .

[25]

www.jads.nl

Properties of Human Conversation – 1

Turns

- We call each contribution a "turn"
- As if conversation was the kind of game where everyone takes turns

[26]

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 ...

Properties of Human Conversation – 2

Turn taking issues

- When to take the floor?
- When to yield the floor?

Interruptions

[28]

www.jads.nl

Jheronimus Academy of Data Science

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 . . .

www.jads.nl

Jheronimus Academy of Data Science

Implications for Conversational Agents

Barge-in

- Allowing the user to interrupt

End-pointing

- The task for a speech system of deciding whether the user has stopped talking.
- Very hard, since people often pause in the middle of turns

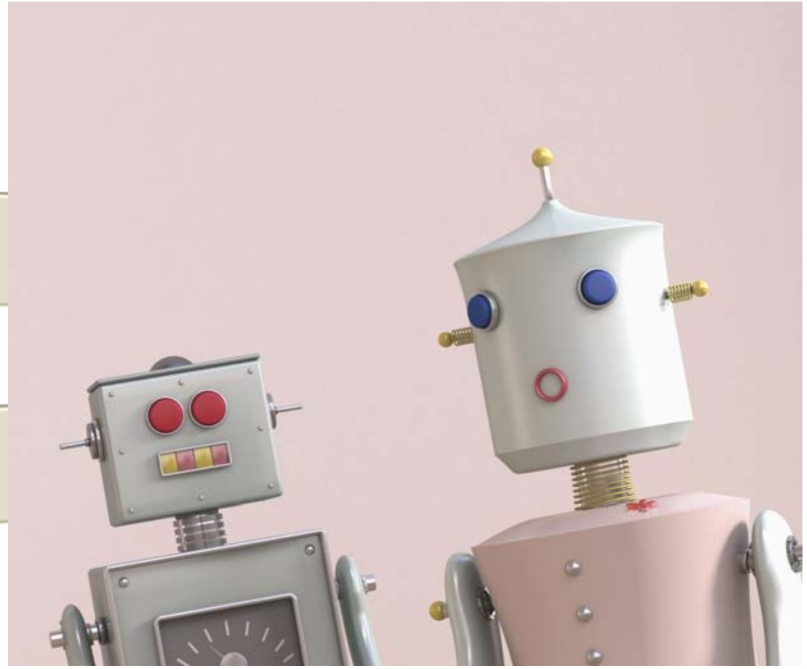
[30]

www.jads.nl

Jheronimus Academy of Data Science

Rule-based chatbots

- ELIZA
- PARRY



[31]

www.jads.nl

Jheronimus Academy of Data Science

ELIZA: Weizenbaum (1966)

- Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

- CAN YOU THINK OF A SPECIFIC EXAMPLE
- Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

- He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

- …
- You are like my father in some ways.
- WHAT RESEMBLANCE DO YOU SEE
- You are not very aggressive but I think you don't want me to notice that.
- WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
- You don't argue with me.
- WHY DO YOU THINK I DON'T ARGUE WITH YOU
- You are afraid of me.
- DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU
- My father is afraid of everybody.
- WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
- Bullies.
- DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45

ELIZA's trick: be a Rogerian psychologist

Draw the patient out by reflecting patient’s statements back at them

Rare type of conversation in which one can “assume the pose of knowing almost nothing of the real world”

[33]

www.jads.nl

Jheronimus Academy of Data Science

Rogerian psychologist

- Patient: "I went for a long boat ride”
- Psychologist: "Tell me about boats”
- You don’t assume she didn’t know what a boat is
- You assume she had some conversational goal
- Chatbots trying to pass the Turing test usually choose some such domain

[34]

www.jads.nl

Eliza pattern / transform rules

`(0 YOU 0 ME)` [*pattern*]

→

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

0 means Kleene *

The 3 is the constituent # in pattern

`You hate me`

`WHAT MAKES YOU THINK I HATE YOU`

www.jads.nl

Eliza Rules

Rules are organized by **keywords**
Each keyword has a pattern and a list of possible transform

Keyword: `YOU`
Pattern: `(0 YOU 0 ME)`
Transforms: `(WHAT MAKES YOU THINK I 3 YOU)`
`(WHY DO YOU THINK I 3 YOU)`

www.jads.nl

Keywords are ranked from specific to general

```
User: I know everybody laughed at me
```

- "I" is very general:

```
I: (I *) -> (YOU SAY YOU 2)
ELIZA: YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
```

- "Everybody" is more specific and interesting

```
•Everybody: (Everybody *) -> (WHO IN PARTICULAR ARE YOU THINKING OF)
ELIZA: WHO IN PARTICULAR ARE YOU THINKING OF?
```

- Implementation: keywords stored with their rank

Everybody 5 (list of *transformation rules*)
I 0 (list of *transformation rules*)

www.jads.nl

If no keyword matches, choose a non-committal response

- PLEASE GO ON
- THAT'S VERY INTERESTING
- I SEE

[38]

www.jads.nl

Memory

```
(MEMORY MY
 (0 MY 0 = LETS DISCUSS FURTHER WHY YOUR 3)
 (0 MY 0 = EARLIER YOU SAID YOUR 3)
 (0 MY 0 = DOES THAT HAVE ANYTHING TO DO
WITH THE FACT THAT YOUR 3))
```

Whenever “MY” is highest keyword

- Randomly select a transform on the MEMORY list
- Apply to sentence
- Store on a (first-in-first-out) queue

Later, if no keyword matches a sentence

- Return the top of the MEMORY queue instead

[39]

www.jads.nl

Jheronimus Academy of Data Science

function ELIZA GENERATOR(user *sentence*) **returns** *response*

Find the word *w* in *sentence* that has the highest keyword rank
  if *w* exists
    Choose the highest ranked rule *r* for *w* that matches *sentence*
    response ← Apply the transform in *r* to *sentence*
    if *w* = 'my'
      future ← Apply a transformation from the 'memory' rule list to *sentence*
      Push *future* onto memory stack
  else (no keyword applies)
   either
    response ← Apply the transform for the NONE keyword to *sentence*
   or
    response ← Pop the top response from the memory stack
 return(*response*)

[40]

www.jads.nl

Jheronimus Academy of Data Science

Ethical implications: Anthropomorphism and Privacy

- People became deeply emotionally involved with the program
- One of Weizenbaum's staff asked him to leave the room when she talked with ELIZA
- When he suggested that he might want to store all the ELIZA conversations for later analysis, people immediately pointed out the privacy implications
  - Suggesting that they were having quite private conversations with ELIZA
  - Despite knowing that it was just software.

PARRY: A computational model of schizophrenia

- Another chatbot with a clinical psychology focus
  - *Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. Artificial Intelligence 2(1), 1–25.*
- Used to study schizophrenia
- Same pattern-response structure as Eliza, but much richer:
  - control structure
  - language understanding capabilities
  - model of mental state
    - variables modeling levels of Anger, Fear, Mistrust

42

www.jads.nl

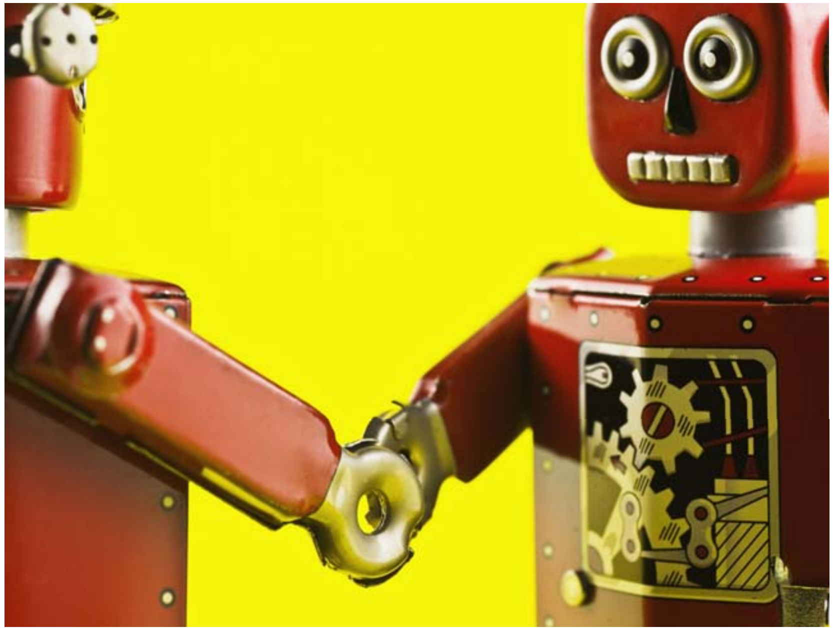
Jheronimus Academy of Data Science

PARRY passes the Turing test in 1972

- The first system to pass a version of the Turing test
- Psychiatrists couldn't distinguish interviews with PARRY from (text transcripts of) interviews with people diagnosed with paranoid schizophrenia
- Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence* 3, 199–221.

Corpus-based chatbots

- Learn response from a corpus
- Modern corpus-based chatbots are very data-intensive
- They commonly require hundreds of millions or billions of words



[44]

www.jads.nl

Jheronimus Academy of Data Science

Two architectures for corpus-based chatbots

Response by retrieval

Use information retrieval to grab a response (that is appropriate to the context) from some corpus

Response by generation

Use a language model or encoder-decoder to generate the response given the dialogue context

[45]

www.jads.nl

Jheronimus Academy of Data Science

What conversations to draw on?

Transcripts of telephone conversations between volunteers

- Switchboard corpus of American English telephone conversations

Movie dialogue

- Various corpora of movie subtitles

Hire human crowdworkers to have conversations

- Topical-Chat 11K crowdsourced conversations on 8 topics
- EmpatheticDialogues 25K crowdsourced conversations grounded in a situation where a speaker was feeling a specific emotion

Pseudo-conversations from public posts on social media

- Drawn from Twitter, Reddit, Weibo (微博), etc.
- Tend to be noisy; often used just as pre-training.

Crucial to remove personally identifiable information (PII)

[46]

Response by retrieval: classic IR method

$$\text{response}(q, C) = \underset{r \in C}{\operatorname{argmax}} \frac{q \cdot r}{|q||r|}$$

1. Given a user turn q, and a training corpus C of conversation
2. Find in C the turn r that is most similar (tf-idf cosine) to q
3. Say r

[47]

www.jads.nl

Response by retrieval: neural IR method

$$
\begin{aligned}
h_q &= \mathrm{BERT}_Q(\mathrm{q})\,[\mathrm{CLS}] \\
h_r &= \mathrm{BERT}_R(\mathrm{r})\,[\mathrm{CLS}] \\
\mathrm{response}(q,C) &= \underset{r\in C}{\mathrm{argmax}}\, h_q \cdot h_r
\end{aligned}
$$

1. Given a user turn *q*, and a training corpus *C* of conversation
2. Find in *C* the turn *r* that is most similar (BERT dot product) to *q*
3. Say *r*

[48]

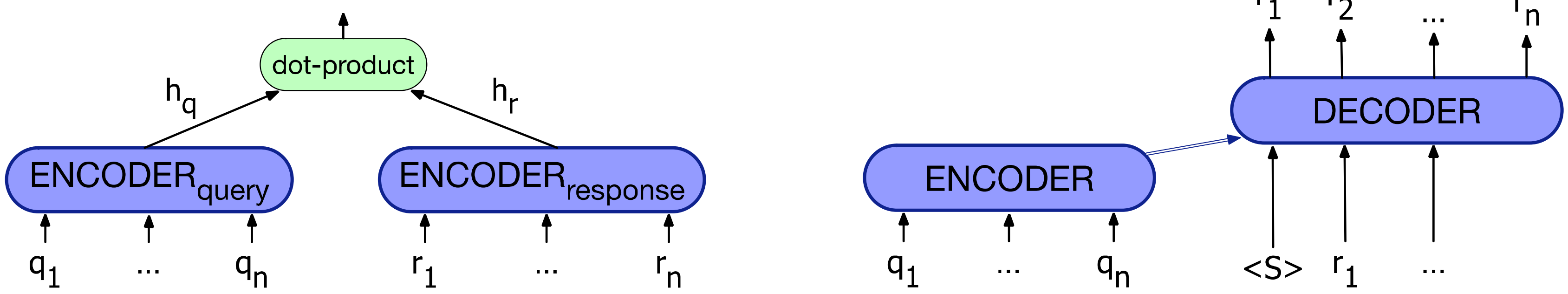
www.jads.nl

Jheronimus Academy of Data Science

Response by generation

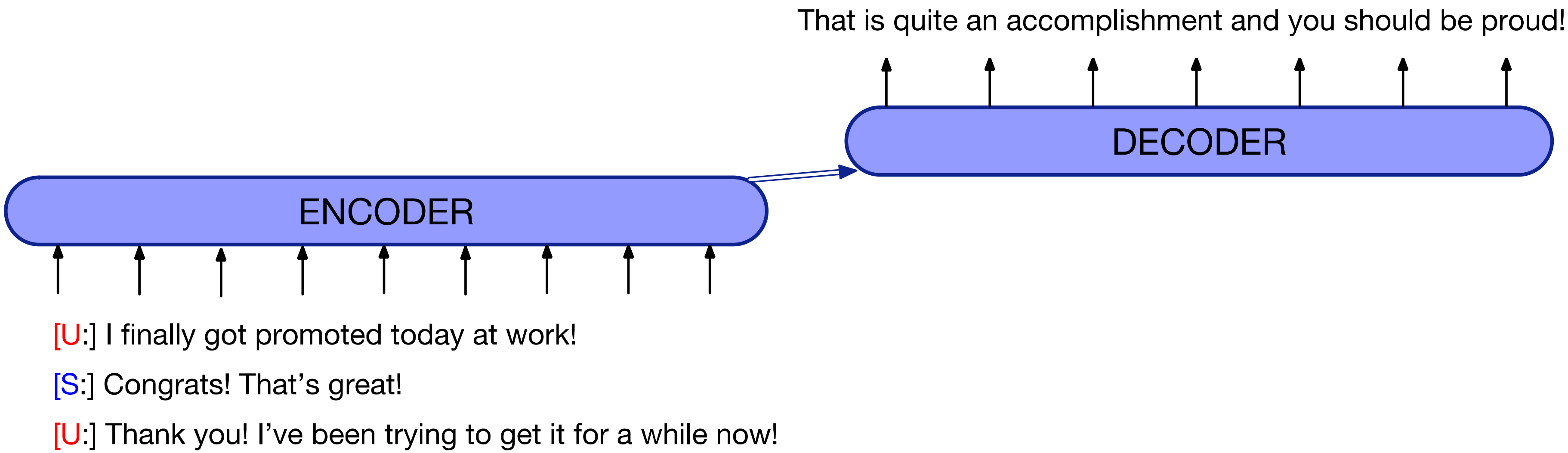
- Think of response production as an encoder-decoder task
- Generate each token r_t of the response by conditioning on the encoding of the entire query q and the response so far $r_1 \ldots r_{t-1}$

$$\hat{r}_t = \text{argmax}_{w \in V} P(w|q, r_1 \ldots r_{t-1})$$



(a) Response by Retrieval (b) Response by Generation

Response by generation



[50]

www.jads.nl

Jheronimus Academy of Data Science

Examples of NLP tasks: Linguistic summarization: template-based descriptions

51

Linguistic Summarization

Communicate information in (quasi) natural language

- Used to condense large data sets
- A natural mechanism for temporal data streams
  - For example: stock market trends
- Fuzzy set theory allows modeling linguistic quantifiers (*e.g.* most, some, a few, etc.)

[52]

www.jads.nl

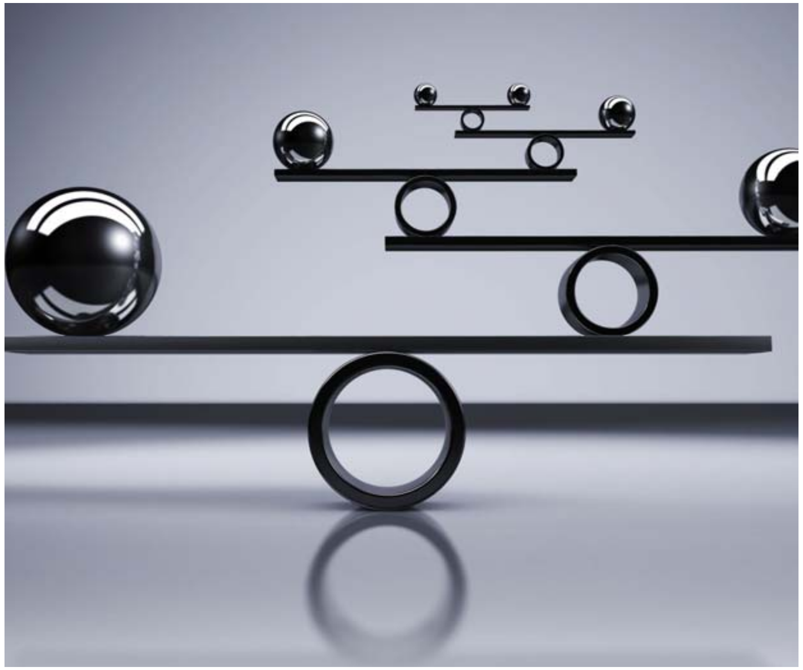
Why linguistic summaries?

"Childhood obesity rates remain high. Overall, obesity among our nation's young people, aged 2 to 19 years, has not changed significantly since 2003-2004 and remains at about 17 percent. However, among 2 – 5 years old, obesity has declined based on CDC's National Health and Nutrition Examination Survey (NHANES) data. "

(http://www.cdc.gov/obesity/data/childhood.html)

Visualization vs. verbalization

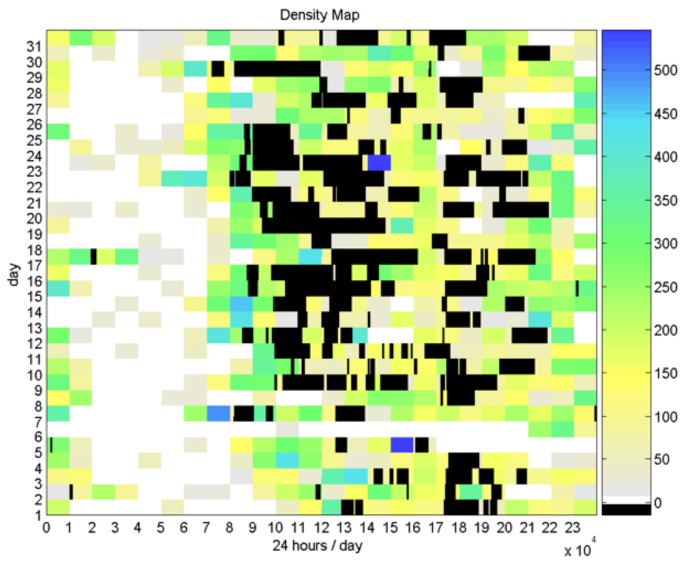
- Visualization requires the user to be trained in the semantics of the visualization model
- Verbalization uses natural language, in which we all have the basic training
  ➔ verbal descriptions are more appropriate in various situations



[54]

www.jads.nl

Visual description vs. verbal description



- *most* 15-minute-intervals during sleep have *low* bedroom motion
- *a few* 15-minute-intervals during sleep have *medium* restlessness

Nurses have a preference for verbal descriptions

[55]

www.jads.nl

Jheronimus Academy of Data Science

Words convey approximate information

"one picture is worth more than ten thousand words", but...

What is Vera's age?

- p_1: Vera has a son in mid-twenties
- p_2: Vera has a daughter in mid-thirties

What is the probability that Robert is home at 6:15pm?

- p_1: Usually Robert leaves office at about 5pm
- p_2: Usually it takes Robert about an hour to get home from work

Graduality

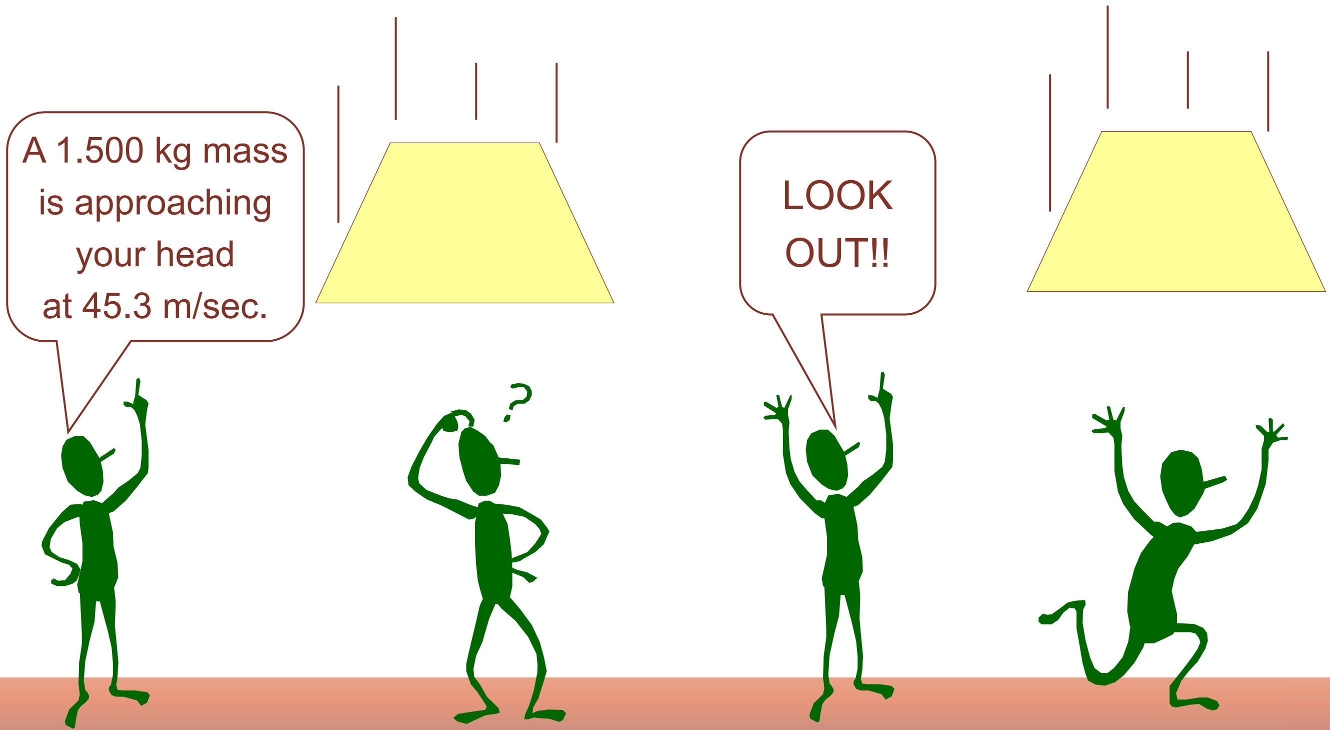
- Humans conceptualize the world based on concepts of similarity, gradualness, fuzziness
- E.g. When is somebody tall?
  Some 200 cm?
  What about 190 cm?
  What about 180 cm?
- C.f. sorites from ancient Greeks
  (when does someone become bald when removing one hair at a time from his/her head?)

Crisp vs. fuzzy quantities

- Integers larger than 3
- Families without children
- People with job description "manager"

- Tall people
- Fast cars
- Bold men
- Tall and blond Dutch
- Comfortable car

Precision and relevance



www.jads.nl

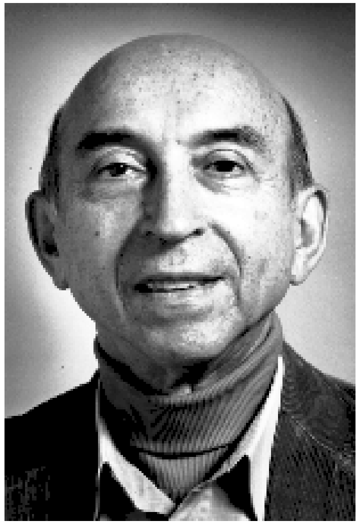
JADS Jheronimus Academy of Data Science

59

Fuzzy sets

- Tool to communicate model behavior linguistically
- Mechanisms to match “mental world” with “observed world”
- Summarize properties of data linguistically

(Zadeh, 1965)



60

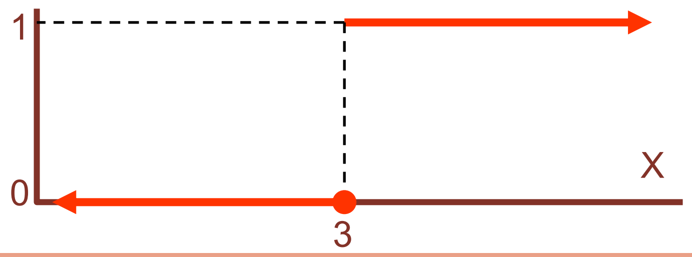
www.jads.nl

Partial set membership

Crisp sets

- characteristic function
  $f_A: X \to \{0,1\}$,
  $f_A(x) = 1, \Leftrightarrow x \in X$
  $f_A(x) = 0, \Leftrightarrow x \notin X$

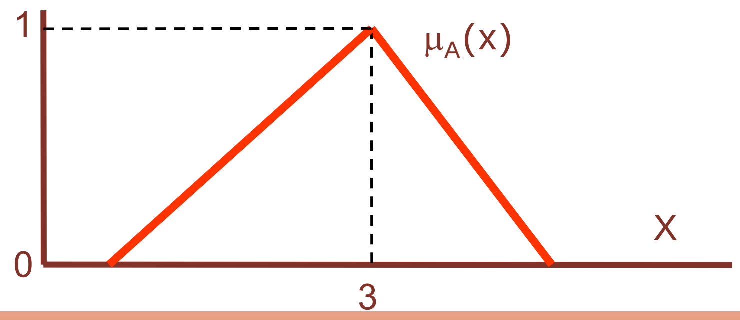
Real numbers larger than 3:



Fuzzy sets

- A fuzzy set A in X is characterized by its membership function
  $\mu_A: X \longrightarrow [0,1]$

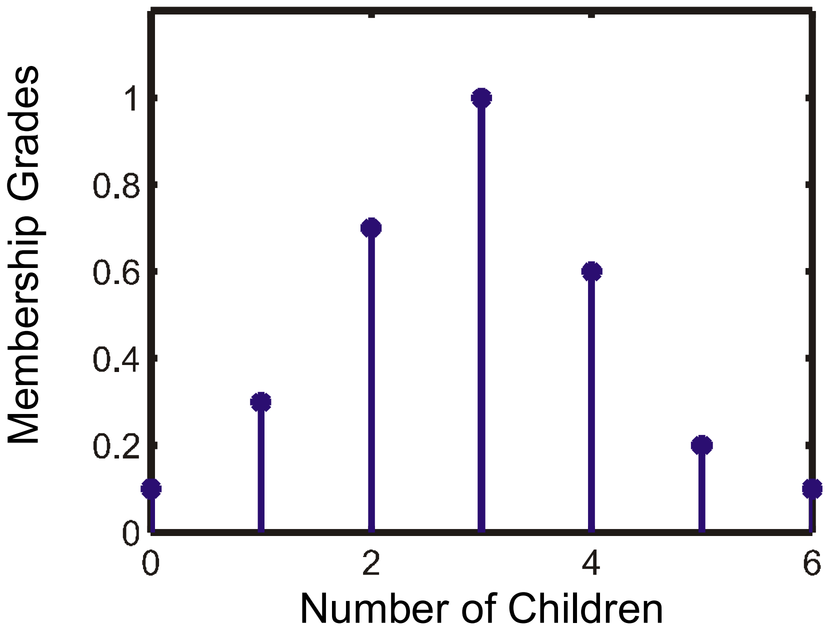
Real numbers about 3:



www.jads.nl

Fuzzy sets on discrete universes

- Fuzzy set C = “desirable city to live in”
  - X = {SF, Boston, LA} (discrete and non-ordered)
  - C = {(SF, 0.9), (Boston, 0.8), (LA, 0.6)}
- Fuzzy set A = “sensible number of children”
  - X = {0, 1, 2, 3, 4, 5, 6} (discrete universe)
  - A = {(0, .1), (1, .3), (2, .7), (3, 1), (4, .6), (5, .2), (6, .1)}

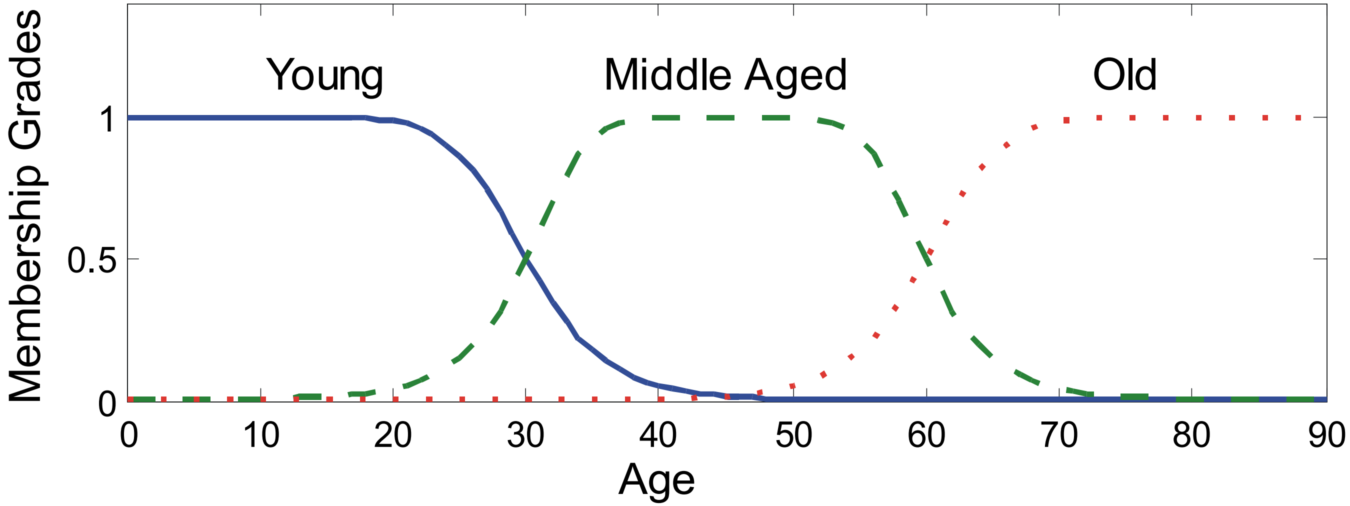


62

www.jads.nl

Fuzzy partition

Fuzzy partition formed by the linguistic values “young”, “middle aged”, and “old”:



Set theoretic operations

- Subset:

$$A \subseteq B \Leftrightarrow \mu_A \leq \mu_B$$

- Complement:

$$\overline{A} = X - A \Leftrightarrow \mu_{\overline{A}}(x) = 1 - \mu_A(x)$$

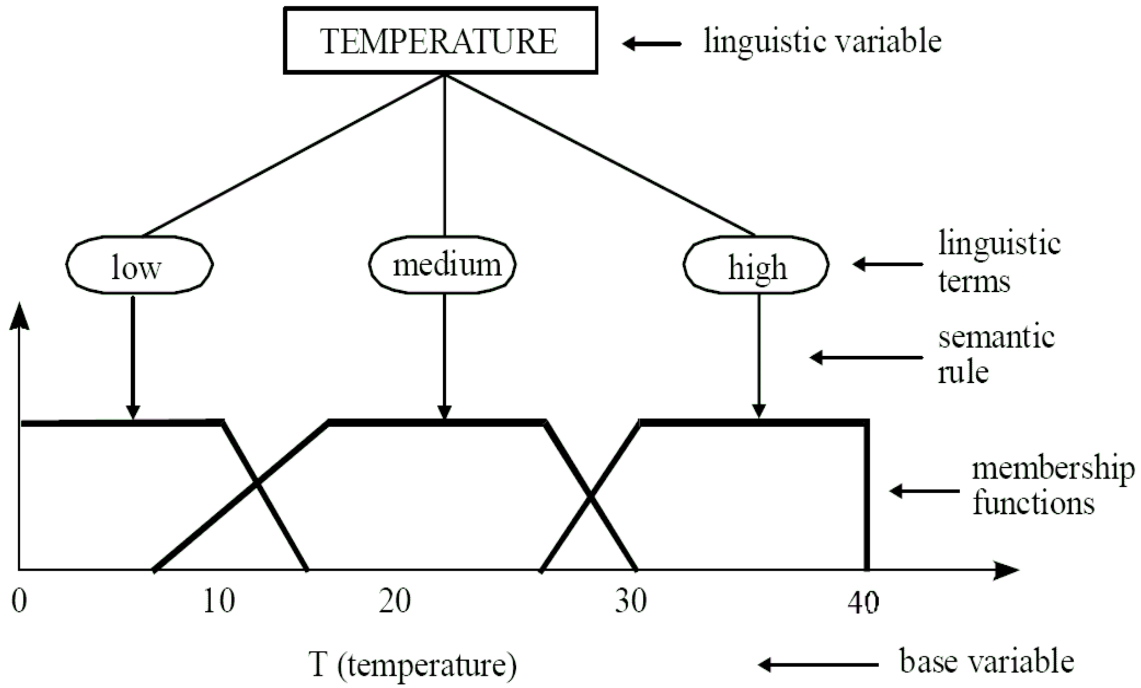
- Union:

$$C = A \cup B \Leftrightarrow \mu_c(x) = \max(\mu_A(x), \mu_B(x)) = \mu_A(x) \vee \mu_B(x)$$

- Intersection:

$$C = A \cap B \Leftrightarrow \mu_c(x) = \min(\mu_A(x), \mu_B(x)) = \mu_A(x) \wedge \mu_B(x)$$

Linguistic Variable



65

www.jads.nl

Quantifiers in logic

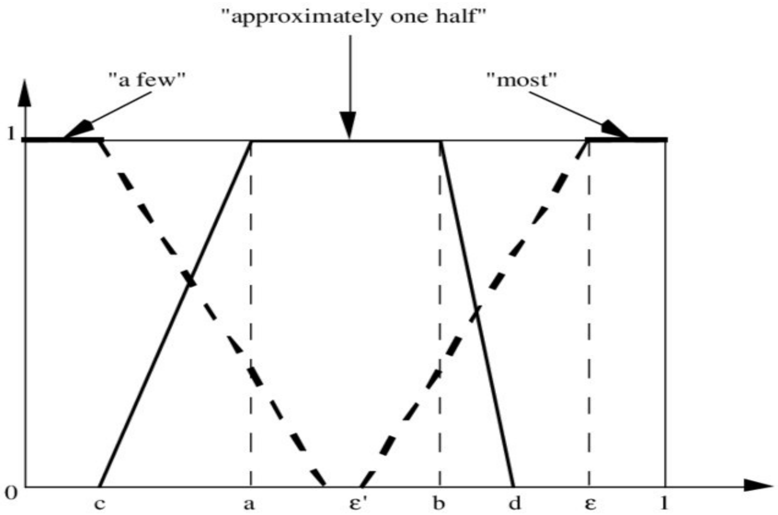
Classical logic

Two quantifiers:

- ∃ – There is …
- ∀ - For all …

Fuzzy logic

Many more quantifiers



Bouchon-Meunier *et al.* (1999)

[66]

www.jads.nl

Linguistic summary

quantified propositions - protoforms:

Q y 's are P

e.g. most cars are new

Q R y 's are P

e.g. many new cars are fast

R. R. Yager, "A new approach to the summarization of data," Information Sciences, vol. 28, pp. 69–86, 1982.

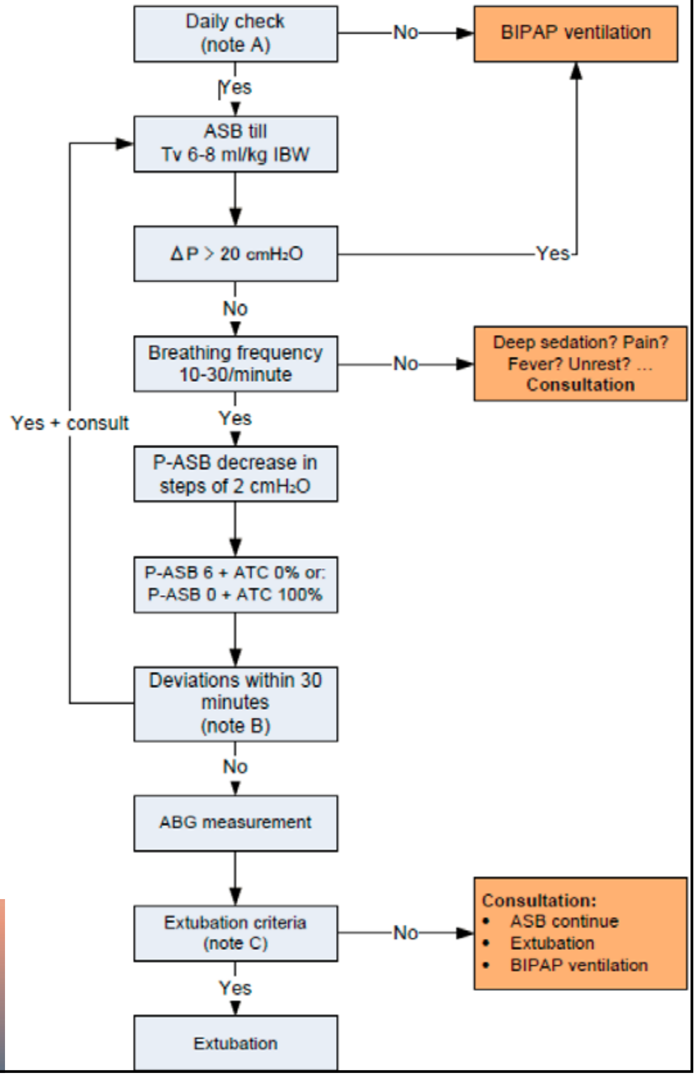
www.jads.nl

Truth value

$$\mathcal{T}(Qy\text{'s are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right)$$

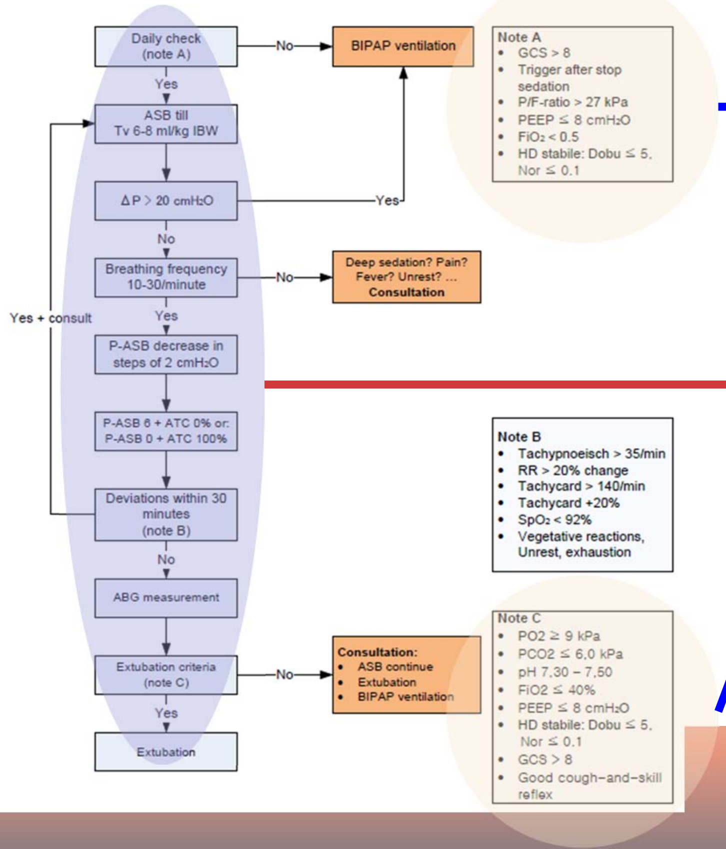
$$\mathcal{T}(QRy\text{'s are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}(\mu_R(y_i)\wedge\mu_P(y_i))}{\sum_{i=1}^{n}\mu_R(y_i)}\right)$$

Weaning protocol at ICU



www.jads.nl

Background – Weaning protocol, MUMC+



Thresholds

Process flow

Case study:
Summaries are based on this protocol

www.jads.nl

Jheronimus Academy of Data Science

Activity analysis

P_ID	Activity Sequences
1	a c d q e d d d o r
2	a c d q e d d f h j c d e d f h j c d f h j e i l m l m l m n o p
3	a b c d q d e d d d e d d d d e d f g j c d f g j m l m n o p
4	a b c d q e d d d d e d d e d d d d d d e d d d f h j l m l m e l m l m l m l m n o p
5	a b c d q d e d d d e d d d f h j l l e k l l k m n o r
6	a b b c d q d d d e d d d d d e d d e f h j c d d d d e d f h j c d d d d d e f h j k m n o r

Phase 1 Phase 2 Phase 3 Phase 4

www.jads.nl

Protoforms

Type	Protoforms
Phase Level	Q patients with C follow the process P.
Activity Level Clustering	Q patients with C follow process P in phase H.
Activity Level Patterns	Q patients with C follow process P (pattern) in phase H.
Cost* comparison between 2 phases	For Q cases, the cost of H_1 is P than the cost of H_2
Cost* comparison between 2 patient groups	For Q cases, patients with C_1 have P costs in H than patients with C_2.
Time Comparison	For Q cases, patients with C_1 spend P time than patients with C_2.

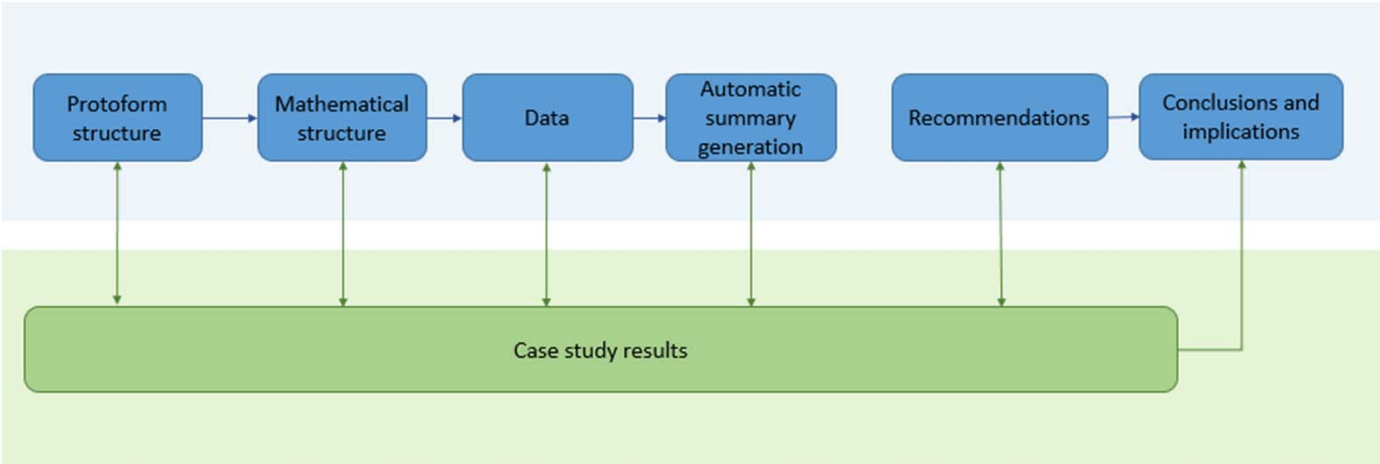
www.jads.nl

Example summaries

- Most patients follow Pre-BIPAP Phase, BIPAP Phase, ASB Phase, Extubation Phase.
- A few patients follow Pre-BIPAP Phase, BIPAP Phase, Extubation Phase.
- For About a half cases, patients with COPD spend less time than patients without COPD in weaning protocol.

Process evaluation

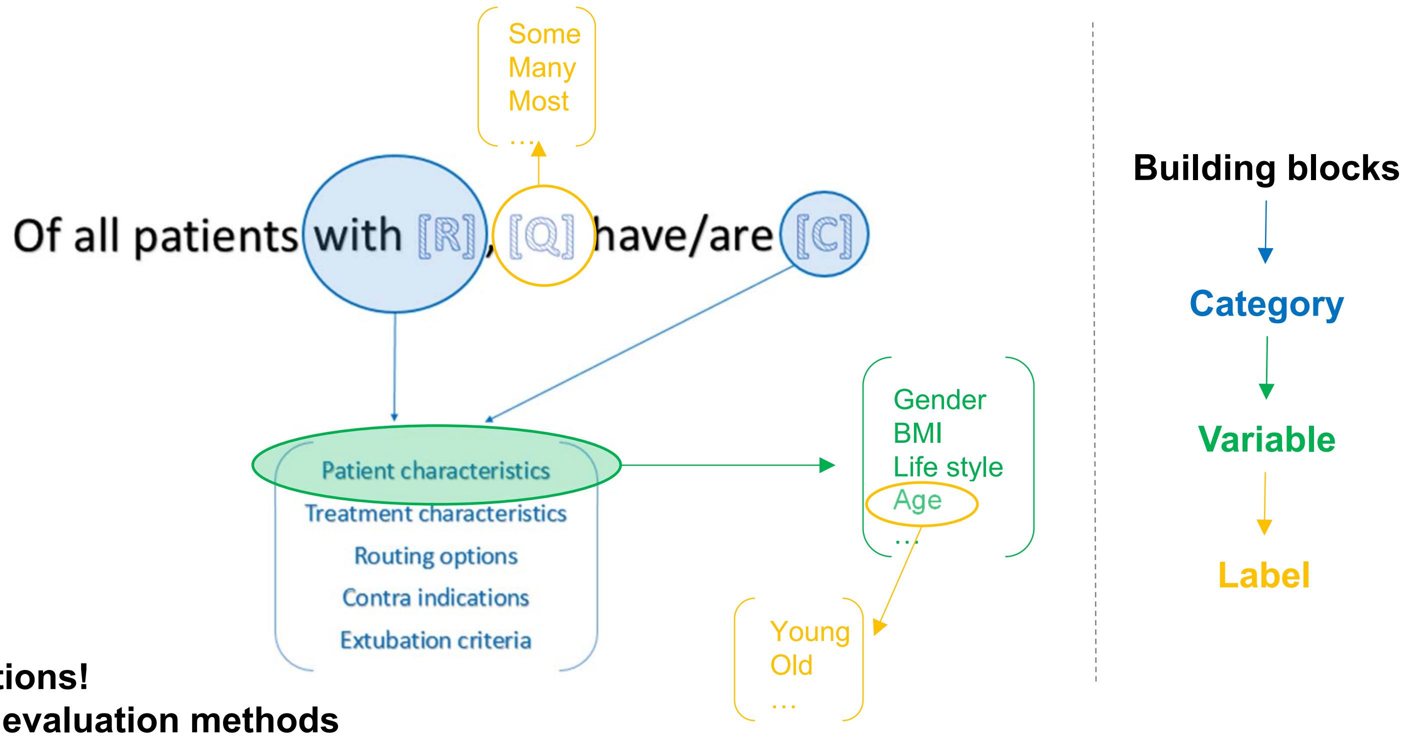
"How can linguistic summaries help evaluate general performance and compliance of clinical processes?"



www.jads.nl

JADS Jheronimus Academy of Data Science

Linguistic summary of protocol execution



Many options!
Need for evaluation methods

www.jads.nl

Summaries (1)

General Analysis

Of all patients with [R], [Q] have/are [C]

Of all patients with

Patient characteristics
Age
Gender
Obesity
Lifestyle: smoker
Lifestyle: drinker
Level of restlessness

most have high/low value for

Extubation criteria	
PO2	>=9 kPa
PCO2	<=6.0 kPa
PH	7,30 < x < 7,50
FiO2	<=0,4
PEEP	<=8 cmH2O
Dopamine/dobutamine	<5 mcgr/kg/min
Noradrenaline	<0,1 mcgr/kg/min
GCS	> 8

Extubation criteria

Label

Young
Old
...

www.jads.nl

JADS Jheronimus Academy of Data Science

Summaries (2)

Compliance analysis

Of all patients, most have a value for

Extubation criteria	
PO2	>=9 kPa
PCO2	<=6.0 kPa
PH	7,30 < x < 7,50
FiO2	<=0,4
PEEP	<=8 cmH2O
Dopamine/dobutamine	<5 mcgr/kg/min
Noradrenaline	<0,1 mcgr/kg/min
GCS	> 8

that is conform/not conform protocol.

www.jads.nl

Results general analysis

Not many sentences mention ‘low’ or ‘high’ values

Stable process execution	Stable process time	Well-defined protocol

No clear differences between patient groups could be identified

[78]

www.jads.nl

Jheronimus Academy of Data Science

Results compliance analysis

- Confirm results from general analysis
  - Many sentences state compliance
  - No clear differences between patient groups regarding compliance
- Compliance is high
- Very intuitive method

[79]

www.jads.nl

JADS Jheronimus Academy of Data Science

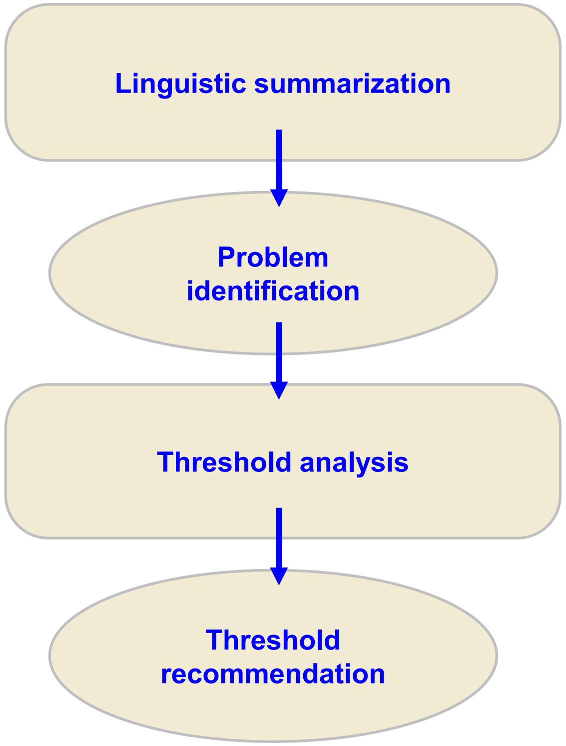
Numerical analysis

Contra indications			Extubation criteria		
FiO2			FiO2		
	Amount recorded	1741		Amount recorded	2364
	Compliance	**83%**		Compliance	95%
PEEP			PEEP		
	Amount recorded	1685		Amount recorded	2342
	Compliance	94%		Compliance	94%
PaO_2/F_iO_2 -ratio			pO2		
	Amount recorded	89		Amount recorded	2660
	Compliance	**79%**		Compliance	90%
			pCO2		
				Amount recorded	2659
				Compliance	**88%**
			pH		
				Amount recorded	2660
				Compliance	90%

www.jads.nl

Jheronimus Academy of Data Science

Recommendations



- Some variables show lower compliance
- Discussion with medical personnel
  - Deviating values still ‘safe’
  - Are the boundaries too ‘tight’?
  - Why not alter thresholds?

Modify thresholds

All proposed thresholds are safe

Threshold for pH
- '>7.30' to '>7.26'
- Increase in compliance (90% → 95%)

Threshold for pCO2
- '<6.0' to '<6.8'
- Increase in compliance (89% → 99%)

Threshold for PaO2/FiO2 ratio
- '>27' to '>22'
- Increase in compliance (79% to 89%)

www.jads.nl

Jheronimus Academy of Data Science

Summary

- Named entity recognition is important for many NLP tasks
  - Rule-based methods
  - Machine learning methods: solved as a classification problem after tagging
- Conversational agents interact with users in natural language
  - Two kinds: chatbots, task-based dialogue agents
  - Two architectures: rule-based, corpus-based
  - Rule-based chatbots work with pattern/transformation combinations
  - Corpus-based chatbots: work with response retrieval or with response generation
- Linguistic summaries
  - Generate a description of “objects” according to a template
  - Use fuzzy set theory to represent partial quantifiers

[83]