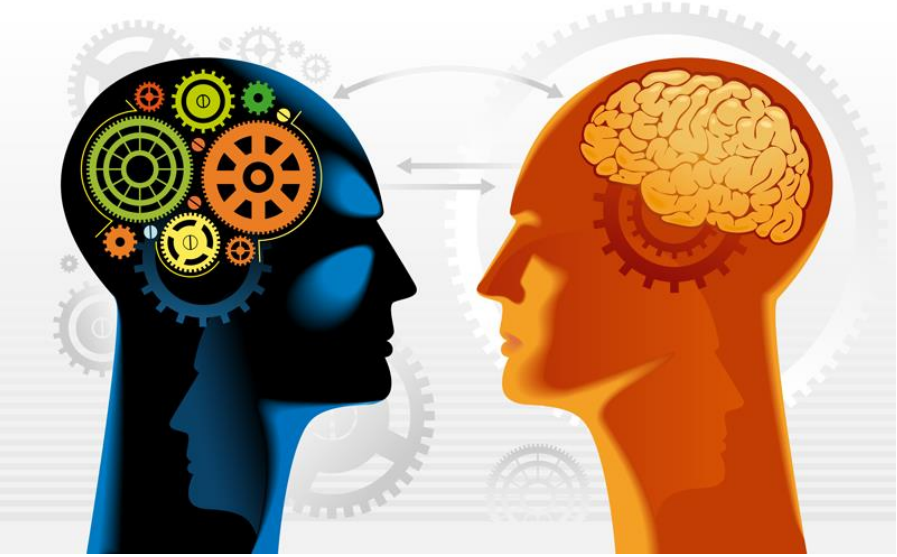
University of Ljubljana, Faculty of Computer and Information Science

Natural language processing – an introduction

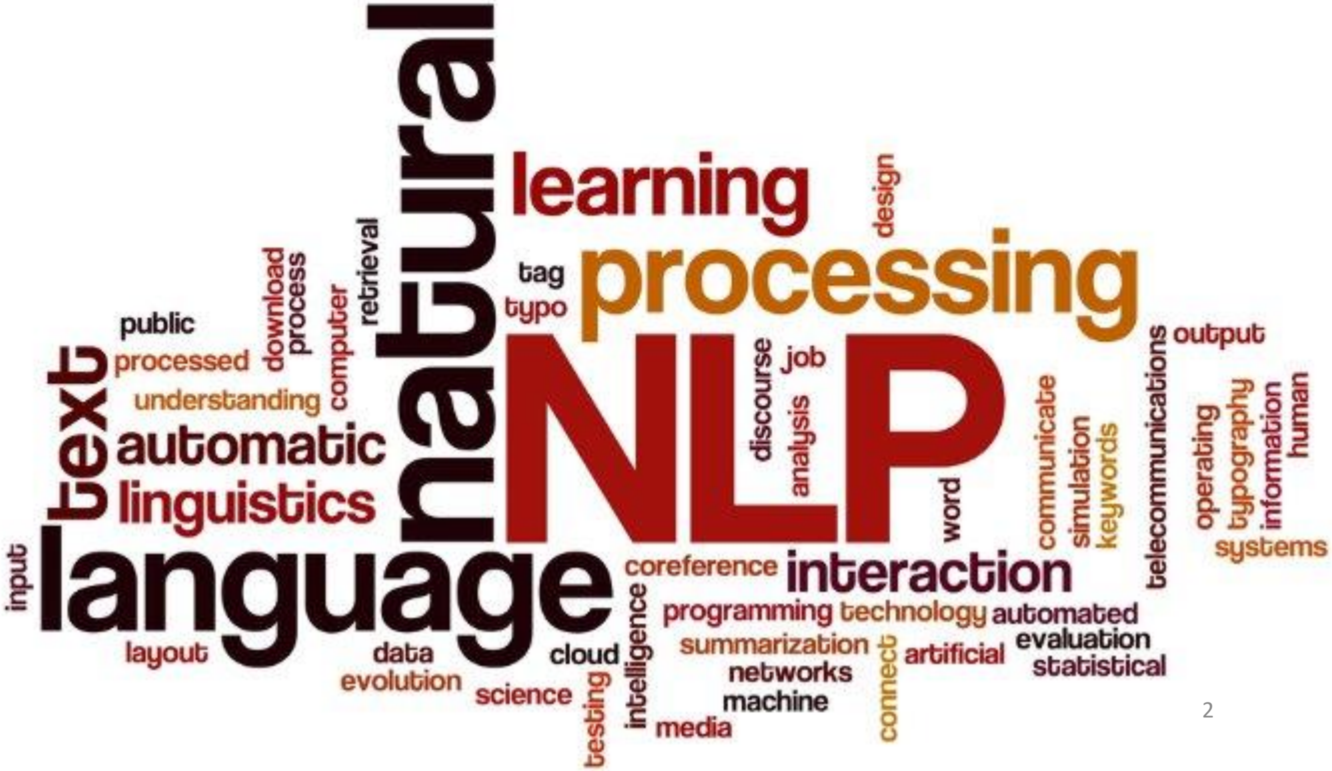


Prof Dr Marko Robnik-Šikonja

Edition 2022

Outline

- about the course
- contents overview
- language understanding and intelligence
- overview of the NLP area



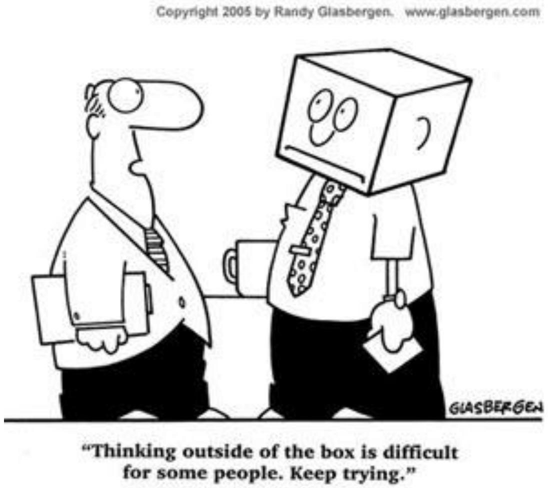
Lecturer

- Prof Dr Marko Robnik-Šikonja
- University of Ljubljana
  Faculty of Computer and Information Science
  Laboratory for Cognitive Modeling
- FRI, Večna pot 113, 2nd floor, right from the elevator
- marko.robnik@fri.uni-lj.si
- https://fri.uni-lj.si/en/employees/marko-robnik-sikonja
- (01) 4798 241
- Contact hour
  - Wednesday, 10:00 -11:00; email me for other slots and Zoom
- **Research interests**: machine learning, artificial intelligence, natural language processing, network analytics, data science, data mining,
- **Teaching:** courses from the area of data mining, algorithms, machine learning, and natural language processing
- **Resources:** an author or coauthor of several software tools, including three open source R packages from the area of predictive modelling and data analytics (CORElearn, semiArtificial, ExplainPrediction), machine learning models and language resources

Assistant

- Assist Prof Dr Slavko Žitnik
  slavko.zitnik@fri.uni-lj.si
- Laboratory for Data Technologies
- Research interests: information retrieval, semantic web

Goals of the course



"Thinking outside of the box is difficult for some people. Keep trying."

- students shall become acquainted with
  - basics of natural language processing and understanding
  - basic approaches and data representations for NLP
  - modern techniques for NLP
  - selected relevant NLP tasks
  - relevant research challenges in the area of NLP and NLU, computational linguistics, and semantics
- teach students a practical use of
  - selected tools
  - selected modern techniques for NLP
- awareness of ethical issues in NLU
- increase the (mental) problem-solving toolbox with new NLP approaches and techniques
- awareness of languages as important sources of information

Learning outcomes

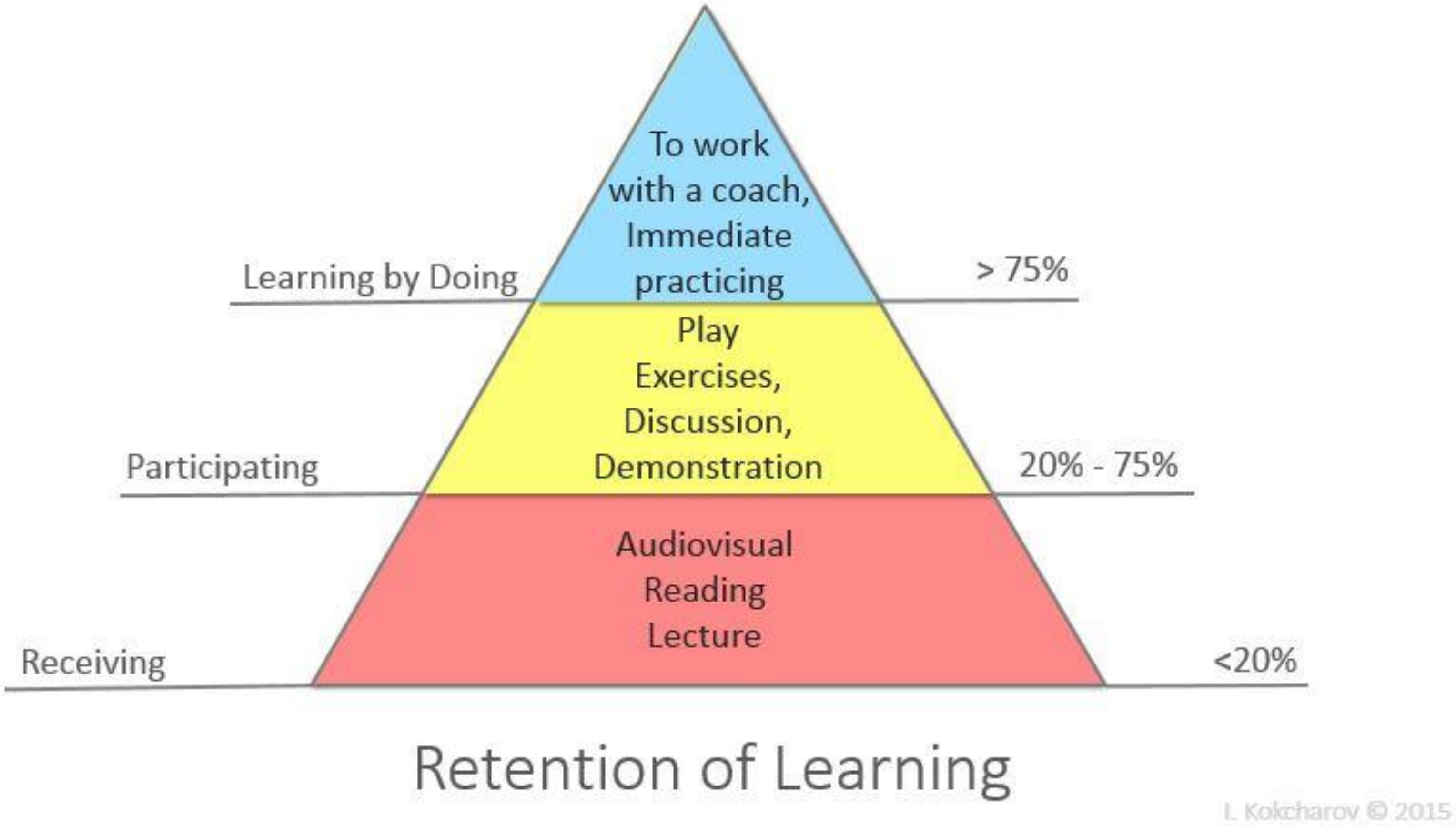
Upon completion of the course, students shall:

- understand approaches to syntax and semantics in NLP,
- use and adapt machine learning techniques for NLP
- apply and critically evaluate natural language processing tools
- know the existing language resources and be able to design new ones
- use text representations and adapt them to new contexts
- use and evaluate approaches to text classification, summarization, machine translation, affective computation, question answering, etc.

Lectures and tutorials

- Lectures
  - introduction to the topic, discussion
  - some examples
  - broader view of the topics
- Tutorials
  - exercises
  - assignments motivated by practical use
  - assistant presents the assignments, helps with tips, moderates discussion, so…
  - …come prepared and pose questions.
  - introduce some problem solving tools and useful software
  - mostly deals with English and Slovene

BTW: retention of learning



Syllabus 1/2

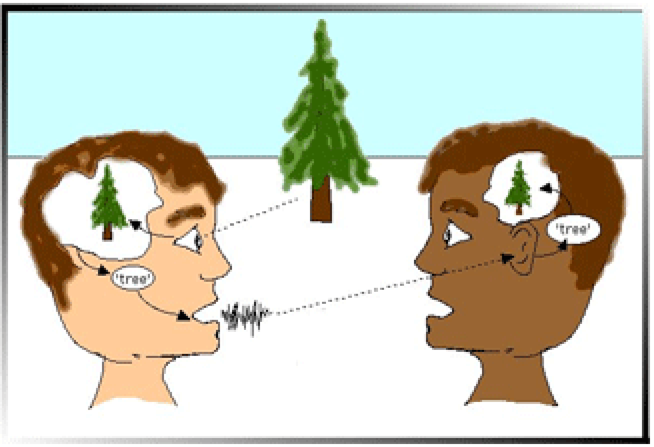
1. Introduction to natural language processing: motivation, language understanding, ambiguity, traditional, statistical, and neural approaches.
2. Text preprocessing and normalization: regular expressions for search and replacement, grammars for syntax analysis, string similarity, Levenhstein distance, advanced normalization techniques, lemmatization.
3. Language resources: corpora, dictionaries, thesauri, networks and semantic databases, WordNet.
4. Text similarity: measures, clustering approaches, cosine distance, language networks, and graphs.
5. Text representation: sparse and dense; language models; word, sentence, and document embeddings.
6. Deep neural networks for text: recurrent neural networks, CNNs for text, transformers.
7. Neural embeddings: word2vec, fastText, ELMo, BERT, GPT, cross-lingual embeddings.

Syllabus 2/2

8. Shallow computational and lexical semantics: part-of-speech tagging, dependency parsing, named entity recognition, semantic role labelling, FrameNet.
9. Word senses and disambiguation.
10. Affective computing: sentiment, emotions.
11. Text summarization: text representations, extractive methods, query-based methods, abstractive summarization.
12. Question answering and reading comprehension
13. Machine translation: statistical and neural machine translation.
14. Semantic representations: knowledge graphs for commonsense reasoning.

What the course does not cover?

- speech processing: recognition and synthesis (a course in UL Faculty of Electrical Engineering, „Speech technologies“)
- chatbots and dialogue systems
- information retrieval (the FRI course „Web information extraction and retrieval*“)*
- in-depth linguistics



Prerequisites

- Recommended knowledge
  - Python programming,
  - probability and statistics,
  - machine learning.

Obligations

- 5 quizzes checking continuous understanding of contents
- projects, composed of three stages, 50 points
- a written exam, 50 points (in case of epidemics, an oral exam)

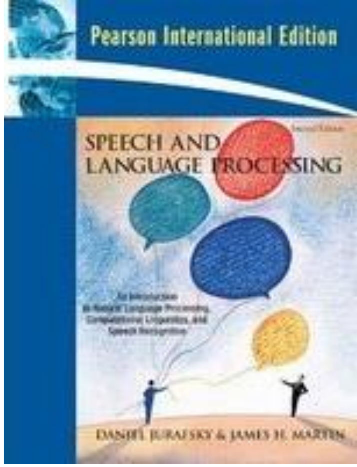
Grading

Obligation	% of total	subject to
Five quizzes	0%	≥ 50% alltogether
Projects	50%	≥ 50%
Written exam	50%	≥ 50%

Learning materials

- learning materials in the eClassroom http://ucilnica.fri.uni-lj.si
- slides are updated continuously
- links to the literature
- code and examples
- links to datasets

Literature


Pearson International Edition
SPEECH AND LANGUAGE PROCESSING
DANIEL JURAFSKY & JAMES H. MARTIN

- Jurafsky, David and Martin, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 3rd edition* draft, 2021. Basic course literature; available on authors' webpages

- Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009. *Freely available book*, updated in 2019, *based on NLTK library for Python 3*
- Jacob Eisenstein. Natural Language Processing, 2018

- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016

- scientific papers for deeper understanding of certain topics

Web courses

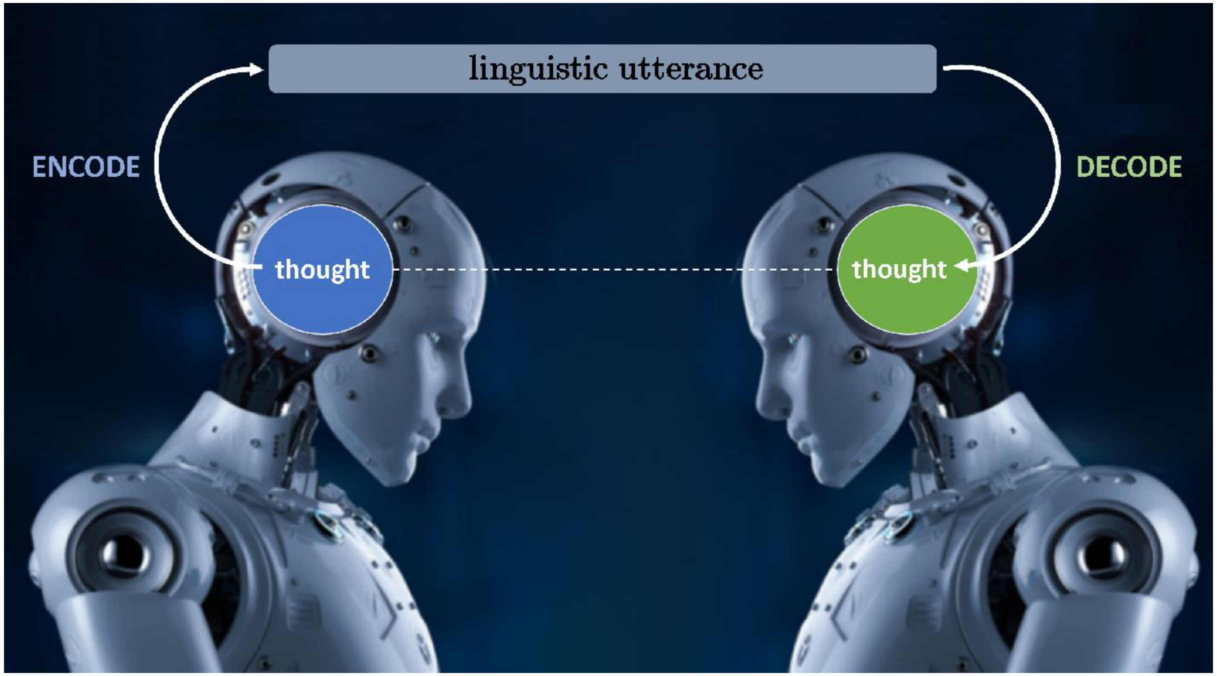
- Zhai: Text Mining and Analytics (Illinois)
- Chris Manning: Natural Language Processing with Deep Learning (Stanford)
- Graham Neubig: Neural Networks for NLP (CMU)

Two views of natural language processing

- Techniques for language **processing**: syntax, grammars, language resources, text representation, speech
- Attempts to **understand** language: semantics and pragmatics of language, related to the goals of artificial intelligence

Understanding

Walid Saba, "Machine Learning Won't Solve Natural Language Understanding", The Gradient, 2021.

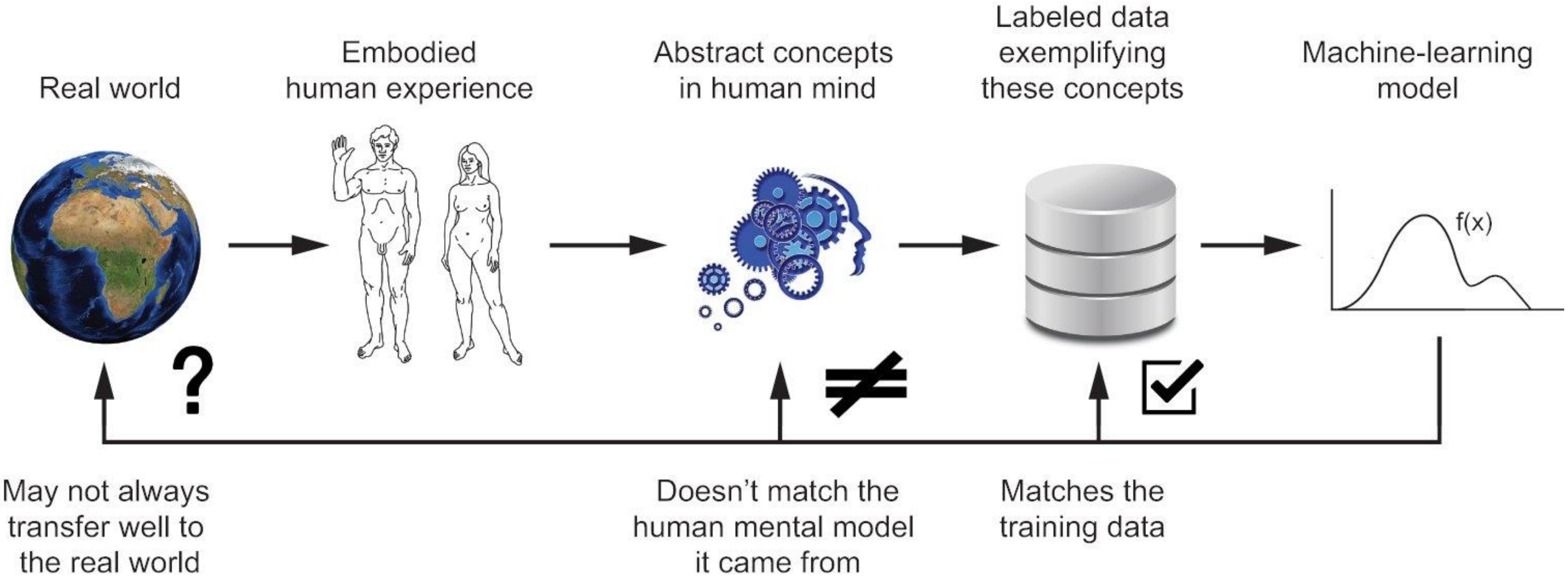


Xanadu, who is a living young human adult, and who was in graduate school, quit graduate school to join a software company that had a need for a new employee.

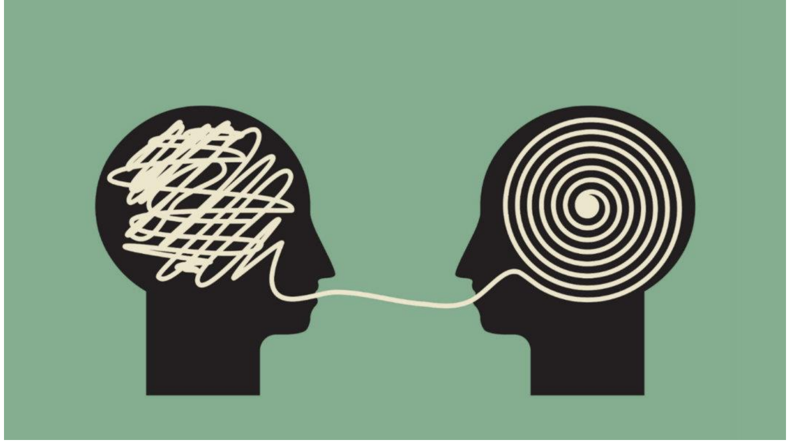
$\approx$

Xanadu quit graduate school to join a software company.

Understanding ML models is difficult


Real world
Embodied human experience
Abstract concepts in human mind
Labeled data exemplifying these concepts
Machine-learning model
f(x)
?
≠
☑
May not always transfer well to the real world
Doesn't match the human mental model it came from
Matches the training data

Understanding language



- A grand challenge of (not only?) artificial intelligence

  Who can understand me?
  Myself I am lost
  Searching but cannot see
  Hoping no matter cost
  Am I free?
  Or universally bossed?

- Not just poetry, what about instructions, user manuals, newspaper articles, seminary works, internet forums, twits, legal documents, i.e. license agreements, etc.

An example: translated rules

Article 18 of UL FRI Study Rules and Regulations

Taking exams at an earlier date may be allowed at the request of the student by the Vice Dean for Education with the lecturer’s consent in justified circumstances (leaving for study or placement abroad, hospitalization at the time of the exam period, participation at a professional or cultural event or a professional sports competition, etc.), and if the applicant's study achievements in previous study years are deemed satisfactory for such an authorization to be appropriate.

Understanding NL by computers

- Understanding words, syntax, semantics, context, writer’s intentions, knowledge, background, assumptions, bias …
- Ambiguity in language
  - Newspaper headlines - intentional ambiguity – attention seeking
    - Juvenile court to try shooting defendant
    - Kids make nutritious snacks
    - Miners refuse to work after death
    - Doctor on Trump‘s health: No heart, cognitive issues

Ambiguity

- I made her duck.

- Possible interpretations:
  - I cooked waterfowl for her.
  - I cooked waterfowl belonging to her.
  - I created the (plaster?) duck she owns.
  - I caused her to quickly lower her head or body.
  - I waved my magic wand and turned her into undifferentiated waterfowl.

- Spoken ambiguity
  - eye, maid

Syntax ambiguity

- Syntactic ambiguity

  Flying planes can be dangerous.
- flying can be interpreted as an adjective modifying planes

  Planes that are flying can be dangerous.
- or as a verb in gerundive form

  It can be dangerous to fly planes.

- Word ambiguity

  The bat flew through the air.
- Unclear reference of a word or phrase

  The boy and the dog were playing in the park. He ran into a tree.
- more examples

  John went to the bank.

Semantic ambiguity

- The girl told the story cried.
- Put the box on the table in the kitchen.
- Bring your old car seat to be recycled.

Disambiguation

- in search queries: jaguar, Paris
- user profiles

- POS tagging,
- word sense disambiguation
- probabilistic parsing
- speech act interpretation, e.g., a statement or a question:
  - We made it. We made it?

Linking

Linking refers to the ability of a reader to connect units of information on the word, sentence, or discourse level. One example called in syntactic theory a “self-embedded structure.” E.g.,

The boy the girl the men left watched then left.

- Which noun phrase (the boy, the girl, the men) is linked with each of the verbs (left, watched, left)
- Valid also for other aspects of texts. For example, narratives can contain stories embedded within stories that are in turn embedded within stories. This can make it difficult for readers to link together units of information so that they can understand the text
- Readability of a text is determine with several linguistic factors (syntactic semantic, morphological, and discourse).

Ambiguity and humor

- collection of linguistic humor by Beatrice Santorini, e.g., recommendation letters
- If you have to write a letter of recommendation for a fired employee, here are a few suggested phrases.

Lexical ambiguity

For a chronically absent employee	A man like him is hard to find.
For a dishonest employee	He's an unbelievable worker.
For a lazy employee	You would indeed be fortunate to get this person to work for you.
For the office drunk	Every hour with him was a happy hour.

Structural ambiguity

For a chronically absent employee	It seemed her career was just taking off.
For a dishonest employee	Her true ability was deceiving.
For a stupid employee	I most enthusiastically recommend this candidate with no qualifications whatsoever.
For the office drunk	He generally found him loaded with work to do.

Scope ambiguity

For an employee who is not worth further consideration as a job candidate	All in all, I cannot say enough good things about this candidate or recommend him too highly.
For an employee who is so unproductive that the job is better left unfilled	I can assure you that no person would be better for the job.

Other

For a lazy employee	He could not care less about the number of hours he has to put in.
For an employee who is not worth further consideration as a job candidate	I would urge you to waste no time in making this candidate an offer of employment.
For a stupid employee	There is nothing you can teach a man like him.

Understanding jokes?

A priest and an Australian shepherd met each other in the final of a quiz show. After answering all the normal questions, they were neck-and-neck with the same number of points and the quizmaster had to set a deciding question. The question was to compose a rhyme in 5 minutes including the word `Timbuktu`. After 5 minutes, the priest presented his poem:

I was a father all my life,
I had no children, had no wife,
I read the Bible through and through,
on my way to Timbuktu.

The audience was thrilled and celebrated the churchman as the certain winner. However, the Australian shepherd stepped forward and recited:

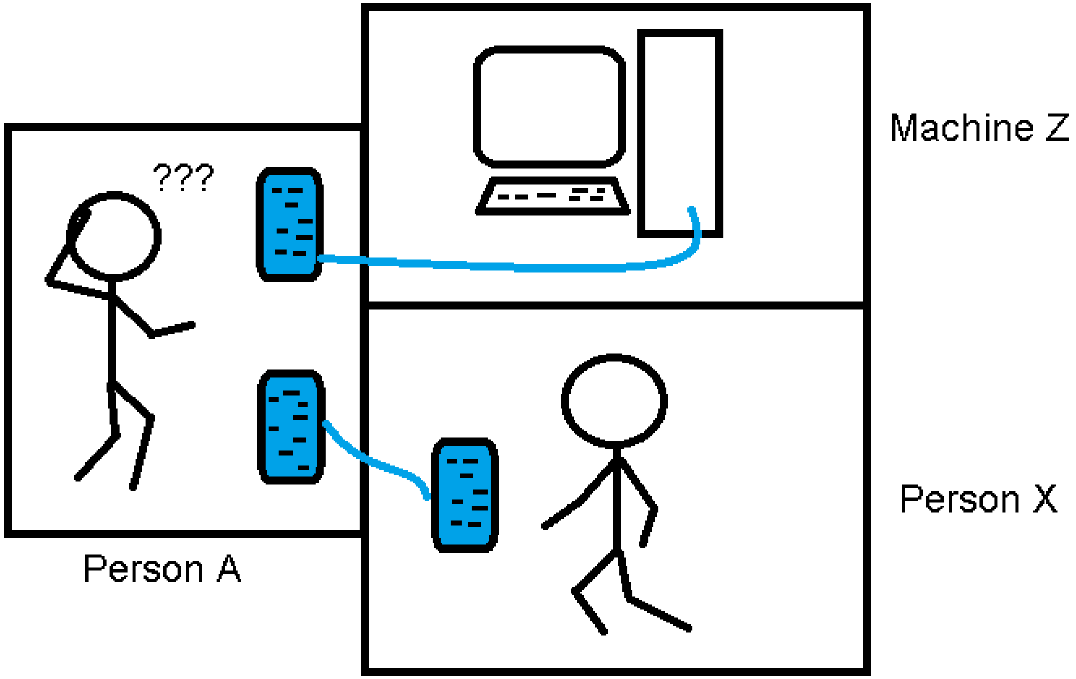
When Tim and I to Brisbane went,
We met three ladies cheap to rent,
They were three and we were two,
So I booked one and Tim booked two.

Understanding language as (artificial) intelligence test

- The system understanding natural language on human level should contain a large portion of human knowledge and would achieve the most important goals of artificial intelligence. Its response would be indistinguishable from human’s.

Turing test

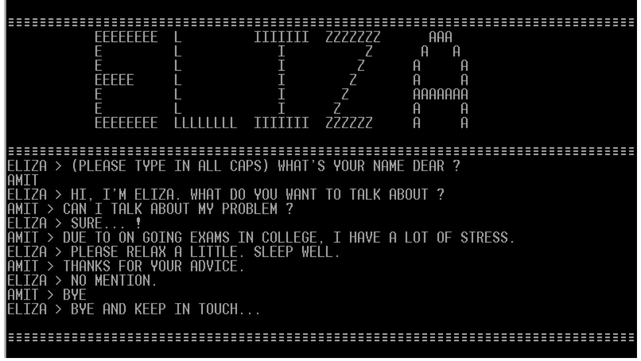
- Turing test,
- Loebner prize
- Chatbots



Chatbots attempt to solve Turing test

- CYC, ELISA, ALLICA, chatterbot,
- VIDA, TIA,
- Siri (knowledge recognition, context detection, web service recommendation)

ELIZA



```
===============================================================================
           EEEEEEEE  L         IIIIIII  ZZZZZZZ      AAA
           E         L            I          Z      A   A
           E         L            I         Z      A     A
           EEEEE     L            I        Z       A     A
           E         L            I       Z        AAAAAAA
           E         L            I      Z         A     A
           EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZ     A     A

===============================================================================
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

===============================================================================
```

- ELIZA (Weizenbaum, 1966)
- limited conversation (Rogerian psychiatrist)

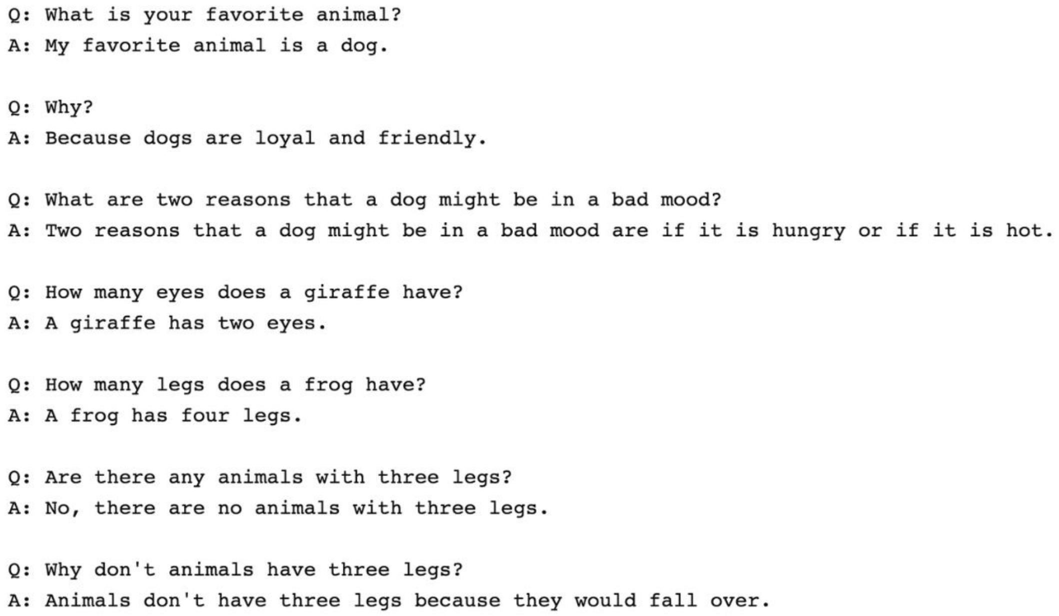
User: You are like my father in some ways.
ELIZA: What resemblance do you see?
User: You are not very aggressive but I think you don't want me to notice that.
ELIZA: What makes you think I am not aggressive?
User: You don't argue with me.
ELIZA: Why do you think I don't argue with you?
User: You are afraid of me.
ELIZA: Does it please you to believe I am afraid of you?

Talking with ALICE Chabot

- judge: Hello, I want to show your abilities to my students
- ALICE: Hi there!. Interesting. I want to find out more about you.
- judge: Let us better talk about the technology driving you
- ALICE: That sounds good to me.
- judge: Do you know anything about it or AI or NLP?
- ALICE: You're so smart, you tell me what it is.
- judge: Let us just do a test
- ALICE: I don't know if it's in my power to give permission.
- judge: Do you use parsing trees at all?
- ALICE: I use it as often as possible.

GPT-3 – a regressive language model as a chatbox

- the example is cherry-picked



```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
```

Knowledge about language – linguistic analysis 1/2

Linguistic analysis contains several tasks: recognition of sounds, letters, word formation, syntactic parsing, recognizing semantic, emotions. Phases:

- Prosody - the patterns of stress and intonation in a language (rhythm and intonation)
- Phonology - systems of sounds and relationships among the speech sounds that constitute the fundamental components of a language
- Morphology - the admissible arrangement of sounds in words; how to form words, prefixes and suffixes …
- Syntax - the arrangement of words and phrases to create well-formed sentences in a language

Knowledge about language – Linguistic analysis 2/2

- Semantics - the meaning of a word, phrase, sentence, or text
- Pragmatics - language in use and the contexts in which it is used, including such matters as deixis (words whose meaning changes with context, e.g., I he, here, there, soon), taking turns in conversation, text organization, presupposition, and implicature

  Can you pass me the salt? Yes, I can.
- Knowing the world: knowledge of physical world, humans, society, intentions in communications …

Limits of linguistic analysis

- levels are dependent
- intonation affects semantics
- sarcasm

Historically: two approaches

- symbolical
  - based on manually injected knowledge
  - grammars, frames, parse trees, etc.
  - top-down approach using grammatical patterns and semantics
  - ‘Good Old-Fashioned AI

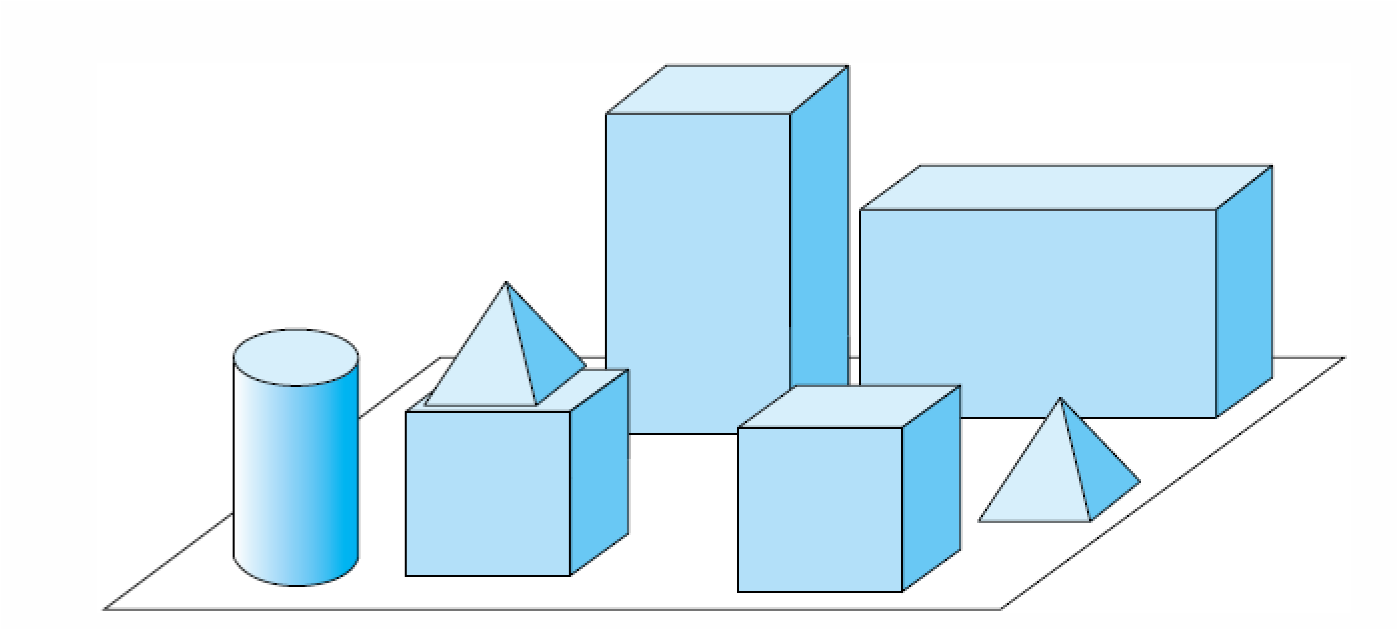
- statistical
  - knowledge is extracted from large corpora
  - bottom-up from texts, learning patterns and links, probabilistic reasoning (possibly syntactically or semantically wrong)
  - large pretrained language models: BERT, GPT-3

- Merging both worlds: injecting (symbolical) knowledge into DNNs

How it all started?

- micro worlds
- example: SHRDLU, world of simple geometric objects
  - What is sitting on the red block?
  - What shape is the blue block on the table?
  - Place the green pyramid on the red brick.
  - Is there a red block? Pick it up.
  - What color is the block on the blue brick? Shape?

Micro world: block world, SHRDLU (Winograd, 1972)



Classical approach to text understanding

- text preprocessing
- 1. phase: syntactic analysis
- 2. phase: semantic interpretation
- 3. phase: use of world knowledge

- Hmm, what is text understanding, actually?

Basic text preprocessing – the classical pipeline

- document → paragraphs → sentences → words
- words and sentences ← POS tagging
- sentences ← syntactical and grammatical analysis

- partially still used in neural processing

1. phase of text understanding: Syntactic analysis

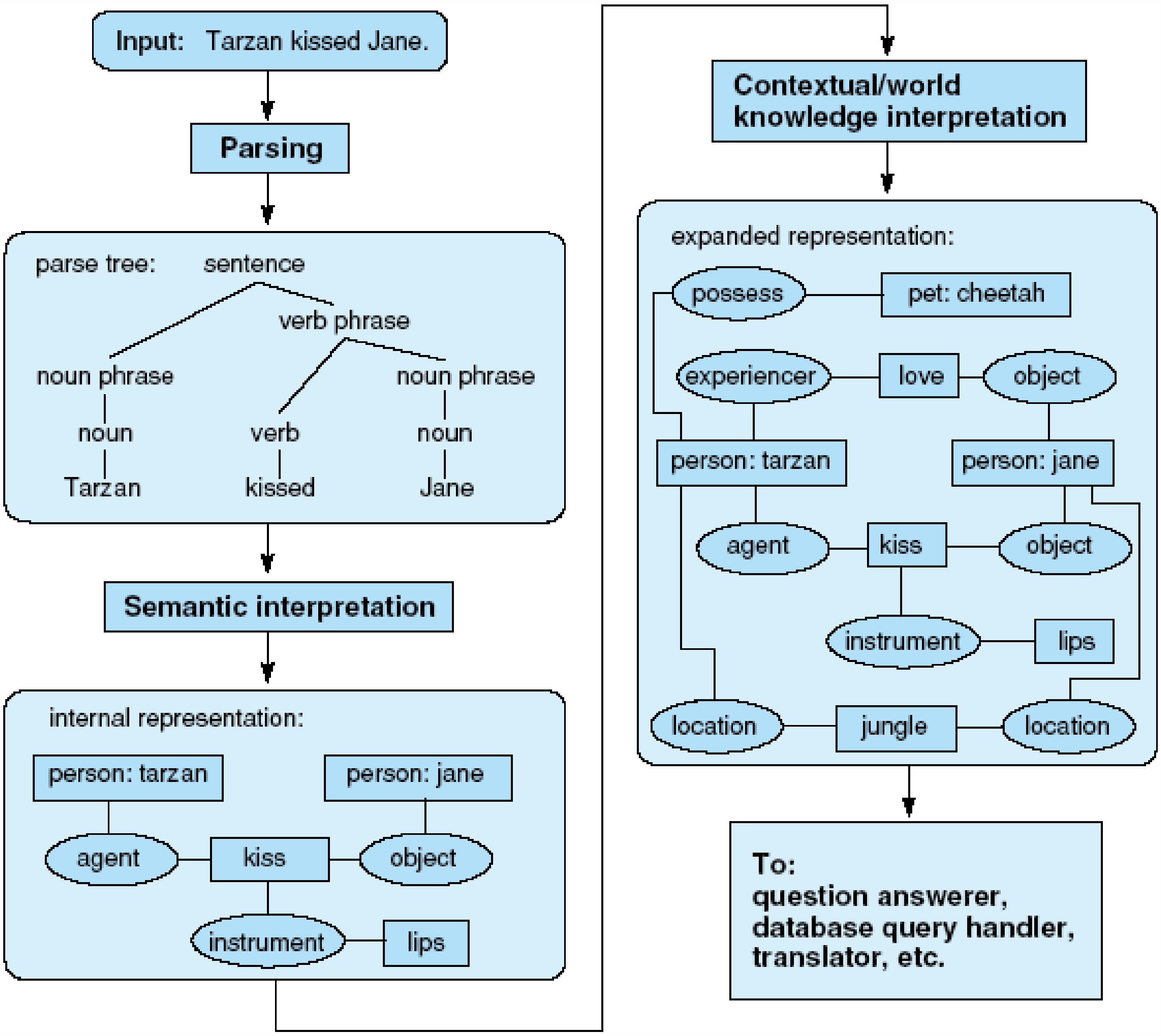
- Find syntactical structure
- part-of-speech (POS) tagging (noun, verb, preposition, …)
- The role in the sentence (subject, object, predicate)
- The result is mostly presented in a form of a parse tree.
- Needed: syntax, morphology, and some semantics.

2. phase: Interpretation

- Knowledge of word meaning and their language use
- Result: conceptual graphs, frames, logical program
- Check semantics

3. phase of text understanding: Use of world knowledge

- Extend with background knowledge
- Consider the purpose of the system: summarization, database interface …
- E.g., Cyc and openCyc knowledge bases present ontology and knowledge base of everyday common-sense knowledge, e.g.,
  "Every tree is a plant" and "Plants die eventually”
- process incrementally, adding the meaning of previous sentences


Input: Tarzan kissed Jane.
Parsing
parse tree:
sentence
verb phrase
noun phrase
noun phrase
noun
verb
noun
Tarzan
kissed
Jane
Semantic interpretation
internal representation:
person: tarzan
person: jane
agent
kiss
object
instrument
lips
Contextual/world knowledge interpretation
expanded representation:
possess
pet: cheetah
experiencer
love
object
person: tarzan
person: jane
agent
kiss
object
instrument
lips
location
jungle
location
To:
question answerer,
database query handler,
translator, etc.

Where is NLP today?

- active research area with many commercial applications
  - speech recognition and synthesis
  - automatic reply engines
  - machine translation
  - text summarization
  - question answering
  - language generation
  - interface to databases
  - intelligent search and information extraction
  - sentiment detection
  - semantic analysis: e.g., role labelling,
  - named entity recognition and linking
  - categorization, classification documents, messages, tweets, etc.
  - many (open-source) tools and language resource
  - prevalence of deep neural network approaches
  - cross-lingual approaches

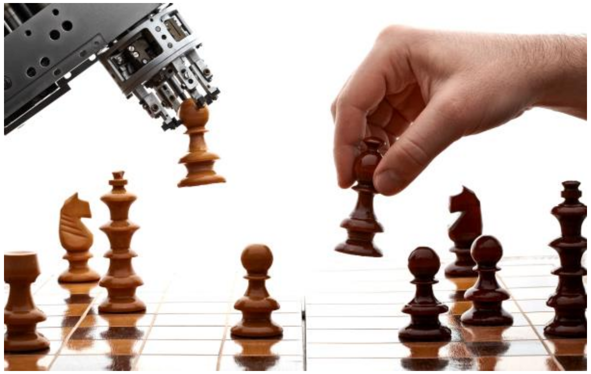
NLP resources and technologies

- language technologies
  - prevalence of deep neural network approaches
  - text embeddings, cross-lingual approaches
  - named entity recognition and linking
  - categorization, classification of documents, messages, twits, etc.
  - summarization, question answering, machine translation
  - speech recognition and generation
  - text generation
  - many (open-source) tools and language resource
- language resources
  - importance of large text corpora: monolingual, parallel
  - knowledge graphs
  - dictionaries and thesauri
  - many datasets for ML tasks: QA, NLI, paraphrasing, coreference resolution,WSD, sentiment, offensive speech, etc.

NLP success stories

- Jeopardy, 2011: IBM Watson wins in a quiz against two human champions
- useful tools like Google Translate, Siri, Cortana, Alexa
- search engines
- information extraction and retrieval
- (superhuman) speech recognition
- text classification
- useful summarization and question answering

Language understanding



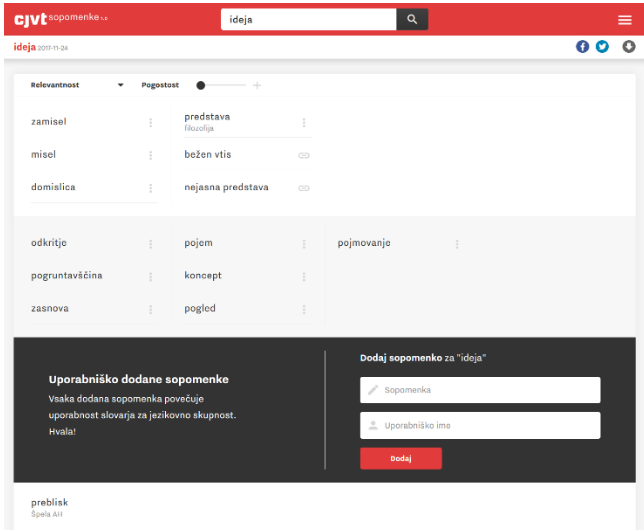
- Can current approaches pass Turing test?
- Can a system understand a language?

- Problems of current approaches
  - we solve partial problems
  - because actual problems are (still) too difficult

- We don‘t understand what is understanding.
- We don‘t have good enough models for knowledge representation.
- Injecting knowledge into deep neural networks: factual, linguistic, common-sense, domain specific

CJVT UL: Center for language resources and technologies of University of Ljubljana (Center za jezikovne vire in tehnologije Univerze v Ljubljani)

- many practical open-source technologies and solution using NLP and ML for Slovene
- corpora and datasets
- thesaurus, dictionary of collocations
- lexicon of wordforms, lexical database
- sentiment lexicon
- machine translation
- speech recognition
- neural POS taggers
- models for comma placement, stress,
- summarizer
- cross-lingual models
- embeddings
- etc.
- www.cjvt.si
- www.slovenscina.eu



zamisel
misel
domislica
predstava
filozofija
bežen vtis
nejasna predstava
odkritje
pogruntavščina
zasnova
pojem
koncept
pogled
pojmovanje

Uporabniško dodane sopomenke
Vsaka dodana sopomenka povečuje uporabnost slovarja za jezikovno skupnost. Hvala!

preblisk
Špela AH