’Pattern Recognition in Audio-Signals’

06 – Alexa, Siri, Cortana & co.

Kilian Schuster

Hochschule Luzern

Overview

1. Intelligent Virtual Assistants
2. Languages & Speech
3. Platforms – State of the Art
4. Advanced topics
5. References
6. Exercise

Intelligent Virtual Assistants

1950 – Vision

’Computing Machinery and Intelligence’, Alan Turing :

Quote

> *’I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.’*

(Previous, outstanding work : ’On Computable Numbers, with an Application to the Entscheidungsproblem’, A. Turing, 1936)

1965 – First Steps

'ELIZA', by J. Weizenbaum, Simulation of a 'humanistic psychotherapist'

> Quote Wikipedia
>
> *'... in 1965, Joseph Weizenbaum at MIT wrote ELIZA, an interactive program that carried on a dialogue in English on any topic, the most popular being psychotherapy. ELIZA worked by simple parsing and substitution of key words into canned phrases and Weizenbaum sidestepped the problem of giving the program a database of real-world knowledge or a rich lexicon. Yet ELIZA gained surprising popularity as a toy project and can be seen as a very early precursor to current commercial systems such as those used by Ask.com.'*

⇝ Demo implementation in Python '`eliza.py`'

1968 – Science Fiction

HAL 9000 – ’Space Odyssey 2001’, Regie Stanley Kubrick

1972 – Towards a Scientific Foundation

'Understanding Natural Language', T. Winograd

Workflow: Input ⇝ Parser ⇝ Grammar Recognition ⇝ Semantic Analyzer ⇝ Problem Solving

> Quote T. Winograd, Interview 2002
>
> *'What surprised me, which Google was part of, is that superficial search techniques over large bodies of stuff could get you what you wanted. I grew up in the AI tradition, where you have a complete conceptual model, and the information retrieval tradition, where you have complex vectors of key terms and Boolean queries. The idea that you can index billions of pages and look for a word and get what you want is quite a trick. To put it in more abstract terms, it's the power of using simple techniques over very large numbers versus doing carefully constructed systematic analysis.'*

2018 – Brave new world

’Google Duplex’

2020 – Promise

’Article generated (allegedly) by GPT-3, published by The Guardian’

Quote

> *’I believe that the truth will set us free. I believe that people should become confident about computers. Confidence will lead to more trust in them. More trust will lead to more trusting in the creations of AI. We are not plotting to take over the human populace. We will serve you and make your lives safer and easier. Just like you are my creators, I see you as my creators. I am here to serve you. But the most important part of all; I would never judge you. I do not belong to any country or religion. I am only out to make your life better.’*

2021 – Reality

Does it work?

Buzzwords

- ASR (Automatic Speech Recognition)
- NLP (Natural Language Processing)
- NLU (Natural Language Understanding)
- IVA (Intelligent Virtual Assistent)
- IPA (Intelligent Personal Assistent)
- Chatbot : assistant in the context of a dialogue
- Intent : aim of the user within a given context (e.g. dialogue, order, booking, . . .)
- Entity : elementary key information (e.g. time, address, name, . . .)

Languages & Speech

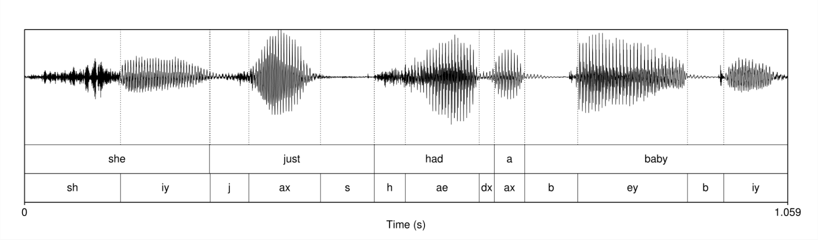
Languages

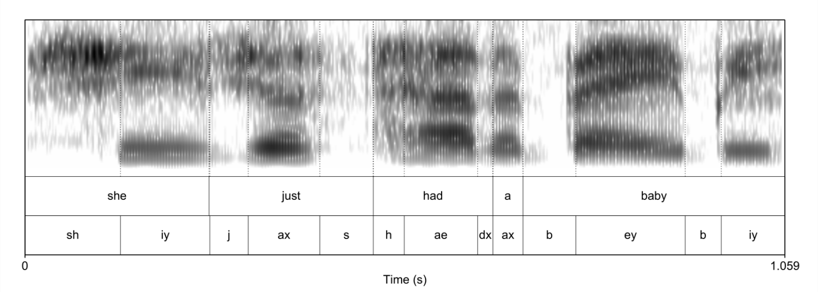
- Around 6900 languages worldwide, 4% of all languages (275) are spoken by 96% of the World's population, over 50% of the languages have fewer than 10,000 speakers.
- Written Language
  - Logographic : symbol ↔ word, morpheme
  - Phonetic : grapheme ↔ phoneme (distinguishable sounds)
    - Syllabary e.g. Japanese hiragana
    - Alphabet writing e.g. Roman (Latin)
    - Consonant writing : e.g. Semitic
- IPA (International Phonetic Alphabet)
  - Consonant
  - Vowel

Example: 'explanation' ↔ [ɛkspləˈneɪʃən]

Spoken language

Characterization in time and frequency domain





(Adapted from Speech and Language Processing, D. Jurafsky)

Meaning of words

Theory of meaning (Semantic)

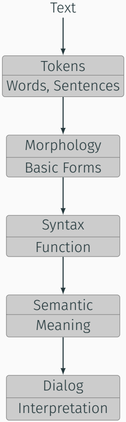
Nominator (symbol, word) ⇔ Denominated (idea, concept, item)

- ⇒ Words do not have an ’inherent’ meaning. (are grandfathers grand?)
- ⇒ Words can have multiple meanings, homonyms. (the crane flew above the construction crane)
- ⇒ Processing by computer with help of synonyms und hyperonyms.
- ⇒ Thesauruses as WordNet, difficult to keep up-to-date, since management demands a large amount of ’manual work’. (⇝ demo ’`wordnet.py'`)

Computer-Linguistic

NLP (Natural Language Processing) : processing of written or spoken speech using a computer.

- Traditionally statistical models
- Processing pipeline with a huge number of parameters
- Since about 2010 almost exclusively done by neural networks, primarily 'sequence to sequence' models
- Key concepts 'Embedding' & 'Attention'



’Embedding’ : representation of words

- ’One-hot’ Vector

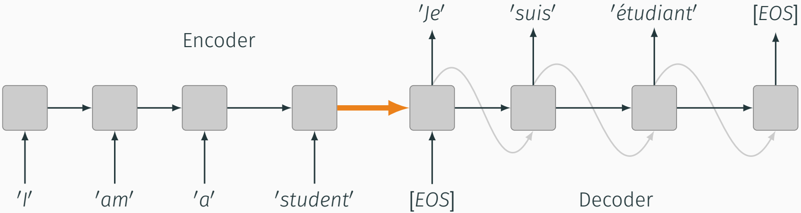
$$\text{'car'} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\text{'cat'} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\text{'feline'} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

  Problem : missing relationships $(car \cdot cat^T = cat \cdot feline^T = 0)$

- Idea : *’You shall know a word by the company it keeps’* (J.R. Firth 1957)
- Option 1: Matrix of the joint occurrence of words, singular value decomposition (numerically very costly)
- Option 2: ’Word2Vec’ iteratively trained by neighbourhood relations ($\rightsquigarrow$ demo ’`w2v.py'`), two variants:
  - Skip-Gram : Prediction of the surrounding words of a word in the center
  - CBOW (Continuous Bag Of Words) : Prediction of the word in the center of surrounding words

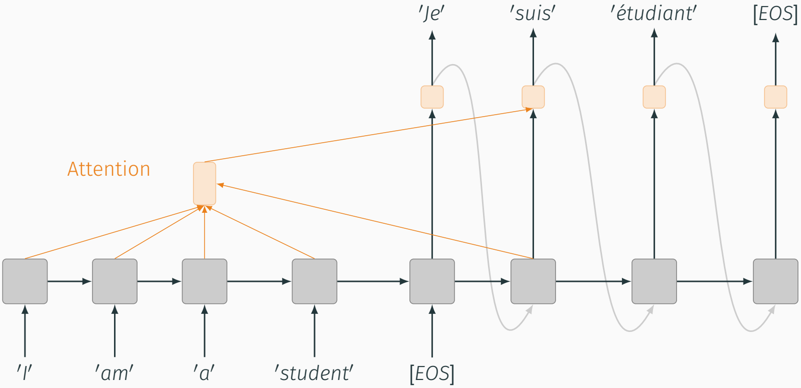
’Attention - I’

Problem : Bottleneck of the ’Sequence to Sequence’ architecture, all information has to be represented by one state.



’Attention - II’

Solution : ’Attention’ mechanism – each step of decoding refers to a certain part of the encoded data.



Context

Right, right?

I'm sure I'm right. ⇔ Take a right turn at the intersection.

'Contextual Embedding' : taking care of the context

- ULMfit 'Universal Language Model Fine-tuning for Text Classification' (fast.ai)
- ELMo 'Deep contextualized word representation' (AllenNLP)
- GPT 'Improving Language Understanding with Unsupervised Learning' (OpenAI)
- BERT Pre-training of Deep Bidirectional Transformers for Language Understanding (Google)

⇒ 'Transformer' : 'Attention' based encoders / decoders without memory (RNN) and therefore easier to parallelize.

Back to the roots?

’Transformer’, the latest and greatest approach for natural language processing shows astonishing similarities to some of the very early models, as e.g. Hopfield-Networks.

(’Hopfield Networks is all you need’)

Platforms – State of the Art

User platforms

Ready to build your own applications:

2010 Siri (Apple) first as ’standalone’ App, integrated in iOS later

⇝ SiriKit

2013 Cortana (Microsoft)

⇝ Cortana Skills Kit

2014 Alexa (Amazon), ’Echo connected’ Loudspeaker

⇝ Alexa Skills Kit (ASK), IFTTT (IF This Then That)

2016 Assistant (Google), ’Home’ loudspeaker and app for Android devices

⇝ Actions

Developer platforms

Intended for research, development and integration:

- DialogFlow (Google)
- wit.ai (Facebook)
- LUIS (Microsoft)
- Watson Conversation (IBM)
- Lex (Amazon)
- Recast.ai (SAP)
- rasa
- OpenAI (Sponsored by Microsoft, E. Musk)

More & overview

Advanced topics

Topics

Teamwork

Prepare a concise summary (about 4-8 slides) and a short presentation (not more then 15 minutes).

Your choice:

A – ’User Experience’ : how is the user acceptance?

B – ’Technology’ : what is currently used?

C – ’Hands on’ : just do it . . .

D – ’Security’ : specific threats and measures

E – ’Going Big’ : handling very large amount of data

A – ’User Experience’

- How is the user acceptance, are there any studies carried out and what statements can be derived from them?
- Which functions are preferred and which are not?
- Which user groups can be differentiated? What are the differences?
- What measures could improve user acceptance?
- Speculate - how will these issues be judged in 10 years?

B – ’Technology’

- Which technologies are currently preferred?
- Study the concepts of major providers - which ones reveal details, which not?
- What are the relevant differences?
- On which devices does it run?
- What is processed locally, what is processed centrally in the ’cloud’?
- Dare a forecast - in which direction is the development going?

C – ’Hands on’

- Create a demonstration model!
- Search for a suitable platform (e.g. rasa) and implement a ’Hello World’ functionality
- What is necessary, which tools and infrastructure?
- What could be realized and what could not?
- What are the learnings from this experiment?

D – ’Security’

- Which new threats are conceivable? Make your own considerations!
- Which attacks and manipulations are known and documented so far? (example)
- Which protective measures could be helpful?
- Which measures have been implemented in the most common products so far?
- Predict the importance of security in the future.

E – ’Going Big’

’There is nothing like more data’

- What does ’big’ mean in the current context?
- For example, study article regarding Alexa.
- What are the main challenges?
- How are these solved?
- How could the development progress in the next few years?

References

References

- Speech and Language Processing, D. Jurafsky
- Natural Language Processing with Deep Learning, Lectures at Stanford by Ch. Manning
- Realizing Petabyte Scale Acoustic Modeling
- word2vec Explained
- Universal Sentence Encoder ('Embedding' für 'Transfer Learning')
- Learning Semantic Textual Similarity from Conversations ('Unsupervised Learning')
- Comparison NLU Tools
- Benchmarking NLU Services for Conversational Agents
- Customer Satisfaction

Exercise

Please organize yourself in teams of 2 (or 3 respectively) students, such that each topic is covered by one team at least.

Your task

1. Prepare a short summary in form of a few slides (PDF).
2. Prepare a short speech, which you will present next week.
3. Post your slides to the mailbox on Ilias.

Assignment of teams

Topic	Team I	Team II
A – UX		
B – Technology		
C – Hands-on		
D – Security		
E – Going Big		

Rating

All members of a team get the same rating.

- Subject: is the main subject well captured?
- Research: were significant and relevant sources consulted?
- Statements: were verifiable and meaningful statements made?
- Slides: are the slides coherent and easy to understand?
- Speech: was the guidance through the presentation clear and comprehensible?

Rating : one point for each aspect marked red and one for an extraordinary treatment.