Aalto University

Modeling conversational language: why it is so hard?

Mikko Kurimo
Department of Signal Processing and Acoustics
Aalto University

Aalto University

Mikko Kurimo

1989-1997 MSc and PhD at Kohonen's CIS lab: **speech recognition with neural networks**

1998-2012 Visiting research fellow in top speech & language labs:

– Research: IDIAP (CH), SRI (USA), ICSI (USA)

– University: Edinburgh, Cambridge, Colorado, Nagoya

2012 Professor in **speech and language processing**

- Teaching **speech recognition and natural language processing**

- Head of Aalto **speech recognition group**

Research topics: speech recognition, adaptation, assessment, diarization, language modeling, audio and video description

Conversational agents have appeared in our phones and homes

Typing-based agents are starting to speak and listen in cars, robots, toys, phones, smart speakers and other devices

Google ASSISTANT

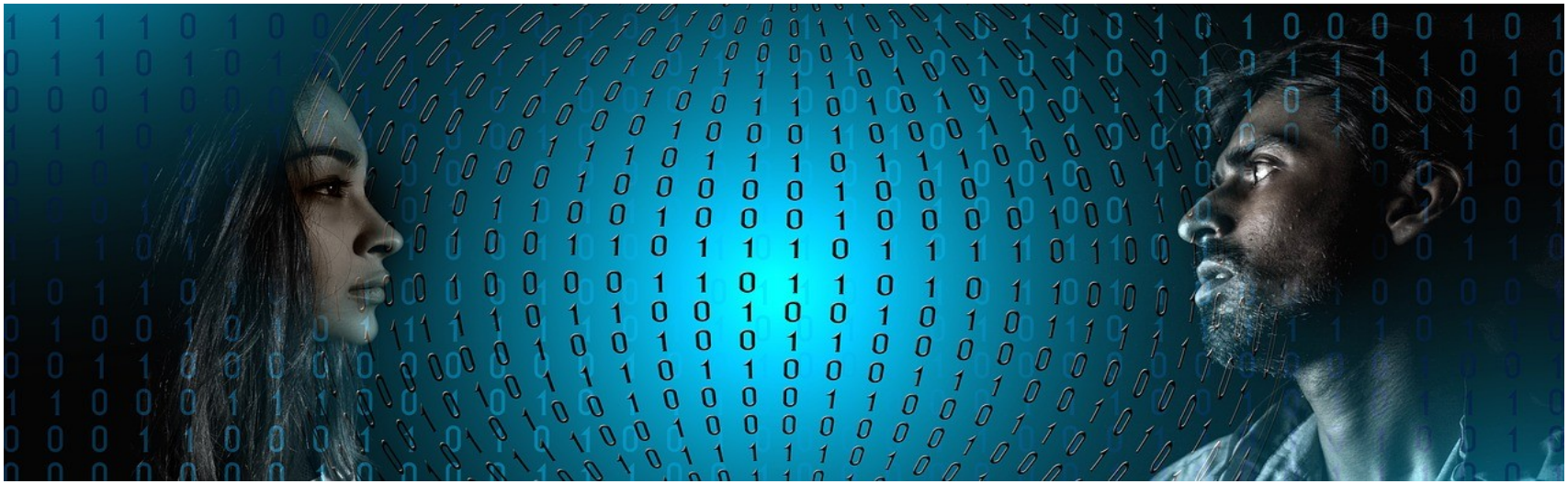
Spoken interfaces for everyday tasks

dictation, captioning, translation, interpretation,
information retrieval, conversational assistants, language learning

skype translator

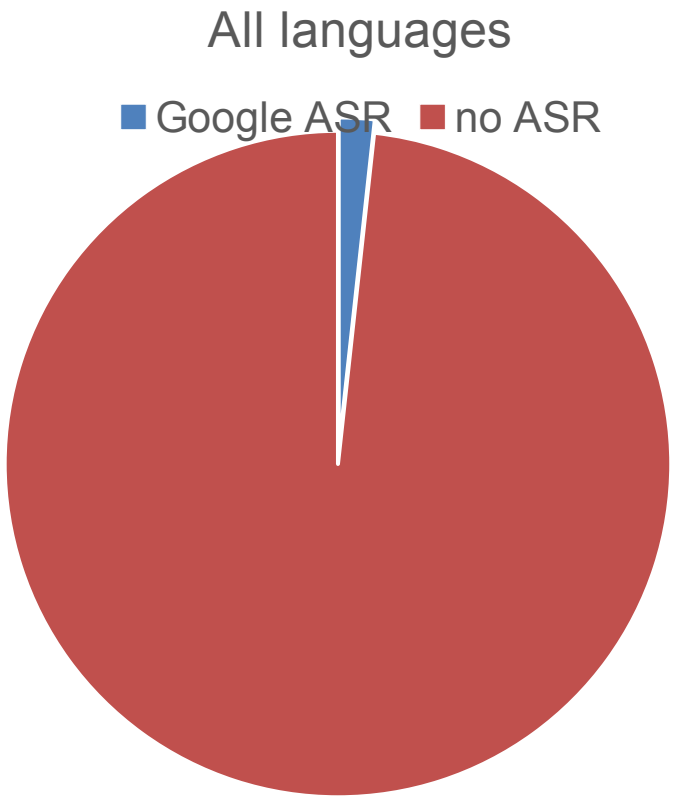
Language is human communication

- Rich communication signal **between humans**
- Human speech is the most complex of all biosignals
- speech => text + emotion,loudness,speed,emphasis,...
- text + *emotion,loudness,speed,emphasis,...* => speech
- How much language “understanding” is needed?
- People perceive the use of language as a sign of “intelligence”



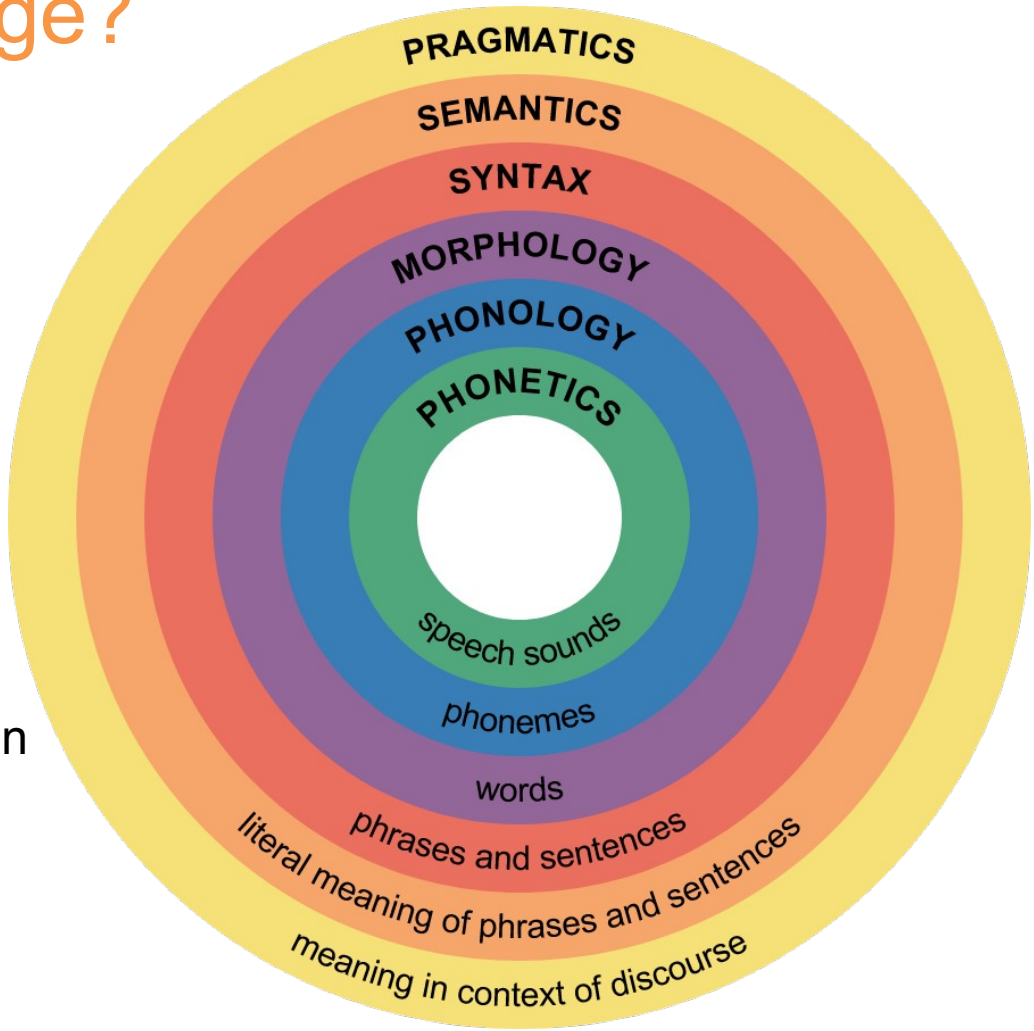
Complexity of natural languages

- 6000+ languages, many dialects
- Each has many words
- Each word is understood slightly differently by each speaker
- Large variety of sentence structures

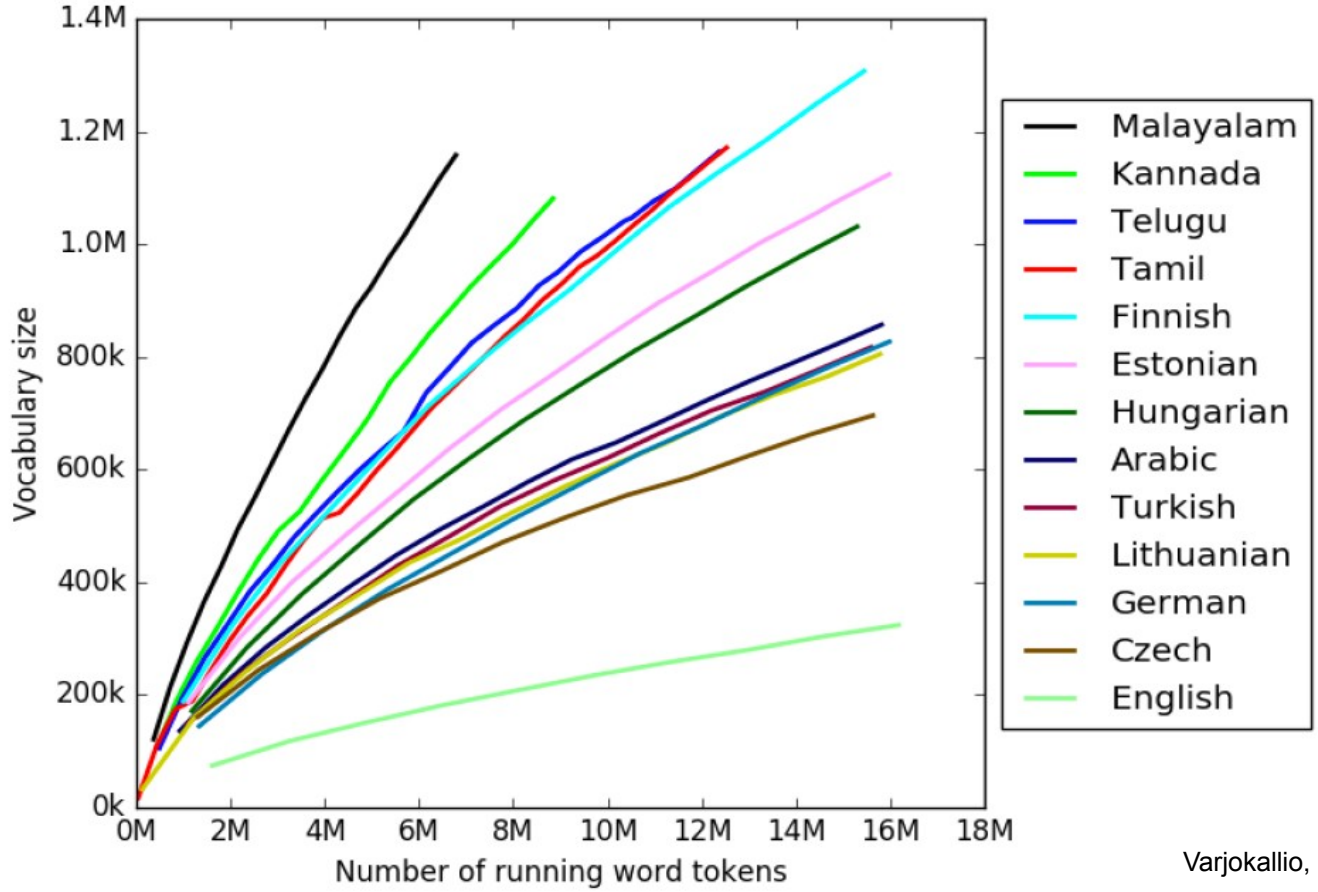


What is in a language?

- Phonetics and phonology:
  - the physical sounds
  - the patterns of sounds
- Morphology: The different building blocks of words
- Syntax: The grammatical structure
- Semantics: The meaning of words
- Pragmatics, discourse, spoken interaction...



Effect of morphology: vocabulary size as a function of corpus size



Varjokallio, Kurimo, Virpioja (2016)

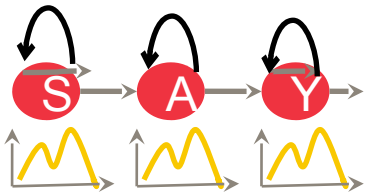
Challenges of natural language

- Understanding the **meaning of words** is subjective:
  - learning language through individual life paths
  - end up having different ways of understanding and producing language
- Many words have several meanings:
  - E.g. “play”, “game”, “window”
- Sentences have several interpretations:
  - E.g. “Big children and adults saw a man with a telescope”

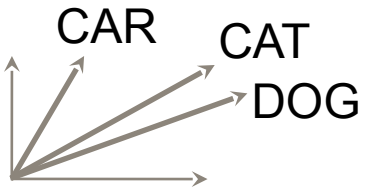
Natural language processing

METHODS

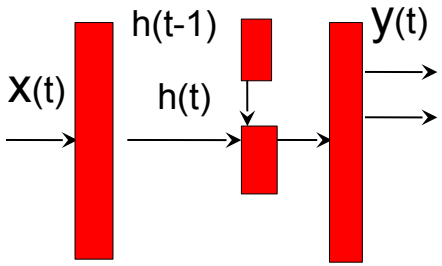
Hidden Markov model



Vector space model



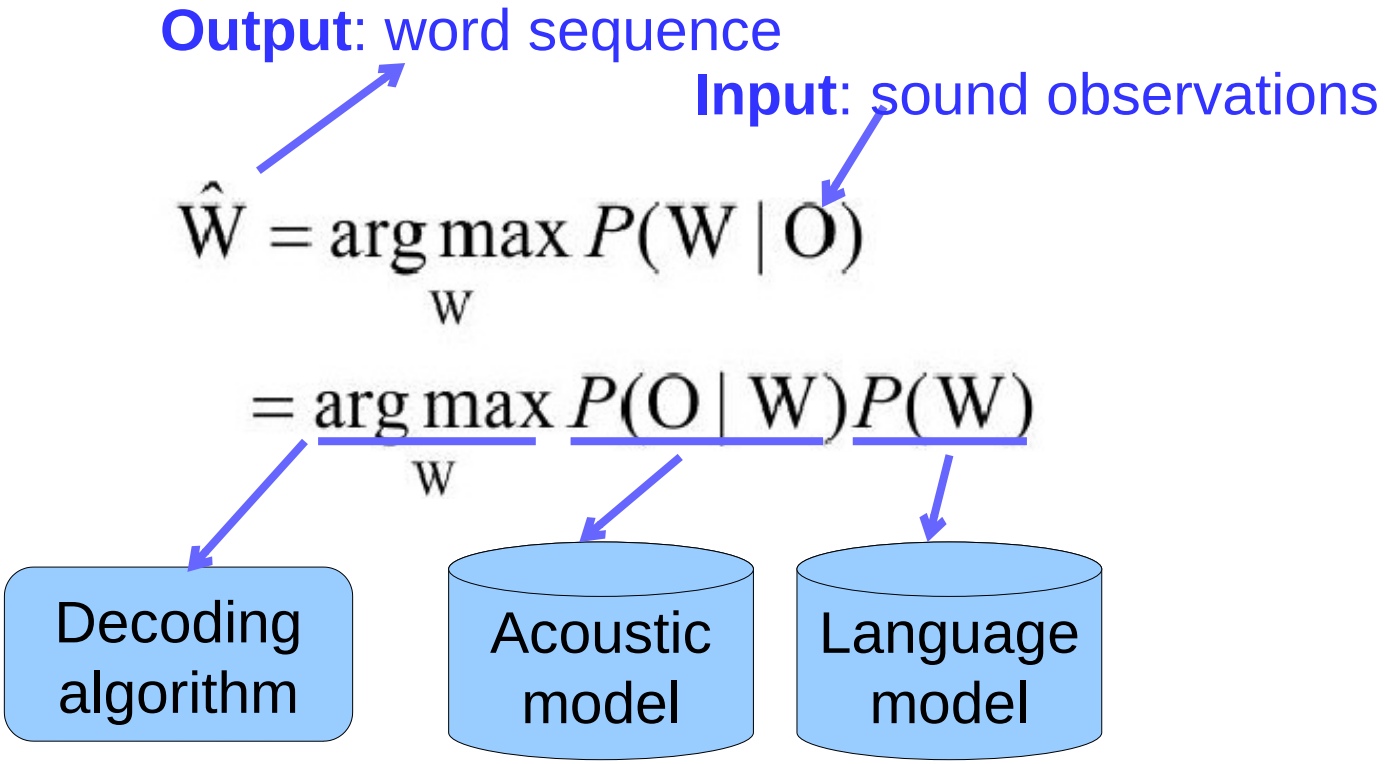
Recurrent neural network



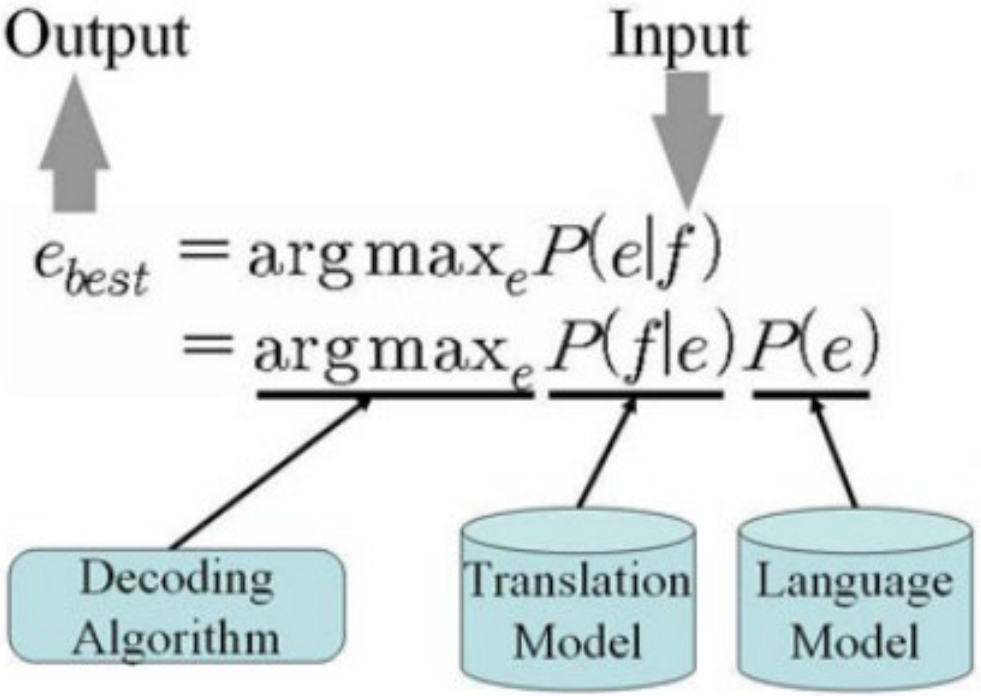
TOOLS

- Speech-to-text
- Text-to-speech
- Machine translation
- Information retrieval
- Named entity recognition
- Sentence parsing
- Topic detection

Speech recognition: large probabilistic models



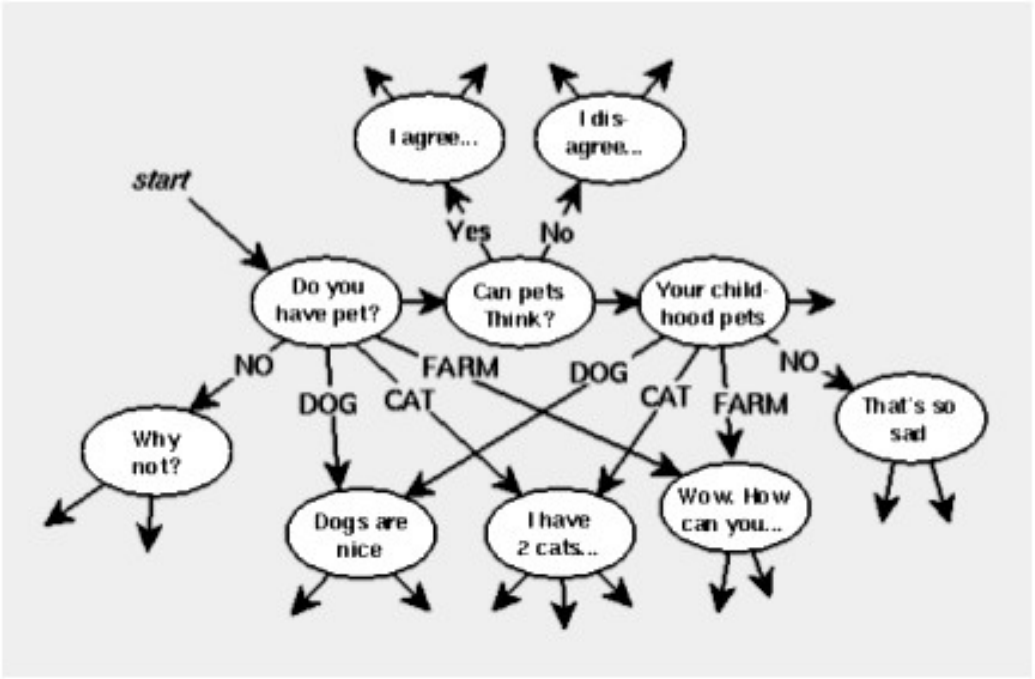
Machine translation: large probabilistic models



isoft.postech.ac.kr

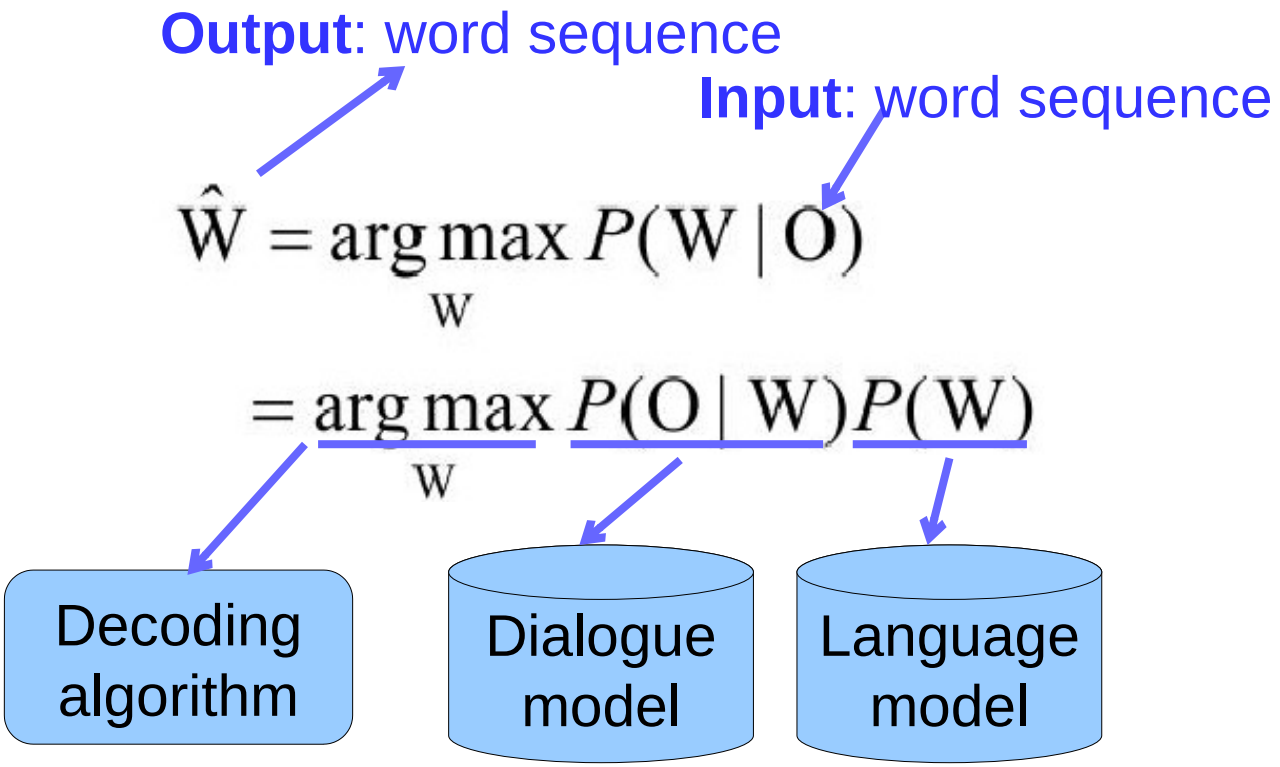
Natural language interfaces: the traditional way

www.zabaware.com



robot-club.com

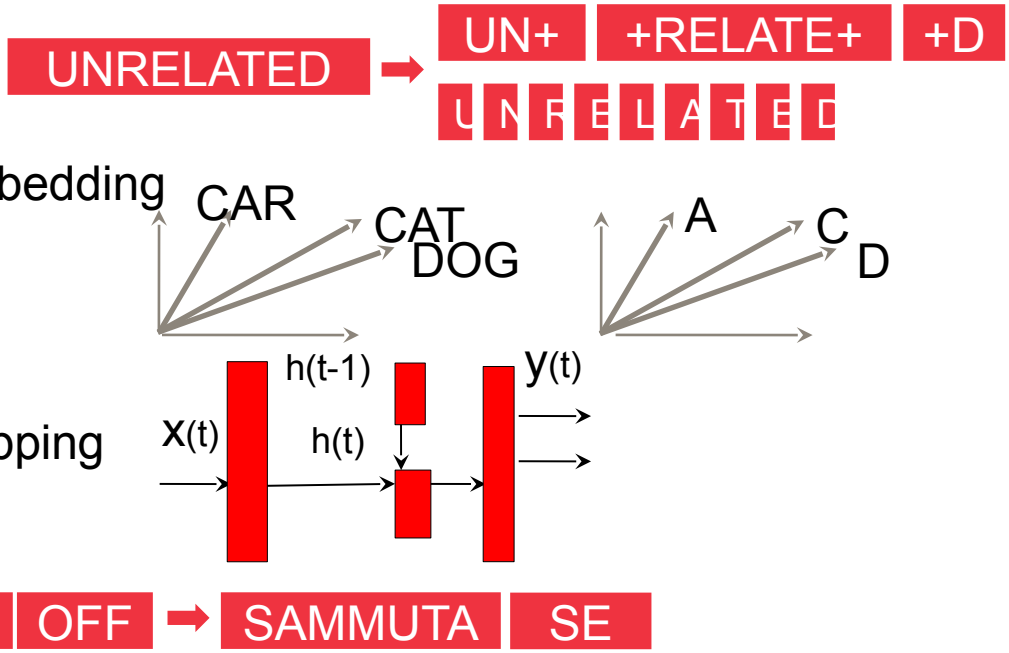
Dialogue generation: a large probabilistic models point of view



A recent revolution in the language modeling approach

- Split language into tokens
- Vector space modeling, embedding
- Representation learning
- Deep & recurrent learning
- Sequence to sequence mapping

=> artificial intelligence



Training problems - and solutions!

- Takes huge amounts of data
- Data should be on target domain, e.g. chats on relevant topics
- Takes huge amounts of computation time and electricity
- May take a long time to converge
- Automatic evaluation metrics

- General model and task model separated
- Pre-trained model of Finnish (e.g. BERT, wav2vec2) trained on general targets (e.g. self-prediction)
- Fine-tuned for specific tasks using annotated target data (e.g. chats)
- Pre-trained models can even be multilingual (monolingual is better)

Curriculum learning (easy sentences first)

Data augmentation by perturbation, synthesis and paraphrasing

Future (unsolved) challenges

Including informal conversations, multimodality, multilinguality, personalization, context sensitivity

- Vad tänker du?
- Jätte kiva show.
- It was amazing.
- Pah, tylsää.

- Who is in the white house now?
- Close the window and return.

Solutions for those challenges are studied in my research group

- Contact: *mikko.kurimo@aalto.fi*
- Publications: http://research.aalto.fi
- Home page: (search: *”Aalto asr home”)*
- Software: (search: *”Aalto asr github”)*
- Demos: (search: *”Aalto asr video”)*

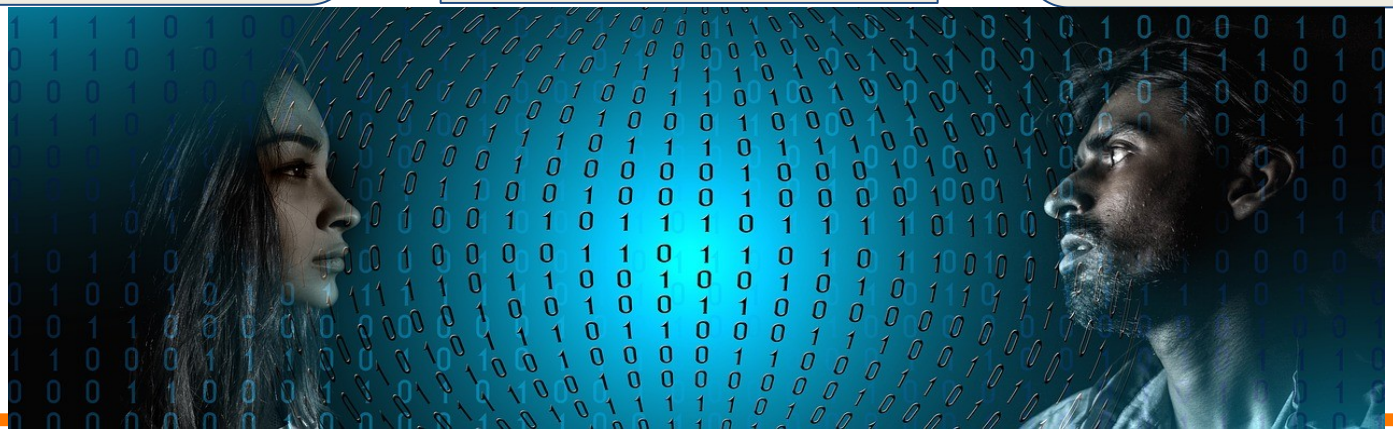
Aalto ASR research group – Our vision

- Studies deep learning methods in **automatic speech recognition** (ASR) and **language modeling** (LM)
- Challenge: Representation and understanding of real-world spoken conversations

Great variation of speakers, styles and languages

Deep learning methods in ASR and LM

Need for understandable and co-operative AI



6 ongoing external funding:

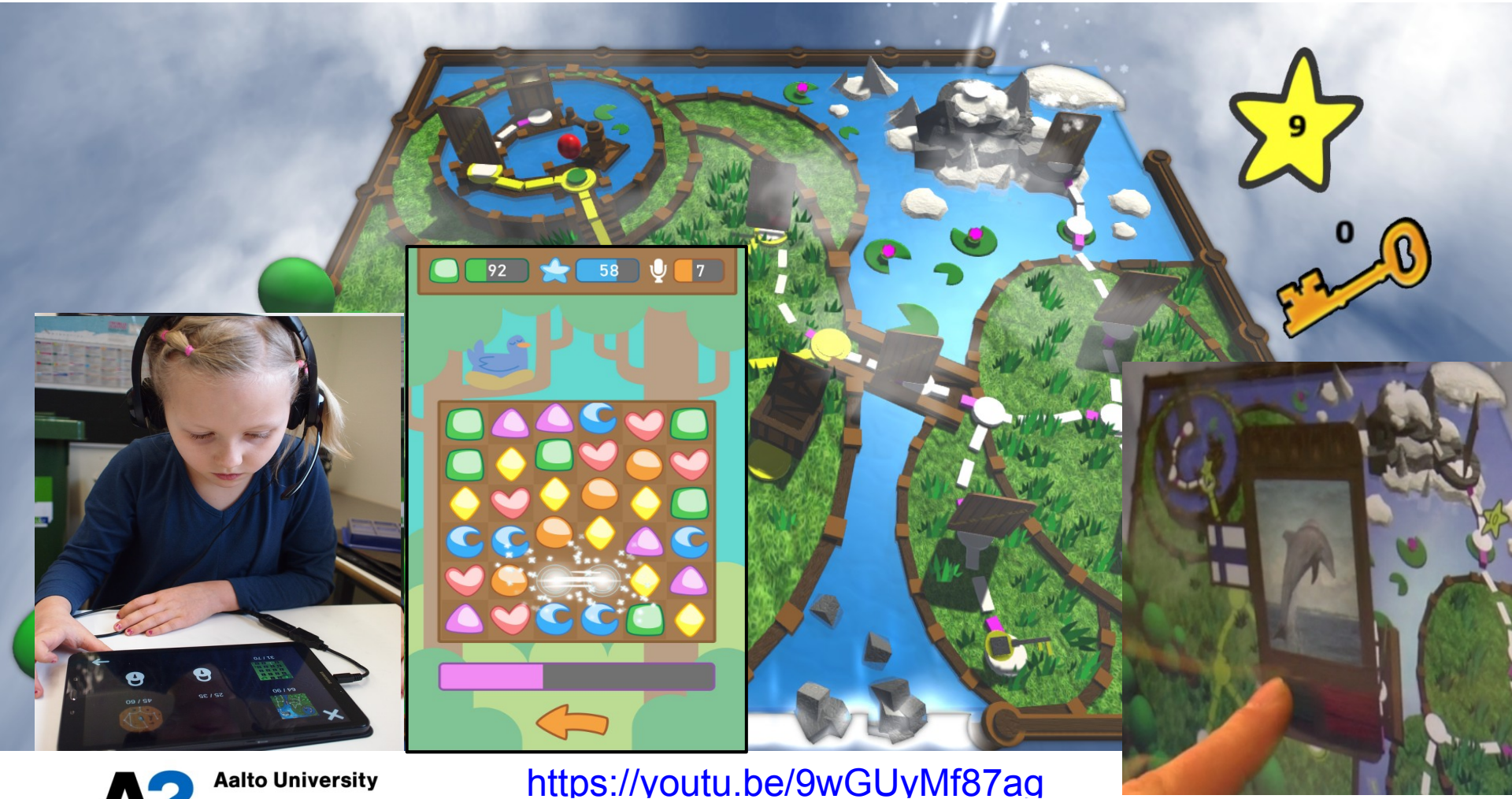
1. Conversational speech (*DonateSpeech/FIN-CLARIN*)
2. Using audiovisual data
   - Analyse old Finnish movies (*MoMaF/AKA*)
   - Combine speech and video recognition (*USSEE/AKA*)
3. Tools for L2 learning
   - Pronunciation games for children (*TEFLON/Nordforsk*)
   - Evaluation of speaking skills (*DigiTala/AKA*)
   - Tools for Aalto's language courses (*Kielibuusti/Gov*)

Online subtitles: for those who do not hear
Challenge: speed, slang, readability

https://www.youtube.com/watch?v=0neezwVilPE

Conversational robots, toys, assistants
Challenge: speed, environment, dialog

Games and language learning: “Say it again Kid”
Challenge: speed, children, training data



Aalto University

https://youtu.be/9wGUyMf87ag
https://www.youtube.com/watch?v=el4De9Q_GYA