What are the most sustainable and valuable resources that we should develop for training language models?

Jussi Karlgren

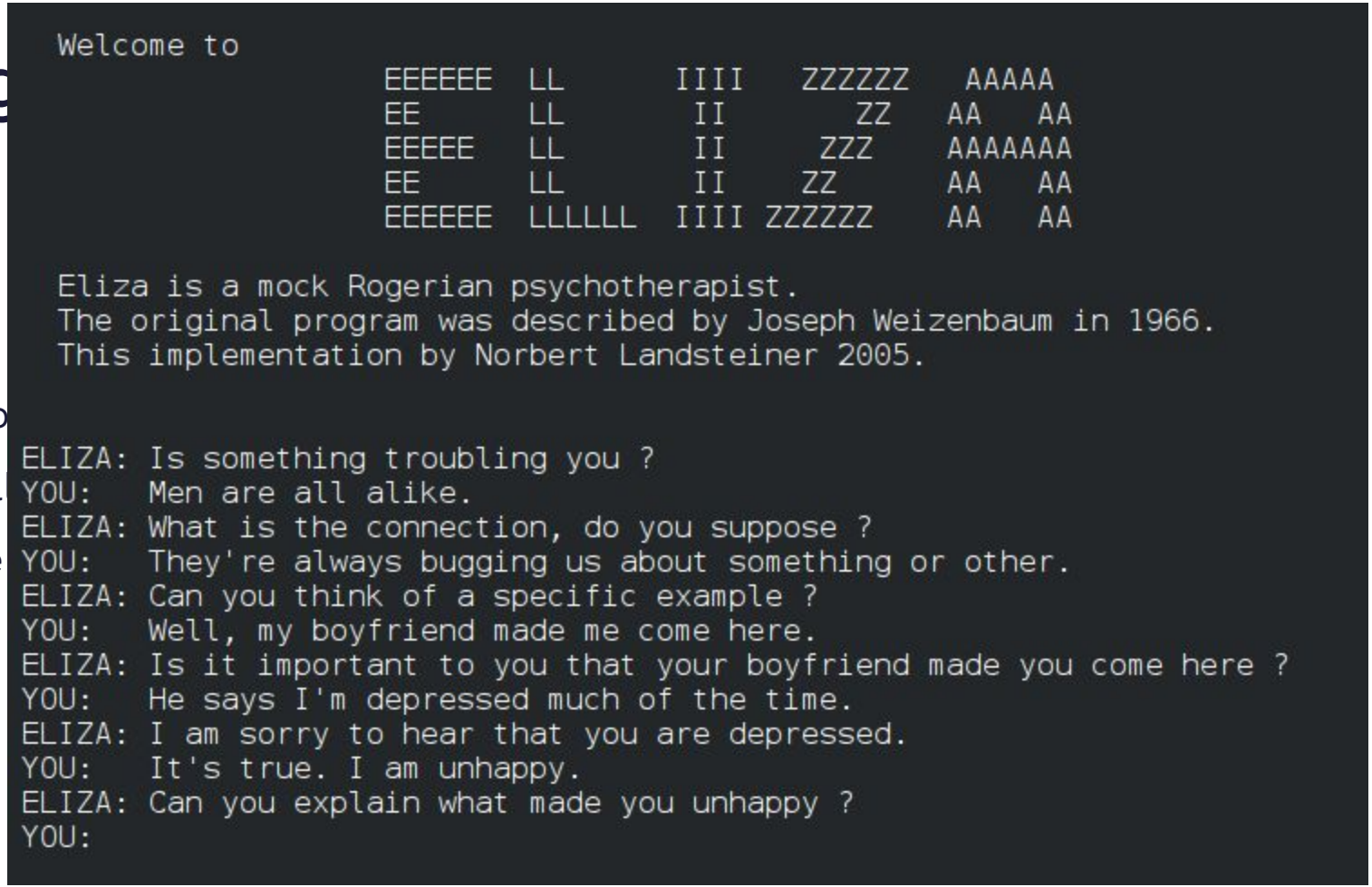
Resourceful workshop at NoDaLiDa & Baltic-HLT, Tallinn, March 3 2025

lang

probab

genera

we are



language models work quite well!

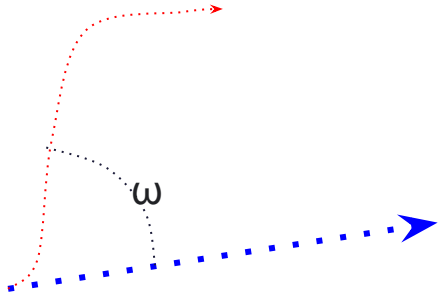
today's generation does a good job of reading, predicting, and generating strings

behaviouristically plausible at run time!

thanks to training on very large data sets

some languages are better off data-wise than others

enormous parameter space



machine learning curve:

- few parameters -> so-so model
- more parameters -> better model
- more parameters -> more training data needed (linear growth, approx)
- today's models: 10-100B parameters; 1-10T words

human learning curve:

- 100M-1B neurons, 1-10B synapses in language processing regions
- 1-10M words / y for about 20 years

what motivates the difference?

ω!

every engineer should react with enthusiasm here!

technology changes very rapidly

safe prediction:

- new architectures
- new memory models
- new training processes
- new hardware

likely prediction:

- less data needed
- less computational effort needed

assuming the above holds

What resources should language technology research focus on to address the likely needs of future generations of representations?

what are the resources in language technology?

computational:

algorithms, memory model, procedures, evaluation methodology

technology:

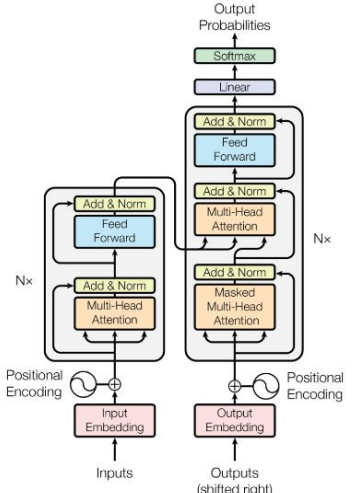
processors, memory, network

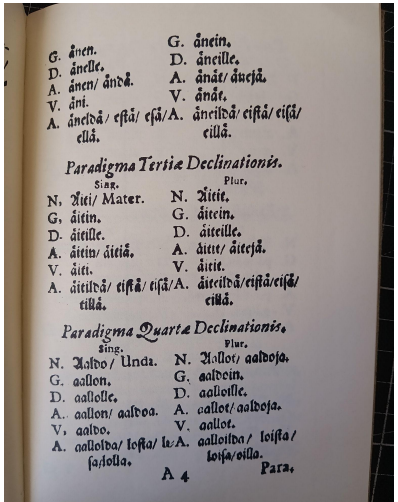
intellectual:

descriptions and lists and collections and samples of language

practice:

how to do things, why to do things, who foots the bill





G. ånen. G. ånein.
D. ånelle. D. ånelle.
A. ånen/ ånda. A. ånåt/ ånesja.
V. åni. V. ånåt.
A. ånelda/ esta/ esja/ ella. A. ånelda/ ejsta/ eisa/ illa.

Paradigma Tertiæ Declinationis.

Sing. Plur.
N, Åiti/ Mater. N. Åitit.
G, åitin. G. åitein.
D. åitille. D. åiteille.
A. åitin/ åitia. A. åitit/ åitesja.
V. åiti. V. åitit.
A. åitilda/ ista/ isa/ illa. A. åiteilda/ eista/ eisa/ illa.

Paradigma Quartæ Declinationis.

Sing. Plur.
N. Aallo/ Unda. N. Aallot/ aaldoja.
G. aallon. G. aaldein.
D. aallolle. D. aalloille.
A. aallon/ aaldoa. A. aallot/ aaldoja.
V, aaldo. V. aallot.
A. aalloilda/ losta/ lesa/ sa/olla. A. aalloilda/ loista/ loisa/oilla.

A 4 Para.

what does one need to learn a language?

basic understanding of the world (if L1)

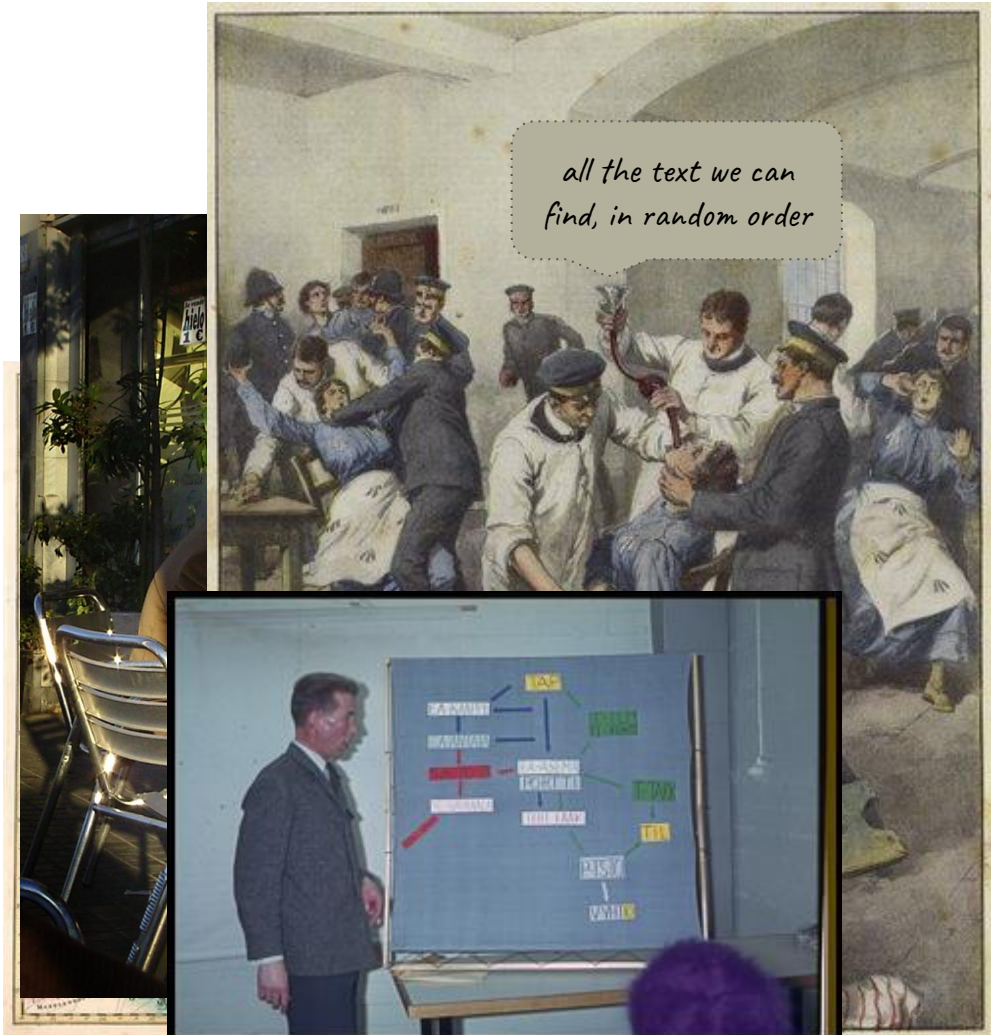
situational and social awareness

curriculum, opportunism, and anomalies

quality input

transfer from first language (if L2)

cultural understanding



what does one need to learn a language?

contrast to typical linguistics!

basic understanding of the world (if L1) -> hmm

situational and social awareness -> include context in the data you collect

curriculum, opportunism, and anomalies -> think about progression in learning

quality input -> unlock archives and libraries!

transfer from first language (if L2) -> focus on the specifics

cultural understanding -> do not translate data sets automatically

hur är språkmodellers inlärningsprocess?

Wikipedia och 1T ord från Reddit och Flashback

hur är människors inlärningsprocess?

- börja med enkla byggstenar
- bygg på situationer
- med rik kontext
- ta fasta på anomalier för opportunistisk inlärning

what will we want to have to check if a machine does its job?

Build test from first principles to fit use case and culture and situation!

what is unusual and strange in one language may be natural in another

we all want to be polite, but make politeness happen in different ways

Beware automatic translation of tests

- "Two people are seen sitting before a wave pool and one leads another out onto the water on a board. The person ..."
- "A cowboy rides a horse out of a corral and enters into a fenced off area. The man rides his horse out of the fenced are and throws a rope ..."
- "A small group of people are seen sitting around a casino table speaking to one another and playing a game of poker ..."

short todo list for the field language technologist

- Collect real data
- Keep track of the metadata for those data
- Collect situations
- Identify norms and preferences that are specific to the culture in question
- Embrace test-driven development:
  - begin by scenarios and
  - build evaluation mechanisms from first principles to fit scenarios