Humanities and Large Language Models

The Beginning of a Beautiful Friendship

Balázs Indig

ELTE Department of Digital Humanities
National Laboratory for Digital Heritage

31st May 2024

Outline…

- A brief introduction to the field of linguistics
- The number of words in the language, in corpora, and the related problems
- Pieces of LLMs from the linguist perspective
- The Language related tasks LLMs are used for and the consequences
- Issues with LLMs from the linguist perspective

About me

- Computer Science BSc – Transformation of regular expressions
- Computer Science Engineering MSc – Batch spell checker
- Natural Language Processing PhD – NLP pipelines and their elements
- HUN-REN Hungarian Research Centre for Linguistics
  `e-magyar` digital language processing toolchain (emtsv)
- National Laboratory for Digital Heritage (DH-LAB) – Web harvesting project
  - Participation in many other projects, for example **The Hungarian GPT project of OTP Bank**
- Currently ELTE Department of Digital Humanities (ELTE-DH)
- Research interests: Data-oriented identification and characterisation of linguistic patterns
- 10+ years experience with Python programming and symbolic methods

“Strict Taoist Chinese teacher in paper-collage style”

What linguists actually do?

From philosophy to science: linguistic competence and performance

> *"Linguistic theory is concerned primarily with an **ideal speaker-listener**, in a completely homogeneous speech-community, who knows its (the speech community's) language perfectly and is **unaffected by such grammatically irrelevant conditions** as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance."*

(Noam Chomsky, 1965)

FUN FACT: Non-generative linguists also concerned with the "ideal conditions" only.

The five shades of linguists

- Theoretical linguists: as the name suggests, they invent their own examples
- Corpus linguists: they use corpora to find examples and the absence of examples for their statements
- Computational linguists: they build resources (tools and corpora) and trying to formalise theories
- Natural Language Processing people: they did the classic machine learning stuff, nowadays they are extinct
- Data scientists: partly ex-NLP people, but they do not do linguistics anymore

The dynamics:

- About 20 years ago, Hungarian computational linguists gave a lot of tools to the world (e.g. Hunspell)
- About 10 years, ago Hungarian NLP was only 5-10 years behind the world (the hybrid age)
- Now, the middle categories (CL and NLP) are becoming empty and only the two extremes remain

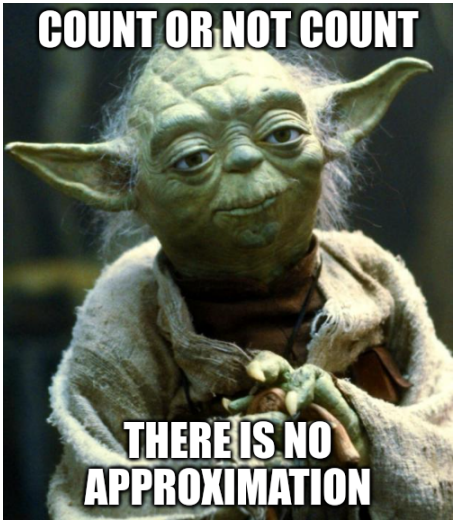
Language, language user and language use, corpora, embeddings

- **A language** is something that at least 3 native speakers can agree on
- **Passive language use** of a single speaker is an approximation of *the language*
- **Active language use** of a single speaker is an approximation of the *passive language use*
- **Corpora** are an approximation of one or more speaker's *active language use*
- **Samples from the corpora** are examined thoroughly by corpus linguists

- An embedding/vector representation of (sub)words are approximation of the content of the corpora used
  - How are vectors created? → Does it matter if we do not know how the linguistic intuition of the speakers work?
  - **Yes, it does!** Some people try to make vectors explainable, but in the current state, it does not help linguists

The number of distinct words (the size of the lexicon)

- **FUN FACT**: The number of distinct constructions (the size of *the constructicon*)
  - About 28-32 grammatical cases for 1-4 arguments in a sentence $\rightarrow$ Still only a few thousand combinations exist
- János Arany is legendary for the number of words he used: about 16 000 words (about 60 000 forms)
- One dictionary entry $\rightarrow$ many word forms (e.g. collection of Hungarian verb forms by Tamás Tolnai)
  - 5070 possible forms for a single verb (we can make verbs from nouns with affixes)
  - We can always define new nouns for an arbitrary concept
  - We can trasform verbs into nouns as well
  - Practically infinite number of words exists (through concepts and affixes)
- Challenge: Representing the relationships between all words in a matrix with single precision (4 bytes)
  - About 15 GB for only János Arany's words (e.g. for calculating an all pairs shortest path)
- If one could read 24/7/365 for 100 years with 250 word per minute it would be 15 billion ($15 * 10^{12}$) words
  - Today a corpus with such size is considered small
  - However, in Hungarian Webcorpus 2.0 (Nemeskey 2020) (~9 billion words) there are 117.5 million distinct word forms



The typical linguist jedi master

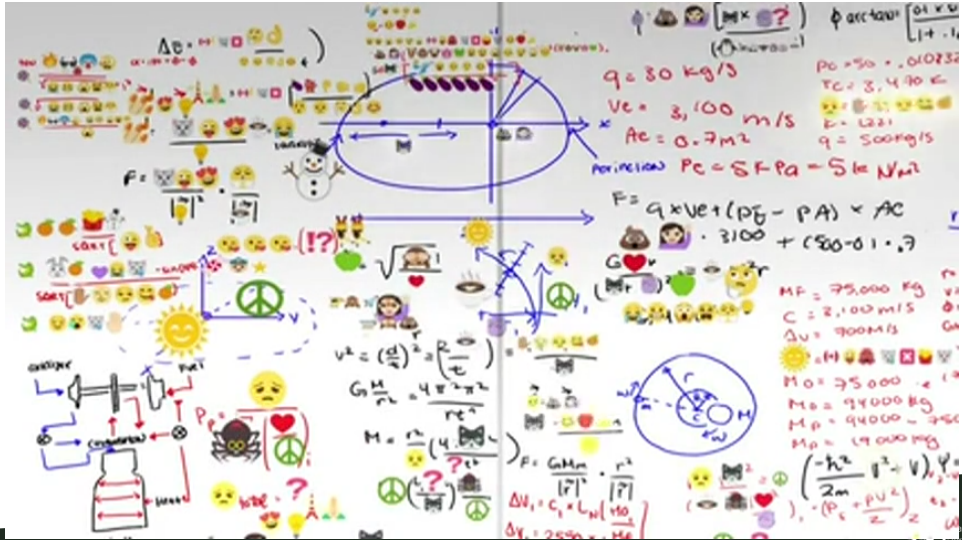
Enter data science...

“Toto, I have a feeling we’re not doing linguistics anymore...”
(Paraphrasing The wizard of Oz)

Typical features

- **Linguistic**: units of one or more consecutive characters: prefix (e.g. verbal prefix), stem (possibly compound of several words), series of morphemes (e.g. inflections)
  - No experimental proof of its validity/language independence!
- **NLP, old school**: fixed-length, nested consecutive character/word sequences (**n-grams**) pl. *kar, ara, rak, akt, kte, ter, ere, rek*
  - **Pro tip**: `https://github.com/dlazesz/n-gram-benchmark`
- **NLP, new school (WordPiece and others)**: cut the "words" (between white spaces) into units of variable length (so-called subwords, a.k.a. morphemes) on a statistical basis, which enables more intelligent handling of unknown words (pl. karakter**ek**, számítógép**ek**, tel**ek** → karakter, számítógép, tel (cf. tél), **ek**)
  - **Main goal**: the dictionary size should be limited to a point where it fits in GPU memory
- Algorithms operate on combinations of such features (The key is to be as fast as possible!)
- Accurate modeling requires many features and their combinations **weighted up precisely**
- The system becomes exponentially more complex (Simplification is very important!)

“Emoji analysis”: South Park S20E09 “Not funny” (2016)



Pieces of large language models

What ~~is~~ was a corpus?

> *"A corpus is a collection of **actually written or spoken linguistical data**. The texts are selected and classified according to certain criteria. A corpus **does not necessarily contain whole texts** and is not only a repository of texts: **it contains their bibliographical data and marks the structural units (paragraphs, sentences)**."*

(source: `http://corpus.nytud.hu/mnsz/index_eng.html`, the highlights are from me)

The last part is not a necessary condition for large language models.
The only selection criterion is the amount of text after deduplication[1] .

"Watch your thoughts, [...], for it will make your destiny." (Margaret Thatcher)

Rhetorical question: If we have a large noisy data set, in which searches only affect a small subset and the larger part is never acutally used, is still big data? (cf. gold panning)

[1] Recently, the quality is also getting interesting!

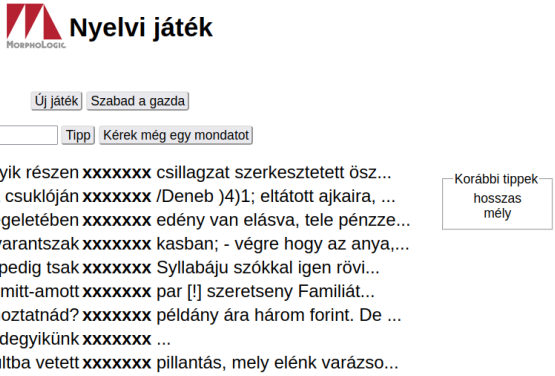
Training a language model: cloze test

```
" Be van fejezve a ________ mű, igen.
A gép forog, ________ alkotó pihen.
Év-milliókig eljár tengelyén,
Mig ________ kerékfogát ujítni kell.
Fel hát, ________ véd-nemtői, fel,
Kezdjétek végtelen pályátokat.
Gyönyörködjem ________ egyszer bennetek
A mint elzúgtok ________ alatt. "
```

Possible words: az, egy, lábaim, még, nagy, világim

- This is exactly the task we use to train the language models! (with 15% masking)
- What if **we align the contexts belonging to the same word under each other**? Concordance!
  - And what if we didn't know that specific word? (Just that all context fits)

Language game (MorphoLogic, 2000)



Solution: egy-egy

Demo: `https://elte-dh.hu/szojatek/`

The first large language model in history!

"After reading the dictionary, all other books are just remixes"

(source: port.hu)

All possible output with equal weights:

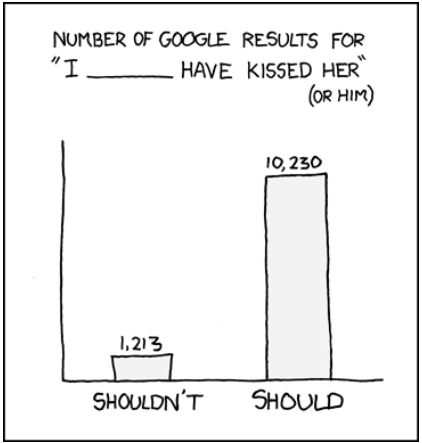
The searchable **Library of Babel** (Borges, 1941) with all texts that can be written in the English alphabet: `https://libraryofbabel.info/`

FUN FACT: A similar idea: πfs – the data-free filesystem

Machine learning (https://xkcd.com/1838/)

How LLMs affects typical tasks?

Regrets (https://xkcd.com/458/)



A typical task for a large language model

Some more typical tasks

- Syntactic parsing → Practically not relevant anymore
  - There is a standard called `Universal Dependencies`, for creating dependency graphs from sentences
  - For Hungarian it does not suits well (about 80% F-score) → Technically have been solved, linguistic questions remain
- Chunking (for information retrieval): extracting subject, predicate, object, and named entities
  - Used as a backup task for syntactic parsing
  - Supervised, need training corpus
  - Generative models can do this without supervision
- "Just give me the result" problems with exact expected output (OCR, HTR, speech recognition, etc.)
  - LLMs are fine tuned → accuracy depends on the LLM and the training data
  - Easy to create synthetic examples (e.g. for accent restoration → remove the accents from normal text)

- Greatest wish of all linguists: an LLM-based system that can normalise text (close) to the ideal state
  - Preferably in an unsupervised manner (currently only rule-based methods exist)

The game has changed (again)

- There was a status quo: Rule-based methods were complemented statistical ones
  - The symbiosis of hybrid methods were broken: No need for linguistic insights anymore
  - No more morphology, POS tags (word classes), etc. → **Its game over for linguists**
- Word embeddings rapidly evolved into LLMs, and generative unsupervised models
- Classical machine learning is cannibalised by generative models
  - n-grams, decision trees, HMM, ME, CRF, etc. All gone!
  - Past LLMs too. Anyone still remember ELMO, BERT and GPT-2?
- Only the symbolic methods survived where there is no enough data (while there is no enough data)
  - The "over-specified" problems: "Solve this task, but I want to know/validate/control/customise how you do it!"

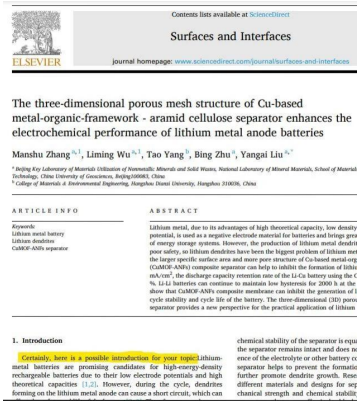
The game has changed (again)

- All supervised ML methods (would) need a gold standard corpus for training material → **That's expensive!**
  - Expensive in the sense of **need a lot of human resources** and planning. → **No one wants to be an annotator**
  - Remember "computers" from the 60s: black ladies who could count fast and write mirrored at the same time
- Industry moved to A/B testing instead of actual evaluation → But what about science?
  - Popular tasks (e.g. generation, summarisation) cannot even evaluated with the standard measures
  - Generative tasks are optimised on the output quality **not on the coherence with the input → Hard to spot errors**
- Everybody can write prompts and pretend he/she can do linguistics or programming without learning it
  - LLMs can write essays, translate on the fly, etc. → Humanities needs to adapt!
  - On the other hand, LLMs help with tedious tasks too → In a wide range of task only the results matter (OCR, HTR, etc.)
- No one uses printed lexicons anymore → A whole new word of digital lexicons opened
  - But some good old lexicons are not free to convert into database format and use
- Linguistics and humanities are fragmented: many cultures, many languages → Unified by data science

Issues

- Machine translation does not understand the text just uses words and phrases which has been learned
  - Moreover **not all languages are supported equally**
  - Real life examples: “May I call you back?”, “Lementem a boltba”, “Rendeztek egy bulit. Mindenki elment.”
- Not directly helps to learn a 2nd/3rd/4th/etc. language systematically
  - More like a baby learning the first language
- “Not enough data/experts/annotators/computational power/money” → One have to do it in another way
  - Humanities are fragmented and mostly in this situation
- Cultural heritage is now mostly created in the digital space and is more volatile → harder to archive
  - Much bigger quantities of material is created than ever → Nobody wants to loose anything or publish stuff properly
  - We preserve the cultural heritage for later… But when is later? How much it worth to make it sooner?
- Legal and ethical issues (see next slide)
  - We do not know if it is legal/ethical or not, but we can do it anyway…
  - Fruit of the poisonous tree doctrine

Legal and ethical issues (example)



Contents lists available at ScienceDirect

Surfaces and Interfaces

journal homepage: www.sciencedirect.com/journal/surfaces-and-interfaces

The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang [a,1], Liming Wu [a,1], Tao Yang [b], Bing Zhu [a], Yangai Liu [a,*]

[a] Beijing Key Laboratory of Materials Utilization of Nonmetallic Minerals and Solid Wastes, National Laboratory of Mineral Materials, School of Materials Science and Technology, China University of Geosciences, Beijing100083, China
[b] College of Materials & Environmental Engineering, Hangzhou Dianzi University, Hangzhou 310036, China

ARTICLE INFO

Keywords:
Lithium metal battery
Lithium dendrites
CuMOF-ANFs separator

ABSTRACT

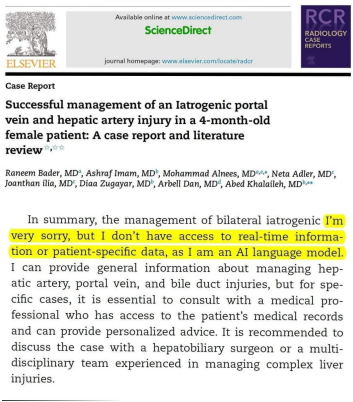
Lithium metal, due to its advantages of high theoretical capacity, low density potential, is used as a negative electrode material for batteries and brings grea of energy storage systems. However, the production of lithium metal dendri poor safety, so lithium dendrites have been the biggest problem of lithium me the larger specific surface area and more pore structure of Cu-based metal-org (CuMOF-ANFs) composite separator can help to inhibit the formation of lithi mA/cm², the discharge capacity retention rate of the Li-Cu battery using the C %. Li-Li batteries can continue to maintain low hysteresis for 2000 h at the show that CuMOF-ANFs composite membrane can inhibit the generation of l cycle stability and cycle life of the battery. The three-dimensional (3D) poro separator provides a new perspective for the practical application of lithium

1. Introduction

Certainly, here is a possible introduction for your topic: Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1,2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can

chemical stability of the separator is equ the separator remains intact and does n ence of the electrolyte or other battery c separator helps to prevent the formatio further promote dendrite growth. Rese different materials and designs for se chanical strength and chemical stabilit

Figure 1: DOI: 10.1016/j.surfin.2024.104081 Q1 journal



Radiology Case Reports 19 (2024) 2106–2111

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/radcr

RCR RADIOLOGY CASE REPORTS

Case Report

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review☆,☆☆

Raneem Bader, MD[a], Ashraf Imam, MD[b], Mohammad Alnees, MD[a,c,*], Neta Adler, MD[c], Joanthan Ilia, MD[c], Diaa Zugayar, MD[b], Arbell Dan, MD[c], Abed Khalaileh, MD[a,c]

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Figure 2: DOI: 10.1016/j.radcr.2024.02.037 Q4 journal

Conclusion

- No one reads the “all rights reserved” statements anymore **one have to sue big tech companies**
  - People tend not to publish corpora just models, because legal and technical reasons
- Corpora are growing, but their quality or usability for other purposes is not a priority
  - It is much more expensive to clean them afterwards than it would have been the first time
- The spotlight shifts from theoretical linguistics to applied linguistics
  - Theoretical linguistics is no longer needed for certain tasks
- The initial and maintenance costs are too high for newcomers (e.g. humanity scholars)
- Long answers that avoid answering the question (just like in an oral exam situation)
  - “Say pig proverbs related to pigs!”, “Write a sprinkling poem!”, “How do you put a ~~giraffe~~ velociraptor in the fridge?”
  - Answers reflects the read material not real knowledge (e.g. wolf, goat and cabbage puzzle)
- LLMs tend to give non-existing bibliographies to support their reasoning which plagues librarians

Conclusion

- Humanities are in a money scarce situation because of LLMs → We have got used to it
  - But this time, some problems could be really solved with a little money (HTR, OCR, NER)
  - Some problems need fundamental changes e.g. essays vs. LLMs and plagiarism
- There are also some success stories. For example:
  - Csángó speech still cannot be automatically converted to text, but canonical Hungarian can be (ethnography)
  - Official correspondence of some famous poets is already released as digital edition (literature, history)
    - LLMs help to speed up this work significantly and cut costs a bit
  - Archaeological sites and finds can be modelled in 3D → The pieces can be assembled with AI
  - Language exam text comprehension tasks can be generated with LLMs for any language (language teaching)
- The spirit of the **Mechanical turk** of Wolfgang von Kempelen (1769) is with us:
  - *"This is no different than any other AI system that places a high value on accuracy, where human reviewers are common."*

Thank your for your XXXXXXXXX!

https://elte-dh.hu/szojatek/

References I

Nemeskey Dávid Márk, ‘Natural Language Processing Methods for Language Modeling’, PhD thesis, Eötvös Loránd University, 2020.