Potentials of AI in the Analysis and Evaluation of Essay-type Tasks

Andrea Palmini, Tunc Yilmaz

Freie Universität Berlin, Center für Digitale Systeme (CeDiS)
Arbeitsbereich E-Learning und E-Examinations (EEE)

VERANSTALTET VON:

BHDL
BERLINER HOCHSCHULNETZWERK
DIGITALE LEHRE

IM RAHMEN EINES PROJEKTES VON:

Netzwerk Landeseinrichtungen für digitale Hochschullehre

GEFÖRDERT VON:

Stiftung Innovation in der Hochschullehre

Agenda

1 **Introduction**
Introduction to Large Language Models (LLMs) with a general overview
ASAG (Automated Short Answer Grading)
Similarity models in general

2 **Transformer Based Models**
Basic definition and purpose
Utilization in text similarity and limitations

3 **The IMPACT Project**
Why relevant to essay type exam evaluations?
Real life examples of text similarity assessment

4 **LLMs for Text Similarity Assessment**
Abilities on top of traditional transformer models
Real life examples of text similarity
Potentials and shortcomings

5 **Discussion**

1 Introduction

What is a large language model?

A large language model is an artificial intelligence algorithm trained on large amount of text data to create a natural language output

- **It uses neural network techniques to process and understand human language**
- **Those techniques are based on the deep learning methodologies, which can detect complex relationships in the text, and also generate text, understanding the semantic and syntactic of a language**

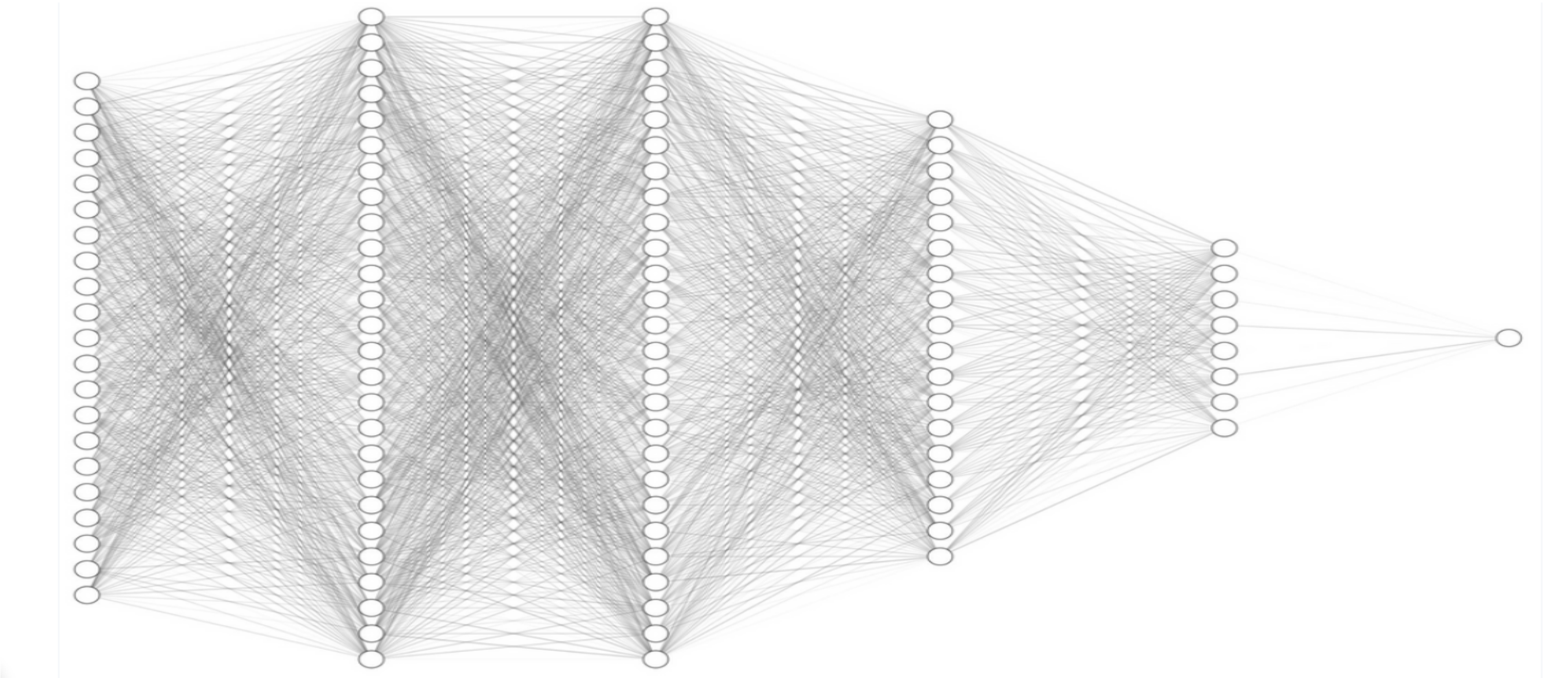
1 Introduction

How does an LLM work?

- **The models are trained on a vast amount of data**
- **Their utility lies on the ability to recognise patterns and relationships they learn from languages in the training phase**
- **This ability is given by their structure: consisting in many layers (feed forward, embedding or attention) which collaborate to process a text and generate an output**
- **The architecture of LLM depends on many factors (computational resource, number of layers, task)**
- **One of the model that revolutionized NLP tasks is the transformer model**

1 Introduction

LLM architecture example



1 Introduction

Transformer Models

Transformer models were introduced in 2017 achieving best performance in different tasks

Key features of this innovative models are:

Attention Mechanism

- helps to focus on important parts of the input (text)
- allows understanding connections between words or elements far from each other (context understanding)

Parallel Processing

-instead of screening input sequentially, the mechanism is employed on all input to handle larger sequences of text

Encoder-Decoder Architecture

- the encoder process the input with the mechanism
- the decoder generate the output sequence based on the encoder representation of the input

1 Introduction

Step-by-step Workflow

- **Pre-training: the models learn to predict new word in a sentence by understanding its surroundings (learning grammar and patterns of reasoning)**

- **Fine-tuning: used to answer specific task (improve performance)**

- **Inference: once trained, given an input text the model generates an answer**

1 Introduction

Examples of tasks

Natural Language Understanding

chatbots engaging in natural conversations, intelligent virtual assistants

Content Generation

creating human-like text such as storytelling or creative writing

Language Translation

multilingual machine translation with better context understanding

Text Summarization

Sentiment Analysis or Classification

1 Introduction

Well-known Transformer Models

BERT

- Bidirectional Encoder Representation from Transformers
- developed by Google
- used for a wide variety of tasks

RoBERTa

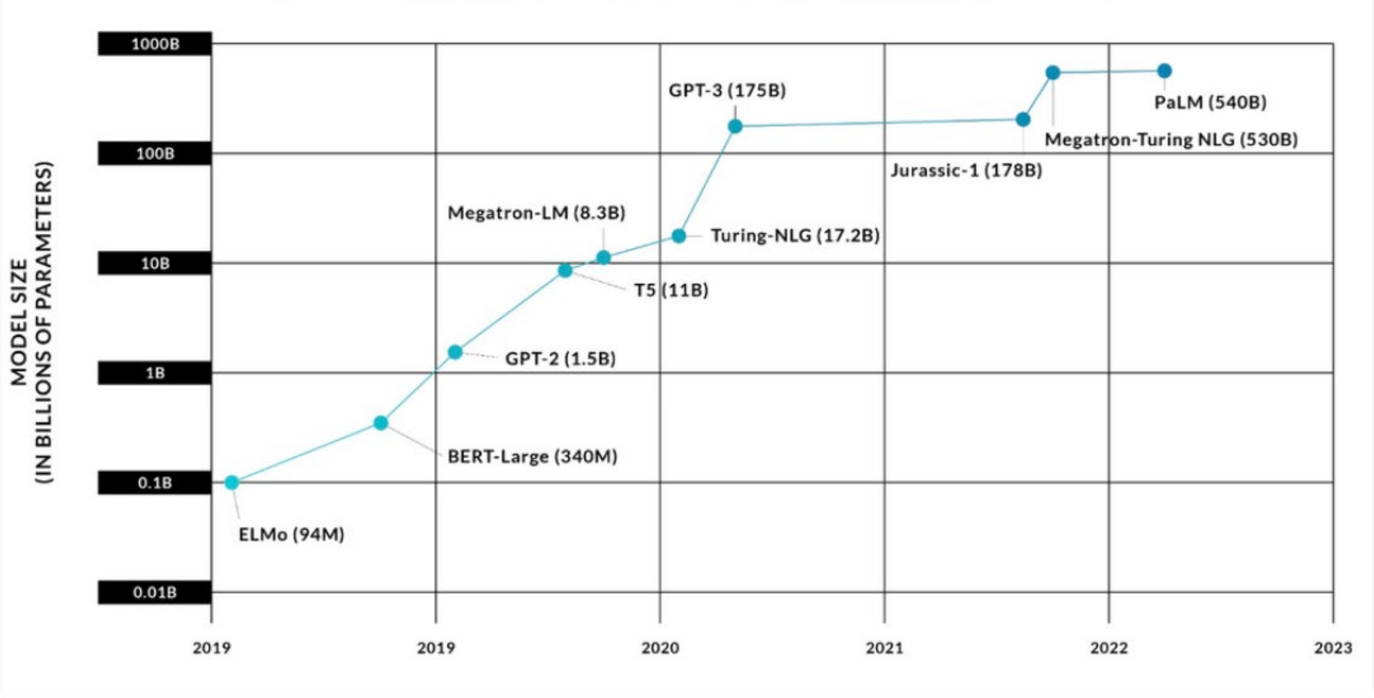
- optimized version of BERT
- developed by Facebook AI
- it tries to optimize the transformer architecture

GPT

- Generative Pre-trained Transformer
- developed by OpenAI and author of the famous Chat GPT

1 Introduction

Development of LLMs in time



1 Introduction

What is Automated Short Answer Grading (ASAG) ?

- Definition: Assessing short (!) answers based on the similarity to model solution texts, by using lexical, syntactic and semantic cues.
- Keywords: “student / learner answer”; “reference / blueprint answer”; “semantic similarity”;
- Scope: More complex than multiple-choice or true/false QA pairs, less complex than long essays with more cohesion and coherence
- Domain: Natural Language Processing (NLP), Learning Analytics (LA), Massive Open Online Course (MOOC) Assignment Assessments

Question
How does water evaporate?

Blueprint Answer
Evaporation occurs when energy (heat) forces the bonds that hold water molecules together to break. When you’re boiling water on the stove, you’re adding heat to liquid water. This added heat breaks the bonds, causing the water to shift from its liquid state to its gaseous state (water vapor), which we know as steam. (US Geological Survey)

Student Answer
Evaporation happens when a liquid substance becomes a gas. When water is heated, it evaporates. The molecules move and vibrate so quickly that they escape into the atmosphere as molecules of water vapor. (National Geographic)

1 Introduction

How does text similarity work in Natural Language Processing (NLP) ?

Bag of Words

The king is dead. | The queen is happy.

≈
%50 similarity

 **proper noun, company name, technology as context**

 **common noun, fruit, nutrition as context**

Word Embeddings


	Age	Alive?	Noun?	...	...	Colour	Location	Gender
King	0.1	1.0	0.8	...	...	0.0	0.0	0.0
Queen	0.1	1.0	0.8	...	...	0.0	0.0	1.0

E(“King”) – E(“Man”) ≈ E(“Queen”) – E(“Woman”)

E(“Germany”) – E(“Berlin”) ≈ E(“Italy”) – E(“Rome”)

The	0.1	0.0	0.1	...	...	0.0	0.0	0.0
king	0.1	1.0	0.8	...	...	0.0	0.0	0.0
is	0.1	0.0	0.1	...	...	0.0	0.0	0.0
dead	0.0	0.0	0.1	...	...	0.0	0.0	0.0

The queen is dead. | The rabbit is dead.

2 Transformer Based Models

The ___ jumped over the fence and escaped.

2 Transformer Based Models

cat
dog
horse
apple

The ____ jumped over the fence and escaped.

2 Transformer Based Models

cat
dog
horse
apple

The ____ jumped over the fence and escaped.

The dark ____ jumped over the fence and escaped.

2 Transformer Based Models

cat
dog
horse
apple

The ___ jumped over the fence and escaped.

cat
dog
horse
apple

The dark ___ jumped over the fence and escaped.

2 Transformer Based Models

The ___ jumped over the fence and escaped.

cat
dog
horse
apple

The dark ___ jumped over the fence and escaped.

cat
dog
horse
apple

Large Datasets

In some recent models, 10 tb of text is not unusual!

Self-Attention

A mechanism to encode how each sentence component is related to others within the context and other dependencies inside a given text. Model parameters are used for this purpose.

Self-Supervision

Huge training data can be obtained without the need for human annotators:
___ horse jumped over the fence and escaped.
The ___ jumped over the fence and escaped.
……

2 Transformer Based Models

How to use for comparing texts ?

Text A

0.1	0.0	0.1	...	...	0.0	0.0	1.0
0.1	0.5	0.8	...	...	0.0	1.0	0.0
0.1	0.5	0.1	...	...	0.0	0.2	0.0
0.0	0.0	0.3	...	...	0.2	0.0	0.2

Vector A

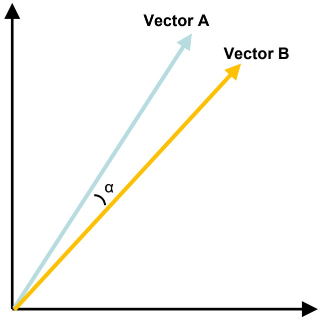
0.1	0.5	0.0	...	...	0.3	0.1	0.0

Text B

0.1	0.0	0.1	...	...	0.2	0.0	0.0
0.1	1.0	0.8	...	...	0.0	0.6	0.3
0.1	0.0	0.1	...	...	0.0	0.0	0.0
0.0	0.7	0.1	...	...	0.0	0.8	0.0
0.4	0.0	0.1	...	...	0.7	0.0	0.0

Vector B

0.1	0.8	0.2	...	...	0.3	0.1	0.1



cos(α) = similarity

Any limitations ?

- What happens if the text comes from a novel context (i.e. unseen in model training before)?
- Could it handle difficult cases, like double negation (not not), or sarcasm?
- Would summarizing longer texts to a fixed vector space cause loss of crucial information?

3 The IMPACT Project

What are the project components?

	I	II
Objective	**Project as Whole** AI Based Implementation of Feedback and Assessment with Trusted Learning Analytics	**Working Package 5** - How to assist professors evaluate and grade exams - Summative assessment
Stakeholders	**Funding** Bundesministerium für Forschung und Bildung	**Joint Project** FUB , GUF, UB, HUB & FeU participate with different roles in different WPs
Project Framework	**Methodology** - Ethical principles - Data protection - Bias and discrimination	**NLP & Statistical Models** How to make use of language models and computational tools to achieve goals

3 The IMPACT Project

Real Life Examples from Transformer Similarity Outputs

Question: 'Jemand sagt „Das ist eine Schweinerei mit dem Kleingedruckten also den Allgemeinen Geschäftsbedingungen . Sie sind verbindlich, selbst wenn ich sie nicht einmal gelesen habe. Und der Verwender kann mich darin auf übelste Weise über den Tisch ziehen. Wie würden Sie das aus rechtlicher Sicht kommentieren?'

Reference Answer: 'Allgemeine Geschäftsbedingungen können tatsächlich auch dann Vertragsbestandteil werden, wenn sie nicht gelesen wurden § 305 Abs. 2 BGB. Aber weil sie nicht gelesen wurden, müssen sie im Wesentlichen fair sein §§ 307 309 BGB.'

Sample Answer	Assigned Grade	Similarity Score
…. Die AGB sind, wie die Person …beschreibt, … sie nicht einmal gelesen hat. Die Regelungen sind laut §305 II Nr.1 und Nr.2 … zur Kenntnis nehmen kann. …. gem. §305c I auch keine überraschenden Klauseln enthalten…. … (§§307 – 309)	10/10	8.4/10
AGB sind, wenn auf sie eindeutig und erkennbar hingewiesen wird, gem. § 305 Abs. II BGB Bestandteil eines Vertrages. … In den Verordnungen zu den Allgemeinen Geschäftsbedingungen gem. 305 ff. … diese geregelt.	4/10	6.9/10
AGB ist das Angebot …, aber wenn eine Partei falsch den Vertrag verstanden hat, kann die Parte laut Par 119 Abs 1 BGB ; den Vertrag anfechten und nach …	0/10	6.6/10

4 LLMs for Text Similarity Assessment

What is new after the release of GPT-series?*

much larger Training Datasets
- multiples of 10tb,
- trained for weeks,
- w/ thousands of GPUs,
- €€€

similar Pre-Training Tasks
next word prediction instead of missing word prediction

billions of Parameters
and counting...
more energy and computing power
↓
more parameters

Fine-Tuning or Alignment
- the model is trained to guess the next word
- how to answer a question?

RL from Human Feedback
- human evaluator assesses the best QA versions
- QA fed back for finetuning)

Multi-Modality
context: "similarity calculation"
↓
under the hood: similarity model

User Data
user happy with a QA pair
↓
use that QA pair in fine-tuning loop

*This video of Andrej Karpathy forms the basis of the information provided in this slide, and much more about LLMs for general audience.

4 LLMs for Text Similarity Assessment

An Example of Text Similarity with GPT3.5*

Question
How does water evaporate?

Text1
Evaporation occurs when energy (heat) forces the bonds that hold water molecules together to break. When you're boiling water on the stove, you're adding heat to liquid water. This added heat breaks the bonds, causing the water to shift from its liquid state to its gaseous state (water vapor), which we know as steam. ([US Geological Survey])

Text2
Evaporation happens when a liquid substance becomes a gas. When water is heated, it evaporates. The molecules move and vibrate so quickly that they escape into the atmosphere as molecules of water vapor. ([National Geographic])

GPT3.5 Assessment:

Text1 and Text2 both describe the process of evaporation, particularly focusing on how water turns from liquid to gas by the application of heat. Both texts convey the fundamental concept that heat causes water molecules to move faster and break free from their liquid state to become water vapor. Here's a breakdown of their similarities and an assigned similarity score:

- **Content:** Both texts discuss how water transforms from a liquid to a gas state through the application of heat, mentioning the increased movement of molecules as a key factor in this process.
- **Key Concepts:** Each text highlights the role of heat (energy) in breaking the bonds between water molecules, leading to evaporation. Text1 provides a more detailed explanation of the process, including the notion of bonds breaking and the specific transition to "water vapor" and "steam." Text2 simplifies the explanation, focusing on the quick movement and vibration of molecules leading to evaporation.
- **Terminology:** Text1 uses terms like "energy (heat)," "bonds," "liquid state," "gaseous state," and "steam," offering a more technical description. Text2 sticks to more general language, making it accessible but less detailed.

Considering these points, the texts are highly similar in their fundamental explanation of evaporation but differ slightly in their complexity and detail level.

Therefore, I would assign a similarity score of:

- **8/10:** This score reflects their strong thematic alignment and shared focus on the basic principles of evaporation, with slight variations in depth and technicality.

*For a German language example with similar reasoning, use this link.

4 LLMs for Text Similarity Assessment

Concerns and Potentials

Concerns

- Bias and ethics: Data that the model trained on will always reflect the bias that
- Privacy and security: Data protection and privacy is paramount in educational institutions
- Computational power and costs: Taking into account growing number of parameters and model sizes
- Hallucinations can be detrimental in the context of learning analytics
- Interpretability: The LLM can output a nice reasoning, but how exactly?

4 LLMs for Text Similarity Assessment

Concerns and Potentials

Potentials

- Prompt Engineering: writing clear instructions and splitting tasks to achieve a certain goal with primitive rules and limitations

"Use only the given text"
"Divide text into two subsections"

- Fine-Tuning: allows application across multiple use cases (e.g. improving similarity for legal texts)

Q: "Müssen die AGB fair sein §§ 307BGB." – A: "Laut §§ 307BGB…"

- Retrieval-Augmented Generation: course material and other information can be used to limit the knowledge base of the LLM while performing tasks

"Use only the relevant course content"

- UX and UI: dialogue style, and other capabilities as well as features give flexibility and control

Student Answer: ….
Blueprint Answer: …..

5 Discussion

5 Discussion

Based on your experience in Learning Analytics and Education: What do you think?

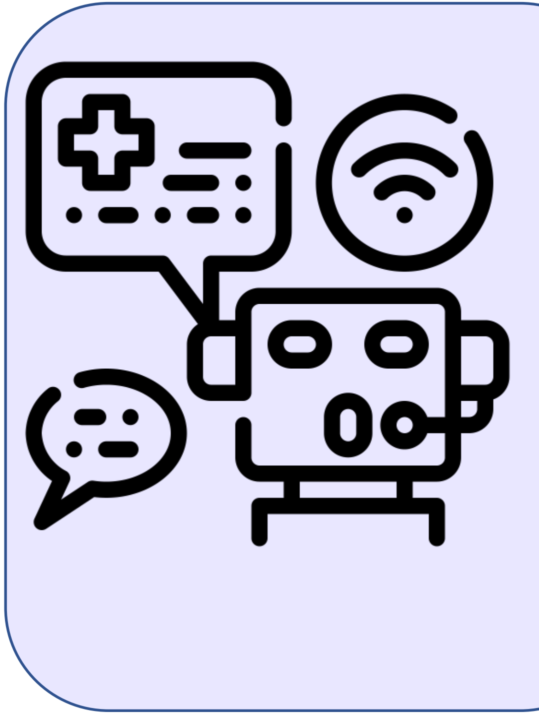
1. What can be some examples of the positives/negatives LLMs may bring, even in a controlled environment?

2. Do language models of any kind hold promise to surpass the quality of human evaluation in exams?

3. What makes a good evaluation of formal and scientific bodies of text?

?
- challenges on examinations
- more relaxed learning opportunities (MOOC etc.)
- technical details of LLMs in the context of examinations

Thank You!



Communication: Andrea Palmini -> andrea.palmini@fu-berlin.de || Tunc Yilmaz -> tunc.Yilmaz@fu-berlin.de

References were given at respective slides with footnotes.

Icons in this presentation are taken from https://icons8.com