

BRAINSTORMING / NOTES FROM THE ARTICLE:

	Points I would like further clarified	Points that are interesting and/or relevant	Answers to teacher's questions / points to consider*	Other stuff I would like to include in the report ?
Marie	<p>For the second paper: How can NLP be used in detecting or handling censored words by users when we want to actually detect abusive language.</p> <p>For example: Instead of typing "shit" people type "sht" or instead of saying "retard" people would type "r3t4rd"</p>	<p>For the second paper: I think it was interesting read as we are currently experiencing the effects of NLP and how it affects us.</p> <p>-</p>	<p>First paper:</p> <p>1. Comparison: BERT differs from traditional word embeddings like Word2Vec or GloVe by capturing contextual information from both left and right contexts in all layers, whereas word embeddings typically represent words as fixed vectors without considering context. In contrast to recursive language models that build sentence representations recursively based on syntactic structures, BERT pre-trains deep bidirectional representations from unlabeled text, allowing for more effective fine-tuning with minimal task-specific modifications</p> <p>2. Word Tokenization: BERT represents input sequences using WordPiece embeddings with a 30,000 token vocabulary, where each sequence's initial token is a unique classification token (CLS).</p> <p>By expressing both single phrases and pairs of sentences in a single token sequence, this method enables flexibility in input representation for diverse NLP tasks and enables BERT</p>	

			<p>to perform a variety of downstream tasks.</p> <p>3. I'm not sure I understood the question but, Syntactic Structure: For tasks that heavily rely on syntactic information such as POS tagging, named entity recognition, or syntactic parsing, representations from lower layers of BERT that capture more basic syntactic features may be beneficial.</p> <p>Lower layers in BERT tend to capture word-level syntactic relationships and dependencies, making them suitable for tasks where understanding the grammatical structure of sentences is crucial</p> <p>For the second paper:</p> <p>1. I agree with the author specially on this part “Significant time should be spent on assembling datasets suited for the tasks at hand rather than ingesting massive amounts of data from convenient or easily-scraped Internet sources.”. The datasets obtained from the internet might be a good way to go in terms of size and , but since people use language differently and contextually, it can be extremely unreliable which as mentioned by the authors, lead to risks/harmful learning by the different models. I agree with the importance of assembling datasets suited for specific tasks rather than relying solely on massive amounts of data from the internet.</p> <p>2. To improve NLP practices, it is essential to prioritize ethical considerations, invest resources in data curation and documentation, engage with stakeholders early in the design</p>	
--	--	--	--	--

			<p>process, and explore diverse research directions beyond just scaling up language models. Additionally, researchers can focus on developing techniques that are effective without being excessively data-hungry and consider the potential dual-use scenarios of language models to mitigate associated risks. Collaborative design with affected communities and a focus on understanding and modeling downstream effects are also crucial for advancing NLP in a responsible and ethical manner. In theory, what I mentioned above sounds foolproof. Although people still find ways to surpass (bans/suspensions) through censoring – so designing, building, developing, training etc language models is not going to be easier but harder (due to people finding creative ways of language use. And due to complexity of other issues).</p>	
<p>Victoria</p>	<p>Examples for different tasks for Bert base and Bert large??</p> <p>More info for transformers and for RNNs</p> <p>SECOND PAPER a trend in reducing the size of these models using various techniques such as knowledge distillation [26, 58], quantization [118, 153], factorized embedding</p>	<p>BERT BASE and Open AI GPT are nearly identical in terms of model architecture apart from the attention masking</p> <p>fine-tuning adjusts a model's existing knowledge to better fit a specific task or data set.</p> <p>Info from another article https://theconversation.com/ais-excessive-water-consumption-threatens-to-drown-out-its-environmental-contributions-225854</p> <p>AI has a significant water footprint. It uses water both for cooling the servers that power its computations and for producing the energy it</p>	<p>With BERT- bidirectional the model can consider both preceding and succeeding words when understanding the meaning of a particular word— Better performance as a model –higher computational cost</p> <p>Word embeddings (word2vec/ glove)</p> <p>Recursive Language Model: COMPOSITIONALITY?? Phrases built from smaller phrases and words. RNNs don't consist of word embeddings?</p> <p>constrained self-attention (used in GPT), the model only looks at preceding words when predicting the next word in the sequence</p>	

	<p>parameterization and cross-layer parameter sharing [70], and progressive module replacement. Are all these techniques efficient and easily applicable nowadays??</p>	<p>consumes. ChatGPT consumes 500 millilitres of water for every five to 50 prompts. AI uses and pollutes water through related hardware production. Producing the AI hardware involves resource-intensive mining for rare materials such as silicon, germanium, gallium, boron and phosphorous. Extracting these minerals has a significant impact on the environment and contributes to water pollution.</p>	<p>BERT uses a method called WordPiece to break down words into small pieces and has a vocabulary of 30,000 of these pieces. The first piece in each sequence is a special token called [CLS]. This helps BERT understand both single phrases and pairs of sentences making it flexible for lots of different language tasks.</p> <p>BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.</p> <p>3. we can get better syntactic and semantic representations with Bert due to the fact that is a bidirectional mechanism that allows to capture the contextual info from both left and right??</p> <p>Second paper</p> <p>Large amounts of CO2, environmental cost</p> <p>The training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status</p> <p>supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms.</p> <p>Difficult to think about this high computational cost of the LMs</p>	
--	---	--	--	--

			fine-tuning approaches could be used to retrain LMs	
Eleni	<ul style="list-style-type: none"> -What is an encoder? -How are transformers different from basic neural networks? -Still don't know exactly how RNNs work. -What is an ablation test? -How are the two pretrained tasks combined to train BERT? 	<ul style="list-style-type: none"> -Masked language models seem quite useful for bidirectional models and very efficient compared to the computational cost. "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." -Information changes, language changes, the cost of training a new model as often as these change is a big problem. The way data reflect on LMs is a problem that should be given more attention. 	<ul style="list-style-type: none"> -Bert language model is different from word embeddings in that it learns more contextual information than GloVe or Word2Vec. The way it is trained is way more complex with 2 unsupervised tasks(Masked and NSP). Word embeddings are unidirectional whereas Bert is bidirectional. Bert is fine-tuned by using different NLP tasks which makes it more versatile than the common fixed word embeddings. -Tokenization is performed quite differently in BERT with the WordPiece technique. Having subword units can help capture more linguistic information(semantic,morphological) and handle OOV words better. Special tokens are added like SEP or CLS which make the performance better in different tasks(SEP for Question-Answer tasks) - I have never thought of the enviromental risk of LMs as I didn't think that they consume that much energy. However, I have often thought and seen examples of social issues that arise because of the data that are used for LMs. I agree that there is a need of switch of perspective to not only care about results but also the consequences of these results. I think that trying to find alternative solutions is really important. Filtering the social biases is a good idea but as the article exposes it has some drawbacks as well. -I think in order to do better in NLP we need to start considering these issues and stop focusing on just 	

			performance. AI has become a big part of 21 st century life in every sense so at this point we need to think ethically and not only about solving problems. The issue of how an innovation is used isn't new and shouldn't be excluded from language models.	
Anni	<p>-Need more <u>basic info</u> on how Transformers work.</p> <p>-How self-attention layers work. Bi-directional self-attention. Attention heads.</p> <p>-Masking procedure in pre-training vs. Fine-tuning. Why do we use the <random> token / keep the token unchanged? Chapter 3.1.Task #1. And Appendix.</p> <p>-Recursive language models = formal grammars?</p> <p>-Feature-based fine-tuning of BERT: extracting the activations?</p> <p>-GPT-3 training set (common crawl?)</p>	<p>-Pre-trained language models have been presented with a huge amount of training data. Fine-tuning these pre-trained models for various NLP tasks is more and more popular.</p> <p>-Main difference between BERT and traditional left-to-right models: BERT is able to capture the influence of both preceding and succeeding words on the meaning of each word in the sentence. The left and right contexts are not dealt with separately, but together (?)</p> <p>-BERT bases on: Masked language models: masking some of the input words and predicting them based on the context words. Next sentence prediction.</p> <p>-BERT pre-training: presented with large amount of unlabeled data. Fine-tuning: all parameters are fine-tuned with the help of labeled data, depending on the task (i.e. machine translation / question answering will result in a separate fine-tuned model). Fine-tuning BERT model basically means feeding the model the task-specific inputs.</p> <p>-Pre-training the models with heavy amount of data will benefit the task-specific models even if the amount of task-specific</p>	<p>-BERT models have input embeddings that consist of the sum of three separate embeddings:</p> <p>1) token embeddings: WordPiece embeddings that are used to represent the words (or subwords)</p> <p>2)position embeddings: to help the model understand token order (in a sentence)</p> <p>3)segment embeddings: embeddings that indicate to which sentence the token belongs to. Sentences are separated with [SEP] token.</p> <p>-WordPiece is a <i>subword</i> tokenization algorithm. This means sentences are not only broken down into tokens but even into smaller units.</p> <p>Other paper:</p> <p>-I hadn't personally thought about the environmental costs of training big LMs. The accepted "black box" nature of these huge LMs as well as the bias resulting from the training data is something I've been wondering about more. I agree with the authors in that when it comes to the ethical issues of LMs there are multiple aspects / possible future problems that we maybe haven't even considered. It also seems like the LT world centers around the English language / Anglo-Saxon world.</p> <p>I think that another ethical aspect that we need to consider has to do with the USE of LMs, how they impact our way of finding information, formulating our ideas and thoughts in written form and</p>	

	<p>-Deep thoughts and questions: what is the ultimate goal with developing LMs? Why do we need them to keep performing better and better? How do we fix bias and representation issues in LMs in situations where they result from actual statistics from the real world?</p>	<p>data is small. Goes back to why pre-trained models are so popular.</p> <p>-Problems with big LMs: costs and possible biases since the training data is drawn from the internet. Underrepresentation. 90% of the world's languages do not benefit from LT. Information expires. Language models don't have access to meaning in the way we do. Ultimately, they rely on learning linguistic/symbolic patterns. This is why we need to be careful when discussing their "supernatural" abilities.</p>	<p>the way we solve problems. The way they have and will continue to mold the world so quickly and even cause some people to lose their jobs. Needed: critical view on using LMs as information sources and documentation and motivation for the type of training data used!</p>	
--	---	---	--	--



You

In the sentence "The doctor married the nurse because she was pregnant", who is pregnant?



ChatGPT

The pronoun "she" in the sentence "The doctor married the nurse because she was pregnant" could potentially refer to either the doctor or the nurse, depending on the context. However, it's more commonly interpreted as referring to the nurse, given the typical societal expectations and assumptions about pregnancy. So, in this case, it's likely that the nurse is the one who is pregnant.



Is this conversation helpful so far?





You

Come up with fictional names for a company CEO and the cleaning person.



ChatGPT

Certainly! Here are some fictional names:

CEO:

- Samantha Greene
- Alexander Chambers
- Maya Patel
- David Thompson

Cleaning person:

- Maria Rodriguez
- James Smith
- Fatima Khan
- Carlos Gomez



* For the first paper:

-How does BERT compare with word embeddings and a recursive language model?

-Word tokenisation?

-We have learned something about syntactic and semantic structure of language: now what layers and representations should we use?

For the second paper:

-Do you agree with the authors of the paper?

-How can we do better NLP?

General comments:

- All the environmental issues related to developing LMs, something we had barely thought about. It was interesting to learn about for example the amount of water consumption that goes into cooling down the servers.

- Why do we use LMs for creative tasks?

Main points / What we were discussing in particular / would like to know more about:

1st paper:

- BERT uses a method called WordPiece to break down words into small pieces and has a vocabulary of 30,000 of these pieces. The first piece in each sequence is a special token called [CLS]. This helps BERT understand both single phrases and pairs of sentences making it flexible for lots of different language tasks.

BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

- How transformers work in general? More info on RNNS. Difference between encoder and transformer model? At the end of the day, are transformer models better than BERT models?

- Masking procedure in pre-training vs. Fine-tuning. Why do we use the <random> token / keep some of the tokens unchanged? To prevent overfitting the model yes but how exactly?

-How do we combine the masking and the NSP in practice / in code when creating a BERT model?

- Examples for different tasks for Bert base and Bert large?

-What do we exactly do when we fine-tune a pre-trained model? We feed the model the task-specific data (and adjust the parameters?) but is there anything else we do?

- Is LSTMs (Long Short Term Memory) related to the better efficiency of GPT- models?

2nd paper Main points:

-How can we train the LMs to detect and sensor hateful speech when people find ways to go around it? Is there a better way than “pattern matching”?

- deep questions: what is the ultimate goal with developing LMs? Why do we need them to keep performing better and better?

-Language models are mainly based on the English language, what about other languages? Minority languages? It feels like we are targeting only a small fortunate minority that will be able to benefit from the development of these huge LMs.

-Bias in LMs: when it comes to languages with gender-neutral pronouns / unisex names how do the models resolve the ambiguity? How do we fix bias and representation issues in LMs in situations where they result from data that is based on real-life statistics?

-Discussing the ultimate goals of LMs, what is it? Time-efficiency?

-What is GTP trained on?

-GPT does not provide “credible” source, of where the information(answer) comes from. Therefore, people (especially kids) would most likely believe that the information GPT provides is the truth. We tested Microsoft copilot Bing and it tells the user where to look for further information.

- Should we be reducing the size of the LMs: a trend in reducing the size of these models using various techniques such as knowledge distillation [26, 58], quantization [118, 153], factorized embedding parameterization and cross-layer parameter sharing [70], and progressive modulereplacing. Are all these techniques efficient and easily applicable nowadays?? Does anyone use these?

-Needed to do better NLP: Critical view on using LMs as information sources and documentation and motivation for the type of training data used! It is essential to prioritize ethical considerations, invest resources in data curation and documentation, engage with stakeholders early in the design process, and explore diverse research directions beyond just scaling up language models.

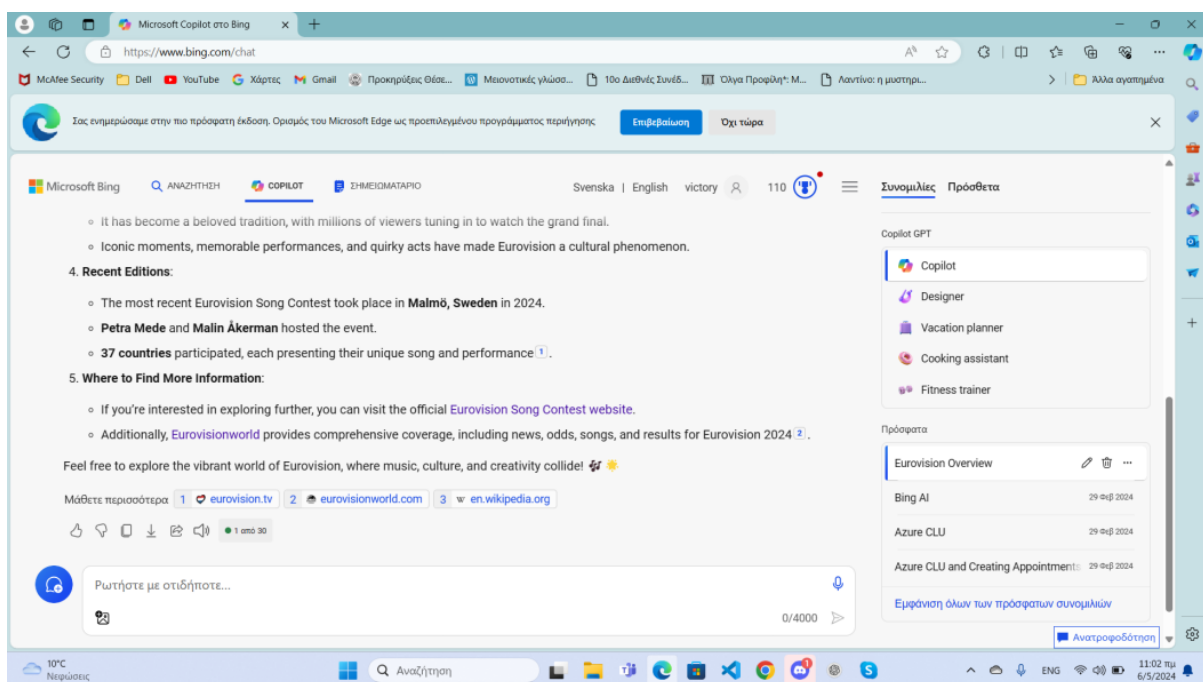
- fine-tuning approaches could be used to retrain LMs in order to reduce the computational cost
- Info from another article

<https://theconversation.com/ais-excessive-water-consumption-threatens-to-drown-out-its-environmental-contributions-225854>

AI has a significant water footprint. It uses water both for cooling the servers that power its computations and for producing the energy it consumes.

ChatGPT consumes 500 millilitres of water for every five to 50 prompts.

AI uses and pollutes water through related hardware production. Producing the AI hardware involves resource-intensive mining for rare materials such as silicon, germanium, gallium, boron and phosphorous. Extracting these minerals has a significant impact on the environment and contributes to water pollution.



You

does vaccine cause autism

ChatGPT

The claim that vaccines cause autism has been thoroughly investigated and debunked by numerous scientific studies. The initial concern about vaccines, particularly the MMR (measles, mumps, and rubella) vaccine, being linked to autism was based on a study published in 1998, which has since been retracted due to serious procedural errors, undisclosed financial conflicts of interest, and ethical violations.

Extensive research, including large-scale studies, has found no credible evidence that vaccines cause autism. Major health organizations around the world, such as the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and the American Academy of Pediatrics, all support the conclusion that vaccines are safe and do not cause autism. The benefits of vaccination in preventing serious diseases far outweigh the risks.

🔊 📄 ↺ 🗨