NorwAI
«Empowering News Summarization with Pre-trained Language Models»
Lemei Zhang & Peng Liu
May 2023
Outline

1 Generative Language Models for Norwegian
2 News Summarization with Instruct Tuning
3 What’s next?
Market Expectations for Interest Rate Hike in February

Steve Evans <steve@evans.co>  
to me

Subject

According to a recent article from CNBC, markets are currently fully pricing in a quarter-point interest rate hike by the Federal Reserve in February. This follows the release of data showing that inflation in the US has slowed in recent months. The article also notes that the Fed has indicated that it plans to maintain its current monetary policy stance until it sees clear signs of inflation rising above its 2% target.

However, the article also notes that some experts are warning that the Fed may be forced to raise interest rates sooner than expected if inflation continues to rise, in order to prevent it from becoming too high. This could have a negative impact on the stock market and the broader economy.

Overall, the market is currently expecting a quarter-point interest rate hike in February, but it is important to keep an eye on inflation and the Fed’s monetary policy stance in the coming weeks.

Best,
Steve

Reply Forward
What is a Language Model?

Gigantic unlabelled training data sets

45 TB dataset

198 GB dataset

Contextualized understanding of words of a language

OpenAI GPT-3

NorwAI GPT-23B

Billions of parameters in transformer model

175 billion parameters

23 billion parameters
• Encoder-Decoder model based on (self-) attention mechanism

• Regardless of sequence length, capture long-term dependencies between input and output and allow more parallelization.

• Basically targeting for translation tasks (WMT’14 EN-GE, EN-FR), it achieved state of the art BLEU score.

Motivation for a large Norwegian GPT Model

Large tech companies, typically based in the United States, are the main driving forces behind commercial LLMs.

GPT-3 Training Dataset Statistics

<table>
<thead>
<tr>
<th>language</th>
<th>number of words</th>
<th>percentage of total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>181014683608</td>
<td>92.64708%</td>
</tr>
<tr>
<td>fr</td>
<td>3553061536</td>
<td>1.81853%</td>
</tr>
<tr>
<td>de</td>
<td>2870869396</td>
<td>1.46937%</td>
</tr>
<tr>
<td>es</td>
<td>1510070974</td>
<td>0.77289%</td>
</tr>
<tr>
<td>it</td>
<td>1187784217</td>
<td>0.60793%</td>
</tr>
<tr>
<td>pt</td>
<td>1025413869</td>
<td>0.52483%</td>
</tr>
<tr>
<td>nl</td>
<td>669055061</td>
<td>0.34244%</td>
</tr>
<tr>
<td>ru</td>
<td>368157074</td>
<td>0.18843%</td>
</tr>
<tr>
<td>ro</td>
<td>308182352</td>
<td>0.15773%</td>
</tr>
<tr>
<td>pl</td>
<td>303813902</td>
<td>0.15550%</td>
</tr>
<tr>
<td>fi</td>
<td>221644679</td>
<td>0.11344%</td>
</tr>
<tr>
<td>da</td>
<td>221551540</td>
<td>0.11339%</td>
</tr>
<tr>
<td>sv</td>
<td>220920577</td>
<td>0.11307%</td>
</tr>
<tr>
<td>ja</td>
<td>217047918</td>
<td>0.11109%</td>
</tr>
<tr>
<td>no</td>
<td>212193289</td>
<td>0.10860%</td>
</tr>
<tr>
<td>zh</td>
<td>193517396</td>
<td>0.09905%</td>
</tr>
</tbody>
</table>
Generative Language Models for Norwegian

**Challenges:**
- Limited availability of Norwegian Text
- Finding sufficient computational resources (Idun)
- Implementing a software stack to make the model accessible

**Goal:** Develop a GPT model that can compete for Norwegian 🇳🇴
NorwAI GPT-23B Training Corpora

- Norwegian data (Dataset size: 139.82 GB)
  - Norwegian mC4 web-crawled corpus: 96.71 GB (9.5 billion tokens)
  - OSCAR web-crawled corpus: 4.96 GB (787 million tokens)
  - Schibsted: 6.9 GB (1.08 billion tokens)
  - National Library dataset: 29.45 GB (4.49 billion words)
  - Social media: Norwegian tweets (602 MB, 180.49 million tokens)
    Norwegian redits (500 MB, 86.74 million tokens)

- Danish Common Crawler Dataset: 16.18 GB (2.5 billion tokens)
- Swedish Common Crawler Dataset: 19.43 GB (2.9 billion tokens)
- German Common Crawler Dataset: 10 GB (1.4 billion tokens)
- English Common Crawler Dataset: 12.37 GB (2 billion tokens)
NorwAI GPT-23B Training Corpora

• Data Pre-processing
  o Data clean pipeline
  o Removing sensitive information (username, email, url, etc.)
  o Removing emoji
  o Removing special characters & symbols
  o Deduplicating within & across the corpuses
### NorwAI GPT Language Model Training

- **NorwAI GPT Models**

<table>
<thead>
<tr>
<th></th>
<th>NorGPT-318M</th>
<th>NorGPT-3B</th>
<th>NorGPT-23B</th>
<th>NorGPT-40B</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Parameters</td>
<td>318.6 Million</td>
<td>2.95 Billion</td>
<td>23.03 Billion</td>
<td>40 Billion</td>
</tr>
<tr>
<td>#Layers</td>
<td>24</td>
<td>32</td>
<td>49</td>
<td>?</td>
</tr>
<tr>
<td>#Attention heads</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>?</td>
</tr>
<tr>
<td>#Hidden unit size</td>
<td>1024</td>
<td>2688</td>
<td>6144</td>
<td>?</td>
</tr>
<tr>
<td>Context window</td>
<td>1024</td>
<td>2048</td>
<td>2048</td>
<td>?</td>
</tr>
<tr>
<td>Training data size</td>
<td>13 GB</td>
<td>196 GB</td>
<td>196 GB</td>
<td>&gt; 196 GB</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>30,000</td>
<td>64,000</td>
<td>64,000</td>
<td>?</td>
</tr>
<tr>
<td>Convergence time</td>
<td>1 week</td>
<td>4 weeks</td>
<td>3-4 months</td>
<td>?</td>
</tr>
</tbody>
</table>
NorwAI GPT Language Model Training

- Training Setting:
  - 28 x NVIDIA A100 SXM 80GB GPUs
  - 20 TB Memory
  - GPT-2 implementation from Microsoft/Megatron-DeepSpeed

- Checkpoint of around (Huggingface version):
  - 1.4 GB (318M Model)
  - 6.2 GB (3B Model)
  - 93 GB (23B Model)
Benchmarking

- NorwAI is developing a comprehensive benchmark set for Norwegian generative language models
  - Translation from English supervised datasets
  - Collection of manually annotated datasets
  - Still work in progress

<table>
<thead>
<tr>
<th>Conversational agent</th>
<th>Norwegian Persona-Chat dataset (26.3 MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text summarization</td>
<td>Norwegian CNN/Dailymail (532 MB)</td>
</tr>
<tr>
<td>Question answering</td>
<td>Norwegian News Summaries (in progress)</td>
</tr>
<tr>
<td></td>
<td>Norwegian Free-Style QA (in progress)</td>
</tr>
</tbody>
</table>
NorwAI GPT Language Model Applications


[3] Social Robot in NorAI.

---

Chatbot

NorGPT

News Generation

Personalized Recommendation

Question Answering

News Summarisation

Social Robot
CNN.COM: CNN reporters and contributors to CNN in Hong Kong gather in front of the U.S. presidential election in South Korea in March, 2013. (Photo: KCNA/Jason K. Smith, AFP Photo)
NorGPT-3B Instruct Tuning

Step 1
Fine-tune the pre-trained model with human prompts

Step 2
Train reward modal (RW) with human ranking

Step 3
Update the fine-tuned LLM with RW from step 2
Output LLM

NorwAI
Step 1: Fine-tune with human feedback

- Norwegian supervised translation dataset:
  - News Summarization dataset from CNN/Dailymail
  - 1.18G, 128,354 prompts in total
  - 1 phase: 18,336 articles
  - 2 phase for ranking: 55,008 articles
  - 3 phase: 55,008 articles

Step 1: Fine-tune with human feedback

LLM
Human Feedback Collection

Step 2

Recruit participants for small samples to get start

Find your participants
Need a different audience? Run a quick survey to recruit your sample.

Amazon Mechanical Turk
Access a global, on-demand, 24x7 workforce

Looking for data labeling solutions to power Machine Learning models?
Amazon SageMaker Ground Truth allows you to easily build and manage your own data labeling workflows and workforce. Or, use Ground Truth Plus, a turnkey data labeling service that provides an expert workforce and manages it on your behalf.

Amazon Mechanical Turk is accessible through both Ground Truth and Ground Truth Plus.
Human Feedback Collection

Step 2

Recruit participants for small samples to get start
Human Feedback Collection

Step 2

Automatic the process

NorBERT

Generated summary

Ground-truth summary

Semantic similarity

Ranking score

Ranking score

NorwAI
Case study

**Norwegian**


**English**

Baseball Hall of Famer Bob Feller died Wednesday night in Cleveland, Ohio, of acute leukemia, according to a statement from the Cleveland Indians. He was 92. Feller was diagnosed with acute myeloid leukemia in August and underwent surgery at the Cleveland Clinic to implant a pacemaker in October. In mid-November, Feller was admitted to hospital with pneumonia. The major leaguer was an icon for Indians fans. “We in Cleveland have been blessed to have him with us these many years.” Indians owner Larry Dolan said. To say he will be missed is an understatement” Feller spent 18 years in a Cleveland Indians uniform. He started 484 games, winning 266 of them. Feller had a lifetime ERA of 3.25, threw three no-hitters, including the only Open Day no-hitter in Major League Baseball history, on April 16, 1940. In 1946, he pitched 36 complete games. “Since 1936 he has been with us. For 75 years he has been a contributing citizen, a model for all athletes and a friend to thousands,” Dolan said. “As it should be in a larger sense, Bob will always be with us. Not on Opening Day, not at Fantasy Camp, not in the press box, but in our hearts.” Indians manager Manny Acta said Feller was a special person. “There has never been a great with such a connection to his original franchise,” Acta said. “When you think of the Cleveland Indians, you think of Bob Feller and vice versa. He was a true patriot and a great Hall of Famer. Boy, he loved the Indians and we all loved him back.”
Case study

<table>
<thead>
<tr>
<th>NorGPT generated summary (before fine-tune)</th>
<th>NorGPT generated summary (after fine-tune)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norwegian</td>
<td>Norwegian</td>
</tr>
<tr>
<td>&quot;Vi vil alltid savne Bob Feller&quot; Faller døde torsdag kveld i Cleveland, Ohio, av akutt leukemi. Han var 92 år gammel. I sin biografi, &quot;The Baseball Hall of Fame,&quot; skrev han at &quot;han var den beste, den mest lojale, den beste fansen og en meget god leder&quot;.</td>
<td>Baseball Hall of Famer Bob Feller dør på akutt leukemi i en alder av 92. Feller ble operert ved Cleveland Clinic for å implantere en pacemaker i oktober. Han ble innlagt på sykehus etter å ha utviklet lungebetennelse. Feller tilbrakte 18 år på uniform.</td>
</tr>
<tr>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>&quot;We will always miss Bob Feller.&quot; Faller died Thursday night in Cleveland, Ohio, of acute leukemia. He was 92 years old. In his biography, &quot;The Baseball Hall of Fame,&quot; he wrote that &quot;he was the best, the most loyal, the best fan and a very good manager&quot;.</td>
<td>Baseball Hall of Famer Bob Feller dies of acute leukemia at age 92. Feller underwent surgery at the Cleveland Clinic to implant a pacemaker in October. He was hospitalized after developing pneumonia. Feller spent 18 years in uniform.</td>
</tr>
</tbody>
</table>

-0.18  Reward score  -0.15
What’s next?

• Generative models such as LLaMA on the LUMI supercomputer

• Separate Bokmål and Nynorsk models

• Streamline the benchmarking workflow

• Evaluation metrics for news summarization
Project Team

Benjamin Kille  Peng Liu  Lemei Zhang  Jon Atle Gulla  Jon Espen Ingvaldsen
Thank you!

Q&A