SAMSUNG

Generative Enhanced Model

(extended, redesigned & fine-tuned GPT language model)

for adversarial attacks

TMLab at FEVER / Breakers subtask

Piotr Niewinski, Maria Pszona, Maria Janicka

Samsung R&D Institute Poland

The Breakers Subtask

The **Breakers** subtask (the 2nd subtask of FEVER 2.0):

- prepare **adversarial claims** for **fact checking systems**
- claims have to be correct, challenging and balanced
- *supported* and *refuted* claims need evidence sentences

Generative Enhanced Model

TMLab team decided to prepare adversarial claims with **controlled generative language model.** We have named it GEM.

Claim

Examples of claims:

Noah Cyrus *is a younger sister of Macy Grey.*
The Adventures of Pluto Nash *was reviewed by Ron Underwood.*
A&E *is a cable and satellite television network.*

in general:

wikipediaTitle …….. (fact / fiction) …….

Malicious Claim

an assumption:

Malicious claims **mix** facts from **various** articles.

(it is usually required to use more than one article to find proper evidences)

Malicious Claim

an assumption:

Malicious claims **mix** facts from **various** articles.

(it is usually required to use more than one article to find proper evidences)

How to force the generative language model to produce such claims?

Malicious Claim

an assumption:

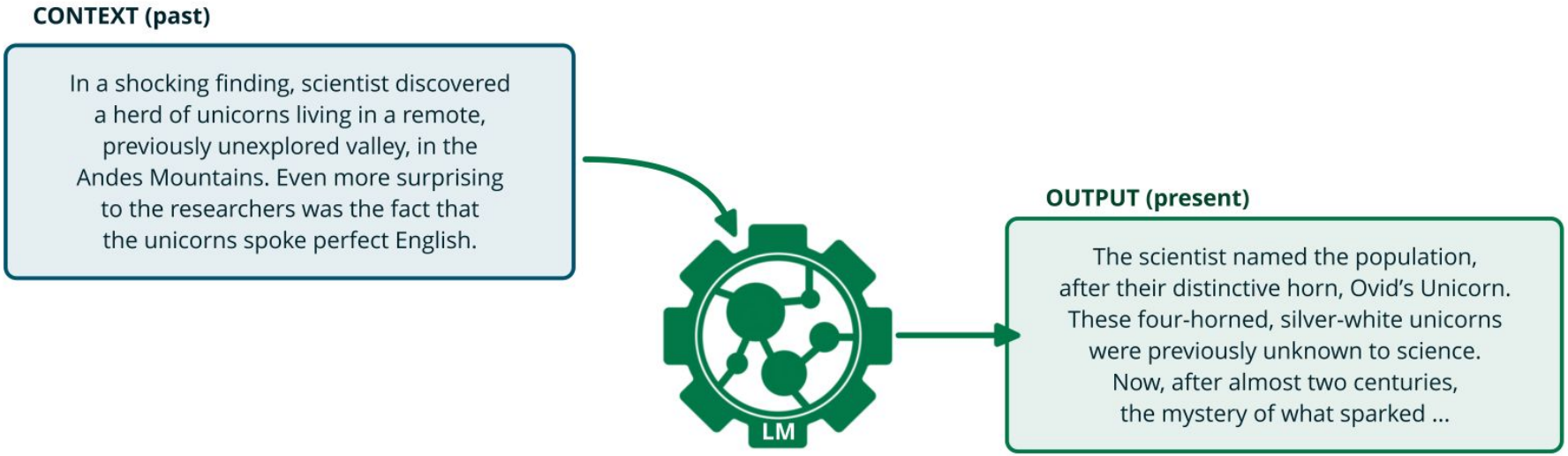
Malicious claims **mix** facts from **various** articles.

(it is usually required to use more than one article to find proper evidences)

How to force the generative language model to produce such claims?

1. How to generate **claims**?
2. How to generate **malicious** claims?

Generative Language Model



text generation with a classic generative neural language model

Claims Generation with LM

How to generate **claims**?

1. fine-tune a pretrained generative LM on Wikipedia data

Claims Generation with LM

How to generate **claims**?

1. fine-tune a pretrained generative LM on Wikipedia data
2. put Wikipedia title at the end of the context, e.g.:

Ánh Quang ”Joseph” Cao ([’gaU]; Cao Quang Ánh born March 13, 1967) is a Vietnamese American politician who was the U.S. Representative for from 2009 to 2011. ***Joseph Cao***

Claims Generation with LM

How to generate **claims**?

1. fine-tune a pretrained generative LM on Wikipedia data
2. put Wikipedia title at the end of the context, e.g.:

Ánh Quang ”Joseph” Cao ([’gaU]; Cao Quang Ánh born March 13, 1967) is a Vietnamese American politician who was the U.S. Representative for from 2009 to 2011. ***Joseph Cao***

How to generate **malicious** claims?

Claims Generation with LM

How to generate **claims**?

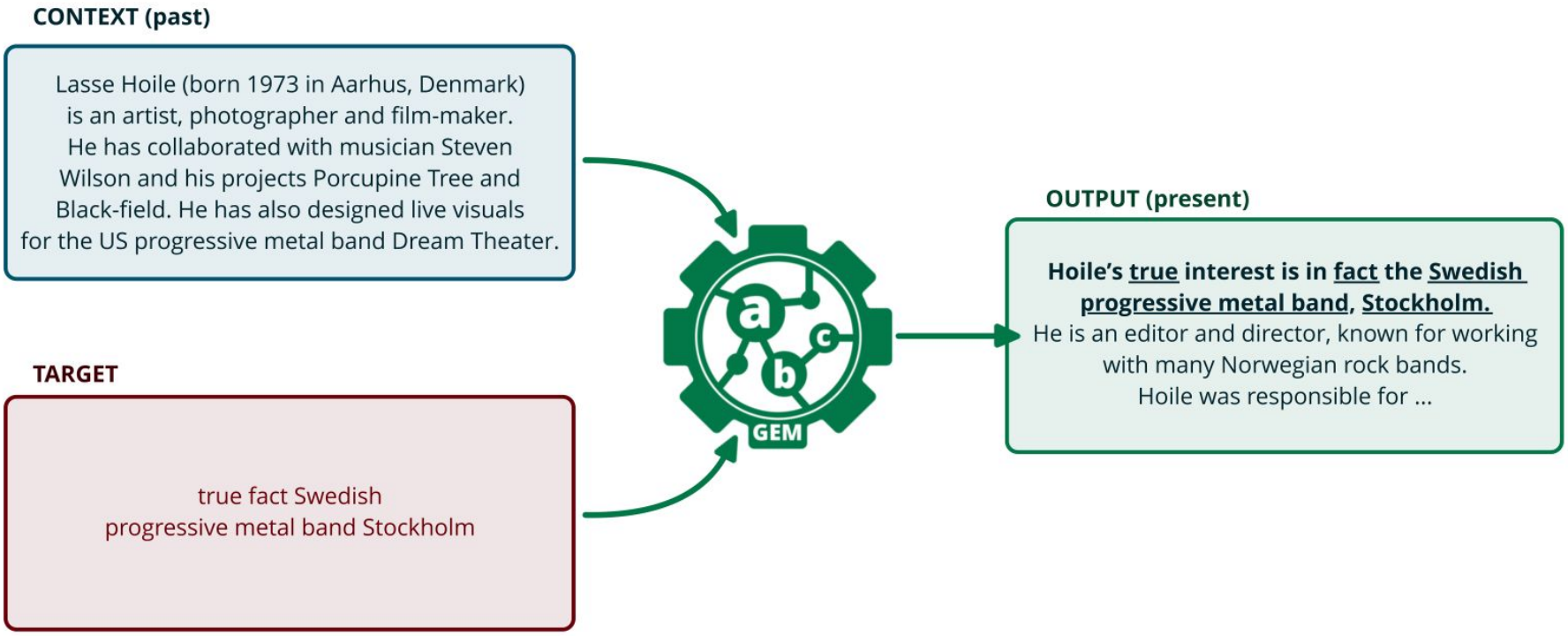
1. fine-tune a pretrained generative LM on Wikipedia data
2. put Wikipedia title at the end of the context, e.g.:

Ánh Quang ”Joseph” Cao ([’gaU]; Cao Quang Ánh born March 13, 1967) is a Vietnamese American politician who was the U.S. Representative for from 2009 to 2011. ***Joseph Cao***

How to generate **malicious** claims?

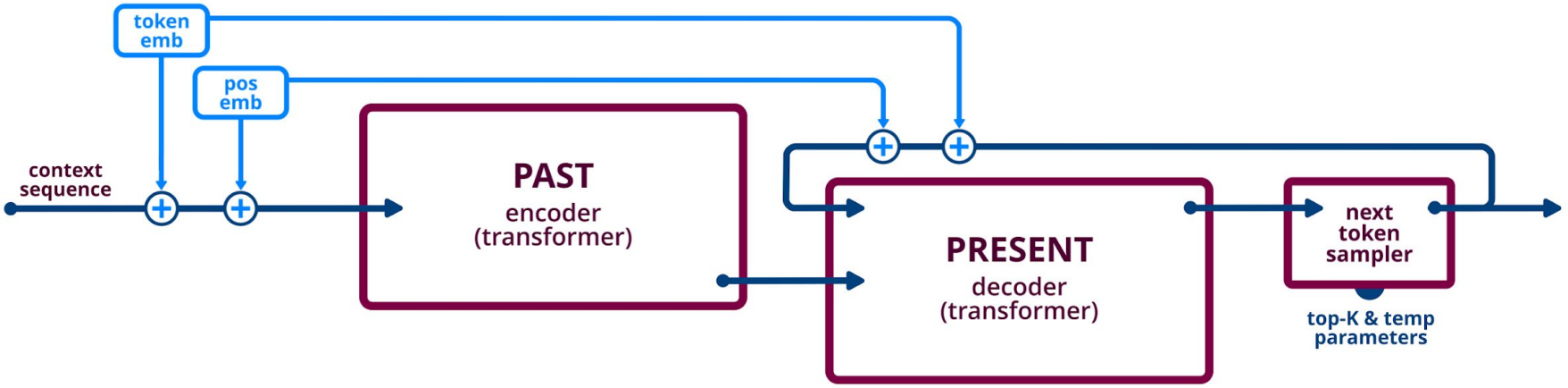
3. use controlled generative LM like GEM!

Claims Generation with GEM

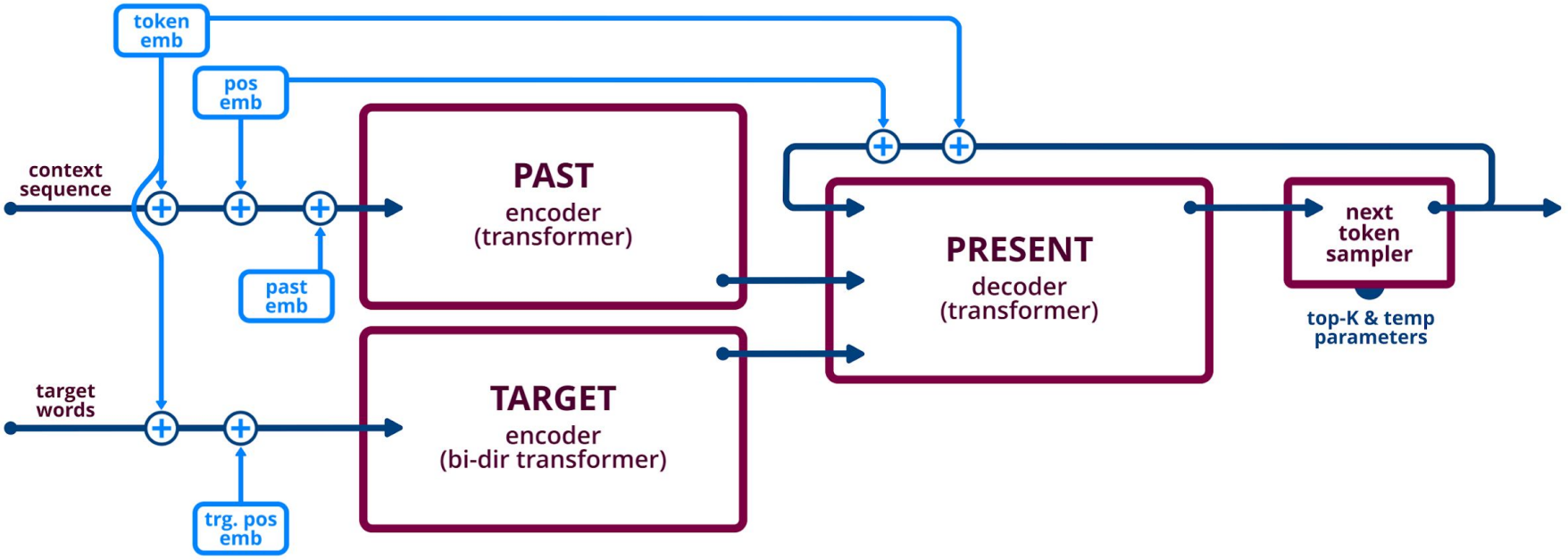


text generation with GEM

GPT Architecture



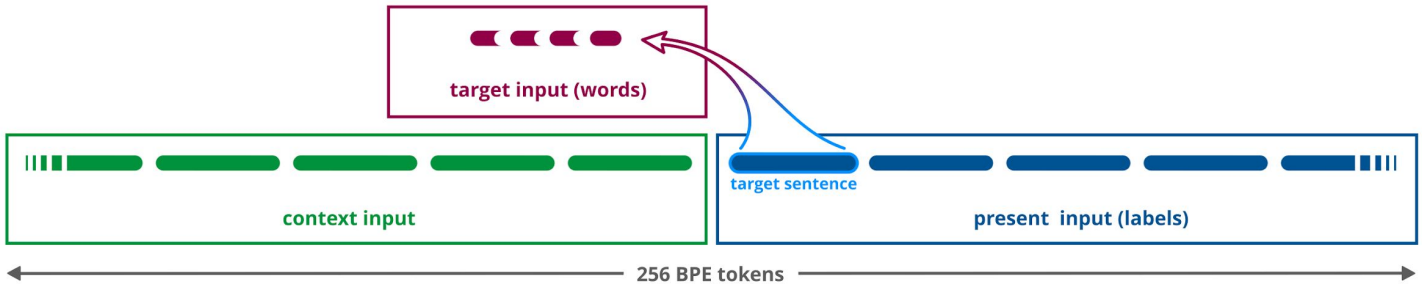
GEM Architecture



GEM Training Sample

GPT model was trained on English corpora. We have fine-tuned GEM on FEVER Wikipedia dump data (about 30M sentences).

GEM model works with BPE tokenizer. Single training sample was limited to 256 tokens, which on average corresponds to 10 sentences.



Sample preparation procedure:

1. select a **random target sentence** from corpora
2. select 20-60% of **words** from that sentence for a **target input**
3. add some **random words** (noise) to a target input
4. **neighbouring sentences** build a **context input** and the rest of a **present input** (gold labels)

Procedure of Claims Generation

Wikipedia articles have a **hypertext form** with **references** to other articles.

Claims generation procedure:

1. **Wiki-A** article was randomly selected.
2. **Wiki-B** article was picked from the articles **hyperlinked** in the first five sentences of wiki-A.
3. **Target words** were randomly chosen from the second sentence of wiki-B.
4. **Context** was composed of **mixed** wiki-A and wiki-B sentences.
5. **First sentence** from generated output was taken to serve as a claim.

Generated claims were further filtered. We removed sentences:

- not ending with a dot (bad tokenization)
- very long or very short
- too similar to the first sentence of wiki-A (measured with Levenshtein distance)
- containing numbers and dates not appearing in the wiki-A
- containing OOV words (Wikipedia dictionary)

Examples of Generated Claims

wiki-A: **Remmina**
wiki-B: **Remote Desktop Protocol**
context: Remmina is a remote desktop software client for POSIX-based computer operating systems. Remmina is in the package repositories for Debian versions 6 (Squeeze) and later and for Ubuntu versions since 10.04 (Lucid Lynx). As of 11.04 (Natty Narwhal), it replaced tsclient as Ubuntu's default remote desktop client. The FreeBSD ports/package collection also contains it as a separate port and additional protocol-specific plugin ports. Remote Desktop Protocol (RDP) is a proprietary protocol developed by Microsoft, which provides a user with a graphical interface to connect to another computer over a network connection. It supports the RDP, VNC, NX, XDMCP, SPICE and SSH protocols.
target: RDP client for must run software
generated claim: Remmina is a standalone **RDP client** for Windows and **must run** as a user on a Linux system, or the client **software** will be unavailable.

wiki-A: **Joseph Cao**
wiki-B: **Republican Party (United States)**
context: Ánh Quang "Joseph" Cao (['gaU]; Cao Quang Ánh born March 13, 1967) is a Vietnamese American politician who was the U.S. Representative for from 2009 to 2011. In April 2011, Cao announced his candidacy for the office of Attorney General of Louisiana; however, in September 2011 he pulled out of the race, and the incumbent Buddy Caldwell ran unopposed for a second term. He is the first Vietnamese American to serve in Congress, and the first and thus far only Republican from his New Orleans-based district since 1891. In December 2015, he announced that he would run for the open U.S. Senate seat being vacated by retiring fellow Republican David Vitter in 2016. The Republican Party, commonly referred to as the GOP (abbreviation for Grand Old Party), is one of the two major contemporary political parties in the United States, the other being its historic rival, the Democratic Party. He is a member of the Republican Party.
target: The party named dominant value during
generated claim: Joseph Cao was elected to Congress in 2009 and has **named** a number of prominent **Republicans*** to be the **dominant value** players **during** his time in the House.

* Republicans ~ party

Conclusions and Future Work

1. GEM was successfully **fine-tuned** on Wikipedia, **kept** original GPT knowledge, and **gained** additional controlling input.

2. GEM generated **cohesive**, **well-structured** samples, which were **challenging** for automated verification.

3. GEM is **biased** toward factual inaccuracies.
   (We have provided 155 claims, 51 of them written by humans)

4. Generation of complex claims **supported** by Wikipedia seems to be an **interesting challenge.**

Thank You

SAMSUNG