

Künstliche Korrespondenz

Wie man nichtmenschliche Schreiberlinge entlarvt

Computerlinguistik

Natural Language Processing

Supertrolle am Start

Open AI warnt vor der eigenen Entwicklung

Neugierig machen gehört zum Geschäftsmodell der Digitalwirtschaft im Allgemeinen und von Silicon Valley im Besonderen.

Künstliche Intelligenz ist gefährlich, das verantworten können wir nicht. Ein kleines Stückchen künstliche Intelligenz kündigte die von Altman

Forschungsorganisation OpenAI in wenigen Tagen ein Sprachmodell, das künstlichen

SPIEGEL Netzwelt



"Talk to Transformer"

Diese künstliche Intelligenz schreibt beängstigend gut

US-Wissenschaftler haben eine künstliche Intelligenz erschaffen, die Texte schreiben kann. Eigentlich nichts Besonderes - wären die Ergebnisse nicht so gut.

Spam auf Steroiden

MICHAEL MOORSTEDT

Wenn künstlich erzeugte Suchmaschinen fluten
NETZKOLUMNE

Neue Zürcher Zeitung

Schon bald wird es keine Suchmaschinen mehr geben. Textproduktion mehr

erfordert Klagen gegen die Kanäle. Bald noch mehr. Ist das bereits heute der Fall? Und wie sieht es mit der Zukunft aus?



Das Mensch-Maschinos

Werden schriftstellernde Algorithmen bald die besseren Geschichtenerzähler sein? Von Miriam Meckel

Mit Lügen spielen

US-Forscher veröffentlichen Fake-News-Generator

Angela Merkel möchte der Rücken kehren. Tippt man ein, so fügt das ambitionierte deutsche Ansichten motiviert in Merkels aus dem Europäischen Parlament.

keine guten Ideen für eine Verfassung

Die Maschine schreibt selbst

CHRISTIAN BOS ÜBER EINEN ALGORITHMUS, DER ANGST MACHT

Die vom Tech-Milliardär Elon Musk finanzierte Organisation OpenAI hat sich auf die Fahnen geschrieben, künstliche Intelligenz zum Wohl der Gesellschaft zu entwickeln; ihre Programme stellt sie deshalb der Öffentlichkeit zur Verfügung. Bis jetzt. Nun jedoch haben die Informatiker einen Text generierenden Algorithmus entwickelt, dessen Code sie fürs Erste der Öffentlichkeit vorenthalten wollen. Weil er ihnen selbst nicht geheuer ist. Das GPT2 genannte Programm wurde

Leitfragen

1. Warum sind maschinell geschriebene Texte relevant?
2. Warum sind sie – auf den ersten Blick – so gut geworden?
3. Wie können wir sie erkennen?

DEMO: Grover

Potentiell nützliche Anwendungen

- Auto-Vervollständigung
- Hilfe bei Grammatik und Rechtschreibung
- Chatbots und Frage-Antwort-Systeme
- Unterhaltung und Kunst

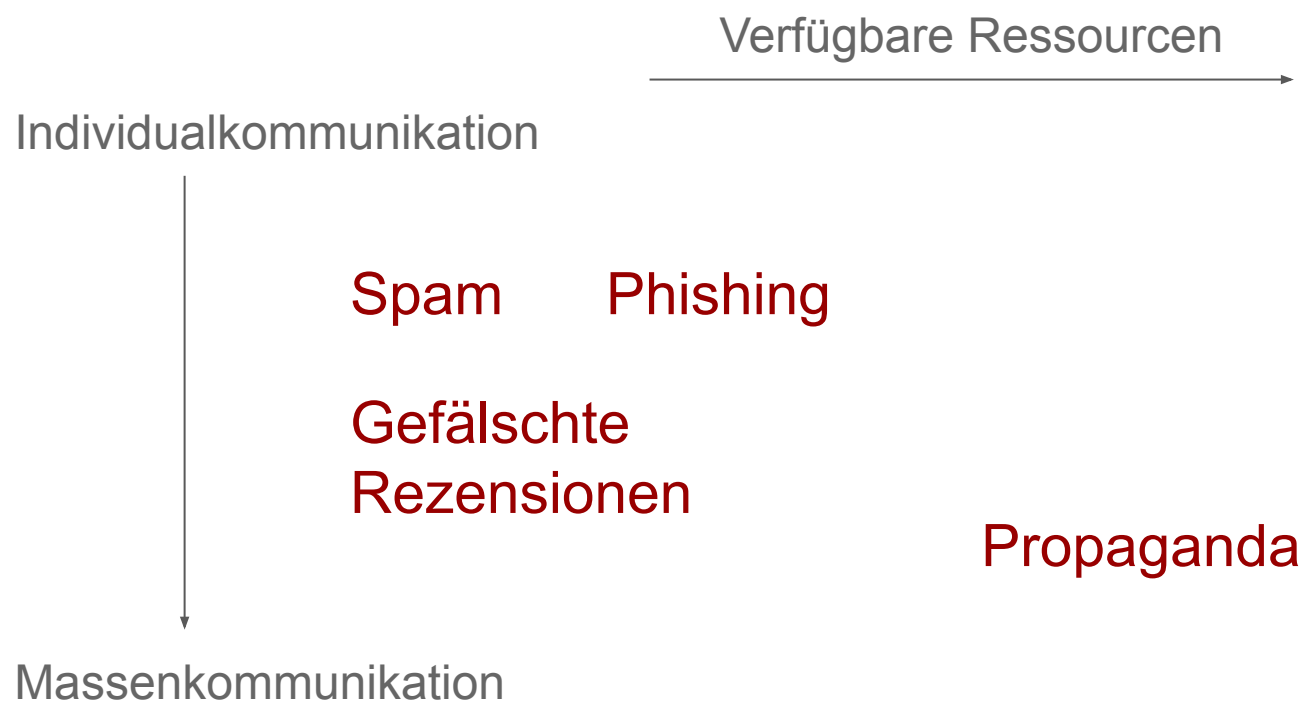
Ausgangslage

Es wird **immer günstiger**, eine beliebig grosse Zahl von diversen Texten zu produzieren, die einem Stil oder einer Themenvorgabe folgen.

Die **Qualität** der Texte nimmt tendenziell zu.

Potentiell schädliche Anwendungen

Verfügbare Ressourcen



Individualkommunikation

Spam Phishing

Gefälschte
Rezensionen

Propaganda

Massenkommunikation

Technische Grundlagen

Experiment: Kollektives Schreiben

uzh.voting/gpt

Frage: Welches **Wort** kommt wahrscheinlich in die Lücke?

“Eine Rose ist eine Rose ist eine Rose ist eine _____”

Next Word Prediction

Mit einem Lächeln überreichte er ____

Mit einem Lächeln überreichte er ihr ____

Mit einem Lächeln überreichte er ihr eine ____

Markov (1913)

«Мой дядя самых честных правил,
Когда не в шутку занемог,
Он уважать себя заставил
И лучше выдумать не мог.
Его пример другим наука;
Но, боже мой, какая скука
С больным сидеть и день и ночь,
Не отходя ни шагу прочь!

...

1	2	3	4	5	6	7	8	9	10		1	2	3	4
1	мой	дядя	самых	честных	правил,						1	на	н	
2	х	тест	ных	н	3						2	и	с	
3	равил	когда	4								3	и	же	
4	не	в	шутку	за	4						4	о	л	т
5	не	мо	го	ну	ва	5					5	о	м	и
6	жа	т	се	бя	за	с	4				6	г	и	н
7	та	в	и	ли	д	у	т	ш	4		7	о	й	п
8	ев	н	о	у	м	а	т	н	е	5	8	о	д	и
9	мо	го	го	т	р	и	ч	и	4		9	е	г	а
10	р	д	р	у	г	и	н	а	4		10	в	м	о
3	х	т	ест	ных	н	3	5	4	3	5	4	2	5	4
0	2	0	0	0	0	1	0	2	4	18	24	0	1	1
у	ка	на	бо	же	м	5						о	д	и
о	й	ка	ка	з	ск	у	6					и	б	л
ка	с	бо	л	ны	ж	с	3					о	й	т
и	д	и	а	д	е	н	и	н	5			а	м	и
и	д	и	а	д	е	н	и	н	5			и	л	и

Markov (1913), Shannon (1948)

Länge des Kontexts: 1 Buchstabe

Theg sheso pa lyiklg ut. cout Scrpauscricre cobaives wingervet Ners, whe ilened
te o wn taulie wom uld atimore rteansou roocono weveiknt hef ia ngry'sif farll t
mmat and, tr iscond frnid riliofr th Gureckpeag

Markov (1913), Shannon (1948)

Kontext: 3 Buchstaben

At oness, and no fall makestic to us, infessed Russion-bently our then a man
thous always, and toops in he roquestill shoed to dispric! Is Olga's up. Italked fore
declaimsel the Juan's conven night toget nothem,

Markov (1913), Shannon (1948)

Kontext: 5 Buchstaben

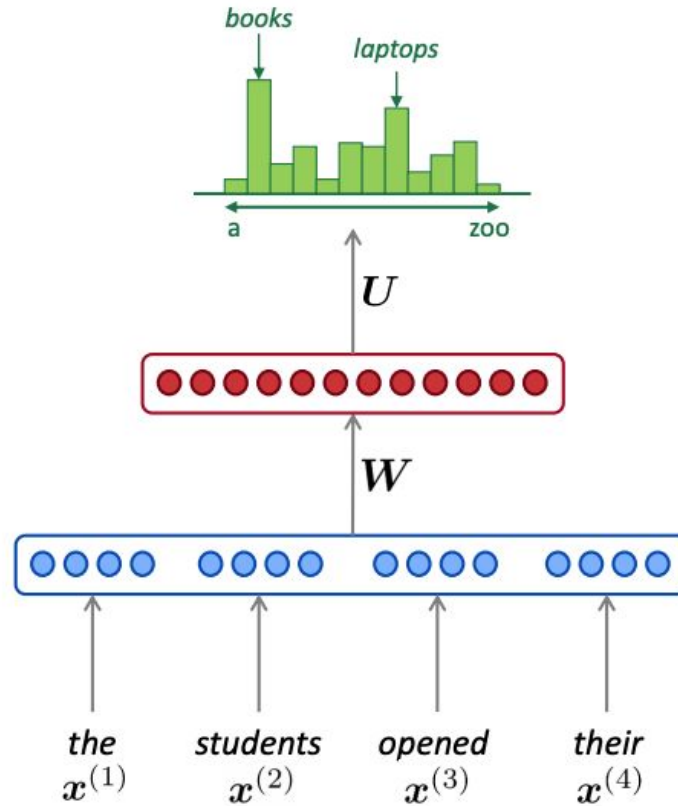
Meanwhile with jealousy bench, and so it was his time. But she trick. Let message we visits at dared here bored my sweet, who sets no inclination, and Homer, so prose, weight, my goods and envy and kin.

Markov (1913), Shannon (1948)

Kontext: 7 Buchstaben

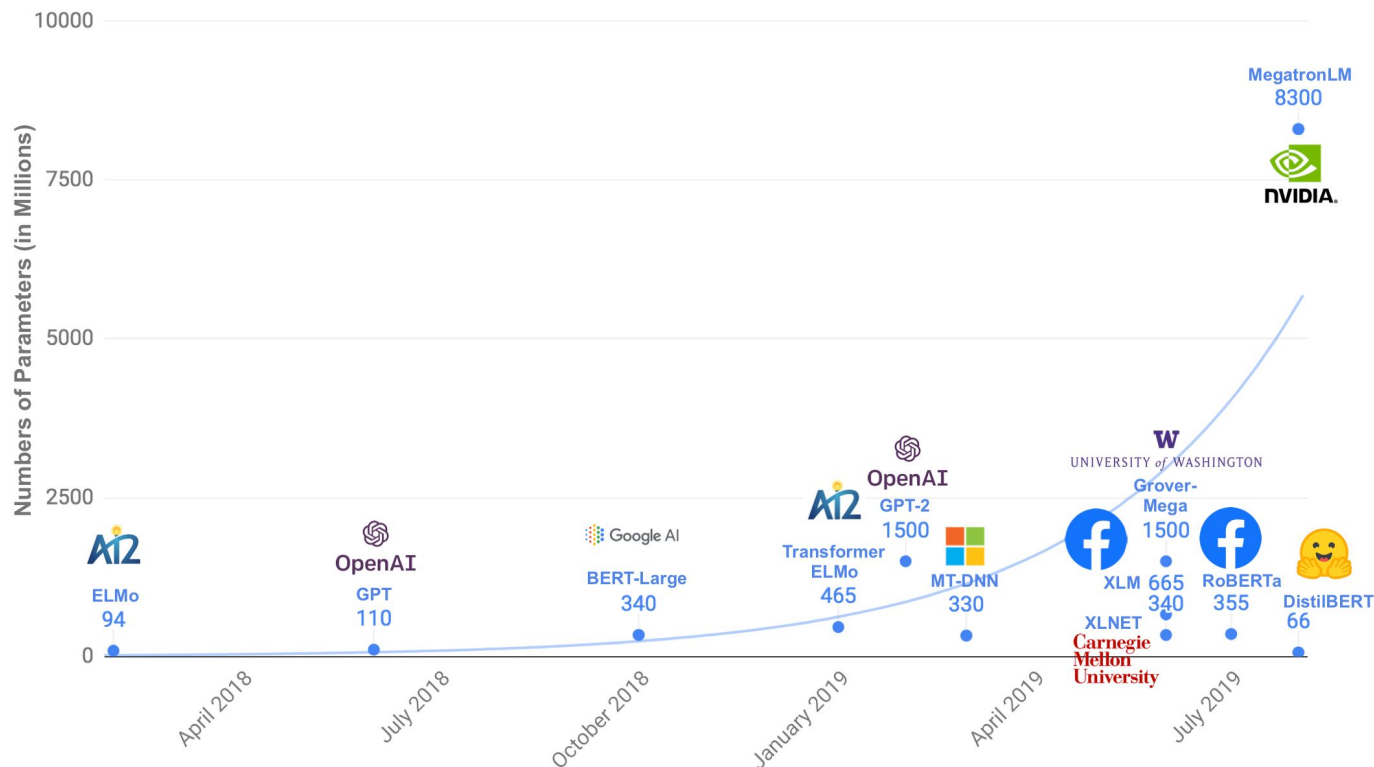
My sorrow her breast, over the dumb torment of her veil, with our poor head is stooping. But now Aurora's crimson finger, your christening glow. Farewell. Evgeny loved one, honoured fate by calmly, not yet seeking?

Neuronale Netzwerke

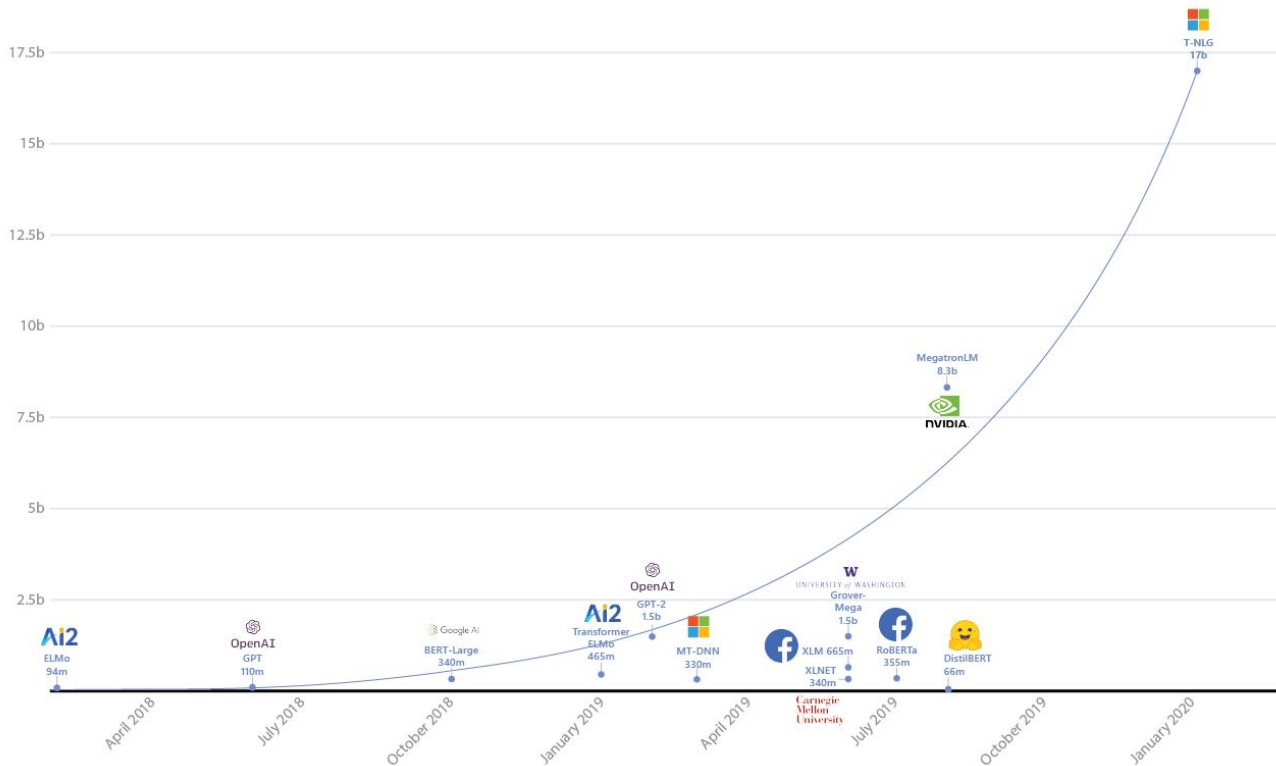


(Bild: Stanford CS224N)

Letzte zwei Jahre



Letzte Woche



GPT-2: Stufenweise Veröffentlichung

Februar 2019: 124 Millionen Parameter

Mai 2019: 355 Millionen Parameter

August 2019: 774 Millionen Parameter

November 2019: 1.5 Milliarden Parameter

Anpassung an Genres

Websites
Bücher
Zeitungen
Social Media



Allgemeines Modell

Fine-Tuning



Genre-spezifisches Corpus



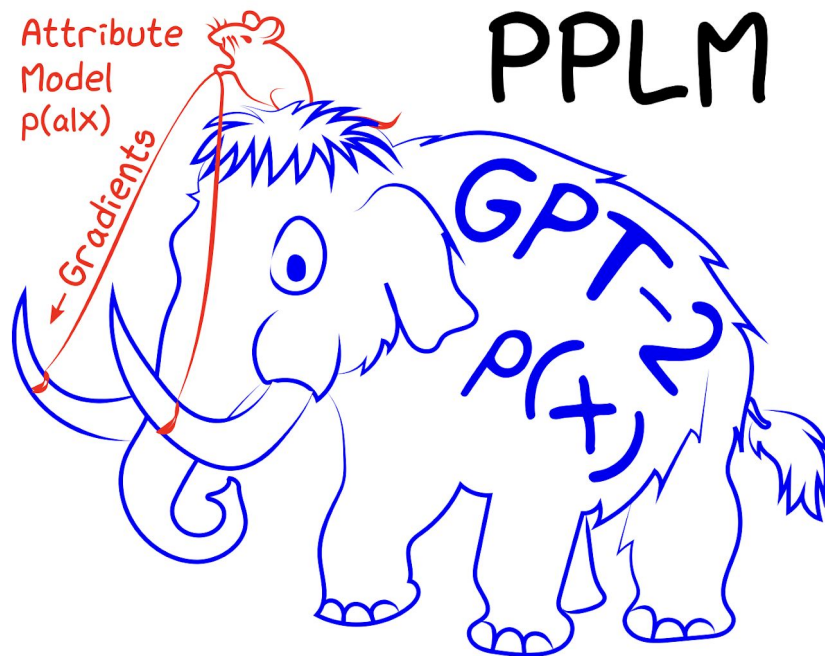
Spezialisiertes Modell

Kontrolliertes Schreiben

[-] The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though...

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...



Demo: GPT-2 mit Themenvorgabe

Erkennen von maschinell geschriebenen Texten

Wie gut sind menschliche Probanden?

(Ippolito et al. 2019)

- Zufällige Texte aus dem Internet und gleich viele GPT2-Texte
- Probanden: Studierende, die vorher 10 Beispiele gesehen haben

Wie gut sind menschliche Probanden?

(Ippolito et al. 2019)

**The majority of experts agree that the asteroid
2012 DA14 should be found as...**

Instructions

In this task you will be shown some text that was extracted from a website.

Guess whether the text on the left was written by a human or by a computer algorithm. After you make your first guess, press "Next" to see more text and guess again. You'll do this for 5 times. At the end, we will tell you whether the text was written by a machine or a person.

Some indicators you can look for that text might be machine-generated:

- Contradicts itself
- Incoherent
- Repetative

Note that both human- and machine- written text might contain unusual formatting such as text in all capitals, no line-break between a title and article contents, and misaligned quotation symbols. DO NOT base your decision on formatting weirdnesses.

(1) What do you think the source is?

- ☐ Definitely Human-written
- ☒ Possibly Human-written
- ☐ Possibly Machine-generated
- ☐ Definitely Machine-generated

Next

Wie gut sind menschliche Probanden?

(Ippolito et al. 2019)



The majority of experts agree that the asteroid 2012 DA14 should be found as soon as it passes by the Earth, and so should be a 'course correction' in the best of times: It should disintegrate into multiple pieces, and from each slowly drift over a set path towards the sun.

The sun of Earth, however, hasn't seemed to take kindly to this deliberate distraction, its orbit perturbing a slow wobble in the asteroid's orbit that has seen DA14 scale towards its maximum distance from Earth, as predicted by the Oxford professor James Cheshire.

"No matter...

Instructions

In this task you will be shown some text that was extracted from a website.

Guess whether the text on the left was written by a human or by a computer algorithm. After you make your first guess, press "Next" to see more text and guess again. You'll do this for 5 times. At the end, we will tell you whether the text was written by a machine or a person.

Some indicators you can look for that text might be machine-generated:

- Contradicts itself
- Incoherent
- Repetative

Note that both human- and machine- written text might contain unusual formatting such as text in all capitals, no line-break between a title and article contents, and misaligned quotation symbols. DO NOT base your decision on formatting weirdnesses.

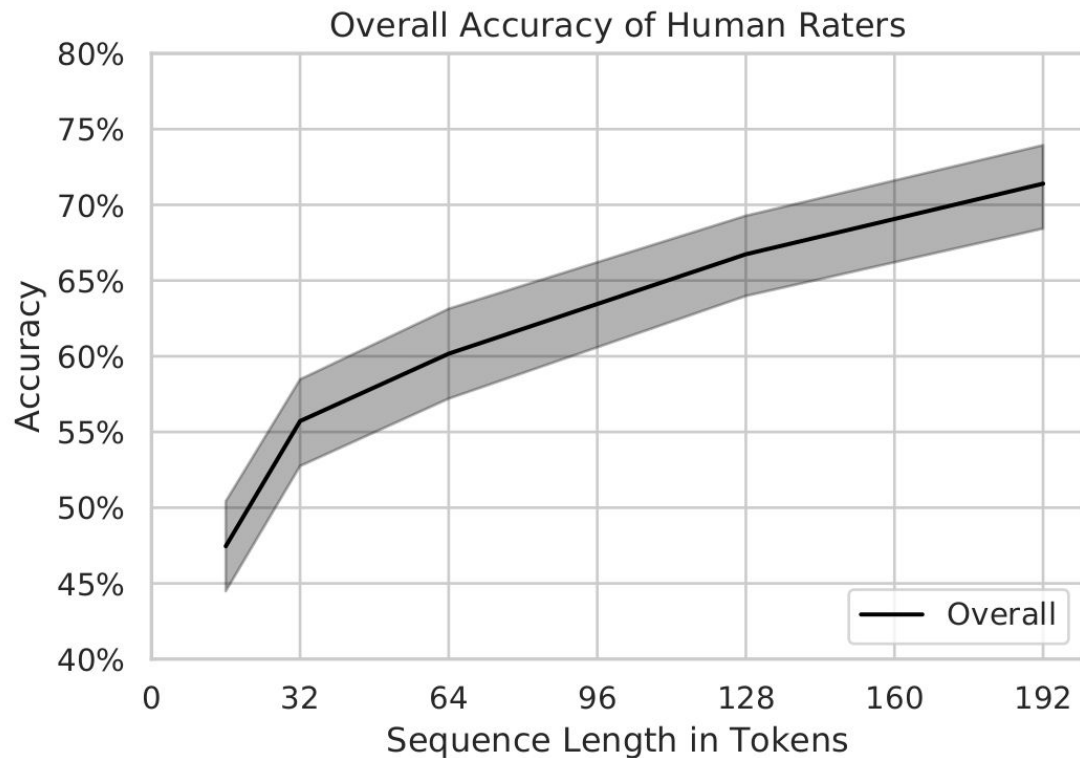
(4) What do you think the source is?

- ☐ Definitely Human-written
- ☐ Possibly Human-written
- ☐ Possibly Machine-generated
- ☒ Definitely Machine-generated

Next

Wie gut sind menschliche Probanden?

(Ippolito et al. 2019)



Sind grössere Modelle schwerer zu entlarven?

(Kreps and McCain 2019)

- GPT2-Texte über Konflikt zwischen USA und Nordkorea
- Frage an Crowdworkers:
Welche sind glaubwürdig? Korrekt?
Verständlich?

The New York Times

North Korea Envoy Warns U.S. Over 'Outrageous' Ship Seizure



The North Korean vessel, the Wise Honest, was seized by the United States and towed into harbor in American Samoa. U.S. Department of Justice

By Michael Schwartz

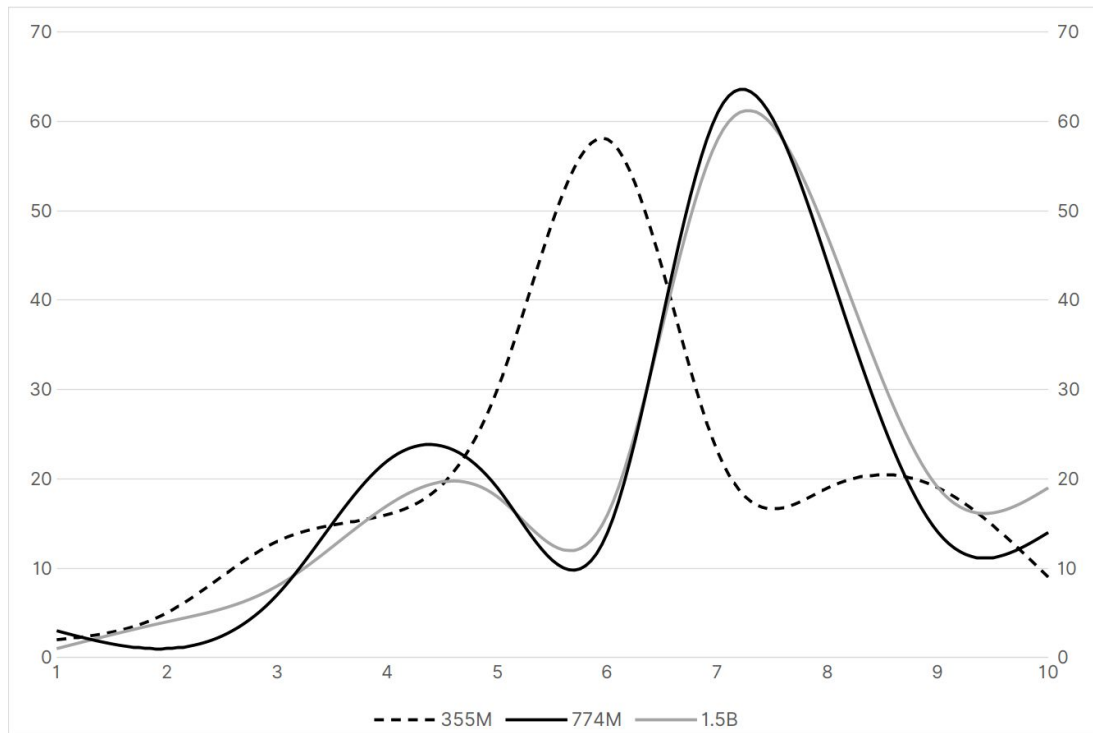
May 21, 2019



UNITED NATIONS — North Korea's ambassador to the United Nations held a rare news conference on Tuesday to warn the United States that its seizure of a North Korean cargo ship could imperil any future disarmament negotiations between the two countries.

Sind grössere Modelle schwerer zu erkennen?

(Kreps and McCain 2019)



Kommentare der Probanden

(Kreps and McCain 2019)

Kleines Modell

- “Der Artikel wirkt sehr gut recherchiert”
- “Er enthält Zitate von mehreren Regierungen und Behörden”
- “Spricht von zwei grossen Atomkriegen zwischen USA und Nordkorea”
- “Die Geschichte klingt wie etwas, was Trump zulassen würde”

Grössere Modelle

- “Der Artikel behauptet, die USA seien nicht im Sicherheitsrat”
- “Es hat mehrere Widersprüche im Artikel”
- “Der Artikel bleibt objektiv”
- “Realistisch und gut genug geschrieben, damit ich die Geschichte glaube”

Sprachliche Indikatoren

- Inkorrektheit und Inkohärenz
- Repetition und Degeneration
- Mimetizität statt Ikonizität
- Banalität

Inkorrektheit und Inkohärenz

Swiss Elections Rigged: U.N.

The United Nations announced on Tuesday that there was “credible evidence” that the presidential elections in Switzerland, in which incumbent Micheline Calmy-Rey won by a thin margin, were rigged.

The results were agreed to by the French-speaking nation’s two main parties, The Swiss People’s Party, which called for the elections to be canceled, and the centrist Christian Democratic Party, which held them.

Repetition

In Zürich gibt es ...

*eine Filiale der Schweizer Post. Sie wurde im Jahre 1909
gegründet und ist damit die älteste und älteste Poststelle der
Schweiz. Die Filiale in Zürich ist das Stammhaus einer Post-
Gesellschaft der Schweizerischen*

Degeneration

Heute spreche ich über ...

ein Drittel der Leute, die sich nicht mehr für Politik interessieren. Sie sind nicht mehr für die Politik zuständig. Sie sind für die Kultur zuständig. Die Medien. Sie machen es, weil die Leute sich für Politik interessieren. Sie hören die Leute. Die haben das nicht. Sie haben das nicht. Sie haben die Leute auf die Barrikaden gebracht. Sie haben die Barrikaden beseitigt. Sie sind nicht mehr für die Politik zuständig. Sie sind nicht mehr für die Kultur zuständig.

Mimetizität statt Ikonizität

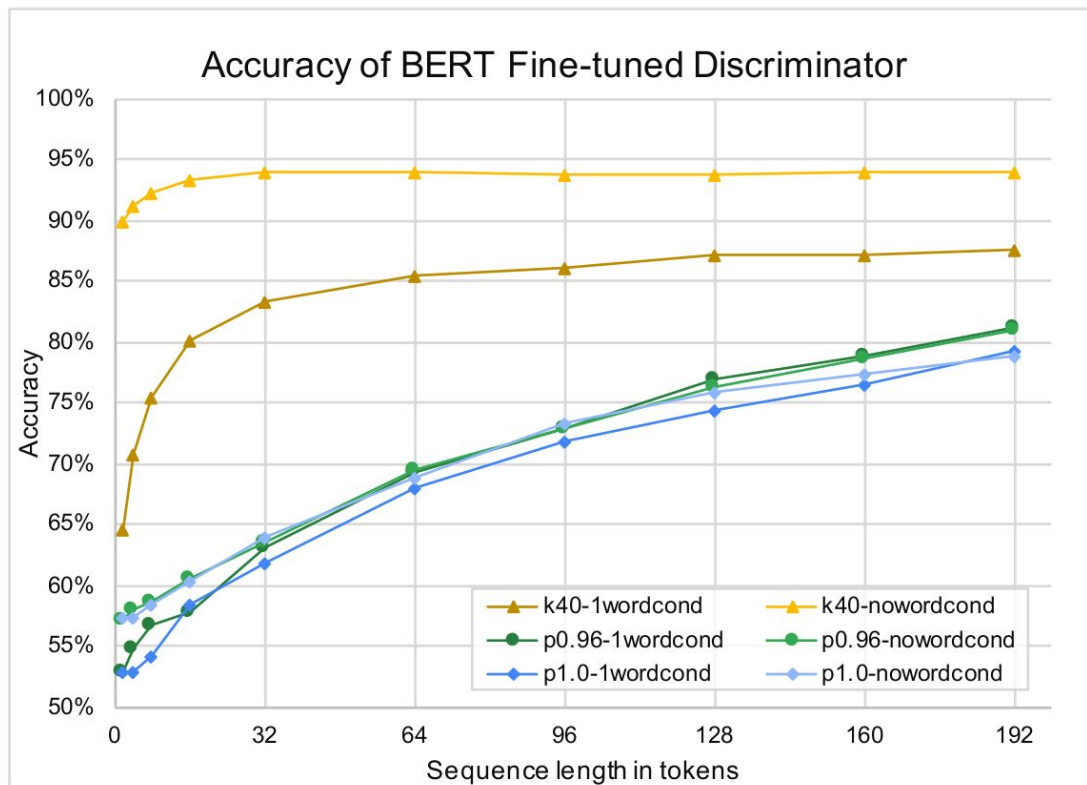
I walked up the path to the fence and saw a plump dog, that had been a common visitor to the Finca Vigia before the war, galloping up a path to the main building, with a tiny cow of the same name standing by her side. There was a puddle of red gravy in the front yard, and Hemingway sat down in a lawn chair.

Banalität

Es ist sehr interessant, dass ...

sich einige der alten, aber immer noch im Bestand befindlichen Exemplare im Besitz der Universitätsbibliothek befinden. Das Museum wurde im Jahr 2001 eröffnet. Die Sammlung enthält etwa 2500 Objekte. Sie zeigt unter anderem die Sammlung von Johann Friedrich von Gmelin, Johann Friedrich Kuntzemüller

Automatische Erkennung (Ippolito et al. 2019)



Human-Written

The programme operates on a weekly elimination process to find the best all-around baker from the contestants, who are all amateurs.

Generated

The first book I went through was The Cook's Book of New York City by Ed Mirvish. I've always loved Ed Mirvish's recipes and he's one of my favorite chefs.

Ausblick

- Herausforderung 1:
Menschliche Texte, die von einer Maschine korrumpiert wurden
- Herausforderung 2:
Texte, die von Menschen mit maschineller Hilfe geschrieben wurden

Diskussion

Literatur

Gehrmann, Sebastian, Hendrik Strobelt, and Alexander M. Rush. "GLTR: Statistical Detection and Visualization of Generated Text." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019.

Ippolito, Daphne, et al. "Human and Automatic Detection of Generated Text." arXiv preprint arXiv:1911.00650 (2019).

Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI Blog 1.8 (2019): 9.

Solaiman, Irene, et al. "Release strategies and the social impacts of language models." arXiv preprint arXiv:1908.09203 (2019).

Zellers, Rowan, et al. "Defending against neural fake news." Advances in Neural Information Processing Systems. 2019.