KI generierte Texte erkennen

Dürrmeier,Wagner

19. Januar 2024

Eingrenzungen

- Behandeln von flüssigen Texten
- Konzentration auf den Lehrkontext
- nur generative Modelle bis Mitte 2023, da für spätere Modelle noch nicht genügend Studienlage vorhanden ist
- bei allen Erkennungsmethoden und Merkmalen gilt grundlegend: je länger/größer der Text desto genauer

Funktion

- **Natural Language Processing:** komplette Maschine
- **Natural Language Understanding:** zieht sich die benötigten Informationen aus eingegebenen Text
- **Natural Language Generation:** nutzt ein Language Model (LM) um Antwort in natürlicher Sprache zu generieren

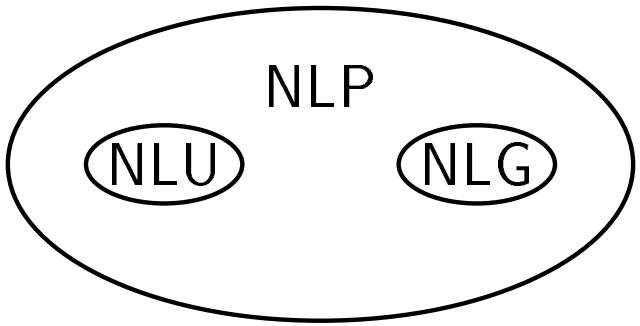


Abbildung: nach Maddula; 2022; *Medium* „NLP, NLU, and NLG"

Arten von LMs

- **neuronale LMs:**[1] neuronale Netze und Maschine Learning
- **wissensbasierte LMs:**[1] linguistisches Wissen
- **statistische LMs:**[1] Auftrittswahrscheinlichkeiten und Wahrscheinlichkeitsverteilungen

[1] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]

statistische Erstellungsmethoden

- $P(w)$ beschreibt Wahrscheinlichkeit, dass auf ein Wort Token w folgt
- V ist die Menge aller möglich folgenden Tokens
- **Top-k-Sampling:**[2] Auswahl von k Tokens, die zusammenaddiert eine möglichst hohe Wahrscheinlichkeit ergeben ($\Sigma w \in V | P(w)$) und zufällige Auswahl aus der erstellten Menge

[2] Rieke; 2023; [10.13140/RG.2.2.35567.00163.]

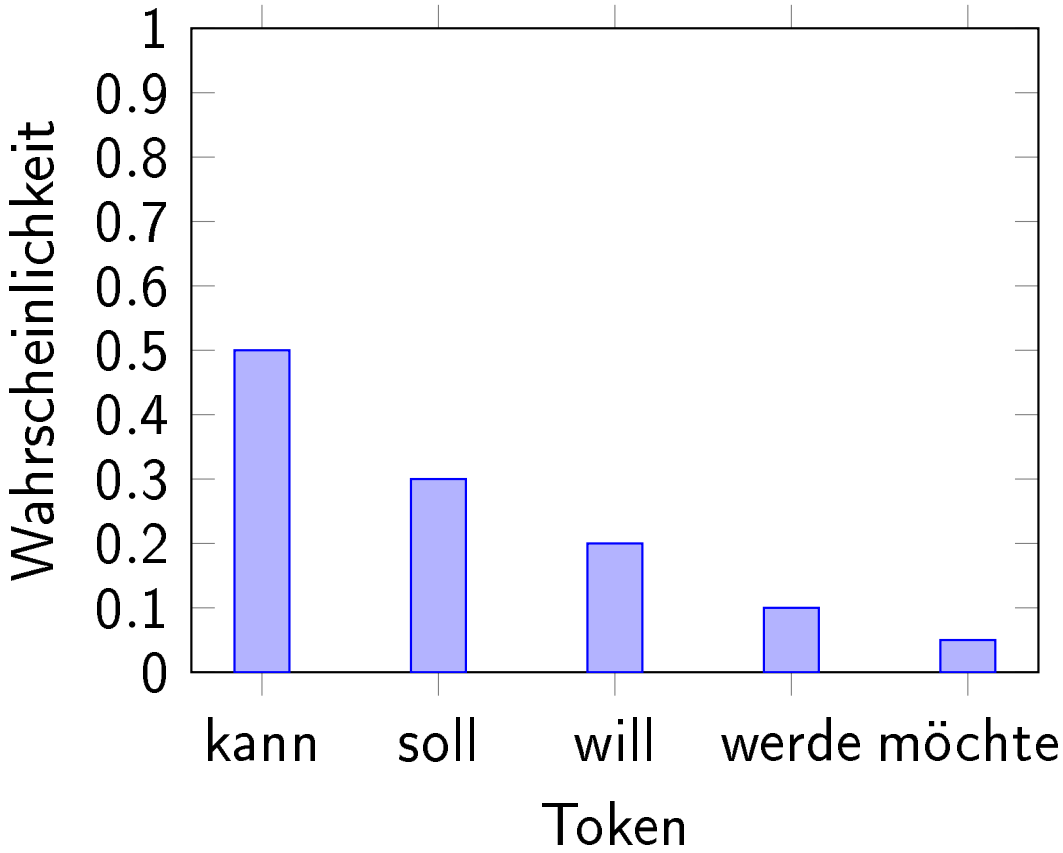
statistische Erstellungsmethoden

- $P(w)$ beschreibt Wahrscheinlichkeit, dass auf ein Wort Token w folgt
- V ist die Menge aller möglich folgenden Tokens
- **Top-k-Sampling:**[2] Auswahl von k Tokens, die zusammenaddiert eine möglichst hohe Wahrscheinlichkeit ergeben $(\Sigma w \in V | P(w))$ und zufällige Auswahl aus der erstellten Menge
- **Top-p-Sampling:**[2]möglichst kleine Menge von Tokens, die den Wert p nicht überschreiten $(\Sigma w \in V | P(w) \geq p)$ und zufällige Auswahl aus der erstellten Menge

[2] Rieke; 2023; [10.13140/RG.2.2.35567.00163.]

Beispiel Top-k-Sampling[3]

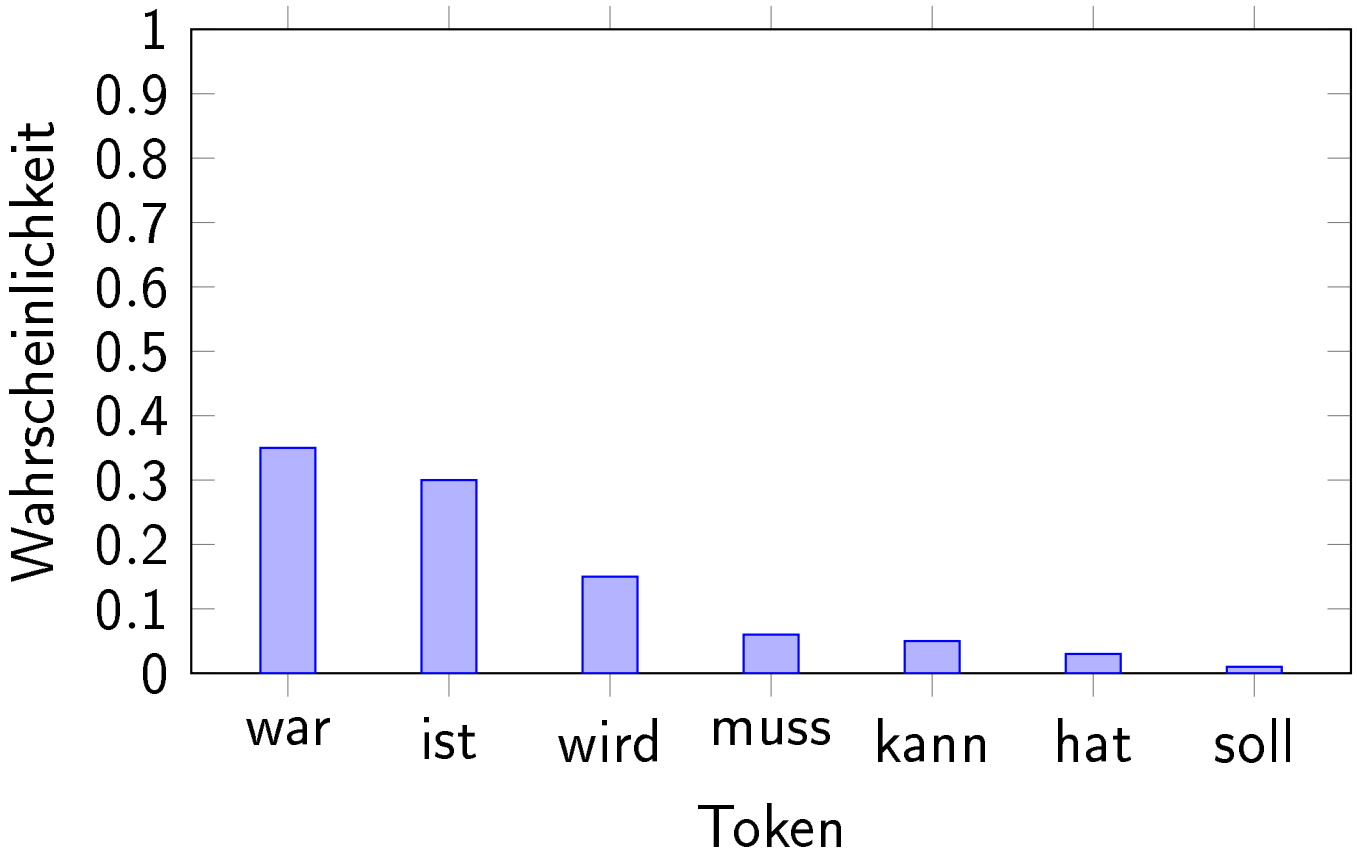
- Top-k-Sampling; Auswahl des nächsten Tokens nach „Ich“
- mit $k := 3$



[3] nach Rieke; 2023; [10.13140/RG.2.2.35567.00163.]

Beispiel Top-p-Sampling[4]

- Top-p-Sampling; Auswahl des nächsten Tokens nach „Es“
- mit $p := 0.9$



[4] nach Rieke; 2023; [10.13140/RG.2.2.35567.00163.]

Merkmale 1

- **Häufigkeitsmerkmale:**[5] bei Menschen eine Verteilung nach dem Zipfschen Gesetz

$$f \approx \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

(mit $s \in \mathbb{R} | s \geq 1$ und mit der Häufigkeit f eines Tokens des Ranges k aus N unterschiedlichen Tokens)[6]

[5] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[6] Zipf; 1949; „Human behavior and the principle of least effort."
[7] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]

Merkmale 1

- **Häufigkeitsmerkmale:**[5] bei Menschen eine Verteilung nach dem Zipfschen Gesetz

$$f \approx \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

(mit $s \in \mathbb{R} | s \geq 1$ und mit der Häufigkeit f eines Tokens des Ranges k aus N unterschiedlichen Tokens)[6]

- **Geläufigkeitsmerkmale:**[7] Konsistenz und Kohärenz des Textes; Maß dafür ist z. B. der Flesch-Index

[5] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[6] Zipf; 1949; „Human behavior and the principle of least effort."
[7] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]

Merkmale 2

- **linguistische Merkmale:**[8]
  - Messen von Phrasalverben und vergleichen mit Auflösung der Koreferenzen
  - Messen der Verteilung von Part-of-Speech-Tags und Named-Entity-Tags

[8] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[9] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[10] Vogelgesang, et.al.;Uni Hohenheim; 2023

Merkmale 2

- **linguistische Merkmale:**[8]
  - Messen von Phrasalverben und vergleichen mit Auflösung der Koreferenzen
  - Messen der Verteilung von Part-of-Speech-Tags und Named-Entity-Tags
- **komplexe phrasale Merkmale:**[9] idiomatische Phrasen bei Menschen häufiger; neuere NLPs verwenden diese inzwischen aber auch

[8] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[9] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[10] Vogelgesang, et.al.;Uni Hohenheim; 2023

Merkmale 2

- **linguistische Merkmale:**[8]
  - Messen von Phrasalverben und vergleichen mit Auflösung der Koreferenzen
  - Messen der Verteilung von Part-of-Speech-Tags und Named-Entity-Tags
- **komplexe phrasale Merkmale:**[9] idiomatische Phrasen bei Menschen häufiger; neuere NLPs verwenden diese inzwischen aber auch
- **grundlegende Textmerkmale:**[9]Häufigkeit von Satzzeichen, Länge von Texten und Absätzen sowie Fehleranfälligkeit

[8] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[9] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[10] Vogelgesang, et.al.;Uni Hohenheim; 2023

Merkmale 2

- **linguistische Merkmale:**[8]
  - Messen von Phrasalverben und vergleichen mit Auflösung der Koreferenzen
  - Messen der Verteilung von Part-of-Speech-Tags und Named-Entity-Tags
- **komplexe phrasale Merkmale:**[9] idiomatische Phrasen bei Menschen häufiger; neuere NLPs verwenden diese inzwischen aber auch
- **grundlegende Textmerkmale:**[9]Häufigkeit von Satzzeichen, Länge von Texten und Absätzen sowie Fehleranfälligkeit
- **zeitabhängige Merkmale:**[10] Aussagen über den Wissensstand einer KI hinaus sind vage und ungenau bzw. falsch

[8] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[9] Crothers, Japkowicz, Viktor; 2023; [arXiv:2210.07321v4]
[10] Vogelgesang, et.al.;Uni Hohenheim; 2023

Detektoren

- Detektion auch in Kombination von Mensch und Maschine möglich, aber bisher zu wenig Studienlage
- Vorstellen der zwei erfolgreicheren maschinellen Möglichkeiten

Maschinen

- **Zero-Shot-Ansatz:**
  - kleiner NLG wird auf Erkennungsmerkmale trainiert [11]
  - Beispiel: GROVER für Fake-News-Erkennung
  - vergleichsweise ineffizient[12]

[11] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[12] Solaiman, Brundage, et al.; 2019
[13] Liu, Ott, et al.; 2019; [arXiv:1907.11692]

Maschinen

- **Zero-Shot-Ansatz:**
  - kleiner NLG wird auf Erkennungsmerkmale trainiert [11]
  - Beispiel: GROVER für Fake-News-Erkennung
  - vergleichsweise ineffizient[12]
- **Fine-Tuning-Ansatz:**[13]
  - NLP mit Texten von Menschen und KIs als Trainingsdatensatz
  - Beispiel: RoBERTa für ChatGPT-2 Erkennung
  - im Allgemeinen sehr effizient (ja nach Trainingssatz)

[11] Crothers, Japkowicz, et al.; 2022; [arXiv:2203.07983]
[12] Solaiman, Brundage, et al.; 2019
[13] Liu, Ott, et al.; 2019; [arXiv:1907.11692]

Menschen

- Entscheidung wird auf Basis von semantischen Ungenauigkeiten getroffen[14]
- mit Experten erhält man eine Erkennungsgenauigkeit von ca. 74%[14]
- nicht trainierte Menschen erkennen im Regelfall nicht KI-Texte[15]

[14] Ippolito, Duckworth, et al.; *ACL 2020 Annual Conference*
[15] Albrecht; 2023; *API*

Probleme

- **Überarbeitung:**[16]
  - Überarbeiten von KI geschriebener Texte durch Menschen im Nachhinein
  - linguistisch basierte Merkmale fallen dadurch weg
  - KIs fehlen so wichtige Merkmale

[16] Ippolito, Duckworth, et al.; *ACL 2020 Annual Conference*
[17] Vogelgesang, et al.; Uni Hohenheim; 2023
[18] Satzung Universität Tübingen; 2023
[19] Albrecht; 2023; *API*

Probleme

- **Überarbeitung:**[16]
  - Überarbeiten von KI geschriebener Texte durch Menschen im Nachhinein
  - linguistisch basierte Merkmale fallen dadurch weg
  - KIs fehlen so wichtige Merkmale
- **Rechtssicherheit:**[17]
  - Detektionsmethoden liefern keine zuverlässige Aussagen
  - also selbst bei vermuteter KI-Autorenschaft kein Grund zum Nicht-Bestehen
  - Universität Tübingen verbietet in ihrer Satzung[18] die Nutzung von KIs im Lehrkontext $\rightarrow$ können durch fehlende Beweisbarkeit nicht durchsetzen
- Verbessern der Situation in Zukunft nicht zu erwarten[19]

[16] Ippolito, Duckworth, et al.; *ACL 2020 Annual Conference*
[17] Vogelgesang, et al.; Uni Hohenheim; 2023
[18] Satzung Universität Tübingen; 2023
[19] Albrecht; 2023; *API*

Lösung

- KIs zulassen und Aufgaben komplex genug machen (z.B. ausdrückliche kritische Auseinandersetzung mit einem Thema fordern)[20]
- KIs ausschließen, indem man neue Themen abfrägt, und erörtern lässt[21]

[20] Albrecht; 2023; *API*

[21] Vogelgesang, et al.; Universität Hohenheim; 2023

Danke

Abschließend möchte ich betonen, dass die Welt der AI-generierten Texte manchmal so raffiniert ist, dass selbst mein Kaffeekocher neidisch wäre. Aber im Ernst, die Erkennung solcher Texte ist entscheidend für die digitale Hygiene unserer Informationen. Vielen Dank für eure Aufmerksamkeit! Wenn ihr jetzt denkt, dass euer Laptop heimlich Romane schreibt, stehe ich bereit, eure Fragen zu entschlüsseln.

-ChatGPT 3.5