SPRACHMODELLE IM NATURAL LANGUAGE PROCESSING

VON FRANZISKA MEYER

INHALTE

- Natural Language Processing (NLP)
  - Beispiele
  - Prozess
- Sprachmodelle
- Entwicklung der Anwendungen

NATURAL LANGUAGE PROCESSING

NATURAL LANGUAGE PROCESSING

- zu dt.: Natürliche Sprachverarbeitung
- oft abgekürzt als NLP
- Teilgebiet der Linguistik, Informatik und Artificial Intelligence
  - Aufgabenbewältigung, für welche menschliches Denken erforderlich wäre

NATURAL LANGUAGE PROCESSING

- Definition:

  „Natural Language Processing behandelt die Verarbeitung der natürlich gesprochenen Sprache, so dass diese verstanden, interpretiert und darauf reagiert werden kann.“

- Ziel: Einwandfreie Kommunikation zwischen Mensch und Computer

NATURAL LANGUAGE PROCESSING

- Komplexität:
  - Wortbedeutungen sind nicht immer eindeutig
  - Kontext muss erfasst werden
  - Erkennbarkeit von Sarkasmus, Ironie, rhetorische Fragen
- Deshalb: Nutzung von Maschine Learning

BEISPIELE VON NLP

- E-Mail Filter
- Smart Assistants
- Suchergebnisse
- Predictive Texts
- Übersetzungen

NLP – INPUT UND OUTPUT

- Input: gesprochene Sprache (muss erst in Dokument umgewandelt werden)
- Prozess: ein annotiertes Dokument entsteht
- Output: eine geschriebene oder gesprochene Ausgabe erfolgt (über Sprachgenerierung)

ABLAUF DES NLP

- Tokenisierung
- Lexikalische/Morphologische Analyse (Stemmen/Lemmatisieren)
- Syntaktische Analyse (Parsing)
- Semantische Analyse (Named Entity Recognition)
- Diskursanalyse
- (Pragmatische Analyse)

BEGRIFFLICHKEITEN

- Token
- Morphologie
- Syntax
- Semantik
- Diskurs
- Pragmatik

TOKENISIERUNG

- Tokenisierung entspricht einer Unterteilung in logische Segmente.

Dies ist ein Beispiel

Dies | ist | ein | Beispiel

LEXIKALISCHE/MROPHOLOGISCHE ANALYSE

- Die interne Struktur der Tokens wird analysiert und die lexikalische Klasse zugewiesen
  - Stemming: Wörter werden in ihre Basisform gebracht

    „affection“, „affected“, affecting“, „affects“ - „Affect“
  - Lemmatization: Wörter werden auf ihre Bedeutung zurückgeführt

    „besser“ - „gut“

SYNTAKTISCHE ANALYSE

- Ermittlung der Satzstruktur über lexikalische Klasse
  - Ableitung der strukturellen Funktionen (Subjekt, Objekt, etc.)
  - Wie funktioniert ein Wort innerhalb eines Satzes?

SEMANTISCHE ANALYSE

- Überprüfung auf Bedeutung mittels Named-Entity-Recognition
  - Zuweisung von Entitäten zu bestimmten Gruppen

    „Microsofts (Organisation) Gründer Bill Gates (Person) war in New York (Ort).“

DISKURSANALYSE

- Beziehung zum vorigen Satz wird erkannt
  - z.B. Frage-Antwort Konstrukt, birgt Bedeutung in nächstem Satz

PRAGMATISCHE ANALYSE

- Reinterpretation der Aussage, sodass eigentliche Bedeutung erkannt wird
- erfordert Wissen aus der realen Welt

SPRACHMODELLE

SPRACHMODELLE

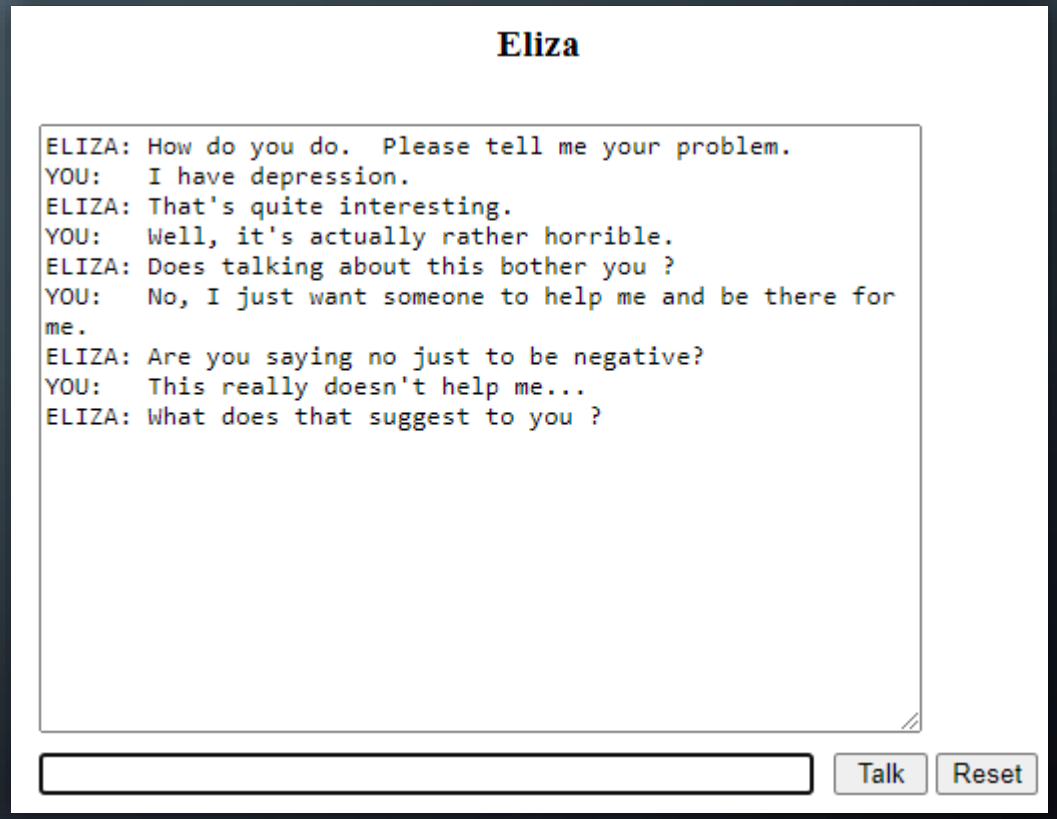
- Definition:

  „Ein Sprachmodell besteht aus einem Korpus an Textdokumenten mit dem es trainiert wurde, die Sprache zu analysieren.“

- Dient einer Auswertung oder Generation von Sprache hinsichtlich des Inputs
- Betrachtung: Entwicklung von Sprachmodellen

ELIZA - VOR DEN SPRACHMODELLEN

- ELIZA ist kein Sprachmodell, sondern „Simulation eines Psychotherapeuten“ (entspricht eher einem Chatbot)
- Basiert auf Thesaurus und vorgefertigten Antwortkonstrukten



RULE-BASED NLP

- Beginn der 1980er: handgeschriebene Regeln zum Training von Modellen
- Entspricht einer kontextfreien Grammatik

  Artikel -> Nomen |Adjektiv
- Nutzt reguläre Ausdrücke

  String „. „ entspricht Satzende

RULE-BASED NLP

- Noch in Benutzung, um Trainingsdaten vorzubereiten
- Nicht unbedingt universell anwendbar

STATISTISCHE SPRACHMODELLE

- Basieren darauf, die Wahrscheinlichkeit eines Wortes auf eine bereits existente Wortfolge zu berechnen

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

- Es werden Wörter x zu Folgewörtern von Wörtern w, wenn sie häufiger als Nachfolger von w in den Daten in Korpus auftreten

STATISTISCHE SPRACHMODELLE

- Unigram: Ein Unigram betrachtet nur sich selbst, kein Verhältnis zu vorigem Wort, es würden zufällige häufige Wörter gewählt

```
that, or, limited, the
```

- Bigram/Trigram: Bigramme betrachten ein, Trigramme zwei Wörter vor dem künftigen Input

```
outside, new, car, parking, lot, of, the, agreement, reached
```

STATISTISCHE SPRACHMODELLE

- N-Gram: beliebig viele Wörter hintereinander ketten für Wahrscheinlichkeitsberechnung
  - Aber:
    - Anspruchsvolle Berechnung
    - Begrenzter Korpus

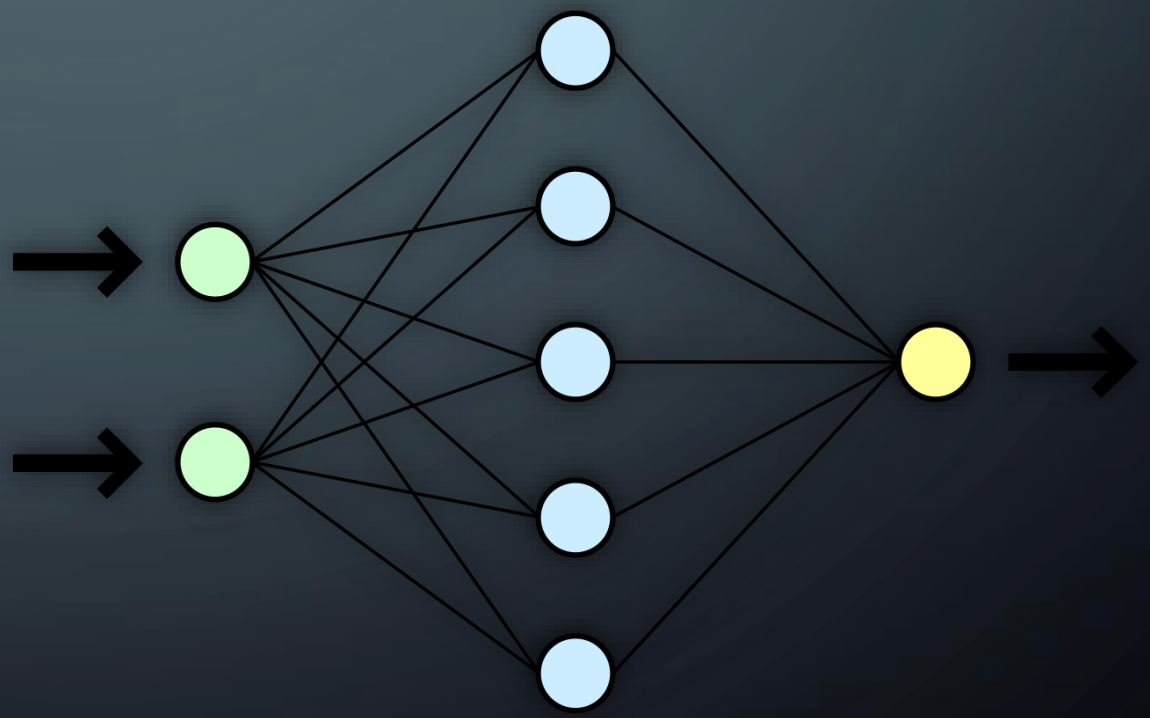
NEURONALE NETZWERKE

- Deep Learning effektiv für Sprachmodellentwicklung
  - Orientierung am menschlichen Gehirn, welches selbstständig lernt

- Maschine lernt von selbst, kein menschliches Eingreifen von Nöten

NEURONALE NETZWERKE

- Neuronale Netze bestehen aus künstlichen Neuronen, welche mit gewichteten Verbindungen vernetzt sind
- Korrelieren Entscheidungen, so tritt Lernprozess ein
- Mehr Schichten = komplexere Sachverhalte



NEURONALE NETZWERKE

- Clustering: ungelabelte Daten werden nach Ähnlichkeiten geordnet und Unregelmäßigkeiten betrachtet
- Klassifikation: für bereits gelabelte Daten, Zusammenhänge zwischen Daten und Labels soll erkannt werden
- Es lassen sich durch Neuronale Netzwerke ebenfalls Vorhersagen treffen

BERT – MLM/NSP

- Neuerung aus 2018 von Google
- BERT ist kein Sprachmodell, nutzt aber eine Neuerung
- Bidirektionales Verhalten – gesamter Input wird auf einmal erfasst

BERT – MLM/NSP

- Masked Language Modeling (MLM)

  „Versteckte“ Wörter können ermittelt werden

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

- Next Sentence Prediction (NSP)

  Sätze können auf ihre Abfolge aufeinander geprüft werden

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

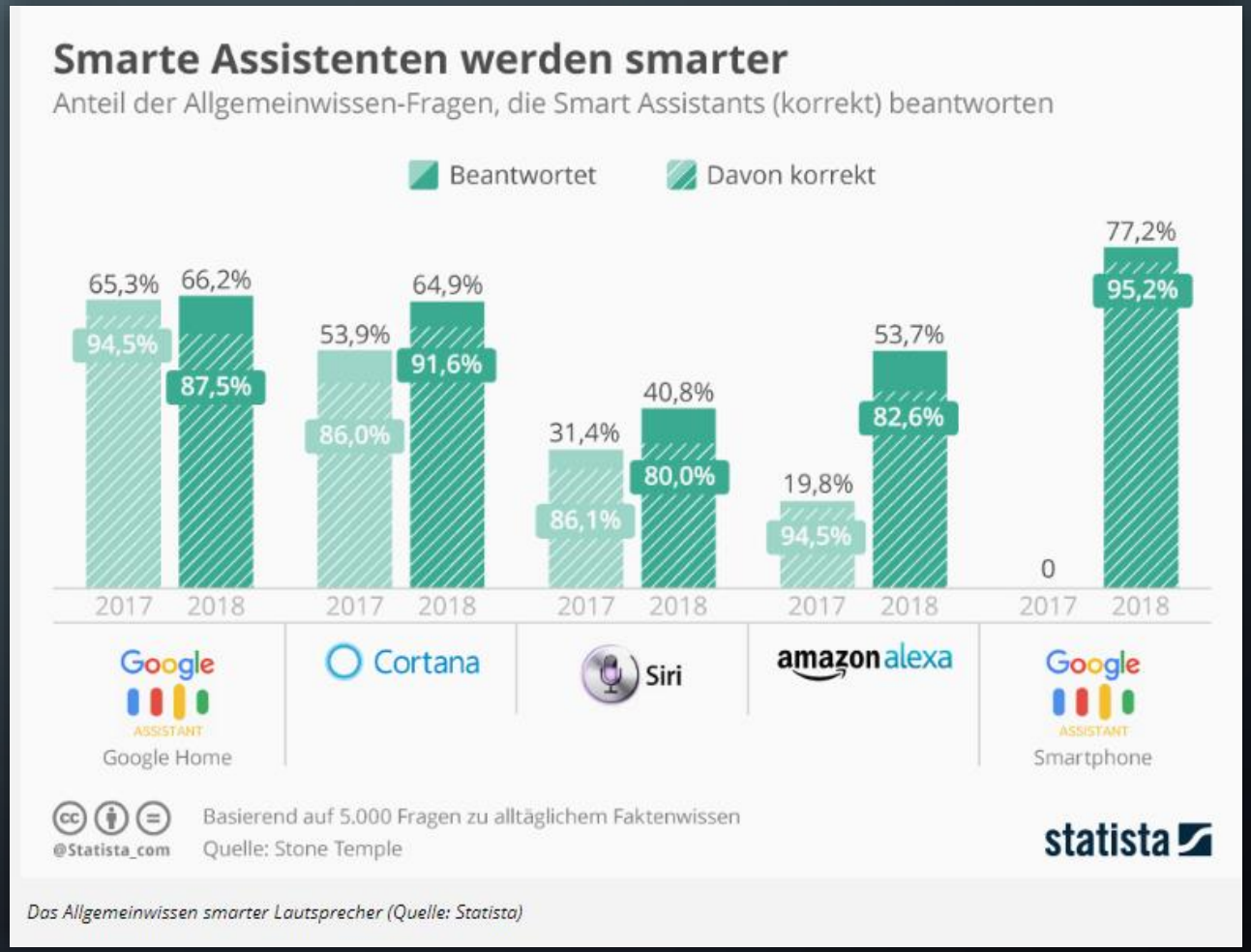
LERNFORTSCHRITT DER ANWENDUNGEN

LERNFORTSCHRITT DER ANWENDUNGEN

- Nicht nur Sprachmodelle entwickeln sich
- NLP Anwendungen lernen konstant weiter

LERNFORTSCHRITT DER ANWENDUNGEN

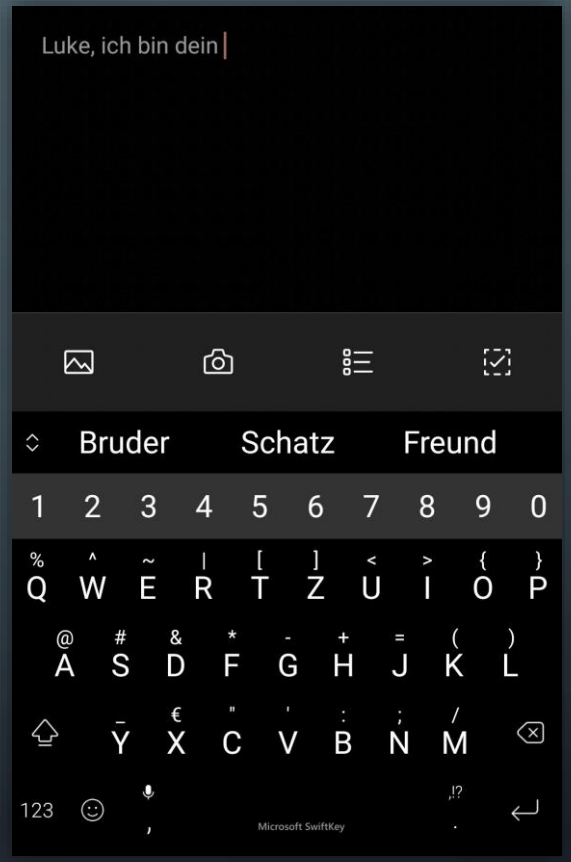
- Smartassistenten verstehen uns immer besser

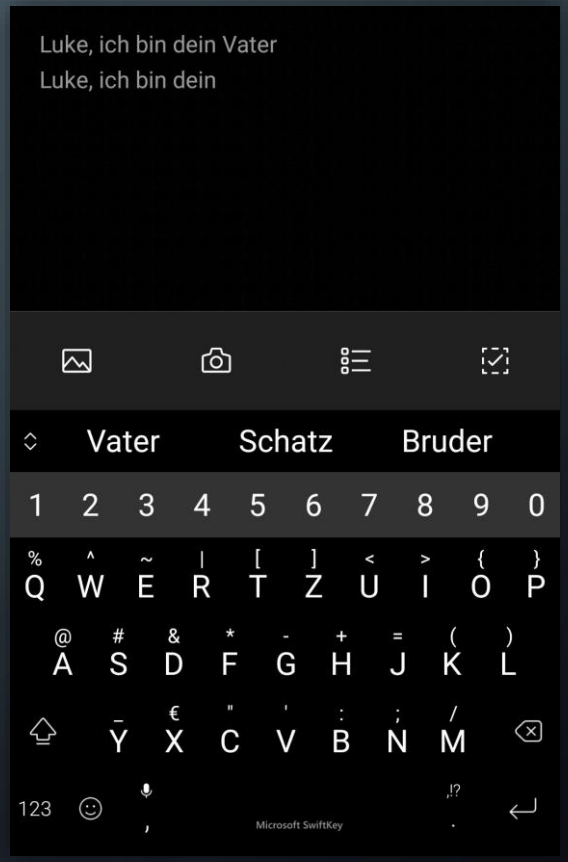


Das Allgemeinwissen smarter Lautsprecher (Quelle: Statista)

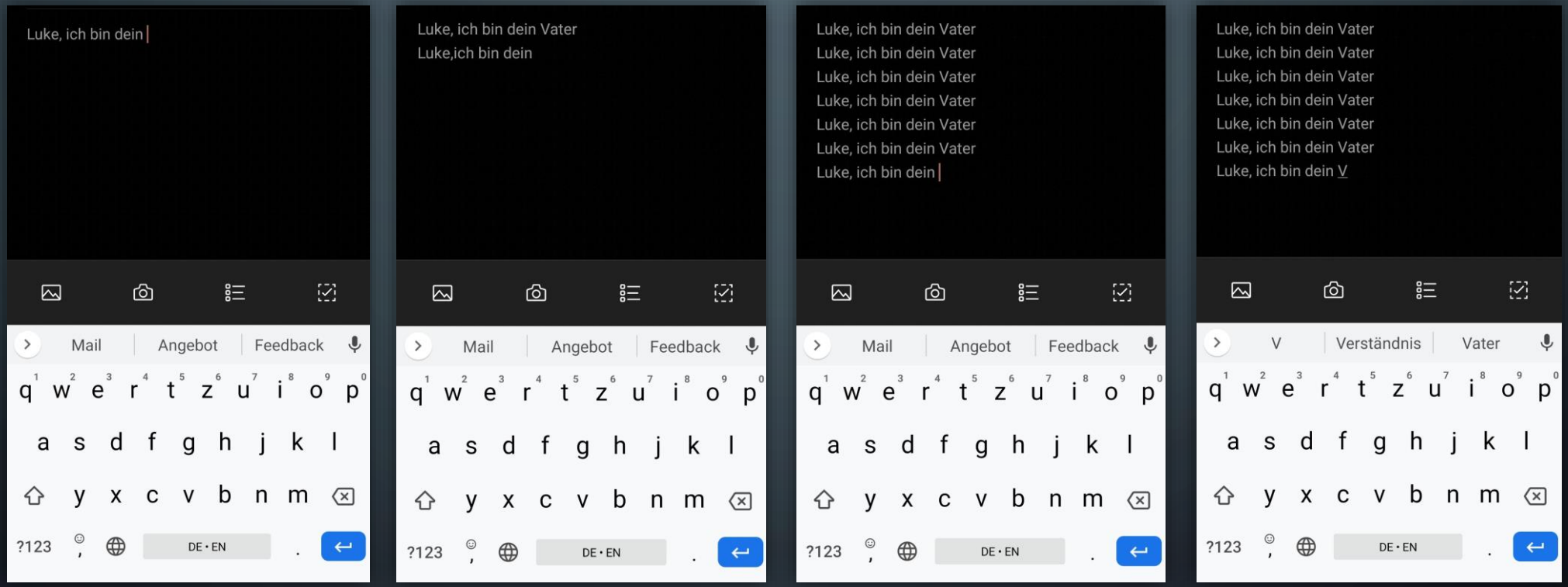
LERNFORTSCHRITT DER ANWENDUNGEN

- Predictive Texts von vers. Keyboards lernen unterschiedlich (schnell)





LERNFORTSCHRITT DER ANWENDUNGEN



QUELLEN

- www.bigdata-insider.de/was-ist-natural-language-processing-a-590102/
- www.sem-deutschland.de/inbound-marketing-agentur/online-marketing-glossar/natural-language-processing/
- www.datenbanken-verstehen.de/lexikon/natural-language-processing/
- www.tableau.com/learn/articles/natural-language-processing-examples
- en.wikipedia.org/wiki/Natural_language_processing
- de.wikipedia.org/wiki/Computerlinguistik
- www.digitalconnection.de/connected-customer/das-jahrzehnt-der-digitalen-sprachassistenten/
- www.xenonstack.com/blog/evolution-of-nlp/
- rstudio-pubs-static.s3.amazonaws.com/69717_91e500e66f784451a5126c405ceaa738.html
- lifehacker.com/how-predictive-keyboards-work-and-how-you-can-train-yo-1643795640
- de.wikipedia.org/wiki/ELIZA
- www.youtube.com/watch?v=5ctbvkAMQO4
- www.youtube.com/watch?v=fOvTtapxa9c&

QUELLEN

- www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm
- de.wikipedia.org/wiki/Computerlinguistik
- web.stanford.edu/class/cs124/lec/languagemodeling.pdf
- www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/
- en.wikipedia.org/wiki/Language_model
- towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
- www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html
- www.bigdata-insider.de/was-ist-deep-learning-a-603129/
- www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/
- pathmind.com/wiki/neural-network
- deepai.org/machine-learning-glossary-and-terms/attention-models
- www.masswerk.at/elizabot/

VIELEN DANK FÜR DIE AUFMERKSAMKEIT!