Was versteht man unter natürlicher Sprachverarbeitung (NLP)?

Machine-Learning-Lösungen kennenlernen

Definition: Natürliche Sprachverarbeitung

Natürliche Sprachverarbeitung (Natural Language Processing, NLP) ist eine Art von künstlicher Intelligenz (KI), die sich damit befasst, wie Computer und Menschen durch menschliche Sprache miteinander interagieren können. Mit NLP-Techniken können Computer unsere natürlichsten Kommunikationsformen – Sprache und geschriebenen Text – analysieren und verstehen und uns auf die gleiche Weise antworten.

Natürliche Sprachverarbeitung ist eine Unterdisziplin der Computerlinguistik. Computerlinguistik ist eine fachrichtungsübergreifende Spezialisierung, die Informatik, Sprachwissenschaften und künstliche Intelligenz miteinander kombiniert, um menschlichen Sprache unter Datenverarbeitungsaspekten zu untersuchen.

Geschichte der natürlichen Sprachverarbeitung (NLP)

Die Geschichte der natürlichen Sprachverarbeitung reicht zurück bis in die 1950er Jahre, als Computerwissenschaftler erstmalig versuchten, Maschinen beizubringen, menschliche Sprache zu verstehen und zu produzieren. Im Jahr 1950 formulierte Alan Turing seinen berühmten Turing-Test, der menschliche Sprache mit von Computern generierter Sprache vergleicht, um herauszufinden, welche Variante lebensechter klingt. Damals begannen Forscher außerdem, die Möglichkeit zu studieren, ob Computer für Übersetzungen eingesetzt werden können.

Während der ersten zehn Jahre bestand NLP größtenteils auf regelbasierter Verarbeitung. Zu Beginn der 1960er Jahre hatten die Forscher neue Wege entwickelt, um menschliche Sprache mit semantischer Analyse, Teilsatzmarkierung und Parsing zu analysieren. Außerdem entwickelten

sie die ersten Korpora. Dabei handelt es sich um große maschinenlesbare Dokumente, die mit linguistischen Informationen versehen sind und zum Trainieren von NLP-Algorithmen eingesetzt werden.

In den 1970er Jahren wurde erstmals statistische NLP als Alternative zu regelbasierten Herangehensweisen eingesetzt. Dabei wurden statistische Modelle zur Analyse und Generierung von Texten in natürlicher Sprache verwendet.

In den 1980er Jahren wurden effizientere Algorithmen zum Trainieren von Modellen und zum Verbessern der Modellgenauigkeit entwickelt. Daraus entstanden letztendlich die Machine-Learning-Algorithmen in der NLP. Beim Machine Learning werden große Mengen an Daten verarbeitet, um Muster zu finden, die anschließend häufig für Vorhersagen eingesetzt werden.

Deep Learning, neuronale Netze und Transformationsmodelle haben die NLP-Forschung von Grund auf verändert. Mit dem Aufkommen von Deep Neural Networks, der Erfindung von Transformationsmodellen und dem sogenannten „Attention-Mechanismus“ entstanden Technologien wie BERT und ChatGPT. Der Attention-Mechanismus kann beispielsweise mehr als nur ähnliche Schlüsselwörter zu Ihrer Abfrage finden. Er gewichtet die einzelnen verbundenen Begriffe anhand ihrer Relevanz. Diese Technologie steckt

hinter einigen der interessantesten NLP-Technologien, die derzeit verwendet werden.

Wie funktioniert die natürliche Sprachverarbeitung?

Natürliche Sprache kann auf verschiedene Arten verarbeitet werden. Bei der KI-basierten NLP werden Machine-Learning-Algorithmen und -Techniken eingesetzt, um menschliche Sprache zu verarbeiten, zu verstehen und zu generieren. Bei der regelbasierten NLP wird eine Reihe von Regeln oder Mustern erstellt, um Sprachdaten zu analysieren und zu generieren. Die statistische NLP nutzt statistische Modelle, die aus großen Datensätzen gewonnen werden, um Sprache zu analysieren und Vorhersagen zu treffen. Die Hybrid-NLP besteht aus einer Kombination dieser drei Herangehensweisen.

Der KI-basierte NLP-Ansatz ist momentan die beliebteste Variante. Wie auch bei anderen datengestützten Lernansätzen ist es beim Entwickeln von NLP-Modellen erforderlich, Textdaten vorab zu verarbeiten und den Lernalgorithmus sorgfältig auszuwählen.

Schritt 1: Vorverarbeitung der Daten
Bei diesem Prozess werden Textdaten bereinigt und vorbereitet, um sie mit dem NLP-Algorithmus analysieren zu können. Einige wichtige Vorverarbeitungstechniken sind

beispielsweise das Textmining, bei dem große Mengen an Texten erfasst und in Daten unterteilt werden, und die **Tokenisierung**, bei der Texte in kleinere Einheiten aufgebrochen werden. Als Einheiten können beispielsweise Satzzeichen, Wörter oder Sätze verwendet werden. Ein **Stoppwortfilter** ist ein Tool, das häufig verwendete Wörter und Artikel entfernt, die für Analysen meist wenig hilfreich sind. **Wortstammerkennung und Lemmatisierung** ermitteln die Stammform von Wörtern, um deren Bedeutung leichter identifizieren zu können. Bei der **Wortartmarkierung** werden Nomen, Verben, Adjektive und andere Wortarten in einem Satz markiert. **Parsing** analysiert die Struktur von Sätzen und die Beziehungen der einzelnen Wörter zueinander.

Schritt 2: Algorithmusentwicklung
Bei diesem Prozess werden NLP-Algorithmen auf die vorverarbeiteten Daten angewendet. Dabei werden nützliche Informationen aus dem Text extrahiert. Natürliche Sprachverarbeitung eignet sich unter anderem für die folgenden Aufgaben:

- Die **Standpunktanalyse** ermittelt die emotionale Stimmung oder das Gefühl in einem Text. Dabei werden Wörter, Sätze und Ausdrücke als positiv, negativ oder neutral markiert.
- Bei der **Erkennung benannter Entitäten** werden benannte

Entitäten wie etwa Personen, Orte, Daten und Unternehmen identifiziert und kategorisiert.

- Die **Themenmodellierung** gruppiert ähnliche Wörter und Sätze, um die wichtigsten Themen in einer Sammlung von Dokumenten oder Texten zu identifizieren.
- Die **maschinelle Übersetzung** verwendet Machine Learning, um Texte automatisch von einer Sprache in eine andere zu übersetzen. Sprachmodelle können die Wahrscheinlichkeit von Wortfolgen in einem bestimmten Kontext vorhersagen.
- Die **Sprachmodellierung** wird zur Autovervollständigung, für automatische Korrekturen und Sprache-in-Text-Anwendungen verwendet.

Zwei wichtige Unterbereiche der NLP sind das **natürliche Sprachverständnis (Natural Language Understanding, NLU)** und die **natürliche Sprachgenerierung (Natural Language Generation, NLG)**. Das NLU versetzt Computer in die Lage, menschliche Sprache mit ähnlichen Tools zu verstehen, wie sie auch von Menschen eingesetzt werden. Die Computer versuchen, Nuancen der menschlichen Sprache wie etwa Kontext, Absichten, Gefühle und Mehrdeutigkeiten zu erkennen. Die NLG befasst sich damit, menschliche Sprache aus einer Datenbank oder einer Reihe von

Regeln zu generieren. NLG hat das Ziel, Text zu produzieren, der für Menschen möglichst leicht verständlich ist.

Vorteile der natürlichen Sprachverarbeitung

Die natürliche Sprachverarbeitung bietet unter anderem die folgenden Vorteile:

- **Natürlichere Kommunikation:** NLP ermöglicht eine natürlichere Kommunikation mit Such-Apps. NLP kann sich an unterschiedliche Stile und Stimmungen anpassen und komfortablere Kundenerlebnisse bieten.
- **Effizienz:** NLP kann zahlreiche Aufgaben automatisieren, die normalerweise von Menschen erledigt werden müssen. Einige Beispiele sind Textzusammenfassungen, Überwachung von Social Media und E-Mails, Spam-Erkennung und Textübersetzungen.
- **Inhaltskuratierung:** NLP kann besonders relevante Informationen für einzelne Nutzer anhand von deren Vorlieben identifizieren. Durch das Verständnis von Kontext und Schlüsselwörtern lässt sich die Kundenzufriedenheit steigern. Besser durchsuchbare Daten können die Effizienz von Such-Tools verbessern.

Welchen Herausforderungen steht die natürliche Sprachverarbeitung gegenüber?

Bei der NLP gibt es immer noch zahlreiche ungelöste Herausforderungen. Menschliche Sprache ist uneinheitlich und oft mehrdeutig, wobei die jeweilige Bedeutung vom Kontext abhängt. Programmierer müssen ihren Anwendungen diese Eigenheiten von Grund auf beibringen.

Homonyme und Syntax können Datensätze verwirren. Und selbst die beste Standpunktanalyse kann Sarkasmus und Ironie nicht immer identifizieren. Menschen brauchen Jahre, um diese Nuancen zu unterscheiden, und selbst dann ist es oft nicht einfach, den Ton einer Textnachricht oder einer E-Mail zu erkennen.

Texte werden in unterschiedlichen Sprachen veröffentlicht, während NLP-Modelle in spezifischen Sprachen trainiert werden. Vor der Integration in Ihre NLP müssen Sie eine Spracherkennung durchführen, um die Daten nach Sprachen zu sortieren.

Unspezifische und zu allgemeine Daten führen dazu, dass die NLP die Bedeutung von Texten weniger genau verstehen und vermitteln kann. Für spezifische Themenbereiche sind mehr Daten erforderlich als die meisten NLP-Systeme zur Verfügung

haben, um zielsichere Aussagen zu treffen. Dies gilt insbesondere für Fachbereiche, die von aktuellen und sehr spezifischen Informationen abhängen. Neue Forschungsarbeiten wie etwa der ELSER (Elastic Learned Sparse Encoder) versuchen, dieses Problem zu beheben und relevantere Ergebnisse zu liefern.

Beim Verarbeiten von personenbezogenen Daten müssen außerdem Datenschutzbedenken berücksichtigt werden. In Branchen wie etwa dem Gesundheitswesen könnte die NLP Informationen aus Krankenakten extrahieren, um Formulare auszufüllen und Gesundheitsprobleme zu identifizieren. Diese Arten von Datenschutzbedenken, Datensicherheitsproblemen und potenziellen Verfälschungen erschweren den Einsatz der NLP in bestimmten Bereichen.

In welchen Geschäftsbereichen lässt sich die natürliche Sprachverarbeitung anwenden?

Die NLP eignet sich für vielerlei Geschäftsanwendungen:

- **Chatbots und virtuelle Assistenten:** Die Nutzer können eine Unterhaltung mit Ihrem System führen. Diese Tools werden oft im Kundenservice eingesetzt. Sie können Nutzer auch durch komplizierte Workflows führen oder beim

Navigieren von Sites oder Lösungen helfen.

- **Semantische Suche:** Diese Variante wird oft im E-Commerce-Bereich eingesetzt, um Produktempfehlungen zu generieren. Sie dekodiert den Kontext von Schlüsselwörtern mit Suchmaschinenanalysen und wissensbasierten Suchfunktionen. Dabei wird die Absicht der Nutzer interpretiert, um möglichst relevante Empfehlungen zu liefern.
- **NER:** Identifizieren von Informationen in Texten, um Formulare auszufüllen oder Texte besser durchsuchbar zu machen. Bildungseinrichtungen können damit beispielsweise von Schülern oder Studenten geschriebene Texte analysieren und Bewertungen automatisieren. Funktionen wie Text-zu-Sprache und Sprache-zu-Text können außerdem für mehr Barrierefreiheit sorgen und die Kommunikation für Menschen mit körperlichen Einschränkungen erleichtern.
- **Textzusammenfassung:** Forscher können große Dokumente über Branchen hinweg zu präzisen, leicht verständlichen Texten zusammenfassen. Die Finanzbranche nutzt diese Funktion, um Nachrichten und Social Media zu analysieren und Markttrends vorherzusagen. In Behörden und im Rechtswesen werden damit wichtige

Informationen aus Dokumenten extrahiert.

Wie sieht die Zukunft der NLP aus?

ChatGPT und generative KI versprechen grundlegende Veränderungen. Mit der Verfügbarkeit von Technologien wie ChatGPT zeichnen sich neue Anwendungsbereiche für die NLP ab. Vermutlich erleben wir bald Integrationen mit anderen Technologien wie etwa Spracherkennung, Bildverarbeitung und Robotik, um fortschrittlichere und ausgefeiltere Systeme zu entwickeln.

Mit neuen und personalisierten NLP-Anwendungen können Computer einzelne Nutzer außerdem besser verstehen und ihre Antworten und Empfehlungen entsprechend anpassen. NLP-Systeme, die mehrere Sprachen verstehen und generieren können, sind ein wichtiger Wachstumsbereich für international tätige Unternehmen. Dazu kommt, dass NLP-Systeme immer besser darin werden, menschlich klingende Texte zu generieren: Sie klingen im Lauf der Zeit immer menschlicher.

Erste Schritte mit NLP und Elastic

Seit Version 8.0 des Elastic Stack können Sie PyTorch-Modelle in

Elasticsearch hochladen und moderne NLP-Funktionen im Elastic Stack bereitstellen, inklusive Features wie etwa der Erkennung benannter Entitäten und Standpunktanalysen.

Der Elastic Stack unterstützt derzeit Transformationsmodelle, die der standardmäßigen BERT-Modellschnittstelle entsprechen und den WordPiece-Tokenisierungsalgorithmus verwenden.

Hier finden Sie Informationen zur aktuell mit Elastic kompatiblen Architektur:

- **BERT**
- **BART**
- **DPR Bi-Encoder**
- **DistilBERT**
- **ELECTRA**
- **MobileBERT**
- **RoBERTa**
- **RetriBERT**
- **MPNet**
- **SentenceTransformers Bi-Encoder mit den oben genannten Transformationsarchitekturen**

Mit Elastic und NLP können Sie Informationen extrahieren, Texte klassifizieren und mehr Suchrelevanz für Ihr Unternehmen bereitstellen.
Erste Schritte mit NLP und Elastic

NLP-Ressourcen (teils nur auf Englisch verfügbar)

- **Ausführliche Informationen zum Thema natürliche Sprachverarbeitung (NLP)**
- **NLP bereitstellen: Texteinbettungen und Vektorsuche**
- **Einführung in NLP-Modelle und die Vektorsuche**

BEI ELASTIC ARBEITEN

Freie Stellen

Karriereportal

PRESSE

Pressemitteilungen

Nachrichtenartikel

PARTNER

Partner finden

Partner-Login

Zugriff beantragen

Partner werden

VERTRAUEN UND SICHERHEIT

EthicsPoint-Portal

Sicherheit und Datenschutz

ECCN Report

Ethics-E-Mail

INVESTOR RELATIONS

Ressourcen für Investoren

Governance

Finanzen

Aktie

EXCELLENCE AWARDS

Bisherige Gewinner

ElasticON Tour

Sponsor werden

Alle Events