# Die fünf wichtigsten Use Cases für NLP (Natural Language Processing)

"Jede hinreichend [fortgeschrittene](https://de.wikiquote.org/wiki/Fortschritt" \o "Fortschritt) [Technologie](https://de.wikiquote.org/wiki/Technik" \o "Technik) ist von [Magie](https://de.wikiquote.org/wiki/Magie" \o "Magie) nicht mehr zu unterscheiden." - Clarkes Drittes Gesetz aus Profiles of the Future

Für Magie wurden ohne Frage die ersten Computer gehalten, die in Sekundenschnelle Berechnungen durchführen konnten, für die ein Mensch ungleich länger braucht. Ein paar Jahrzehnte später durchdringt der Computer alle Ebenen des menschlichen Alltags, sei es privat oder geschäftlich: Der PC ist zum unentbehrlichen Assistenten für sämtliche Prozesse, die sich in Form von Daten abbilden lassen geworden. Diese Entwicklung wurde getragen von kontinuierlichen Verbesserungen (z.B. Rechenleistungsgewinne bei zeitgleicher Minitiarisierung der Hardware, siehe Moore’s Law), die neue impulsgebende Innovationen, u.a. Big Data & AI, möglich machten. Ein weiterer „Magiemoment“ im kollektiven Gedächtnis ist der Erstkontakt mit einem Sprachassistenten wie Siri (Apple), Cortana (Microsoft) oder dem Google Assistant. Hinter dieser Technologie verbirgt sich ein Forschungsgebiet namens NLP („Natural Language Processing“). Während NLP schon seit der Fünfziger Jahre Informatiker beschäftigt, fristete es auf der Anwendungsseite ein kaum sichtbares Dasein, da häufig Datenmengen zu groß und angewandte Modelle zumeist nicht ausgereift waren. Dies änderte sich Anfang der 2010er, als sich NLP zum integralen Bestandteil der Machine Learning Landschaft etablierte und eine neue Modellgeneration künstlicher neuronaler Netzwerke Anwendungen möglich machte, die zuvor unerreichbar erschienen.

# Was ist NLP?

Hierbei versteht man unter "Natural Language" eine Sprache, die für die alltägliche Kommunikation von Menschen verwendet wird; Sprachen wie Englisch und Deutsch. Im Gegensatz zu künstlichen Sprachen - wie Programmiersprachen und mathematischen Notationen - entwickeln sich natürliche Sprachen im Laufe der Zeit weiter und lassen sich nur schwer mit expliziten Regeln erfassen.

**NLP bezeichnet das Forschungsgebiet, das sich mit der Nutzbarmachung natürlicher Sprache durch Computermanipulation beschäftigt.**

Der Anwendungsbereich von NLP ist breit gefächert: So reicht er zum Beispiel vom einfachen Zählen von Worthäufigkeiten, um verschiedene Schriftstile zu vergleichen bis hin zu komplexen Algorithmen, die als Ziel das komplette semantische "Verstehen" natürlicher Sprache haben. Technologien, die auf NLP basieren, finden immer mehr Einzug in digitale Technologien des Alltags. Beispielsweise haben Smartphones eine prädiktive Texterkennung, die das Tippen deutlich komfortabler macht, Web-Suchmaschinen ermöglichen den Zugang zu Informationen, die in unstrukturiertem Text eingeschlossen sind und maschinelle Übersetzung ermöglicht es, in Chinesisch geschriebene Texte abzurufen und auf Französisch zu lesen.

Man schätzt, dass der Großteil „relevanter Information“ im Internet nicht in strukturierten Datenbanken in Zahlenform vorliegt, sondern in unstrukturierter, natürlicher Sprache.

*Aus diesem Grund kommt von einer höheren Warte betrachtet dem Gebiet NLP unter anderen Forschungsbereichen im Bereich Big Data & AI (z.B. Bilderkennung, Forecasting, …) eine Schlüsselrolle zu: Mithilfe von NLP wird in Zukunft die Menge nutzbarer Daten dramatisch steigen. Dies ist entscheidend für sämtliche andere Big Data Technologien, da diese derart designed sind, dass ihre Nutzbarkeit stark von der zur Verfügung stehenden Datenmenge abhängt.*

Abgesehen von dieser exponierten Stellung von NLP, ermöglicht NLP es, eine Vielzahl praktischer Probleme elegant zu lösen, und schafft so Raum für mehr Effizienz und Innovation. Um Ihnen eine Idee davon zu geben, wozu man mit NLP heutzutage in der Lage ist, werden Ihnen hier fünf wichtige Anwendungsfälle vorgestellt. Es sei bemerkt, dass viele dieser Anwendungen aufeinander aufbauen, was für NLP typisch ist.

# Information Retrieval: Named Entity Recognition (NER):

## Worum geht es?

Vergleichen Sie die folgenden Satzpaare:

1. „Bert ist super! Wir kennen uns schon seit den Achtzigern und gehen durch dick und dünn.“ (Bert = Person)
2. „BERT setzt neue Maßstäbe für eine Vielzahl von NLP-Tasks. Seine Universalität in Kombination mit seiner Performance verspricht eine flexible, erfolgreiche Anwendung. “ (BERT = Sprachenmodell)

Für den Menschen ist vollkommen offensichtlich, dass es sich hierbei um zwei verschiedene „Berts“ handelt (auch abgesehen von Groß- und Kleinschreibung). Diese scheinbar einfache Unterscheidung ist für den Computer allerdings alles andere als trivial. Ein Entity Recognition Modell könnte die beiden Sätze wie folgt kennzeichnen:

1. „**[Bert]** *(Person)* ist super! Wir kennen uns schon seit den [**Achtzigern]** *(Zeit)* und **[gehen durch dick und dünn]** *(Phrase)*.“
2. „**[BERT]** *(Machine Learning Modell)* setzt neue Maßstäbe für eine Vielzahl von **[NLP]** *(Machine Learning Feld)*-Tasks. Seine Universalität in Kombination mit seiner Performance verspricht eine flexible, erfolgreiche Anwendung.“

## Anwendungsbereiche

Entity Recognition ist immer dann von Nutzen, wenn man Informationen über die erkannten Entitäten hat oder sammeln möchte oder aber wenn man die Beziehungen zwischen verschiedenen benannten Entitäten extrahieren möchte. Außerdem handelt es sich bei Entity Recognition um einen klassischen Anwendungsfall von NLP, der häufig Grundlage für andere Anwendungen ist. Das Erfassen relevanter Informationen in Textdokumenten basiert zum Beispiel auf Entity Recognition: Wenn man Textdokumente nach seinen Inhalten klassifizieren möchte, sind distinkte Entitäten häufig ein wichtiger Faktor. Eine schriftliche Bewerbung ist beispielsweise dadurch gekennzeichnet, dass Bildungseinrichtungen, Fähigkeiten u.ä. aufgelistet werden, während bei einer Rechnung üblicherweise eine Artikelnummer, ein Mehrwertsteuersatz und ein Nettobetrag zu finden sind. Abhängig von der Anwesenheit solcher Entitäten kann man gut zwischen Rechnung und Bewerbung unterscheiden. Andere Anwendungsfälle sind tabellarische, automatische Zusammenfassungen von Texten sowie intelligente Textsuche.

## Technologie

Zur Bestimmung der Entitäten werden klassischerweise statistische Verfahren oder aber regelbasierte Algorithmen verwendet. In jüngster Zeit werden auch NER-Task spezifische Machine Learning Algorithmen benutzt, sowie allgemeinere Language Models (dazu später mehr).

# Information Retrieval: Sentiment Analysis

## Worum geht es?

Aus der Kommunikationswissenschaft ist bekannt, dass der Großteil der Kommunikation sich nicht daraus speist, *was* gesagt wird, sondern *wie*. Das *wie* betrifft bekanntermaßen vor allem das Nonverbale, allerdings lässt häufig schon die Wortwahl eines Textes Rückschlüsse auf die Stimmung und Position des Autors zu. Sentiment Analysis nähert sich dieser Problematik, indem es systematisch die Stimmung analysiert und quantifiziert.

## Anwendungsbereiche

Eine prominente Anwendung findet Sentiment Analysis in der Finanzwelt, wo Algorithmen durchgängig das Internet nach neuen Informationen bezüglich börsennotierter Unternehmen absuchen. Pessimistische wie optimistische Stimmen werden durch den Sentiment Analysis – Algorithmus aufbereitet und werden zum entscheidenden Input für Trade-Bots. Ein Trade-Bot, dem aufgrund eines erfolgreichen Sentiment Analysis – Algorithmus mehr Daten zur Verfügung steht, hat das Potential, seinen Konkurrenten einen Schritt voraus zu sein.

In Onlineshops findet Sentiment Analysis häufig Anwendung in intelligenten Recommendersystemen. Wenn der Kunde ein Produkt betrachtet, werden ihm zusätzlich weitere passende Produkte angeboten, die die Schwächen ausmerzen, welche in den Rezensionen des angeklickten Artikels benannt werden.

Bezüglich Markenmarketings ist es besonders wichtig, die Wahrnehmung der Marke durch die Kunden zu erfassen. Insbesondere soziale Medien erweisen sich dazu als ergiebig. Gepaart mit der Analyse von Review Seiten, Blogs und fachkundigen Foren ergibt sich so breites Meinungsfeld, das durch Sentiment Analysis erfasst wird. Aus diesen Daten erkennbar gemachte Muster/Trends können eine wichtige Entscheidungsgrundlage in einem datengetriebenen, modernen Unternehmen sein.

## Technologie

Üblicherweise wird dem Text ein Vektor zugeordnet werden, der numerisch die Gefühlslage in je nach Anwendungsfall relevanten Dimensionen (z.B. „positiv-negativ“ oder „hilfsbereit-abweisend + „freundlich-gereizt“ oder „bullish-bärig“) erfasst. Dazu können verschiedene Algorithmen benutzt werden. In simplen Modellen wird der Text in n-grams, also in Wortfolgen, zerlegt und diese Wortfolgen nach statistischen Verfahren ausgewertet. Fortgeschrittenere Algorithmen berücksichtigen auch die grammatikalische Struktur der Sätze, sodass dem Sentiment ein Subjekt, das „fühlt“, und ein Objekt, das das Sentiment hervorruft, zugeordnet werden kann. Die derzeit performantesten Modelle sind neuronale Netzwerke, die auf großen Textsätzen trainiert werden.

# Sprachgenerierung

## Worum geht es?

Unter Sprachgenerierung versteht man die Synthese von Sprache, um Informationen in menschliche Sprache zu überführen. Durch Sprachgenerierung können also harte Informationen in natürliche Sprache verpackt werden und dem Endnutzer in verständlicher Form weitergegeben werden.

## Anwendungsbereiche

Ein solches Szenario ist an vielen Stellen denkbar: Das automatisierte Schreiben von Berichten, die Generierung von Texten auf der Grundlage der Analyse von Einzelhandelsverkaufsdaten, die Zusammenfassung von elektronischen Krankenakten, die Erstellung von textlichen Wettervorhersagen aus Wetterdaten und sogar die Erstellung von Witzen sind mögliche Anwendungen. Darüber hinaus kommt es auch bei anderen NLP Anwendungsfällen wie Chatbots und Q&A-Systemen zur Sprachgenerierung.

## Technologie

Üblicherweise lässt sich das Sprachgenerierungsproblem durch die folgende Aufgabenfolge abbilden:

* Inhaltsbestimmung: Welcher Inhalt soll durch den Text transportiert werden?
* Dokumentenstrukturierung: In welcher Abfolge sollen die Informationen erscheinen?
* Aggregation: Ähnliche Informationen werden nebeneinander platziert (z.B. durch Aufzählungen).
* Wortwahl: Welche Worte sind am passendsten, um die Informationen zu beschreiben?
* Identifikatorenerkennung: Entitäten im Text werden manchmal auf verschiedene Weise paraphrasiert, z.B. durch ein Pronomen oder eine Beschreibung, damit es nicht zu Wortwiederholungen kommt.

und letztlich:

* Realisierung: Das korrekte Zusammenführen obiger Ergebnisse unter Berücksichtigung der Grammatik und Orthographie.

Diese Arbeitsschritte wurden bisweilen meist von einzelnen Algorithmen bearbeitet, allerdings setzt sich auch hier der Trend durch, ein einziges Machine Learning Modell (meist ein sog. rekursive neuronale Netzwerk (RNN)) zu benutzen, das sich dieser strikten Aufgabenteilung entledigt. Hierzu wird ein entsprechender Textkorpus, der auf die Aufgabe angepasst ist, dem Modell als Lernbeispiel übergeben. Wenn das Ziel beispielsweise die Generierung von Zusammenfassungen ist, bewertet das Modell im Trainingsvorgang eine Datenbank vieler Texte und ihrer Zusammenfassungen. Anhand dieser Beispiele lernt das Modell, Zusammenfassungen von bisher ungesehenen Texten zu erstellen. Je nach Architektur des Netzwerkes lassen sich darüber hinaus noch andere Funktionalitäten ableiten. So erlauben die meisten Modelle den Benutzern die Kontrolle über die Bedeutung der generierten Sätze. Durch die Wahl eines Vektors von Merkmalen auf Satzebene ist es möglich, das Netzwerk zu instruieren, wie zum Beispiel: "Sagen Sie etwas Gutes über einen Data Science und NLP in etwa dreißig Worten."

# Question Answering Systems

## Worum geht es?

Ein Frage-Antwort-System („QA-System“) zielt darauf ab, auf einzeln gestellte Fragen die korrekte Antwort zu finden. Die Fragen können ganz unterschiedlicher Natur sein; neben „Wer, Wie, Was“ sind beispielsweise auch semantisch eingeschränkte Fragen („Unter der Bedingung, dass ...“) möglich.

## Anwendungsbereiche

Als Standalone-System halten QA-Systeme Einzug in die Industrie durch IBM Watson und in der Wissenschaft durch Wolfram Alpha. Häufiger sind sie allerdings als Bestandteil von Chatbots wie Siri und Co zu finden, um (kontextlose) Wissensfragen zu beantworten.

## Technologie

Zur Beantwortung der Frage wird auf eine Vielzahl von strukturierten (Datenbanken) und unstrukturierten (Wikipedia, Wissenschaftsportale, usw.) Quellen zugegriffen. Letztere werden in einem Vorbereitungsschritt mithilfe anderer NLP-Verfahren nutzbar gemacht. Ein Frage-Antwort System eignet sich besonders für sehr spezifische Probleme, da der QA-Algorithmus auf bereitgestellte Datenbanken trainiert werden kann. Somit kann das System unkompliziert für bestimmte Fragestellungen maßgeschneidert werden.

**„Je spezifischer die Frage, desto einfacher lassen sich Antworten finden.“**

Allgemeiner teilt man Fragen in zwei Kategorien ein:

* Closed Domain Fragen: Diese Fragen betreffen ein eingrenzbares Fachgebiet und sind von ihrer Struktur einfach aufgebaut, sodass klar ist, in welchen Datenbanken die relevante Informationen zu finden sind, die zum Beantworten der Frage herangezogen werden. Beispiel:
* Open Domain Fragen: Im Gegenzug ist hier nicht klar, in welchen Datenbank die relevanten Informationen zu finden sind. Dadurch sind solche Fragen ungleich komplexer, zumal häufig „Weltwissen“ erforderlich ist, um diese zu beantworten.

Ein QA-System, das auf sämtliche Open und Closed Domain Fragen sinnvolle Antworten geben kann, wird von einigen als „General AI“ bezeichnet. Eine solche künstliche Intelligenz ist noch nicht am Horizont – eine Mehrheit in der Forschungscommunity geht davon aus, dass eine reine inkrementelle Verbesserung durch mehr Daten und komplexere Modelle sich eine General AI nicht realisieren lässt, sondern dass es neuartige, innovative Modellarchitekturen von Nöten sind.

# Chatbots

## Worum geht es?

Während QA Systeme lediglich einzelne Fragen beantworten können, können Chatbots in einen interaktiven Informationsaustausch mit dem Benutzer gehen. Sie verstehen die natürliche Sprache des Benutzers, merken sich die Historie der Konversation, und reagieren entsprechend. Darüber hinaus können Chatbots, je nach Architektur, auch Closed & Open Domain Fragen beantworten. Damit kann der Scope von Chatbots deutlich größer als von QA-Systemen sein.

Das Startsignal zur theoretischen Beschäftigung mit Chatbots wurde 1950 durch den berühmten Turings Test gegeben: Hierbei schickt ein Mensch eine beliebig große Anzahl beliebiger Fragen in Textform an einen ihm nicht sichtbaren Agenten. Dieser Agent ist entweder selbst ein Mensch, oder aber ein Computer, der ein Chatbotprogramm ausführt. Ist die Testperson nach der Bewertung aller vom Agenten zurückgegebenen Antworten unschlüssig darüber, ob die Antworten von einem Computer oder einem Menschen kommen, gilt der Turing Test als bestanden. Bislang konnte noch kein Chatbot den Test bestehen – allerdings wurden große Fortschritte bezüglich des Turing Tests gemacht, wenn man das Anwendungsszenario [ähnlich wie bei QA-Systemen] beschränkt. So hat Google 2018 mit Google Duplex eindrücklich einen Chatbot vorgeführt, der in der Lage ist, am Telefon einen Frisörtermin zu vereinbaren, und dabei ununterscheidbar von einem Menschen agiert.

## Anwendungsbereiche

Es gibt vielfältige Einsatzmöglichkeiten im Kundenservice – je nach Komplexität des Szenarios von der automatischen Vorbereitung (z.B. Sortierung) eines Kundenauftrags hin zur kompletten Abwicklung einer Kundenerfahrung.

## Technologie

Über die Jahre sind im Wesentlichen vier Chatbotarchtikturen erprobt worden: In einer ersten Generation wurde die Anfragen auf bekannte Muster geprüft und entsprechend angepasste vorgefertigte Antworten ausgegeben („pattern matching“). Etwas ausgefeilter ist das sogenannte „grounding“, wo durch Named Entity Recognition (s.o.) aus Wissensbibliotheken (z.B. Wikipedia) extrahierte Informationen in einem Netzwerk organisiert werden. Ein solches Netzwerk hat den Vorteil, dass nicht nur eingetragenes Wissen abgerufen werden kann, sondern, dass auch nicht registriertes Wissen durch die Netzwerkstruktur inferiert werden kann. Beim „searching“ werden direkt Fragen-Antwortpaare aus dem Konversationsverlauf (oder aus davor registrierten Logs) zum Suchen einer passenden Antwort herangezogen. Beim etabliertesten Ansatz werden Machine Learning Modelle angewandt, um dynamisch passende Antworten zu generieren („generative models“).

# Warum NLP jetzt an Fahrt gewinnt

„Daten sind das Öl des 21. Jahrhunderts.“ (-Anonym)

Diese prominente Behauptung ist zugegebenermaßen etwas plakativ, aber wir bei STATWORX beobachten gemeinsam mit unseren Kunden im Alltag, dass sie keinesfalls aus der Luft gegriffen ist – ganz im Gegenteil.

Insbesondere durch NLP als Schlüsseltechnologie zur Nutzbarmachung vieler verschiedener Arten von Datenquellen, wird die Menge an Daten signifikant ansteigen. Darüber hinaus werden die obig beschriebenen Anwendungen in der nächsten Zeit immer mehr den Endbenutzer erreichen und die Grenzen zwischen Menschen und Maschinen im Alltag weiter auflösen, da die Integration von State-of-the-art-NLP-Technologien in den letzten zwei Jahren signifikant erleichtert wurde. Dies ist dadurch begründet, dass eine äußerst heterogene Landschaft anwendungsspezifischer Algorithmen zunehmend durch globalere Modelle ersetzt wird. Diese sogenannten „Language Models“ werden auf riesigen Textkorpora von den Big Players wie Facebook, Google und Microsoft über Wochen vortrainiert und können auf eine Vielzahl unterschiedlicher Downstream-Tasks (Anwendungen) nachjustiert werden. In vielen Anwendungen (u.a. Sentiment Analysis und Named Entity Recognition) werden in einem solchen Setup beste Ergebnisse erzielt. Die Vorteile dieses Ansatzes besteht darin, dass sich die vortrainierten Modelle herunterladen lassen und nur die Feinjustierung für die spezifische Anwendung durch zusätzliche Daten selbst gemacht werden muss. Folglich lassen sich mittlerweile mit wenig Entwicklungsaufwand performante NLP-Anwendungen entwickeln. Über das zurzeit erfolgreichste Language Model („BERT“ bzw. „roBERTa…“) lesen Sie in diesem Whitepaper. [Link, Dominiques Beitrag]. Wir von STATWORX haben jahrelange Erfahrung in der Entwicklung von maßgeschneiderten NLP-Lösungen und stehen Ihnen gerne zur Beratung zur Verfügung.

A screenshot of a cell phone

Description automatically generated

Chatbotlife magazine