E-BOOK

Mehr Gespräch und mehr Action

Eine Datenstruktur für konversationelle KI

NetApp

Inhalt

Keine Lust mehr auf Smalltalk? Es geht auch mehr.

NLP steht für „Natural Language Processing“ (Verarbeitung natürlicher Sprache) und wird auch als Computerlinguistik (CL) und *konversationelle KI* bezeichnet.

Wie auch immer Sie diese KI-Systeme bezeichnen möchten, haben sie Eines gemeinsam: sie sprechen wie Menschen, verstehen Kontext und geben intelligente Antworten. Möglich wird dies durch erhebliche Fortschritte im Bereich von Deep Learning, wodurch sich KI-Systeme von reinen Transaktionsgesprächen wegbewegen und eine immer natürlichere Sprache beherrschen.

Deep Learning macht nicht nur KI benutzerfreundlicher, sondern auch intensive Kenntnisse in den Bereichen Linguistik und regelbasierten Techniken im Backend überflüssig. Deep Learning stößt die Tür zu NLP-Lösungen für Branchen mit komplexer Spezialsprache auf, so etwa für Behörden, Finanzdienstleistungen, Gesundheitswesen, Biowissenschaften, Automobilbranche, Fertigungsindustrie und den Einzelhandel.

Daten sind der Schlüssel für bessere Unterhaltungen
Die zugehörigen KI-Modelle können äußerst umfangreich und komplex sein. Für sie müssen Unmengen an Daten in Windeseile bewegt werden. Eine NLP-Infrastruktur kann daher nur erfolgreich funktionieren, wenn Sie folgende Eigenschaften aufweist:

1. Dreht den Datendurchsatz hoch
2. Macht die Pipeline frei
3. Sie ist reaktionsschnell wie ein Geistesblitz

NLP: Nicht mehr nur für Chatbots

NLP ist die neue globale Sprache – von intelligenten Assistenten bis hin zu Suchmaschinen –, der Sie an zahlreichen Stellen begegnen. Auch an Orten, an denen man nicht unbedingt damit rechnet.

Beurteilung der Bonität

NLP kann im Rahmen der Bonitätsprüfung anhand von Standort, Aktivität in sozialen Medien, Browserverhalten usw. zur Ermittlung von Score-Werten eingesetzt werden.

Suche nach passenden Probanden für klinische Studien

Probanden für klinische Studien zu finden, kann ziemlich schwierig sein. Meist liegt das daran, das potenzielle Kandidaten nichts von den laufenden Studien wissen. Mit NLP können Forschende automatisch den Studien die richtigen Patienten zuordnen.

Strafverfolgung

Polizeibehörden ermitteln mithilfe von NLP Tatmotive. Damit tragen sie dazu bei, dass die Sicherheit steigt, die Gewaltrate sinkt und die Polizeiarbeit stärker auf die Nöte und Wünsche der Bevölkerung eingeht.

Fahrzeugwartung

NLP erleichtert Fahrzeughaltern die Wartung ihrer Automobile. Statt sich durch dicke Bedienungsanleitungen zu wälzen, können sie dem Fahrzeug einfach Fragen stellen wie „Was ist das für eine Warnleuchte?“ oder „Wie kann ich die Sicherung wechseln?“.

Flugzeugreparatur

NLP hilft Mechanikern dabei, Informationen aus extrem umfangreichen Wartungshandbüchern so zusammenzufassen, dass sie die von den Piloten gemeldeten Probleme besser verstehen.

1. Den Datendurchsatz hochdrehen

Für NLP sind gigantische Datenmengen erforderlich. Wenn Sie sich klar machen, wie viele Wörter seit den Anfängen der Menschheit weltweit gesprochen wurden, bekommen Sie eine ungefähre Ahnung.

NLP muss in der Lage sein, Sprach-Input zu verarbeiten, zu verstehen und in einer riesigen Datenbibliothek Verweise zu suchen, um innerhalb von Millisekunden eine passende Antwort zu geben.

Diese Anforderung ist angesichts der Komplexität der menschlichen Sprache eine besondere Herausforderung. Neben den Unmengen an Regeln und Ausnahmen muss das Modell auch idiomatische Nuancen, Sarkasmus und Humor erkennen. Bei einigen branchenspezifischen Modellen kommt eventuell auch Wissen zu einer bestimmten Domäne, einem bestimmten Unternehmen oder bestimmten Produkten hinzu.

Daher sind die Modelle für konversationelle KI auf Millionen oder gar Milliarden Parameter angewachsen. In der Regel gilt: Je mehr Daten, desto genauer das Modell. Modelle dieser Größe zu trainieren, kann selbst mit den leistungsfähigsten Frameworks für maschinelles Lernen und Deed Learning mehrere Wochen Computing-Zeit in Anspruch nehmen.

Google Translate

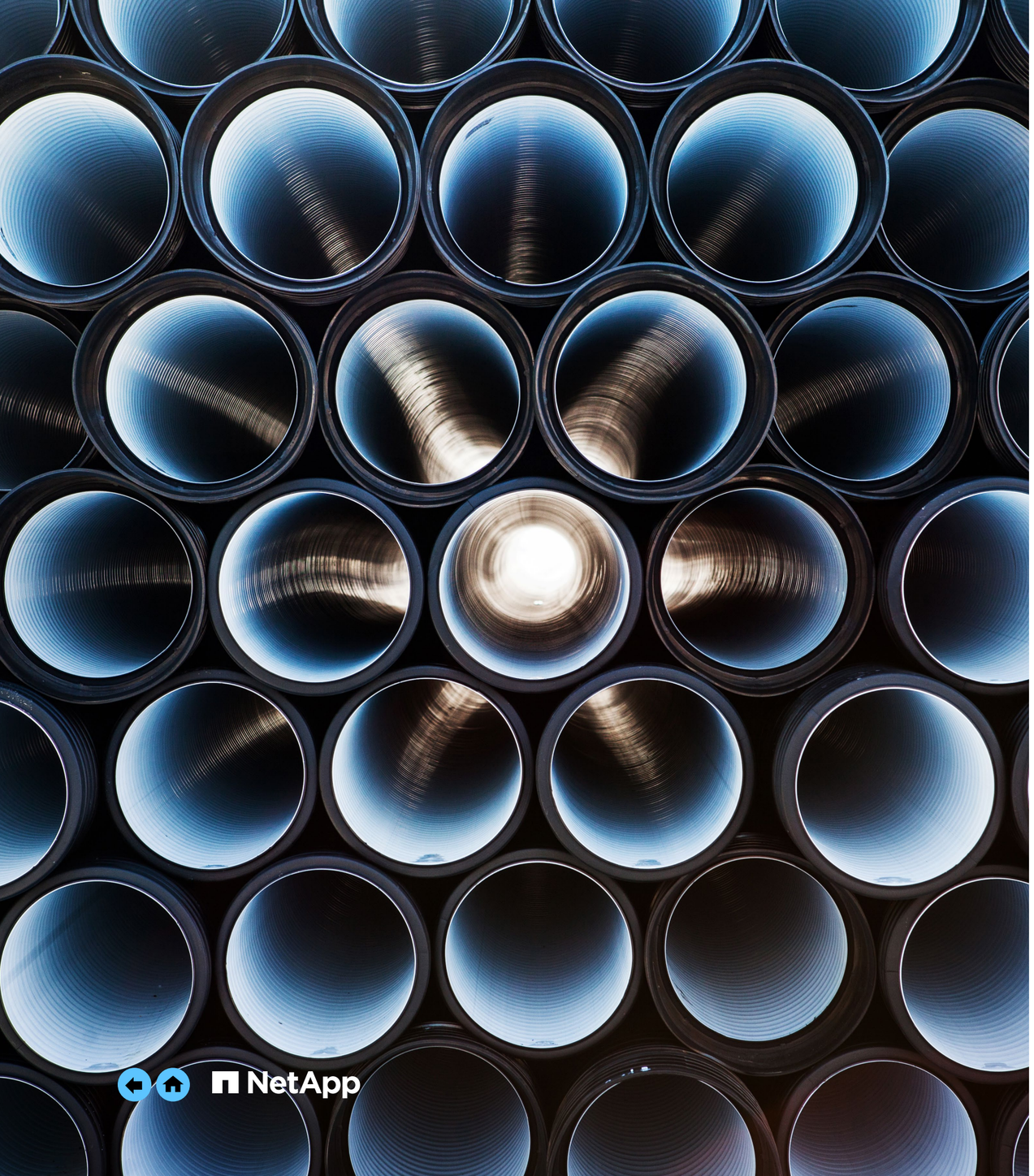
Google Translate unterstützt über 100 Sprachen und versucht mittels Crowdsourcing, Übersetzungen und Modell-Training für Sprachen mit beschränkten Trainings-Korpora (quellsprachlichen Datensätzen) zu testen und zu optimieren. Google Translate verarbeitet jeden Tag 140 Milliarden Wörter. Das ist *Tag für Tag* die Arbeit von 70 Millionen menschlichen Übersetzern.

Google BERT

Google BERT ist ein weit verbreitetes NLP-Modell mit 340 Millionen Parametern. BERT ist deswegen ein Durchbruch im Bereich des NLP, weil es keine Sprachschnittstelle für Transaktionen mehr darstellt (wie etwa Verzweigungsalgorithmen für Telefonate), sondern in der Lage ist, richtige Unterhaltungen zu führen. Das System kann Texte lesen und Fragen mit extrem hoher Genauigkeit beantworten.

BioMegatron

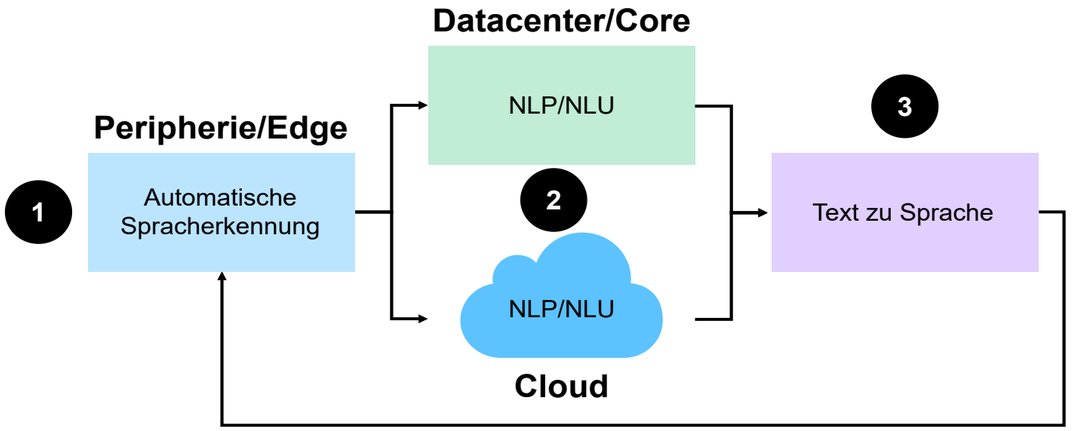
BioMegatron ist das größte jemals trainierte Sprachmodell auf der Basis biomedizinischer Transformer. Es verfügt über bis zu 1,2 Milliarden Parametervarianten und wurde auf der Grundlage von 6,1 Milliarden Wörtern aus PubMed, einer Sammlung von Abstracts und vollständigen Artikeln aus Fachpublikationen, trainiert.



2. Die Pipeline frei machen

Schnelle und effektive NLP erfordert eine Datenpipeline für das gesamte Ecosystem von der Aufnahme und Erkennung bis zur Sprachsynthese. Daten müssen schnell und ungehindert die einzelnen Pipelinephasen durchlaufen, damit die Sprachverarbeitung in Echtzeit möglich wird.

Eine typische NLP-Pipeline setzt sich aus 3 Phasen zusammen:



In einer modernen NLP-Infrastruktur werden an Edge-Standorten pro Tag mehrere Terabyte an Daten erfasst. Wenn der Zugriff auf diese Daten durch eine Silo-Infrastruktur eingeschränkt wird, kratzt Deep Learning nur an der Oberfläche.

3. Reaktionsschnell wie ein Geistesblitz

Damit KI menschliche Sprache replizieren kann, muss das System so schnell sein wie das menschliche Gehirn – oder noch schneller. Je größer das Modell, desto länger die Verzögerung zwischen einer Benutzerfrage und der KI-Antwort. Damit die Antwort natürlich wirkt, müssen sämtliche Berechnungen in einem Zeitfenster von 300 Millisekunden erfolgen.

Dieser Prozess gliedert sich in mehrere Schritte:

1. Sprache des Benutzers in Text umwandeln.
2. Bedeutung des Textes verstehen.
3. Nach der im Zusammenhang besten Antwort suchen.
4. Die Antwort in gesprochener Sprache vortragen.

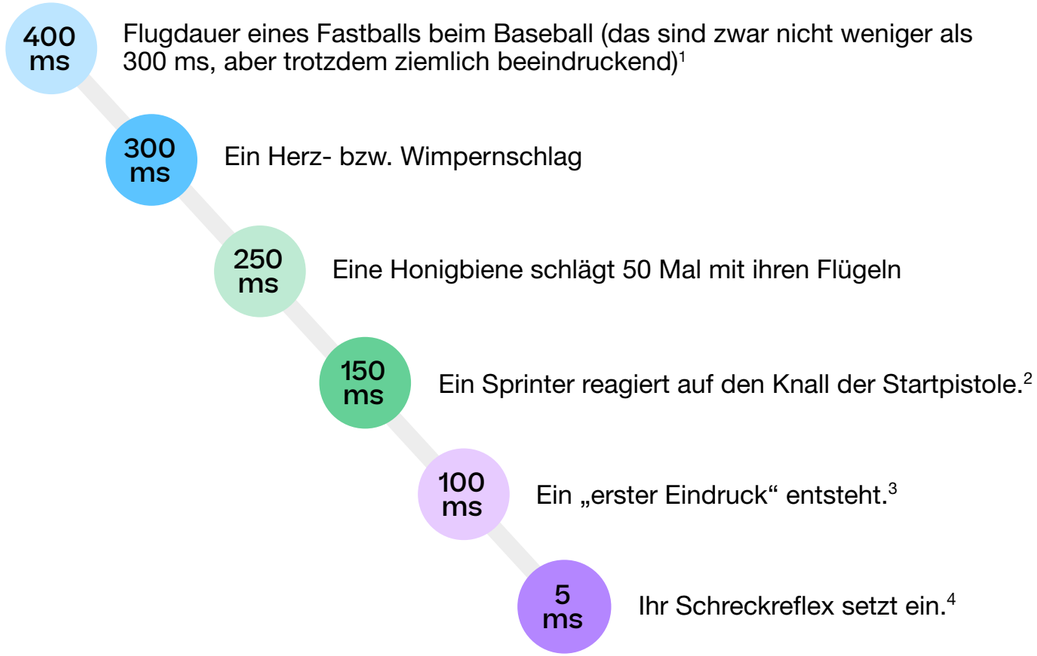
Angesichts dieser anspruchsvollen Latenzanforderungen müssen die Entwickler von konversationeller KI häufig Kompromisse eingehen. Ein qualitativ hochwertiges und überaus komplexes Modell braucht eventuell länger als ein weniger sperriges Sprachverarbeitungsmodell, das schnell Ergebnisse hervorbringt, dafür aber weniger nuanciert zu antworten vermag.

Ein Sprachassistent unterbricht die Unterhaltung vielleicht manchmal mit einem Einschub wie „Lassen Sie mich kurz nachschauen“ oder mit Geräuschen, die eine peinliche Stille überdecken sollen – ähnlich, wie es ein unsicherer Gesprächspartner auch tun würde. Die ideale konversationelle KI – das Ideal und zentrale Ziel bei NLP – ist so ausgereift, dass sie Fragen von Personen richtig versteht, und so schnell, dass sie die Fragen zügig und in natürlicher Sprache beantworten kann.

Was bedeutet „schnell“ bei Unterhaltungen?

Die Latenz bei NLP liegt in der Regel unter 300 Millisekunden, also 0,3 Sekunden für eine Antwort in Echtzeit. Wie schnell ist das? Sehr schnell.

300 Millisekunden oder weniger dauern Dinge wie diese:



NetApp spricht Ihre Sprache

Mit NetApp ONTAP AI auf Basis von NVIDIA DGX-Systemen und mit der Cloud verknüpften NetApp All-Flash-Storage-Systemen können riesige moderne Sprachmodelle so trainiert und optimiert werden, dass schnelle Inferenzen möglich sind. Eine Data Fabric auf der Basis von NetApp vereinfacht das Datenmanagement in der KI-Datenpipeline zwischen Peripherie, Datacenter und Cloud:

- **NetApp Lösungen für KI** beseitigen Engpässe und sorgen für eine effizientere Datenerfassung, eine Beschleunigung von KI-Workloads und eine reibungslosere Cloud-Integration.
- **NetApp Lösungen für einheitliches Datenmanagement** ermöglichen, Daten reibungslos und kostengünstig in der Hybrid-Multi-Cloud-Umgebung zu bewegen.
- **Das erstklassige NetApp Partner-Ecosystem** bietet vollständige technische Integration mit führenden KI-Anbietern, Channel-Partnern, Systemintegratoren, Software- und Hardwareanbietern und Cloud-Partnern. Gemeinsam entstehen intelligente, leistungsstarke und vertrauenswürdige KI-Lösungen, mit denen Sie Ihre Geschäftsziele erreichen.
- **NetApp Professional Services** helfen Ihnen mit ihrer Fachkompetenz, die Komplexität zu reduzieren sowie dei Möglichkeiten und den Erfolg Ihrer KI zu steigern.

NetApp ist übrigens laut IDC MarketScape weltweit führender Anbieter von dateibasiertem Scale-out-Storage.[5] Das ist ein wichtiger Punkt, denn bei daten- und dateiintensiven Workloads, kommt es genau darauf an.

Machen Sie Ihre Data Scientists glücklich

5 ×

5-mal mehr Daten durch die KI-Pipeline leiten

<60 Sekunden

Datensätze in Sekunden statt in Stunden oder Tagen kopieren

~20 Minuten

KI-Infrastruktur in ca. 20 Minuten konfigurieren dank Integration in Ansible

NetApp Retail Assistant: Blaupause für den Erfolg

NetApp und NVIDIA haben mit Jarvis, einem End-to-End Framework von NVIDIA für die Entwicklung konversationeller KI-Servies, einen Virtual Retail Assistant entwickelt, der Input in gesprochener und geschriebener Sprache akzeptiert und Fragen zum Wetter, zu Interessengebieten und Preisen beantwortet. Dafür stellt der Assistent eine Verbindung zur Weatherstack-API, der Yelp Fusion-API und dem eBay Python-SDK her. Mehr dazu hier.

Der NetApp Retail Assistant (NARA) basiert auf:

- **NVIDIA Jarvis**: Jarvis bietet GPU-beschleunigte Services für konversationelle KI mit einer End-to-End-Pipeline für Deep Learning, die für niedrige Latenzzeiten optimiert ist.
- **NetApp ONTAP AI**: In dieser bewährten Architektur werden NVIDIA DGX-Systeme und NetApp All-Flash-Storage miteinander kombiniert. ONTAP AI sorgt für eine zuverlässige Optimierung des Datenflusses und ermöglicht damit das Training und die Ausführung komplexer Konversationsmodelle, ohne dass die Latenzobergrenzen überschritten werden.
- **NVIDIA NeMo**: Als Python-Toolkit bietet NeMo Möglichkeiten für die Erstellung, das Training und die Feinabstimmung von GPU-beschleunigten konversationellen KI-Modellen. Mit NeMo lassen sich Modelle entwickeln, die sich durch benutzerfreundliche APIs auszeichnen und Anwendungen für die automatische Spracherkennung (Automatic Speech Recognition, ASR), die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) und die Umwandlung von Text in Sprache (Text-to-Speech, TTS) bieten.

Reicht das zum Thema NLP?

Gut erkannt. Was steht als Nächstes an? Gespräche mit Wildtieren? Wir können Eichhörnchen nicht das Sprechen beibringen. Wir können Ihnen aber zeigen, wie Sie die richtige KI-Infrastruktur für NLP entwickeln.

Weitere Informationen zu NetApp Lösungen für KI:

- NetApp AI
- ONTAP AI
- NetApp Lösungen für NLP

Fragen? Sprechen Sie noch heute mit einem KI-Spezialisten von NetApp.

1. O'Neill, Shane. Real-time bidding: What happens in 200 milliseconds? Nanigans.
2. Welsh, Tim. Exactly how long does it take to think a thought? The Christian Science Monitor. 1. Juli 2015.
3. Wargo, Eric. How Many Seconds to a First Impression? Association for Psychological Science. 1. Juli 2006.
4. Wise, Jeff. What Is the Speed of Thought? New York Magazine. 19. Dezember 2016.
5. Potnis, Amita. IDC MarketScape: Worldwide Scale-Out File-Based Storage 2019 Vendor Assessment. IDC. Dezember 2019.

Info zu NetApp

In einer Welt voller Generalisten beweist sich NetApp als Spezialist. Wir haben ein Ziel fest im Blick: Ihr Unternehmen darin zu unterstützen, Ihre Daten optimal zu nutzen. NetApp bringt die Datenservices, denen Sie vertrauen, in die Cloud und die Einfachheit und Flexibilität der Cloud in Ihr Datacenter. Selbst bei höchsten Ansprüchen lassen sich die branchenführenden NetApp Lösungen in unterschiedlichsten Kundenumgebungen und den weltweit führenden Public Clouds einsetzen.

Als Cloud- und Daten-orientierter Softwareanbieter stellt nur NetApp alle Technologien bereit, mit denen Sie Ihre eigene maßgeschneiderte Data Fabric aufbauen, Ihre Clouds vereinfachen, Ihre Public Clouds anbinden und so die richtigen Daten, Services und Applikationen sicher bereitstellen können – immer und überall.

Weitere Informationen erhalten Sie unter **www.netapp.de**