Sprache, Spracheingabe, Text & Übersetzung mit Google Cloud APIs

M.Sc. Onur Yilmaz

Angewandte Künstliche Intelligenz

Schriftliche Ausarbeitung - Cloud Computing

Fachhochschule Südwestfalen

Gutachter: Prof. Dr. Giefers

7. August 2023

Inhaltsverzeichnis

Einleitung

Im Rahmen dieser Arbeit werden verschiedene Technologien und Anwendungen im Bereich Sprache und Textverarbeitung vorgestellt, die auf den Diensten von Google Cloud basieren, einschließlich der *Cloud Speech API*, der *Cloud Translation API*, der *Natural Language API* und der *Text-to-Speech API* [1].

Im Abschnitt über die *Grundlagen* wird eine Einführung in die Konzepte der API, des Cloud Computing und die Relevanz von Sprachtechnologien in der heutigen Zeit gegeben.

Der Abschnitt *Spracherkennung und -transkription* fokussiert sich auf die *Cloud Speech API*, die die Transkription von Audio in Text ermöglicht, und die Methoden zur Messung und Verbesserung der Sprachgenauigkeit.

In der *Sprachübersetzung* wird die *Cloud Translation API* behandelt, die den Prozess der Übersetzung von Texten in verschiedene Sprachen ermöglicht.

Der Bereich *Textanalyse* befasst sich mit der *Natural Language API*, die Techniken zur Klassifizierung von Text in Kategorien und zur Analyse von Entitäten und Sentiments bietet.

Im Abschnitt *Sprachsynthese* wird die *Text-to-Speech API* vorgestellt, die die Erzeugung synthetischer Sprache ermöglicht.

Die Arbeit dient nicht nur als theoretischer Überblick, sondern bietet auch praktische Einblicke und Anleitungen zur Verwendung dieser Tools. Dabei werden unterschiedliche Schwierigkeitsgrade und Themenbereiche abgedeckt, um einen umfassenden Einblick in die Möglichkeiten der Sprach- und Textverarbeitung mit Google Cloud zu bieten.

1 Grundlagen

1.1 Was ist eine API

Eine API (*Application Programming Interface*) ist eine Schnittstelle, die es verschiedenen Softwareanwendungen ermöglicht, miteinander zu kommunizieren. Es handelt sich im Wesentlichen um eine Reihe von Regeln und Protokollen, die von den Entwicklern befolgt werden müssen, um auf die Funktionen eines Softwareprodukts zuzugreifen. APIs können in verschiedenen Formen existieren, wie z.B. Web-APIs, die über HTTP-Kommunikation arbeiten, oder als Bibliotheken und Frameworks für spezifische Programmiersprachen.

1.2 Erstellen eines API-Schlüssels

Ein API-Schlüssel ist ein eindeutiger Identifikator, der verwendet wird, um eine Anwendung zu authentifizieren, die auf die Funktionen einer API zugreifen möchte. Dieser Schlüssel dient als eine Art "Passwort", das sicherstellt, dass nur autorisierte Anwendungen Zugang zur API haben. Die genauen Schritte zum Erstellen eines API-Schlüssels können je nach Anbieter der API variieren, aber in der Regel beinhalten sie das Anmelden bei einem Entwicklerportal, das Erzeugen eines neuen Schlüssels über eine Benutzeroberfläche und das Kopieren dieses Schlüssels in die Anwendung, die die API nutzen wird.

1.3 Erstellen und Aufrufen der API-Anfrage

Um eine API-Anfrage zu erstellen, muss ein Entwickler eine HTTP-Anfrage (normalerweise GET oder POST) an die URL der API senden, zusammen mit allen erforderlichen Parametern und dem API-Schlüssel. Die Anfrage kann auch einen Anforderungskörper enthalten, der zusätzliche Daten zur Verarbeitung an die API sendet. Nach dem Senden der Anfrage wird die API diese verarbeiten und eine Antwort zurückgeben, normalerweise in Form einer JSON- oder XML-Datei, die die angeforderten Daten oder das Ergebnis der Verarbeitung enthält.

1.4 Interaktion mit der Google Cloud API

Der Prozess der Interaktion mit der Google Cloud API kann in vier grundlegende Schritte unterteilt werden:

1. Der Entwickler sendet eine API-Anfrage von seiner Anwendung aus. Diese Anfrage enthält die erforderlichen Parameter und den API-Schlüssel. Ein Beispiel für eine solche Anfrage könnte folgendermaßen aussehen:

   ```
   GET /api/v1/resource?param=value&api_key=YOUR_API_KEY
   ```

2. Die Anfrage wird an die Google Cloud API gesendet, die auf einem Server von Google gehostet wird. Dies geschieht im Hintergrund durch die Netzwerkinfrastruktur und erfordert in der Regel keinen spezifischen Code.

3. Die Google Cloud API verarbeitet die Anfrage. Dies kann beinhalten, dass sie auf Daten in einer Datenbank zugreift, Berechnungen durchführt oder andere Funktionen ausführt. Auch dies geschieht auf dem Server und ist für den Entwickler in der Regel nicht sichtbar.

4. Die Google Cloud API sendet eine Antwort zurück an die Anwendung des Entwicklers. Diese Antwort enthält die Daten oder das Ergebnis, das der Entwickler angefordert hat. Eine typische Antwort könnte folgendermaßen aussehen:

   ```
   {
   "status": "success",
   "data": {
   // The requested data or result of the operation
   }
   }
   ```

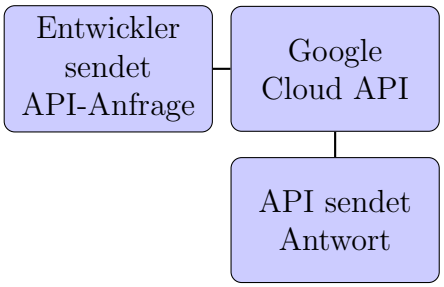

Abbildung 1: Interaktion mit der Google Cloud API

2 Spracherkennung und -transkription

Die Spracherkennung und -transkription, auch als *Automatic Speech Recognition* (ASR) bekannt, ist ein wichtiger Bereich der künstlichen Intelligenz und der Signalverarbeitung. In diesem Abschnitt werden wir uns damit beschäftigen, wie gesprochene Sprache in geschriebenen Text umgewandelt werden kann. Dieser Prozess wird als Transkription bezeichnet.

2.1 Transkription von Sprache zu Text

Die Transkription von Sprache zu Text ist das Verfahren, bei dem Audiosignale, die menschliche Sprache enthalten, analysiert werden, um den entsprechenden Textinhalt zu erzeugen. Dies kann in vielen Anwendungen nützlich sein, von automatischen Untertiteln für Videos bis hin zur sprachbasierten Interaktion mit virtuellen Assistenten.

2.2 Schritte im Speech-to-Text Prozess

Der Prozess der Sprache-zu-Text-Transkription kann in mehrere Teilschritte zerlegt werden, darunter:

1. **Vorverarbeitung oder Signalverarbeitung:** Hier wird das rohe Audiosignal in ein geeignetes Format für die weitere Verarbeitung umgewandelt. Dies kann das Filtern von Hintergrundgeräuschen oder das Aufteilen des Signals in kleinere Segmente, so genannte "Frames", umfassen.

2. **Merkmalsextraktion:** In diesem Schritt werden relevante Merkmale aus dem vorverarbeiteten Signal extrahiert. Diese Merkmale können die Lautstärke, den Ton und die Geschwindigkeit der Sprache umfassen.

3. **Akustische Modellierung:** Hier wird das extrahierte Merkmalssignal mit einem akustischen Modell verglichen, das auf der Grundlage von Trainingsdaten erstellt wurde. Dieses Modell kann verwendet werden, um die wahrscheinlichsten Phonemsequenzen für das gegebene Signal zu bestimmen.

4. **Sprachmodellierung:** In diesem Schritt wird die erkannte Phonemsequenz in Worte und Sätze übersetzt. Dies geschieht mit Hilfe eines Sprachmodells, das Informationen über die Wahrscheinlichkeiten verschiedener Wort- und Satzstrukturen enthält.

5. **Dekodierung:** Der letzte Schritt besteht darin, die wahrscheinlichste Wortsequenz zu finden, die das erkannte Phonemsignal und das Sprachmodell erklärt. Dieser Schritt kann als ein Optimierungsproblem angesehen werden, bei dem das Ziel darin besteht, die Wortsequenz zu finden, die die Gesamtwahrscheinlichkeit maximiert.

2.3 Messung und Verbesserung der Sprachgenauigkeit

Die Genauigkeit einer Sprach-zu-Text-Transkription kann auf verschiedene Arten gemessen werden, und es kann nützlich sein, mehrere Metriken zu verwenden, abhängig von den spezifischen Anforderungen. Eine gängige Methode, die oft als Standard für Vergleiche herangezogen wird, ist die *Word Error Rate* (WER). Die WER misst den Anteil der falsch transkribierten Wörter im gesamten Datensatz. Dies bedeutet, dass eine niedrigere WER eine höhere Transkriptionsgenauigkeit anzeigt.

Im Kontext der ASR-Genauigkeit wird häufig der Begriff *Grundwahrheit* (Ground Truth) verwendet. Grundwahrheit bezeichnet die 100% genaue (typischerweise menschliche) Transkription, mit der die Genauigkeit der maschinellen Transkription verglichen wird.

- **Einfügefehler (I):** Dies bezieht sich auf Situationen, in denen Wörter, die nicht in der ursprünglichen, korrekten Transkription (der "Grundwahrheit") vorhanden sind, vom Spracherkennungssystem in das transkribierte Ergebnis eingefügt werden.
- **Substitutionsfehler (S):** Dies tritt auf, wenn Wörter, die sowohl in der ursprünglichen Transkription als auch im transkribierten Ergebnis vorhanden sind, vom System nicht korrekt transkribiert werden. Hierbei wird ein Wort aus der Grundwahrheit durch ein anderes Wort im transkribierten Text ersetzt.
- **Löschfehler (D):** Dies bezieht sich auf Wörter, die in der Grundwahrheit vorhanden sind, aber im transkribierten Text des Spracherkennungssystems fehlen.

Die Formel zur Berechnung der WER ist:

$$\text{WER} = \frac{S + D + I}{N}, \tag{1}$$

wobei N die Anzahl der Wörter in der Grundwahrheit (dem korrekten Text) ist.

2.4 Google Cloud Speech API

In diesem Abschnitt lernen wir nun verstehen wie eine Audiodatei zur Transkription an die Speech-to-Text API gesendet wird.

Erstellen einer API-Anforderung

Um eine Anforderung an die API zu senden, erstellen wir eine `request.json` Datei. Diese Datei enthält zwei Objekte: `config` und `audio`.

Im `config` Objekt teilen wir der Speech-to-Text API mit, wie die Anforderung verarbeitet werden soll. Der Parameter `encoding` gibt an, welche Art von Audiokodierung wir verwenden, und `languageCode` gibt die Sprache des Audiomaterials an.

Im `audio` Objekt übergeben wir der API die URI der Audiodatei, die in diesem Fall in Google Cloud Storage gespeichert ist.

Aufrufen der Speech-to-Text API

Wir senden die Anforderung zusammen mit dem API-Schlüssel an die API. Die Antwort der API wird in einer Datei namens `result.json` gespeichert. Diese enthält das transkribierte Audio sowie eine Konfidenzangabe, die angibt, wie sicher die API ist, dass sie das Audio korrekt transkribiert hat.

3 Sprachübersetzung

3.1 Erkennung und Übersetzung von Texten

4 Textanalyse

4.1 Klassifizierung von Text in Kategorien

4.2 Entitäten- und Sentimentanalyse

5 Sprachsynthese

5.1 Erzeugung synthetischer Sprache

Literatur

[1] Google LLC. Language, speech, text, & translation with google cloud apis, 2023.