Eine KI die 100 Sprachen versteht

Andre Soblechero Salvado

Kurzgefasst

1. **Das Ziel** ist es ein Modell zu trainieren, welches ähnliche Vektoren für Texte erzeugt, welche wiederum unabhängig von der Sprache das Gleiche aussagen.

2. Es wird ein **großer Datensatz** verwendet, welcher aus Übersetzungen von alltäglichen aber auch rechtlichen Texten besteht.

3. Es wir ein **Modell trainiert**, welches einen Satz in einer Sprache entgegennimmt und in eine andere Sprache übersetzt.
Dieses vorgehen nennt sich fachlich Machine Translation (MT)

Wozu?

Was wäre wenn Sie eine Aufgabe in einer Sprache lernen und dann diese Aufgabe in 100 unterschiedlichen Sprachen ausführen könnten?

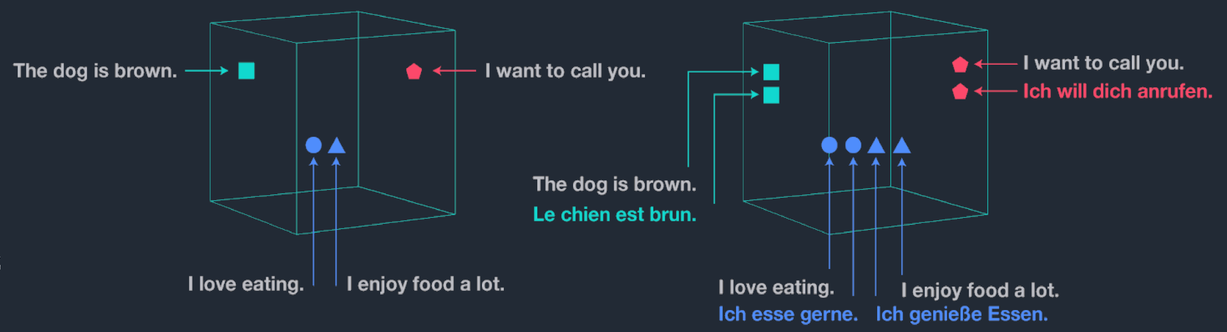
Durch diese Möglichkeit müssten nur noch Daten in einer Sprache existieren um ein Machine Learning Modell zu trainieren.

Zusätzlich müssen nicht unzählige Variationen von beispielsweise „Wie geht es dir?“ in den Trainingsdatenpunkten existieren, da dass Modell für ähnliche Texte ähnliche Vektoren erzeugt.

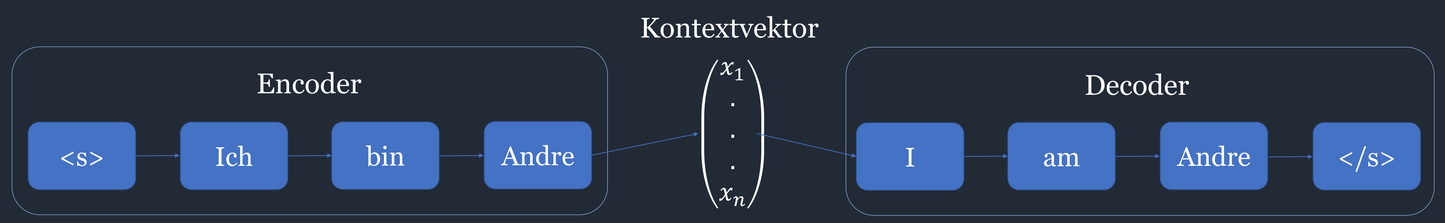
Ziel

Das Ziel ist es, wie rechts visualisiert, Texte in unterschiedlichen Sprachen, welche aber eine ähnliche Aussage besitzen durch Vektoren, welche nahe beieinander liegen, darzustellen.

Praktisch sind dies im Gegensatz zur Visualisierung meist viel mehr Dimensionen.



Modell



Der **Encoder** nimmt einen Text entgegen, welcher verarbeitet wird und gibt einen Kontextvektor aus.

Der **Decoder** nimmt den Kontextvektor entgegen um daraus die Übersetzung Wort für Wort, bis </s> auftaucht, zu generieren.

Durch dieses Training wird das Modell dazu gezwungen im **Kontextvektor** alle für die Übersetzung notwendigen Informationen zu speichern. Dazu gehören die Repräsentationen der Eingangstexte und somit mehr oder weniger deren Aussage.

Daten

Der verwendete Datensatz besteht aus 50 Millionen Übersetzungen von 100 unterschiedlichen Sprache ins englische. Unter den 100 Sprachen sind auch Sprachen wie Uighurisch vertreten von welcher generell wenig Ressourcen existieren.

Referenzen

3 Dimensionales-Bild: https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/

LASER Paper, Grundlage dieser Ausarbeitung: https://arxiv.org/abs/1812.10464

HAW Logo: https://iconape.com/wp-content/files/gh/192683/svg/192683.svg