

Verbesserte Sprachübersetzung für Schweizerdeutsch mit einem hybriden Dynamic-Window-Ansatz

Die Forschung arbeitet seit mehreren Jahren intensiv an Speech Translation Systemen, die schweizerdeutsche Sprache zu deutschem Text übersetzen sollen. Das Fehlen einer offiziellen Grammatik sowie die Vielzahl an Dialekten stellen dabei grosse Herausforderungen an ein Speech Translation System.

Die meisten Systeme für Schweizerdeutsch werden aktuell auf satzweiser Übersetzung trainiert und ausgewertet. In der Realität sind Audioaufnahmen aber meistens mehrere Minuten lang und beinhalten mehr als einen Satz / Sprecher. Mit einem System, das darauf trainiert wurde, einen Satz zu übersetzen, kann man deswegen nicht direkt längere Sequenzen ohne Qualitätseinbussen übersetzen.

Mit genau diesem Problem beschäftigt sich diese Arbeit. Mittels synthetischer Datensätze werden längere Szenarien simuliert. Um die Audiodatei aufzuteilen, verfolgt die erhaltene Pipeline einen Sliding Window Ansatz. Wir zeigen, dass durch intelligentes Preprocessing mit Algorithmen wie Speech Activity Detection oder Speaker Diarization die Übersetzungsqualität auf längeren Sequenzen um bis zu 125% auf 58.19 BLEU Punkte gesteigert werden kann. Dies ist schon sehr nahe an der theoretischen Referenz, bei der jeder Satz einzeln übersetzt wird und damit 59.69 BLEU Punkte erreicht werden. Bei noch schwierigeren Sequenzen ohne Sprechpause, bei denen die Audiodatei nicht mehr einfach aufgeteilt werden kann, belegen wir, dass ein dynamischer Window Ansatz die Übersetzungsqualität verdoppelt. Auf dem komplexeren Podclub Datensatz, der aus Podcastaufnahmen besteht, verbessert sich der BLEU Score mit einer Kombination aus Preprocessing und dynamischem Window Ansatz, dem hybriden Dynamic Window Ansatz, um 64% auf 41.57 Punkte. Eine qualitative Analyse zeigt aber, dass das Übersetzungssystem noch nicht genügend gute Übersetzungen liefert, die für automatische Transkriptionen ausreichen.



Diplomierende

Lukas Bamert
Kevin Kläger

Dozierende

Jan Milan Deriu
Mark Cieliebak

Bild klein 1.