Zürcher Hochschule
für Angewandte Wissenschaften

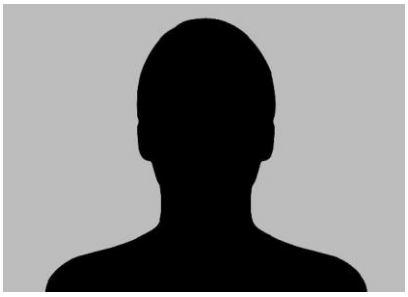
School of Engineering

Untersuchung von Pre-Training von Wav2Vec2 auf schweizerdeutschen Dialekten für Sprachübersetzung und Klassifizierung

Sprachen und Dialekte mit geringen Datenmengen, wie z.B. Schweizerdeutsch, erfordern Systeme, die über zahlreiche Sprachen hinweg verallgemeinern können, um moderne Sprachübersetzungs- und Erkennungsanwendungen zu entwickeln.

Diese Arbeit testet die Fähigkeiten des Transformer-basierten, vortrainierten, sprachübergreifenden Wav2Vec2-XLS-R Modells auf schweizerdeutschen Korpora. Wir werten das System anhand eines Sprachübersetzungssystems von Schweizerdeutsch auf Standarddeutsch als auch eines Klassifikationssystems, welches Dialekte vier Regionen zuweist, aus. Wir wenden zusätzlich 2100 Stunden unlabelled schweizerdeutsche Daten in einem pre-training Verfahren auf das Modell an, um die Auswirkungen dieser Daten auf das bereits vortrainierte System zu untersuchen. Das Ergebnis dieser Arbeit ist ein Übersetzungssystem, das 18,08% WER und 68,86 BLEU auf dem 'SNF' Testkorpus und 68,05 BLEU auf dem 'SDS-200' Test-Split erreicht. Es belegte den ersten Platz im '2nd Swiss German Speech to Standard German Text' SwissText shared task mit 68,1 BLEU auf dem privaten Evaluationssplit.

Im Klassifikationsexperiment, welches Schweizer Dialekte in vier verschiedene Regionen kategorisiert hat, wird ein gewichteter F1-Score von 0,49 erreicht, wobei die Ostschweizer Region den besten F1 mit 0,68 erzielt. Wir haben gezeigt, dass die Verwendung zusätzlicher pre-train Daten in dieser Grössenordnung beim XLS-R-Modell für die Sprachübersetzung nicht von Vorteil ist, sich aber bei der Klassifizierung positiv auswirken kann. Mithilfe einer Diskussion für zukünftige Forschungsansätze hoffen wir, dass das Interesse für dieses Themengebiet steigt.



Diplomierende
Patrik Randjelovic
Samuel Jason Stucki

Dozierende
Jan Milan Deriu
Mark Cieliebak

Bild klein 1.