October 3rd 2023

Wie gelangt man von Low-Resource-Herausforderungen zu High-Tech-Lösungen?

Florian Schottmann, Head of Research

textshuttle.

textshuttle.

Warum Textshuttle? Für Unternehmen Registrieren Anmelden

Englisch (erkannt)

We are Textshuttle.

Sprache suchen

- ✓ Deutsch
- Französisch
- Italienisch
- Rätoromanisch
- Schweizerdeutsch Beta

Plus 15 weitere Sprachen in unserer Business-Lösung

Sprachform

Feedback

DOC PPT
Ziehen Sie eine Datei hierhin

Hochladen

19 / 7500

21

↓ Über uns

textshuttle.

Warum Textshuttle? Für Unternehmen Registrieren Anmelden

Englisch (erkannt) ↔ Schweizerdeutsch Beta Sprachform

Textshuttle allows you to translate into Swiss German.

Textshuttle ermöglichts Ihne, is Schwizerdütsche z übersetze.

DOC PPT
Ziehen Sie eine Datei hierhin
Hochladen

54 / 7500

61

Feedback

↓ Über uns

Herausforderungen mit Schweizerdeutsch in maschineller Übersetzung

1. Sehr wenige Trainingsdaten öffentlich zugänglich (~20.000 parallele Segmente zu Hochdeutsch für verschiedene schweizerdeutsche Dialekte)
2. Unklar, wie Modelle trainiert werden können, die konsequent in einen Dialekt übersetzen
3. Performance ist schwer automatisch zu evaluieren

Herausforderung #1: Knappheit der Trainingsdaten

Herausforderung #1: Knappheit der Trainingsdaten

Wahl der Trainingsmethode

Erster Praxistest:

Finetuning eines bestehenden Modells vs. Training eines neuen Modells (From Scratch)

Herausforderung #1: Knappheit der Trainingsdaten

Wahl der Trainingsmethode

Erster Praxistest:

Finetuning eines bestehenden Modells vs. Training eines neuen Modells (From Scratch)

Methode	Durchschn. DA-Wert ↑	Durchschn. z-Wert ↑
Finetuning (bestes Modell)	79.365	0.051
From Scratch (bestes Modell)	**85.107**	**0.246**
...		

Herausforderung #1: Knappheit der Trainingsdaten

Experimente mit verschiedenen Datenmengen

	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23

Herausforderung #1: Knappheit der Trainingsdaten

Experimente mit verschiedenen Datenmengen

	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23

→ Schlussfolgerung: Nur die öffentlich zugänglichen Daten zu verwenden ist aussichtslos.

Herausforderung #1: Knappheit der Trainingsdaten

Experimente mit verschiedenen Datenmengen

	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23
Öffentlich verfügbare Daten + erstellte Übersetzungen	**14.8**	**0.49**

Herausforderung #1: Knappheit der Trainingsdaten

Experimente mit verschiedenen Datenmengen

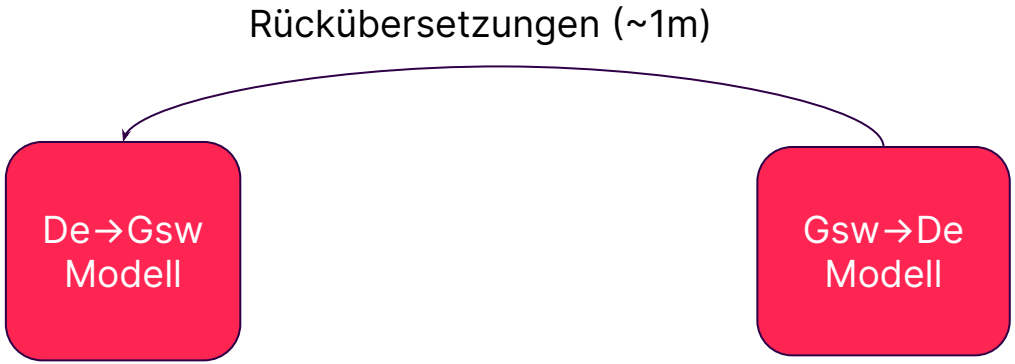
	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23
Öffentlich verfügbare Daten + erstellte Übersetzungen	14.8	0.49
Öffentlich verfügbare Daten + erstellte Übersetzungen + Rückübersetzungen	**16.3**	**0.53**

Herausforderung #1: Knappheit der Trainingsdaten

Erstellung von Rückübersetzungen

Herausforderung #1: Knappheit der Trainingsdaten

Erstellung von Rückübersetzungen



Improving Neural Machine Translation Models with Monolingual Data Sennrich et al. (2016)

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Zielzustand: Das Modell basiert auf Wortrepräsentationen, die sich robust gegenüber Abweichungen der Schreibweise eines Wortes verhalten.

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Zielzustand: Das Modell basiert auf Wortrepräsentationen, die sich robust gegenüber Abweichungen der Schreibweise eines Wortes verhalten.

→ Relevante Data Augmentation Methode: BPE-Dropout

Subwords werden mit 10%iger Wahrscheinlichkeit in kürzere Subwords zerlegt

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Hintergrund:

- Trainingsbeispiele werden vor dem Modelltraining in Subwords aufgeteilt
- Maschinelle Übersetzungsmodelle lernen, einen Text Subword pro Subword zu übersetzen
- Sequenzen an Zeichen, die oft nacheinander auftauchen, werden in Subwords zusammengefügt, bis eine Gesamtzahl an Subwords erreicht ist

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Beispiel:

SRC: Übersetzung

→ BPE: [“Über”, “setzung”]

TRG: Übersetzig

→ BPE: [“Über”, “setzig”]

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Beispiel:

SRC: Übersetzung

→ BPE: [“Über”, “setzung”] vs. BPE-Dropout: [“Über”, “setz”, “ung”], [“Ü”, “ber”, “set”, “zung”], …

TRG: Übersetzig

→ BPE: [“Über”, “setzig”] vs. BPE-Dropout: [“Über”, “se”, “tz”, “ig”], [“Übers”, “etz”, “ig”], …

Herausforderung #1: Knappheit der Trainingsdaten

Vorbereitung des Textes für das Modelltraining

Konsequenz: Wortrepräsentationen sind robuster gegenüber Abweichungen der Schreibweise

withdra

BPE	BPE-dropout
aimed	withd
molecules	withdrawal
aromatic	withdraw
specialties	withdrawn
publishers	withdrew

5-Nearest-Neighbours der gelernten Repräsentation von “withdra” im Englischen*

*Weitere Details in BPE-Dropout: Simple and Effective Subword Regularization (Provilkov et al., 2020)

Herausforderung #2: Konsistente Übersetzung in verschiedene Dialekte

Herausforderung #2: Konsistente Dialektübersetzung

Automatische Dialektidentifikation

Ziel:

Herausfinden, zu welchem Dialekt ein schweizerdeutsches Segment gehört

Herausforderung #2: Konsistente Dialektübersetzung

Automatische Dialektidentifikation

Ziel:

Herausfinden, zu welchem Dialekt ein schweizerdeutsches Segment gehört

Lösung:

Training eines Klassifizierungsmodells für schweizerdeutsche Dialekte

Herausforderung #2: Konsistente Dialektübersetzung

Automatische Dialektidentifikation

Ziel:

Herausfinden, zu welchem Dialekt ein schweizerdeutsches Segment gehört

Lösung:

Training eines Klassifizierungsmodells für schweizerdeutsche Dialekte

- Nutzung gecrawlter Daten
- Labels anhand der Textherkunft zugewiesen

Herausforderung #3: Evaluation schweizerdeutscher Übersetzungen

Herausforderung #3: Evaluation der Übersetzungen

Genauigkeit von Evaluationsmetriken

Bereits gezeigte Ergebnisse:

	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23
Öffentlich verfügbare Daten + erstellte Übersetzungen	14.8	0.49
Öffentlich verfügbare Daten + erstellte Übersetzungen + Rückübersetzungen	**16.3**	**0.53**

Herausforderung #3: Evaluation der Übersetzungen

Genauigkeit von Evaluationsmetriken

Bereits gezeigte Ergebnisse:

	BLEU	chrf++
Öffentlich verfügbare Daten	2.2	0.23
Öffentlich verfügbare Daten + erstellte Übersetzungen	14.8	0.49
Öffentlich verfügbare Daten + erstellte Übersetzungen + Rückübersetzungen	**16.3**	**0.53**

Zwischen diesen Systemen besteht ein spürbarer Qualitätsunterschied.

Herausforderung #3: Evaluation der Übersetzungen

Genauigkeit von Evaluationsmetriken

Was danach kam:

	BLEU	chrf++
System 1	16.3	0.53
System 2	17.3	0.54
System 3	16.8	0.54
...	...	...

Nach gezielter Auswertung sahen wir, dass sich die Systeme oft auch in Iterationen verbessert haben, in denen die Metriken keine Verbesserung anzeigen.

Herausforderung #3: Evaluation der Übersetzungen

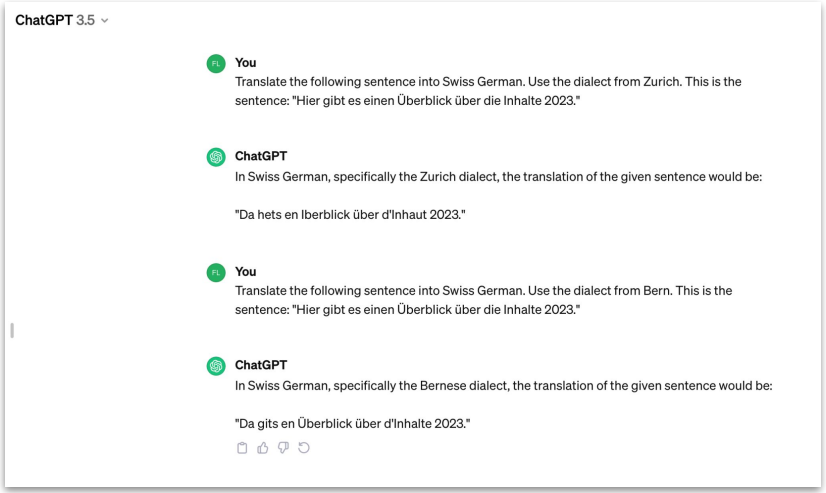
Genauigkeit von Evaluationsmetriken

Herkömmliche Evaluationsmetriken für maschinelle Übersetzungen sind oft ungeeignet, um zwischen zwei konkurrierenden MT-Systemen für Schweizerdeutsch zu entscheiden.

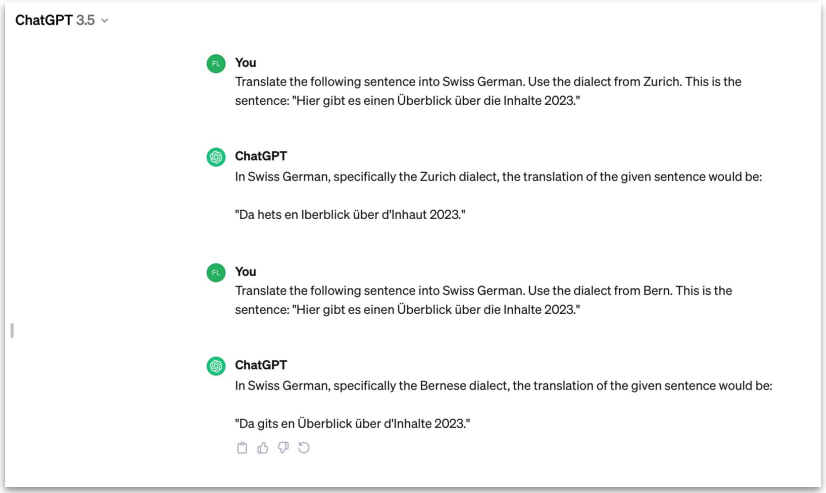
	Pairwise Accuracy
BLEU	0.695
chrF++	0.712
COMET-22	0.712
COMET QE	0.627
Ours	**0.864**

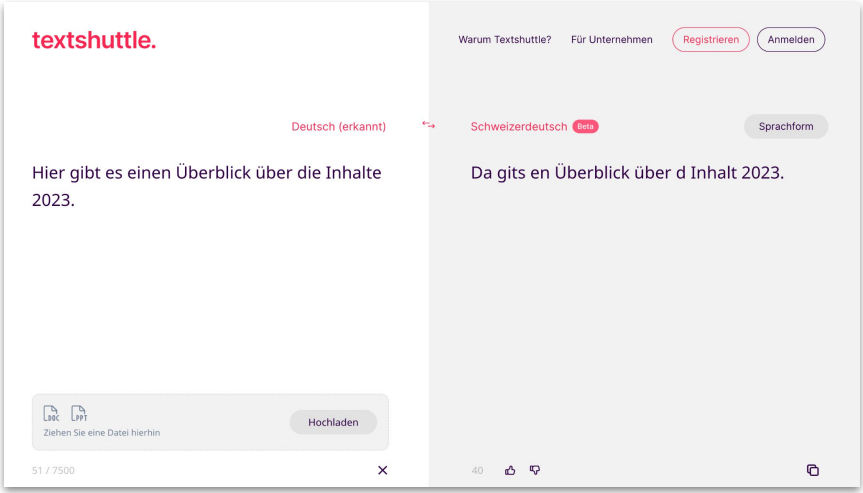
Ausblick

Outlook


ChatGPT 3.5
You
Translate the following sentence into Swiss German. Use the dialect from Zurich. This is the sentence: "Hier gibt es einen Überblick über die Inhalte 2023."
ChatGPT
In Swiss German, specifically the Zurich dialect, the translation of the given sentence would be:
"Da hets en Iberblick über d'Inhaut 2023."
You
Translate the following sentence into Swiss German. Use the dialect from Bern. This is the sentence: "Hier gibt es einen Überblick über die Inhalte 2023."
ChatGPT
In Swiss German, specifically the Bernese dialect, the translation of the given sentence would be:
"Da gits en Überblick über d'Inhalte 2023."

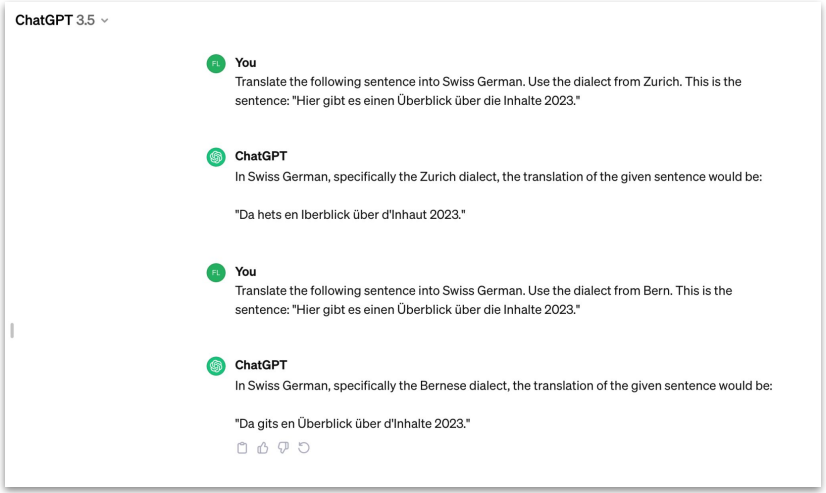
Outlook

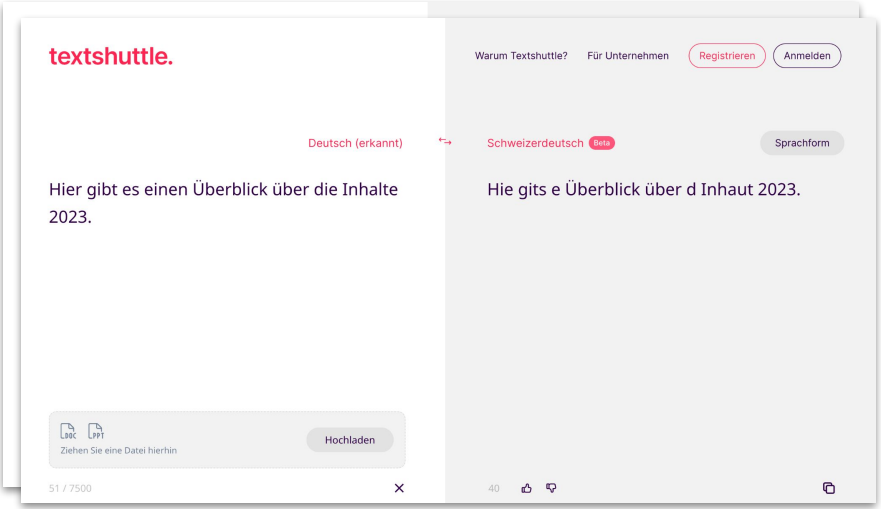




Zurich dialect

Outlook





Bernese dialect

Textshuttle
Schaffhauserstrasse 339
8050 Zürich
Schweiz

+41 (0)44 500 73 45
welcome@textshuttle.com

Danke
Gracias
Merci
Dziękuję
Tack