IP5 - NLP Schweizerdeutsch

Fabio Strappazzon, Matthias Ernst

Inhalt

[Einführung 3](#_Toc497652749)

[Zusammenfassung 3](#_Toc497652750)

[Problemstellung 3](#_Toc497652751)

[Vorgehen 3](#_Toc497652752)

[NLP 4](#_Toc497652753)

[Daten 5](#_Toc497652754)

[Tasks 6](#_Toc497652755)

[Task 1 6](#_Toc497652756)

[Satzbewertung 6](#_Toc497652757)

[Ausfilterung 6](#_Toc497652758)

[Wortalignierung 6](#_Toc497652759)

[Fehlerkorrektur 6](#_Toc497652760)

[Ergebnis 6](#_Toc497652761)

[Task 2 7](#_Toc497652762)

[Ergebnis 8](#_Toc497652763)

[Schlussfolgerung 9](#_Toc497652764)

# Einführung

## Zusammenfassung

In dieser Arbeit werden Übersetzungen und Transkriptionen, die auf der Website [www.dindialaekt.ch/](http://www.dindialaekt.ch/) erfasst wurden, mithilfe von NLP-Tools analysiert und verarbeitet.

## Problemstellung

Auf der Website [www.dindialaekt.ch/](http://www.dindialaekt.ch/) können Benutzer unter Anderem zwei Tasks erledigen:

1. Transkription schweizerdeutschen Audiofiles
2. Übersetzung von schweizerdeutschem Text auf Hochdeutsch

Zu jedem Audiofile und jedem schweizerdeutschen Ausgangstext gibt es mehrere (etwa 3 bis 7) zusammengehörende Transkriptionen / Übersetzungen.

Aus den so gewonnenen Transkriptionen und Übersetzungen sollen nun über Wortalignierung mögliche Schreibweisen für schweizerdeutsche Ausdrücke (Task 1) und mögliche schweizerdeutsch Übersetzungen für hochdeutsch Ausdrücke gefunden werden (Task 2).

Die Aufgabe wird dadurch erschwert, dass viele Transkriptionen unvollständig (Mit Platzhaltern in Form von \*\*\* oder ???) oder gar komplett falsch sind. Weiter haben Benutzer bei Task 2 die Freiheit, die Satzstellung frei zu wählen.

## Vorgehen

Beide Tasks erfordern eine ähnliche Herangehensweise: Erst muss ein Algorithmus gefunden werden, der die vorhandenen zusammengehörigen Übersetzungen und Transkriptionen qualitativ bewertet. Dann muss anhand dieser Bewertung entschieden werden, welche Übersetzungen und Transkriptionen brauchbar und unbrauchbar für das weitere Verfahren sind. Mit den verbleibenden Sätzen können nun mit Satz / Wortalignment Tools Gruppen von Ausdrücken gebildet werden, die bedeutungsgleich sind.

# NLP

# Daten

Tabellarische Daten. Für swg2g folgende relevante Spalten:

TASK\_ID: identifiziert zusammengehörige Übersetzungen

INFO: vom Benutzer erfasste hochdeutsch Übersetzung

REF: schweizerdeutscher Originaltext

Auszug als CSV:

VALID;TASK\_ID;TASK\_RUN\_ID;USER\_ID;INFO;REF\_ID;REF

True;2530;16287;2887;das war dann bereits zu einem späteren Zeitpunkt, das war als....;BS:0012;das isch den scho schpööter gsii das isch gsii woo

True;2530;19101;4117;Das war bereits später, das war als;BS:0012;das isch den scho schpööter gsii das isch gsii woo

True;2531;453;1777;und mit neunzehn ist man in die RS eingerückt;BS:0016;und mit nüünze isch men in d eress iigruggt

True;2531;2245;1995;und mit 19 ist man in die RS eingerückt.;BS:0016;und mit nüünze isch men in d eress iigruggt

Für Transcribe folgende relevante Spalten:

TASK\_ID: identifiziert zusammengehörige Transkriptionen

INFO: vom Benutzer erfasste schweizerdeutsch Transkription

Auszug als CSV:

URL;VALID;TASK\_ID;TASK\_RUN\_ID;USER\_ID;INFO;REF

AUDIO;True;1829;24287;2887;Ma het dénn alz zäme glääseni Lüt was me hät chöne zämetriibe allz was Chopf u Loch hégi müessi ez höwe héts aube ghéésse;https://dindialaekt.ch/data/transcribe/SDS\_CD1\_1\_11\_speaker1\_1.mp3

AUDIO;True;1829;30820;3563;ma het denn allz zäme glääsenei lüt was mer het chöne zämetribe alts was chopf u loch heegi müesi ez höve hets aube gheesse;https://dindialaekt.ch/data/transcribe/SDS\_CD1\_1\_11\_speaker1\_1.mp3

AUDIO;True;1830;6158;3019;mer s alles us grückt d erste het dri gschnitte es paar schritt gange e zwetei made e dritt, e virtte bi al zäme drin gis sind;https://dindialaekt.ch/data/transcribe/SDS\_CD1\_1\_11\_speaker1\_2.mp3

AUDIO;True;1830;10807;2218;när isch also alles usgrückt dr erschte het dri gschnitte es paar schritt gange zweite id maade e dritte e vierte bis au zäme drin gsi si;https://dindialaekt.ch/data/transcribe/SDS\_CD1\_1\_11\_speaker1\_2.mp3

# Tasks

## Task 1

### Satzbewertung

Die Bewertung der Sätze beruht auf der Prämisse, dass «gute» Transkriptionen sich ähneln und «schlechte» Transkriptionen sich von den anderen stark unterscheiden.

So könnte beispielsweise die Levenshtein-Distanz (Editierdistanz) verwendet werden: Es wird die Levenshtein-Distanz zwischen jedem Satzpaar berechnet. Derjenige Satz, der zu allen anderen die tiefste Levenshtein-Distanz hat, ist folglich der «ähnlichste» und somit beste Satz.

Das gleiche Prinzip ist anwendbar mit Tools für die automatische Bewertung von Maschinenübersetzungen.

Word Error Rate ist eine Metrik die auf dem gleichen Prinzip wie die Levenshtein-Distanz basiert, jedoch auf Wortebene agiert, anstatt auf Phonem-Ebene.

BLEU ist ein Algorithmus, dessen Bewertung sehr gut mit deren von Menschen korrespondiert. Er ist viel verwendet und hat gute Performance. Integriert in NLTK

Einige Metriken, die BLEU-basiert sind, und versuchen dessen Schwächen zu verbessern: TER, METEOR, NIST

Weiter: LEPOR / hLEPOR

### Ausfilterung

Methoden zum Ausfiltern:

Statistik: Outlier mithilfe von Standardabweichung o.Ä. erkennen. Problem: Daten sind nicht normalverteilt, zu kleine Datenmenge

Fixer Minimalwert: gute Baseline. Für BLEU sollte dieser Wert etwa zwischen 0.3 und 0.6 liegen.

Clustering: Da es zwingend mehrere gute Sätze geben muss, damit diese als gut erkannt werden, können eindimensionale Clustering Algorithmen verwendet werden und dann die höchsten 1-2 Cluster als akzeptiert angesehen werden. Problem: Wahl des Clustering-Algorithmus in Anbetracht der extrem kleinen Datenmenge.

Iteratives Filtern und Bewerten: Solange sich die durchschnittliche Bewertung deutlich verbessert, wird der schlechteste Satz entfernt und alle anderen Sätze ohne den entfernten neu bewertet. Probleme: was ist eine «deutliche» Verbesserung? Die Tendenz der Daten zu gruppieren verfälscht das Ergebnis weiter

### Wortalignierung

Obwohl es sich hierbei um ein Wortalignierungsproblem handelt, ziehen wir nur Satzalignierer in Betracht. Dies aus folgendem Hintergrund:

…

Um Satzaligner für Wortalignierungsprobleme zu verwenden, können Sätze als Abschnitte, Wörter als Sätze und Buchstaben als Wörter angesehen werden.

Satzaligner benutzen eines oder mehrere folgender Informationen: Wörterbuch, Textlänge, Textähnlichkeit. Da es in unserem Fall keine Wörterbücher gibt, können alle Tools, die ein Wörterbuch benötigen, ausgeschlossen werden.

Bleualign ist ein Satzalignierungstool, das versucht, den BLEU Score zwischen den alignierten Sätzen zu maximieren.

Hunalign (wenn ohne Wörterbuch verwendet) aligniert in einer ersten Iteration anhand der Satzlänge. Aus der so gewonnenen Information wird ein Wörterbuch generiert, das in einer zweiten Iteration zur Verbesserung der Alignierung verwendet wird.

### Fehlerkorrektur

### Ergebnis

## Task 2

# Ergebnis

# Schlussfolgerung