IP5 - NLP Schweizerdeutsch

Fabio Strappazzon, Matthias Ernst



Betreuer: Wolfgang Weck, Daniel Kröni

Auftraggeber: Universität Zürich, Institut für Computerlinguistik

Fachhochschule Nordwestschweiz, Studiengang Informatik

Windisch, 07.01.2018

Diese Arbeit ist im Auftrag des Instituts für Computerlinguistik der Universität Zürich entstanden. Das Institut betreibt das Projekt «Citizen Linguistics: Locate that dialect!», in welchem Benutzer spielerisch schweizerdeutsche Texte klassifizieren, transkribieren und übersetzen.

Das Ziel dieser Arbeit ist die Verarbeitung von schweizerdeutschen Transkriptionen, um unterschiedliche Schreibweisen für dieselben Ausdrücke zu finden. Weiter soll es möglich sein, unvollständige Transkriptionen automatisch zu vervollständigen. Im Gebiet des Natural Language Processing ist dies unter dem Begriff der String-, Wort- oder Textalignierung bekannt. Tools und Algorithmen hierfür existieren zwar, müssen aber auf Verwendbarkeit in Bezug auf unsere Datencharakteristiken untersucht werden. Weiter sind Vor- und Nachbearbeitungsschritte (Ausfilterung unbrauchbarer Transkriptionen, qualitative Bewertung von Transkriptionen und Alignierungen, Kombination unabhängiger Alignierungen) nötig.

Wir stellen ein Verfahren vor, das aufgrund mehrerer Transkriptionen in Form eines CSV-Files eine Liste von Wortgruppen generiert. Wörter und Ausdrücke innerhalb derselben Wortgruppe sind dabei unterschiedliche Schreibweisen mit derselben Bedeutung.

**Inhaltsverzeichnis**

[1 Einführung 4](#_Toc503554303)

[1.1 Zusammenfassung 4](#_Toc503554304)

[1.2 Projektkontext 4](#_Toc503554305)

[1.3 Problemstellung 4](#_Toc503554306)

[1.4 Vorgehen 4](#_Toc503554307)

[2 Daten 6](#_Toc503554308)

[2.1 Charakteristiken der Daten 6](#_Toc503554309)

[3 Alignierung von schweizerdeutschen Transkriptionen 7](#_Toc503554310)

[3.1 Satzbewertung 7](#_Toc503554311)

[3.2 Ausfilterung 9](#_Toc503554312)

[3.3 Wortalignierung 13](#_Toc503554313)

[3.4 Filterung des Alignments 13](#_Toc503554314)

[3.5 Fehlerkorrektur / Interpolation 21](#_Toc503554315)

[3.6 Bewertung des Ergebnisses 21](#_Toc503554316)

[4 Ergebnis 22](#_Toc503554317)

[5 Schlussfolgerung 23](#_Toc503554318)

[6 Literaturverzeichnis 23](#_Toc503554319)

# Einführung

## Zusammenfassung

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

## Projektkontext

Im Rahmen des vom Schweizerischen Nationalfods geförderten Agora-Project «Citizen Linguistics: Locate that dialect!» ist die Website [www.dindialaekt.ch/](http://www.dindialaekt.ch/) entstanden.

Auf dieser Website können Benutzer unter Anderem folgenden Task erledigen:

* Transkription schweizerdeutschen Audiofiles
* Identifizieren von Dialekten mittels Lokalisierung auf einer Karte
* Übersetzung von schweizerdeutschen zu hochdeutschen Texten

Im Rahmen dieser Arbeit wollen wir aus den so gewonnenen Transkriptionen weitere Informationen gewinnen.

Zu jedem Audiofile gibt es mehrere (etwa 3 bis 7) zusammengehörende Transkriptionen.

Aus den so gewonnenen Transkriptionen sollen nun über Wortalignierung mögliche unterschiedliche Schreibweisen für schweizerdeutsche Ausdrücke gefunden werden.

Die Aufgabe wird dadurch erschwert, dass viele Transkriptionen unvollständig (Mit Platzhaltern in Form von \*\*\* oder ???) oder gar komplett falsch sind.

## Problemstellung

* Unbrauchbare Transkriptionen ausfiltern
* Gute Transkriptionen erkennen
* \*\*\* Auslassungen interpolieren aus guten Transkriptionen
* Schreibvarianten von Wörtern und Ausdrücken erkennen über String-Alignierung

Daraus ergeben sich uns folgende zwei Fragestellungen:

**Welche Tools funktionieren mit unseren Daten?**

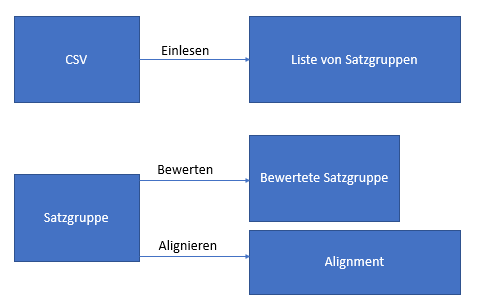
Natural Language Processing (NLP) ist ein etabliertes Gebiet. Für viele Problemstellungen gibt es Tools und Algorithmen. Viele dieser Tools haben jedoch gewisse Voraussetzungen. Beispielsweise benötigen sie einen Textkorpus einer gewissen Grösse oder sie benötigen Wörterbücher der zu untersuchenden Sprachen. Da es im Schweizerdeutschen keine genormte Rechtschreibung und daher auch keine Wörterbücher gibt, sind uns in der Wahl der Tools einige Grenzen gesetzt. Wir wollen also herausfinden, was für Tools grundsätzlich für unsere Art von Daten verwendbar sind.

**Welche und wie viele Daten brauchen wir für gute Resultate?**

Die Tools, die wir als «brauchbar» erkennen, werden jedoch auch Grenzen haben. Für die Betreiber der Website ist wichtig zu wissen, dass für eine Audiodatei ein Minimum and Transkriptionen vorhanden sein muss, um sinnvolle Analysen damit anzustellen. Konkret wollen wir also herausfinden, wie viele Transkriptionen zur selben Audiodatei vorhanden sein müssen, damit die von uns angewendeten Techniken funktionieren.

## Vorgehen

Wir teilen den gesamten Prozess in mehrere Transformationen auf, die teilweise voneinander abhängig sind.



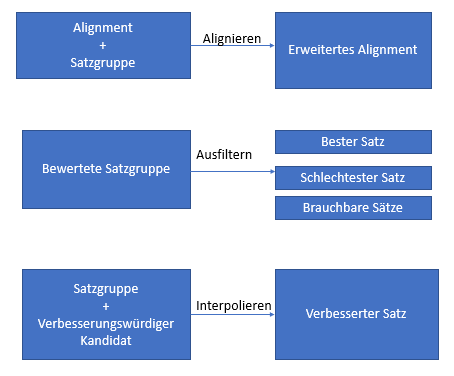


Abbildung 1: Alle Transformationen

Sätze einer Satzgruppe können nach Qualität bewertet werden. Die macht es möglich, den besten und schlechtesten Satz zu identifizieren. Dieser Vorgang ist nötig, um für das Alignment unbrauchbare Sätze auszufiltern.

Der Alignierungsprozess generiert aus einer Satzgruppe eine List von Wort/Ausdrucksgruppen, die bedeutungsgleich sind. Der Alignierungsprozess kann ausserdem eine solche bestehende Liste von Wortgruppen mit den Wörtern einer weiteren Satzgruppe erweitert werden.

Gefilterte Daten sind für den Alignierungsprozess streng genommen nicht zwingend. Da dies jedoch das Ergebnis verbessert, wird es empfohlen. Der Alignierungsprozess kann ausserdem den Satzbewertungsprozess verwenden, um das Ergebnis zu verbessern.

Der Interpolationsprozess findet Auslassungen in Form von \*\*\* in einem Satz und ersetzt sie mithilfe der anderen Transkriptionen derselben Satzgruppe. Dieser Prozess basiert auf dem Alignierungsprozess. Auch hier empfiehlt es sich, mit gefilterten Satzgruppen zu agieren.

# Daten

Tabellarische Daten. Folgende relevante Spalten:

TASK\_ID: identifiziert zusammengehörige Transkriptionen

INFO: vom Benutzer erfasste schweizerdeutsch Transkription

Auszug aus 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>

Unser Tool benötigt die Daten als UTF-8 formatiertes CSV mit ‘;’ als Spaltendelimiter.

## Charakteristiken der Daten

Stand 2017-07-08

2848 Transkriptionen zu 554 verschiedenen Audiodateien.

Pro Audiodatei sind zwischen 1 und 14 Transkriptionen vorhanden. Durchschnitt 5.14, Median 5.

So sehen zwei plausible Transkriptionen aus:

«Äs isch die Zyt wo d'Lüt afänd üsszie för d' Härdöpfel understuen fascht alli hän da un dert än Härdöpfelblätz. Ds Grosis ämel o.»

«Es isch die Ziit, wo d'Liit afend üsziehn fer d'Härdöpfu underztüen Faschd alli hend \*\*\* \*\*\* en Häröpfublätz, z'Grosis emu o.»

Die Zweite beinhaltet eine Auslassung (\*\*\*), da der Benutzer den Ausdruck «da un dert» nicht verstanden hat.

So sieht eine unbrauchbare Transkription aus: «dfbdfgh \*\*\* \*\*\* \*\*\* \*\*\* \*\*\*»

# Alignierung von schweizerdeutschen Transkriptionen

## Satzbewertung

Für jeden Satz aus einer Satzgruppe wollen wir wissen, wie gut er ist. «Gut» in diesem Zusammenhang bedeutet, wie akkurat die Transkription ist. Eine Bewertung pro Satz hilft uns in folgenden Szenarien:

* Ausfiltern von unbrauchbaren Sätzen:

Unbrauchbare Sätze verschlechtern das Alignierungsergebnis. Deshalb wollen wir sie so früh wie möglich erkennen und entfernen. Dieses Verfahren ist im Kapitel Ausfilterung beschrieben.

* Satzpaarbildung bei der Wortalignierung:

Die Information, welcher Satz potentiell der beste ist, hilft im eigentlichen Prozess der Wortalignierung. Dies ist im Kapitel Wortalignierung beschrieben.

Für die Task-ID 1851 (Audio: <https://dindialaekt.ch/data/transcribe/SDS_CD1_1_3_speaker1_4.mp3>) gibt es beispielsweise folgende vier Transkriptionen:

1. élk ék ékl ékl asda
2. Nacher heisis uusenageretschöplet ond \*\*\* zämebonge weder us weder of d'Reite uecheto od'Frocht heisi am Bode usegrächet ond nachere na Huufe geschtossed.
3. nächäär heisis userenangere gschüttlet und Burdine zämme bunge wider us wider ufd Reiti ufetoo ud Frucht heisi am Bode usegrächet und nächäär ane Huufe gschtoosse
4. Nachhäär héysis usenangèrègschüttlet und Puurdine zämepunge wider us wider uft Reyti uechetoo u Pfrucht heysi am Bode usegrächet unt nachhäär ane Huuffe gschtoosse

Sofort ersichtlich ist, dass Transkription 1 nicht gut und in diesem Fall sogar völlig unbrauchbar ist.

Die andern drei Transkriptionen sind zwar brauchbar, weisen jedoch dennoch qualitative Unterschiede auf:

Transkription 2 ist unvollständig: \*\*\* ersetzt hier «Burdine» oder «Puurdine».

Transkription 4 verwendet die Buchstaben è und é, was für schweizerdeutsch eher ungewöhnlich ist.

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

In unserem Beispiel ist Transkription 1 zu allen andern unähnlich. Transkription 2 ist zu Transkription 3 und 4 etwa gleich ähnlich, tendenziell etwas ähnlicher zu Transkription 3, da diese keine è und é verwendet. Transkription 3 und 4 sind sich sehr ähnlich.

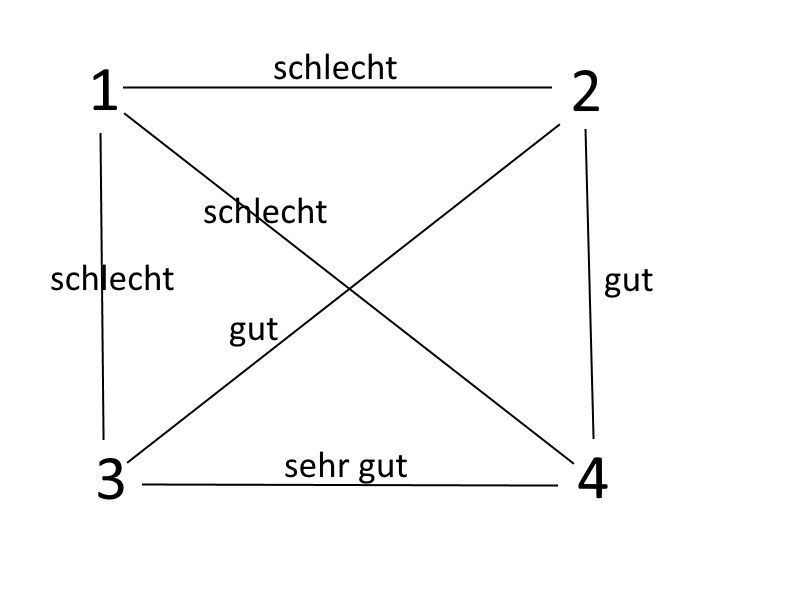
Wenn wir in der Lage sind, die Ähnlichkeit zwischen den Sätzen zu berechnen, können wir also relativ gut entscheiden, wie «gut» eine Transkription ist. 

Abbildung 2: Relationen der Sätze bezüglich Ähnlichkeit

Pro Satz also:

Satz 1: schlecht, schlecht, schlecht.

Satz 2: schlecht, gut, gut.

Satz 3: schlecht, gut, sehr gut.

Satz 4: schlecht, gut, sehr gut.

Folglich brauchen wir ein Verfahren, die Ähnlichkeit zwischen Texten zu bestimmen:

So könnte beispielsweise die Levenshtein-Distanz (Levenshtein, 1966) verwendet werden: Es wird die Levenshtein-Distanz zwischen jedem Satzpaar berechnet und dabei pro Satz die Distanz zu jedem anderen Satz aufsummiert.

Im Bereich der automatischen Bewertung von Maschinenübersetzungen wird eine ähnliche Idee genutzt: «the closer a machine translation is to a professional human translation, the better it is» (Papineni, Roukos, Ward, & Zhu, 2002), je näher eine Maschinenübersetzung an einer professionellen menschlichen Übersetzung ist, desto besser ist sie.

Papineni et al. stellen den Algorithmus «BLEU» vor, der eine Maschinenübersetzung mit einem Set von Referenzübersetzungen vergleicht. Je ähnlicher die Maschinenübersetzung zu den Referenzübersetzungen ist, desto besser wird er bewertet, wobei 0 die schlechteste und 1 die beste Bewertung ist. BLEU hat eine hohe Korrelation zu menschlichen Bewertungen und ist eine beliebte und einfach zu berechnende Metrik.

Wir können dieses Prinzip auf unser Problem anwenden, indem wir jeweils einen der Sätze als zu bewertende Maschinenübersetzung ansehen, und die jeweils anderen Sätze als die Referenzübersetzungen.

Mit diesem Verfahren erhalten wir für unsere vier Sätze folgende Bewertungen, wobei 1 sehr gut und 0 sehr schlecht ist:

Satz 1: 0.00033940394936699897

Satz 2: 0.5970591840896966

Satz 3: 0.8215011858242206

Satz 4: 0.7990890169242502

Satz 3 schneidet am besten ab, dicht gefolgt von Satz 4. Zu Satz 2 besteht ein gewisser Abstand und Satz 1 hat die mit Abstand am schlechteste Bewertung. Dies korreliert schon ziemlich gut mit unserer intuitiven Bewertung.

Für unsere Zwecke eignet sich BLEU also sehr gut. Es gibt einige weitere Metriken, die wir nicht in Betracht gezogen haben:

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

METEOR (Banerjee & Lavie, 2005) und NIST (Doddington, 2002) bauen auf BLEU auf und versprechen einige Kleinigkeiten zu verbessern.

LEPOR (Han, Wong, & Chao, 2012) und die verbesserten Iterationen hLEPOR und nLEPOR, die versprechen einige der Probleme mit BLEU und den darauf aufbauenden Metriken zu lösen.

Sowie eine Vielzahl weiterer Metriken und Verfahren. Da BLEU der de facto Standard ist und wir mit BLEU ausreichend gute Ergebnisse erhalten und nicht erwarten, mit einem anderen Verfahren fundamental bessere Ergebnisse zu erhalten, werden wir diese in dieser Arbeit ignorieren.

All diese Metriken wurden für die Dokumentebene konzipiert. Auf der Satzebene, wie wir sie hier verwenden, liefern sie etwas schlechtere Ergebnisse. Für BLEU wurden deshalb einige sogenannte Smoothing Funktions entwickelt. Smoothing Function 7 liefert die besten Ergebnisse auf der Satzebene. (Chen & Cherry, 2014) wir werden diese deshalb hier verwenden.

|  |  |  |
| --- | --- | --- |
| Satz | Default BLEU | BLEU mit Smoothing Function 7 |
| 1 | 0.00033940394936699897 | 0.0002645904060323999 |
| 2 | 0.5970591840896966 | 0.6781557399224509 |
| 3 | 0.8215011858242206 | 0.9185006575847926 |
| 4 | 0.7990890169242502 | 0.8947321304759088 |

Mit der Smoothing Function ist also der Unterschied zwischen guten und schlechten Sätzen noch etwas grösser geworden. Erwähnenswert ist, dass mit Smoothing Function die Bewertung eines Satzes besser als 1 sein kann.

## Ausfilterung

Als Bereinigungsprozess wollen wir alle Sätze ausfiltern, die für uns keinen sinnvollen Informationsgehalt haben und potentiell das Ergebnis verschlechtern. Da wir bereits ein Verfahren zur qualitativen Satzbewertung haben, können wir diese Ausfilterung auf den Satzbewertungen basieren.

Im obigen Beispiel zur Task-ID 1851 ist Satz 1 mit der Bewertung von 0.0003 der einzige, den wir ausfiltern wollen. Satz 2 mit einer Bewertung von 0.678 ist für uns wertvoll und soll nicht ausgefiltert werden.

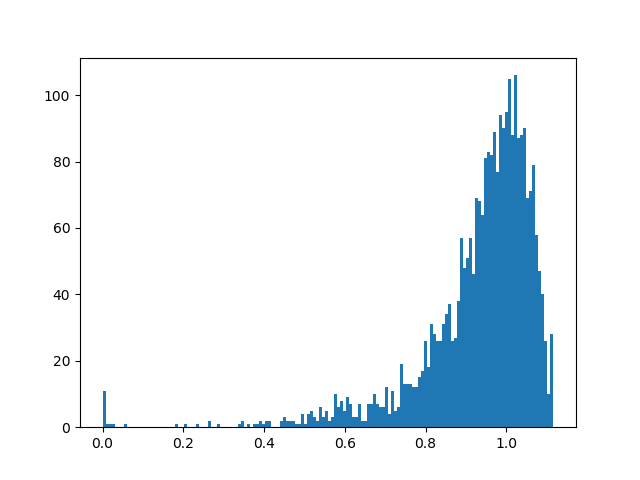


Abbildung 3: Histogramm der Bewertungen aller Sätze

Aus Abbildung 2 wird die Verteilung der Bewertungen ersichtlich. Die meisten Bewertungen liegen zwischen 0.8 und 1.1. Der Durchschnitt liegt bei 0.9332, der Median bei 0.9698. Nur 51 Sätze haben eine Bewertung von unter 0.5 und 15 davon liegen sogar unter 0.1.

|  |  |  |
| --- | --- | --- |
| Bewertung | Satz | Referenzsatz (mit besserer Bewertung) |
| 0.00026 | élk ék ékl ékl asda | nächäär heisis userenangere gschüttlet und Burdine zämme bunge wider us wider ufd Reiti ufetoo ud Frucht heisi am Bode usegrächet und nächäär ane Huufe gschtoosse |
| 0.00896 | thurgauer thurgauer thurgau | dem isch scho so, das mis s Thema sehr interessiert und au imer interessiert hett. aso, ich machs nüd wegem Dschob noch i wäss no gär nid won ig will here nochane. äm- also mi interessierts würkli und es isch kuul. |
| 0.0551 | Altwiibervolk \*\*\* \*\*\* Das fuärät gad achli \*\*\* \*\*\* | Au z Wybervolch cho do nid e so lang schloofe die müend de des ds Belätebrösöm zum z Früestück rüschte und das fueret denn e chli und de mögeds de die eiwer Lüüt ender öppe bis zum Zmittag verliid |
| 0.2042 | Kleider hät er eifsch nebed s'bett gleit \*\*\* \*\*\* \*\*\* | Klaider hät er eifach näb z Bett gworffe so müad ischer gsii und sis Bett isch es bizli ghurz und sini Füass luegend une use |
| 0.3914 | nacher wo ich bueb gsi bin sägese gmäht worde \*\*\* so het mer e sbögli dra gmacht, an d sägese und hets schön aaböglet \*\*\* \*\*\* | daher won ich Bueb gsi bi isch so mit de Sägesse gmeiht worde, neh d Fruchte no schön gschtange isch so hett mer s Beegli dra gmacht, a d Sägesse und hett schön abeeglet und sind d'Schnitter no cho und hei das eweg gno und scheen of d Zettle gleit |
| 0.5744 | Last tragen Jeder trägt eine Last Jeder trägt einen Teil Wäre es etwas zum Verkaufen, so würde es jeder anbieten | Büürdeli träge ei jede träit es Bürdeli ei jede träit e Teil wärs eppis zum Verchräämere, es häts e jede feil. |
| 0.6588 | Ein jeder wollte mehr Land besitzen, .. und sie wurden sich nicht einig. Zuletzt machten sie ab, .... sie wollten an einem bestimmten Tag am frühen Morgen... so wie der Hahn gekräht habe, von jedem Hauptort den Landläufer (Landboten) aussenden.. Und dort, wo die beiden aufeinander treffen, .. sollte für alle Zeit die Grenze sein. | e jede Teil het welle mi Welt ha und si sind eifach nid z'einte worde do heit's jetzt letscht abgmached si welled amene bschtimmte Tag am morged früe so wi dr Guli kräht heig vo jedem Hauptort der Landläufer usschicke und wo dä beed zämechömet söll für all ebig Ziite d'Gränze si |
| 0.7991 | de Bäärgamme mües dä alläg im Herbscht emene \*\*\* übergä. dä miends die Gäischtleche Bihäärdig und de Tokter is Aesse iilade und dem Eerebrediger de Chääs psoorge | de Bärgamme müesti allig im Herbscht emene Wiil d'Chilbi vergä, de miends di Geischtliche, di \*\*\* und dr Dokter is Ässe iilade und dem Ehreprediger dr Chääs bsorge |

Aus dieser Tabelle ist ersichtlich, wie schlechte Sätze bewertet werden. Zwei der hier präsentierte Sätze wurden dabei fälschlicherweise auf hochdeutsch transkribiert, was für uns nicht wünschenswert ist.

Der simpelste Weg, Sätze auszufiltern, ist einen Minimalwert einzuführen. Je tiefer dieser ist, desto mehr unbrauchbare Sätze werden nicht ausgefiltert. Je höher der Minimalwert, desto mehr brauchbare Sätze gehen verloren. Unter 0.3 sind kaum noch brauchbare Sätze zu finden, über 0.7 sind praktisch alle Sätze brauchbar. Als Default schlagen wir einen Wert von 0.5 vor.

Etwas schwieriger wird es in Satzgruppen, in denen es nur 2 Transkriptionen gibt.

|  |  |  |
| --- | --- | --- |
| Bewertung | Satz A | Satz B |
| 0.35 | Also, Sie fragen wann ich geboren bin ? Ich bin am 26. Januar 1912 geboren. | jaa \*\*\*\*\*\*\*wän das ich geboore bi? eem ...ich bin am sächsezwänzigschte Jäner nünzehundertzwölf geboore |
| 0.53 | aber es Bögli für e Chäs üse z'näh, do het Hans Fähri gseit, das sig z'churzes do manglet mer denn o no z'gseh für nes angers - jo das wei mer denn luege, ja | oderes Bögli für è Chääs üüseznäh und det hät hans fäärè gseit das sig z churzes damol luegtemer de scho z gsee fürnes anders |
| 0.65 | jetze chumensch güetlech witer und lüegen uber de abrosch uf ds Dach u gseh - da isch ds zennusch hüt üsgspreiteti gsi | Jetzt bechums güet för hütte, und lüegend öber de äberarsch ufs Dach und gseh, da isch s'Tenusch \*\*\* üsbreiteti gsi |
| 0.80 | Z Hasli hett der Rott-Bueb gläbt Sy Übernae isch offebar derthär cho wiu är als Bueb,wie's früecher Bruuch gsi isch lang e Rock agha het Als Schwinger isch är gförderet worde Und är heig Ross-Iise vo blosser Hand abenand grisse | zHasli hät de Rockbueb gläbt si Uebername isch offebar deet häärchoo wüu är as Bueb, wis früner Bruuch gsi isch lang ne Rock aagha héd Us Schwinger ischèr gförchtet woorde und er häig Rossiise vo bloosser hand abenanggrisse |

Satzgruppen mit nur 2 Transkriptionen scheinen tendenziell schlechter bewertet zu werden. Diese Sätze haben eine durchschnittliche Bewertung von lediglich 0.676.

Auch hier gibt es weitere Ansätze, Outlier auszufiltern, die wir nicht weiterverfolgen konnten:

In der Statistik werden Outlier mittels Durchschnitt und Standardabweichung erkannt. Dieser Ansatz ist für unsere Daten jedoch weniger geeignet, da statistische Methoden oft von normalverteilten Daten und grösseren Datenmengen ausgehen.

Pro Gruppe könnten die Sätze in unterschiedliche Qualitätskategorien klassifiziert werden. Danach könnten die schlechtesten Kategorien verworfen werden. Dies liesse sich als eine Art eindimensionaler Clustering Algorithmus implementieren (Jenks, 1967).

Man könnte versuchen, die durchschnittliche Bewertung innerhalb einer Satzgruppe zu verbessern, indem iterativ der Satz mit der schlechtesten Bewertung ausgeschlossen und der Durchschnitt neu berechnet wird. Sobald sich der Durchschnitt nicht mehr gross ändert, wurden alle schlechten Sätze ausgefiltert. So verlagert sich das Problem zu der Frage, was eine grosse Veränderung des Durchschnitts ist.

Für die Zwecke dieser Arbeit sind wir nicht davon überzeugt, dass einer dieser alternativen Ansätze uns merklich bessere Resultate liefern wird. Weiter macht es unser Ansatz eines einfachen Minimalwerts für einen Benutzer sehr einfach, diesen mit vorhersagbaren Resultaten anzupassen. So können wir dem Benutzer die Möglichkeit geben, den Tradeoff zwischen fälschlicherweise gefilterten und fälschlicherweise ungefilterten Sätzen selbst zu bestimmen.

## Wortalignierung

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

Wortalignierer alignieren Wörter aus unterschiedlichen Sprachen, in denen die Wortstellung völlig unterschiedlich sein können. Satzalignierer gehen generell von der gleichen Satzreihenfolge in beiden Texten aus. In unseren Texten wisse wir, dass die Wortreihenfolge immer die selbe sein wird. Es können jedoch einzelne Wörter fehlen oder Wortpaare zusammengeschrieben werden. Damit sind normale Wortalignierer zu flexibel, wohingegen wir alle Anforderungen für ein Satzalignierungsproblem erfüllt haben.

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. In <http://lium3.univ-lemans.fr/mtmarathon2010/ProjectFinalPresentation/SentenceAlignment/sentence_alignment.pdf> werden einige Satzaligner vorgestellt und verglichen. Die Resultate hierbei sind generell sehr ähnlich, weshalb wir keine fundamental unterschiedlichen Ergebnisse erwarten. Wir haben zwei der vorgestellten Aligner (Bleualign und Hunalign) ausprobiert.

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.

Fastalign ([Chris Dyer](http://www.cs.cmu.edu/%7Ecdyer), [Victor Chahuneau](http://victor.chahuneau.fr), and [Noah A. Smith](http://www.cs.cmu.edu/%7Enasmith). (2013), <http://www.aclweb.org/anthology/N13-1073>) ist ein Wortalignierer. Er bietet jedoch eine Option an, die diagonale Alignierungen bevorzugt. Da sich jedoch das Input- und Outputformat stark von demjenigen von Bleualign und Hunalign unterscheidet und wir davon ausgehen müssen, dass das Resultat nicht fundamental besser ist, haben wir hier Fastalign nicht in Betracht gezogen.

Vorgehen:

Aus einer Satzgruppe wählen wir Paare von Sätzen und lassen das Alignierungstool entscheiden, welche Ausdrücke zusammengehören. So erhalten wir Paare von Ausdrücken, die gleichbedeutend sein sollten. Nun wollen wir diese Ausdruckspaare zu ganzen Gruppen gleichbedeutender Ausdrücke vereinen. Wir sehen also jeden Ausdruck als einen Knoten in einem Graph an. Das Alignierungstool sagt uns, zwischen welchen Knoten eine Kante existiert. Mit jedem Alignment eines Satzpaars werden dem Graph also 0-n neue Knoten und n neue Kanten hinzugefügt. Am Ende dieses Prozesses ist jeder verbundene Teilgraph eine Gruppe gleichbedeutender Ausdrücke.

## Filterung des Alignments

Das Ergebnis dieses Prozesses ist eine Liste von Gruppen von Ausdrücken. Die Ausdrücke einer Gruppe sollten hierbei gleichbedeutend sein. Da wir wissen, dass die Ausdrücke das Resultat von Transkriptionen derselben Audiodatei sind, sollten die Ausdrücke bis auf die Schreibweise identisch sein. Fehlerhafte Transkriptionen, schlechte Alignment-Ergebnisse und fehlerhaftes Zusammenbauen der Ausdrucksgruppen verschlechtern jedoch das Ergebnis. Deshalb müssen wir einen Weg finden, das generierte Alignment qualitativ zu bewerten, um dann ungenügende Ergebnisse auszufiltern.

Algorithmen

**Bleu-Score:**

BLEU ist für die Bewertung längerer Texte gedacht und liefert beim Vergleich einzelner Wörtern kein sinnvolles Ergebnis.

**Levenshtein-Distanz**

Auch Editierdistanz genannt, berechnet die Anzahl an Operationen (Einfügen, Löschen, Ersetzen von Buchstaben) die es braucht, um von Wort A auf Wort B zu kommen.

Beispiel: Geld und Gäld haben eine Distanz von 1.

**Gewichtete Levenshtein-Distanz**

Besser geeignete Version von der Levenshtein Distanz für das Vergleichen von Phonetisch gleich Klingenden Worten. So kann man hier mit der Gewichtung der Operationen auf den Buchstaben eine fairere Editierdistanz berechnen. Beispiel: Gäld und Geld klingen im Schweizerdeutschen sehr ähnlich, würden aber mit der normalen Version eine Distanz von 1 aufweisen. Jedoch hier bei der Gewichtung wird die Substitution von gewissen Zeichenpaaren billiger gemacht, wenn sie akustisch ähnlich sind.

<https://github.com/nadvornix/python-fizzle>

<https://pypi.python.org/pypi/weighted-levenshtein/0.1>

**Damerau-Levenshtein-Distanz**

Die Operationen der Levenshtein Distanz wird um Transponieren erweitert, diese ermöglicht das Vertauschen von zwei nebeneinanderstehenden Buchstaben.

<https://github.com/gfairchild/pyxDamerauLevenshtein>

**Soundex**

Errechnet die Lautähnlichkeit von zwei Strings. Gedacht um englische Namen zu vergleichen. Beispiel: Smith und Smiff.  
Schlechte Unterstützung für Deutsch, keine für CH-Deutsch.

<https://pypi.python.org/pypi/soundex>

**Metaphone**

Generiert Keys basierend auf Lautähnlichkeit eines Wortes. Im Englischen kann man so zwei Wörter vergleichen. Beispiel: Programming und Programmer resultieren beide in PRKRM.

**Double Metaphone**

Generiert Keys auf Lautähnlichkeit eines Wortes wie Metaphone, unterstützt jedoch auch Slawisch, Germanisch, Celtic, Griechisch, Französisch, Italienisch, Spanisch, Chinese und andere Herkünfte.

<https://github.com/dracos/double-metaphone>

**Entscheid über den Algorithmus**

Um mal eine Baseline zu schaffen, wählen wir die Gewichtete Levenshtein-Distanz mit kleinen Anpassungen wie zum Beispiel, dass ä und e tiefer bewertet werden. Eine ausführlichere Gewichtung wäre bestimmt noch besser und genauer, jedoch schätzen wir den Aufwand als zu gross ein. Als Unterstützung verwenden wir aber noch Double-Metaphone. Durch die Kombination von zwei verschiedenen Algorithmen kommen wir auf bessere Ergebnisse.   
Unsere Algorithmen filtern wie folgt die Alignments aus:

|  |  |  |
| --- | --- | --- |
| Metaphone Bewertung | Levensthein Bewertung | Resultat |
| OK | OK | NICHT AUSGEFILTERT |
| OK | FAILED | NICHT AUSGEFILTERT |
| FAILED | OK | NICHT AUSGEFILTERT |
| FAILED | FAILED | AUSGEFILTERT |

**Hier ein paar Beispiele:**

**Params: Levensthein Filter (0.4)**

|  |  |
| --- | --- |
| Satzpaar mit Levenstheinvalue und Metaphone Bewertung | Wird ausgefiltert |
| ('sägäslì', 'sägessli', 0.275, False) | Nein, da Levensthein OK |
| ('de', 'die', 0.3333333333333333, True) | Nein, da beide OK |
| ('wèrdët', 'wäärded', 0.42857142857142855, True) | Nein, da Metaphone OK |
| ('ëm', 'am', 0.1, True) | Nein, da beide OK |
| ('?ù', 'scho', 1.0, False) | Ja, da beide FAILED |
| ('ùn', 'und', 0.6666666666666666, False) | Ja, da beide FAILED |

Zeitpunkt der Bewertung des Alignments

**Pro Satzpaar**

Zwei Sätze werden aligniert und dann werden die daraus resultierenden Wortpaare bewertet. Schlechte Paare werden in diesem Schritt direkt verworfen, siehe Algorithmen oben, die guten in einem Graphen abgelegt. Die Wörter sind die Nodes.

**Pro Satzgruppe**

Alle Sätze einer Satzgruppe werden miteinander aligniert, ein Graphen mit alignierten Wörter wird kreiert. Alle Wörter werden vor dem Einfügen in den Graphen bewertet und ausgefiltert.

**Global**

Alle Sätze werden miteinander aligniert, ein Graphen mit alignierten Wörter wird kreiert. Alle Wörter werden vor dem Einfügen in den Graphen bewertet und ausgefiltert.

**Tester**

**TODO**

**Entscheid über Parameter des Filters**

|  |  |
| --- | --- |
| **Filtervalue für Levenshtein** | **Graphauschnitt** |
| **1** |  |
| **0.8** |  |
| **0.6** |  |
| **0.4** |  |
| **0.3** |  |
| **0.2** |  |
| **0.0** |  |

Durch die Beispiele kann man gut sehen, dass es ein Abwägen zwischen Quantität und Qualität ist. Bei einer Levenshtein Filtervalue von 0.6, hat man in «druf los» und «druflosmäje» in einer Gruppe. Wir stellten uns dann die Frage ob dies für unsere Aufgabe Sinn macht und kamen zum Schluss, dass uns Qualität viel wichtiger ist wie Quantität. Da man in einer Alignierungsgruppe nur das selbe Wort, in anderen Schreibweisen finden möchte und nicht etwas mit einer anderen Bedeutung («druf los» und «druflosmäje» ist nicht das Gleiche). Deshalb haben wir uns dazu entschieden, die Filtervalue für Levenshtein auf streng zu setzen, sprich um die 0.25 herum.

Filtervalue von 1 nimmt alle Wörter und eine Filtervalue von 0 heisst, dass die komplette Entscheidung bei Double-Metaphone liegt.

## Fehlerkorrektur / Interpolation

bad sentence:

Es isch so das mich sThema interessiert hät aso ich mags nöd weg em Job noch, \*\*\* aso mi intressiert würklich und es isch cool\*\*\*

improved sentence:

Es isch so das mich sThema interessiert hät aso ich mags nöd weg em Job noch, äm- aso mi intressiert würklich und es isch cool\*\*\*

improved sentence(with additional word filter):

Es isch so das mich sThema interessiert hät aso ich mags nöd weg em Job noch, äm- aso mi intressiert würklich und es isch kuul.

## Bewertung des Ergebnisses

Um verschiedene Parameter und Tools miteinander vergleichen zu können, muss es einen Weg geben, das generierte Alignment qualitativ zu bewerten.

Wir haben einen Goldstandard für die Alignierung von Wörtern aus Task 1 erstellt, mit diesem werden wir die verschiedenen Methoden zur Alignierung messen und so vergleichen können.

.

Wir haben zum Testen folgende GOLD\_STANDARD\_SET = [2048, 2095, 2374, 1930, 1929] Satzgruppen genommen.

# Ergebnis

# Schlussfolgerung

[Abbildung 1: Alle Transformationen 5](#_Toc503554996)

[Abbildung 2: Relationen der Sätze bezüglich Ähnlichkeit 8](#_Toc503554997)

[Abbildung 3: Histogramm der Bewertungen aller Sätze 10](#_Toc503554998)

# Literaturverzeichnis

Banerjee, S., & Lavie, A. (29. 6 2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, S. 65-72.

Chen, B., & Cherry, C. (6 2014). A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, S. 362-367.

Doddington, G. (24. 4 2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceeding HLT '02 Proceedings of the second international conference on Human Language Technology Research*, S. 138-145.

Han, A.-F., Wong, D., & Chao, L. (12 2012). LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. S. 441-450.

Jenks, G. F. (1967). The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography 7*, S. 186–190.

Levenshtein. (11 1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, S. 707-710.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-j. (12. 7 2002). BLEU: a Method for Automatic Evaluation of Machine Translations. *Proceeding ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, S. 311-318.