h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Construction of the corpus

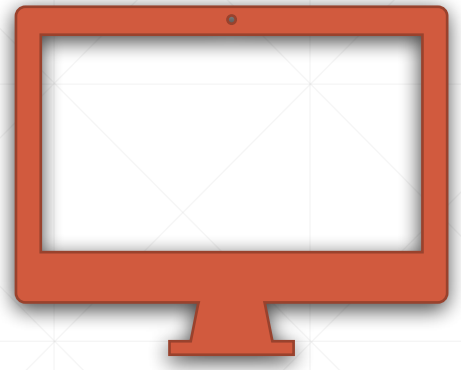
Simple German – German

Mina Schütz – 09.05.2018

Agenda

1. Creating a corpus
2. German > Simple German
3. Monolingual corpus
4. Sentence alignment
5. Simplifying text automatically
6. Steps of creating the corpus
7. Results and further research

1 Creating a corpus



- Plain text
- Needs an ID
- Date of texts
- Size of texts
- Bilingual vs. Monolingual
- Copyright
- Domain
- Consistency for large corpus
- Hypertext

2 German > Simple German

- Text-to-text simplification
- Methods
  - Deletion
  - Rephrasing
  - Reordering
  - Sentence splitting
  - Insertion
  - …

- ➢ Not enough data and research papers for German
- ➢ Exists for English, Spanish, French, Swedish, Brazilian Portuguese
- ➢ **Goal**: statistical machine translation system
- ➢ **Used**: monolingual sentence alignment algorithm
- ➢ Monolingual corpus

3 Monolingual corpus

Parallel

- Mutual translations
- Needs to be aligned L1 -> L2… LN
- Used for automatic translation
- Full text / text samples
- Static / dynamic

Comparable

- Same topic > different resources
- Independently from each other
- 2 or more languages

4 Steps for creating the corpus

1. Crawling through websites
2. Searching for German articles
3. Following their links to Simple German articles
4. Extracting both articles
5. Cleaning data from HTML tags
6. Tokenizing the articles
   - 7.000 sentences
   - 70.000 tokens

4 Steps for creating the corpus

7. Splitting corpus
   - 70 % Training set
   - 10 % Development set
   - 20 % Test set

8. Sentence alignment

9. Clustering the paragraphs

10. Learning mapping rules

5 Sentence alignment

$$\frac{2x\ \text{matches}\ (T1, T2)}{|\,T1\,| + |\,T2\,|}$$

- Matches = Number of alignments between T1 and T2
- Result = 1,0 > best parallel corpus

5 Sentence alignment

German:

In den Osnabrücker Werkstätten (OW) und OSNA-Techniken sind rund 2.000 Menschen mit einer Behinderung beschäftigt.

(“In the Osnabrück factories and OSNA-Techniken, about 2.000 people with disability are employed.”)

Simple German:

In den Osnabrücker Werkstätten und den Osna-Techniken arbeiten zweitausend Menschen mit Behinderung.

(“Two thousand people with disability work in the Osnabrück factories and Osna-Techniken.”)

5 Sentence alignment

German:

Der Beauftragte informiert über die Gesetzeslage, regt Rechtsänderungen an, gibt Praxistipps und zeigt Möglichkeiten der Eingliederung behinderter Menschen in Gesellschaft und Beruf auf.

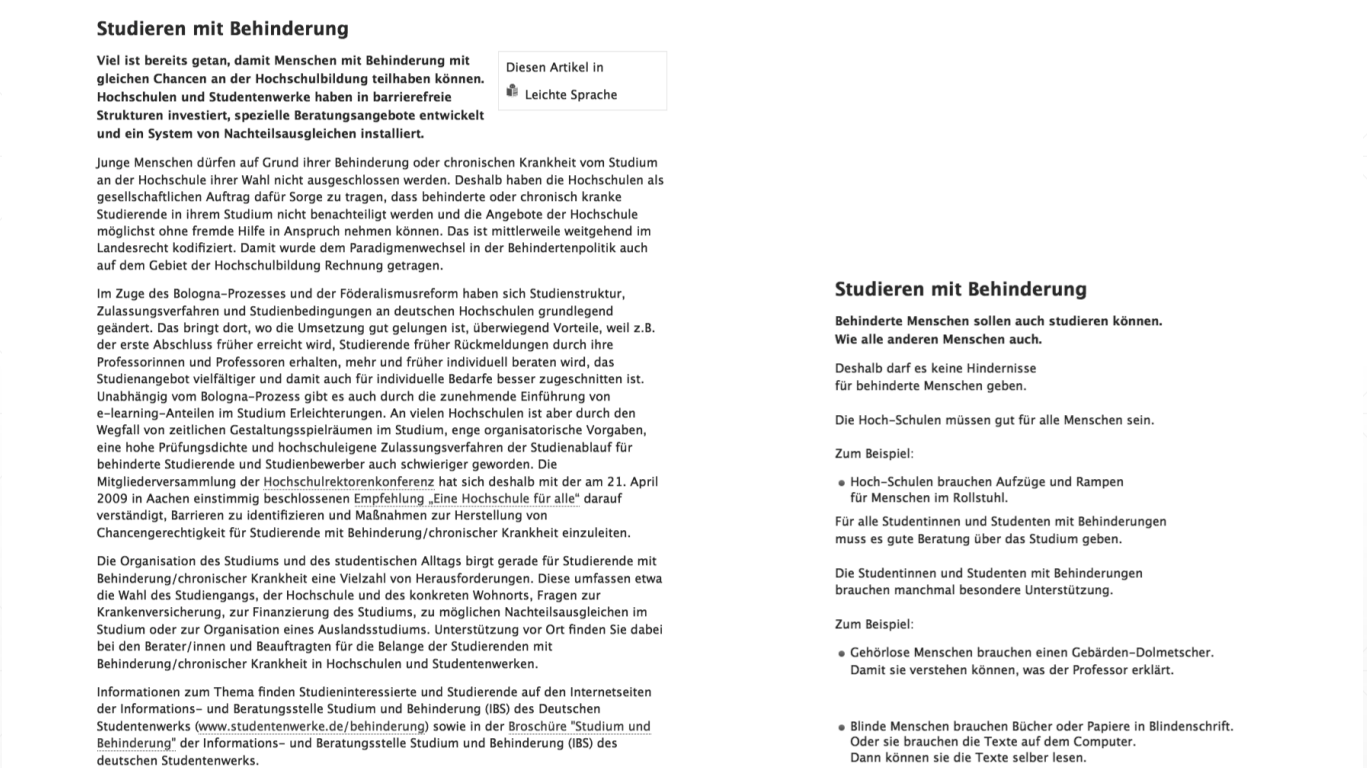
(“The delegate informs about the legal situation, encourages revisions of laws, gives practical advice and points out possibilities of including people with disabilities in society and at work.”)

Simple German:

Er gibt ihnen Tipps und Infos.

(“He provides them with advice and information.”)

5 Sentence alignment



Studieren mit Behinderung

Viel ist bereits getan, damit Menschen mit Behinderung mit gleichen Chancen an der Hochschulbildung teilhaben können. Hochschulen und Studentenwerke haben in barrierefreie Strukturen investiert, spezielle Beratungsangebote entwickelt und ein System von Nachteilsausgleichen installiert.

Diesen Artikel in
Leichte Sprache

Junge Menschen dürfen auf Grund ihrer Behinderung oder chronischen Krankheit vom Studium an der Hochschule ihrer Wahl nicht ausgeschlossen werden. Deshalb haben die Hochschulen als gesellschaftlichen Auftrag dafür Sorge zu tragen, dass behinderte oder chronisch kranke Studierende in ihrem Studium nicht benachteiligt werden und die Angebote der Hochschule möglichst ohne fremde Hilfe in Anspruch nehmen können. Das ist mittlerweile weitgehend im Landesrecht kodifiziert. Damit wurde dem Paradigmenwechsel in der Behindertenpolitik auch auf dem Gebiet der Hochschulbildung Rechnung getragen.

Im Zuge des Bologna-Prozesses und der Föderalismusreform haben sich Studienstruktur, Zulassungsverfahren und Studienbedingungen an deutschen Hochschulen grundlegend geändert. Das bringt dort, wo die Umsetzung gut gelungen ist, überwiegend Vorteile, weil z.B. der erste Abschluss früher erreicht wird, Studierende früher Rückmeldungen durch ihre Professorinnen und Professoren erhalten, mehr und früher individuell beraten wird, das Studienangebot vielfältiger und damit auch für individuelle Bedarfe besser zugeschnitten ist. Unabhängig vom Bologna-Prozess gibt es auch durch die zunehmende Einführung von e-learning-Anteilen im Studium Erleichterungen. An vielen Hochschulen ist aber durch den Wegfall von zeitlichen Gestaltungsspielräumen im Studium, enge organisatorische Vorgaben, eine hohe Prüfungsdichte und hochschuleigene Zulassungsverfahren der Studienablauf für behinderte Studierende und Studienbewerber auch schwieriger geworden. Die Mitgliederversammlung der Hochschulrektorenkonferenz hat sich deshalb mit der am 21. April 2009 in Aachen einstimmig beschlossenen Empfehlung „Eine Hochschule für alle" darauf verständigt, Barrieren zu identifizieren und Maßnahmen zur Herstellung von Chancengerechtigkeit für Studierende mit Behinderung/chronischer Krankheit einzuleiten.

Die Organisation des Studiums und des studentischen Alltags birgt gerade für Studierende mit Behinderung/chronischer Krankheit eine Vielzahl von Herausforderungen. Diese umfassen etwa die Wahl des Studiengangs, der Hochschule und des konkreten Wohnorts, Fragen zur Krankenversicherung, zur Finanzierung des Studiums, zu möglichen Nachteilsausgleichen im Studium oder zur Organisation eines Auslandsstudiums. Unterstützung vor Ort finden Sie dabei bei den Berater/innen und Beauftragten für die Belange der Studierenden mit Behinderung/chronischer Krankheit in Hochschulen und Studentenwerken.

Informationen zum Thema finden Studieninteressierte und Studierende auf den Internetseiten der Informations- und Beratungsstelle Studium und Behinderung (IBS) des Deutschen Studentenwerks (www.studentenwerke.de/behinderung) sowie in der Broschüre "Studium und Behinderung" der Informations- und Beratungsstelle Studium und Behinderung (IBS) des deutschen Studentenwerks.

Studieren mit Behinderung

**Behinderte Menschen sollen auch studieren können.
Wie alle anderen Menschen auch.**

Deshalb darf es keine Hindernisse
für behinderte Menschen geben.

Die Hoch-Schulen müssen gut für alle Menschen sein.

Zum Beispiel:

- Hoch-Schulen brauchen Aufzüge und Rampen
  für Menschen im Rollstuhl.

Für alle Studentinnen und Studenten mit Behinderungen
muss es gute Beratung über das Studium geben.

Die Studentinnen und Studenten mit Behinderungen
brauchen manchmal besondere Unterstützung.

Zum Beispiel:

- Gehörlose Menschen brauchen einen Gebärden-Dolmetscher.
  Damit sie verstehen können, was der Professor erklärt.

- Blinde Menschen brauchen Bücher oder Papiere in Blindenschrift.
  Oder sie brauchen die Texte auf dem Computer.
  Dann können sie die Texte selber lesen.

Figure 1: Comparison of AS and LS article from `http://www.einfach-teilhaben.de`

5 Simplifying text automatically

- **RULE BASED**
  - Replacing words
  - Rephrasing
  - Embedded sentences
  - Passive constructions
- **CORUPS BASED**
  - Machine translation
  - Substitution
  - Reordering
  - Splitting
  - Deletion
- **HYBRID APPROACH**
  - Classifier decides
  - Machine translation
  - Rule based translation

- Remove stop words
- Lowercase every word
- Replace dates
- Replace numbers
- Replace names

Step 1: Pre-Processing

Step 2: Training phase

- Clustering paragraphs of German texts
- Calculating the mappings between the sets
- Cosine similarity > 0
- In each set a sentence has to be aligned to a sentence in the other set
- 200 iterations

Step 3: Testing phase

- Assignment of each paragraph to the cluster it was closed to

- Combination of every German paragraph with all Simple German paragraphs

Results

- Algorithm was not good for German
- German nouns
- Cosine similarity (decomposition / lemmatisation)
- Different web sources for data
- Broad domains

Method	Precision	Recall	F1
Adapted algorithm of Barzilay and Elhadad (2003)	27.7%	5.0%	8.5%
Baseline I: First sentence	88.1%	4.8%	9.3%
Baseline II: Word in common	2.2%	8.2%	3.5%

Table 2: Alignment results on test set

Sources

- http://www.glottopedia.org/index.php/Parallel_corpus (02.05.2018)
- https://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Simple_German_4 (05.05.2018)
- Klaper, D., Ebling, S., & Volk, M. (2013). Building a german/simple german parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 11-19).
- http://www.alpha-archiv.de/fileadmin/PDFs/Qualifizierungsarbeiten/Masterarbeit_Kuhlmann_Copy.pdf (23.04.2018)
- https://ota.ox.ac.uk/documents/creating/dlc/appendix.htm (23.04.2018)
- https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm (24.04.2018)
- Jehat, D. J, Germann, E., Lintner A., & Soland C. (2017). Wahlprogramme in Leichter Sprache – Eine korpuslinguistische Annäherung. In *„Leichte Sprache“ im Spiegel theoretischer und angewandter Forschung* (pp. 229-246).
- Caseli, H. M., Peirera, T. F., Specia, L., Pardo, T.A.S., Gasperin, C. & Aluisio, S.M. Building a Brazilian Portuguese parallel corpus of original and simplified texts. From http://conteudo.icmc.usp.br/pessoas/taspardo/CICLing09-CaseliEtAl.pdf (20.04.2018)
- https://en.oxforddictionaries.com/definition/corpus (17.04.2018)