A New Aligned Simple German Corpus

Vanessa Toborek[1], Moritz Busch[1], Malte Boßert[1], Christian Bauckhage[1,2], Pascal Welke[1,3]
[1]University of Bonn, [2]Fraunhofer IAIS, [3]TU Wien
toborek@cs.uni-bonn.de

Background

- Text in simple language benefits language learners, people with learning difficulties, and children
- Most popular dataset Simple English Wikipedia (Coster and Kauchak 2011)
- In German, there are two forms of simple language: Leichte Sprache (LS) and Einfache Sprache (ES)
- There is no publicly available Simple German resource in any of the above simplifications
- Text simplification requires aligned resources

Not aligned	[AS] Sämtliche Bälle müssen kugelförmig sein.	All balls must be spherical in shape.
	[LS] Die Fahnen sollen bunt sein.	The flags should be colorful.
Partially Aligned	[AS] Diverse öffentliche Verkehrsmittel bieten eine optimale Anbindung an die Hamburger Innenstadt, die Autobahn sowie den Flughafen.	Various means of public transport offer an optimal connection to Hamburg's city center, the highway as well as the airport.
	[LS] Die fahren bis zur Autobahn und zum Flughafen.	They go all the way to the highway and the airport.
Aligned	[AS] Bei Milch ist, falls es sich nicht um Kuhmilch handelt, die Tierart des Ursprungs anzugeben.	For milk, if it is not cow's milk, indicate the animal species of origin.
	[LS] Manchmal ist die Milch nicht von einer Kuh. Dann muss man sagen von welchem Tier die Milch ist.	Sometimes the milk is not from a cow. Then you have to say which animal the milk is from.

Example sentence pairs aligned between Simple German [LS] and German [AS] and their translations. Examples show successfully, partially, and wrongly aligned sentences.

Our Dataset

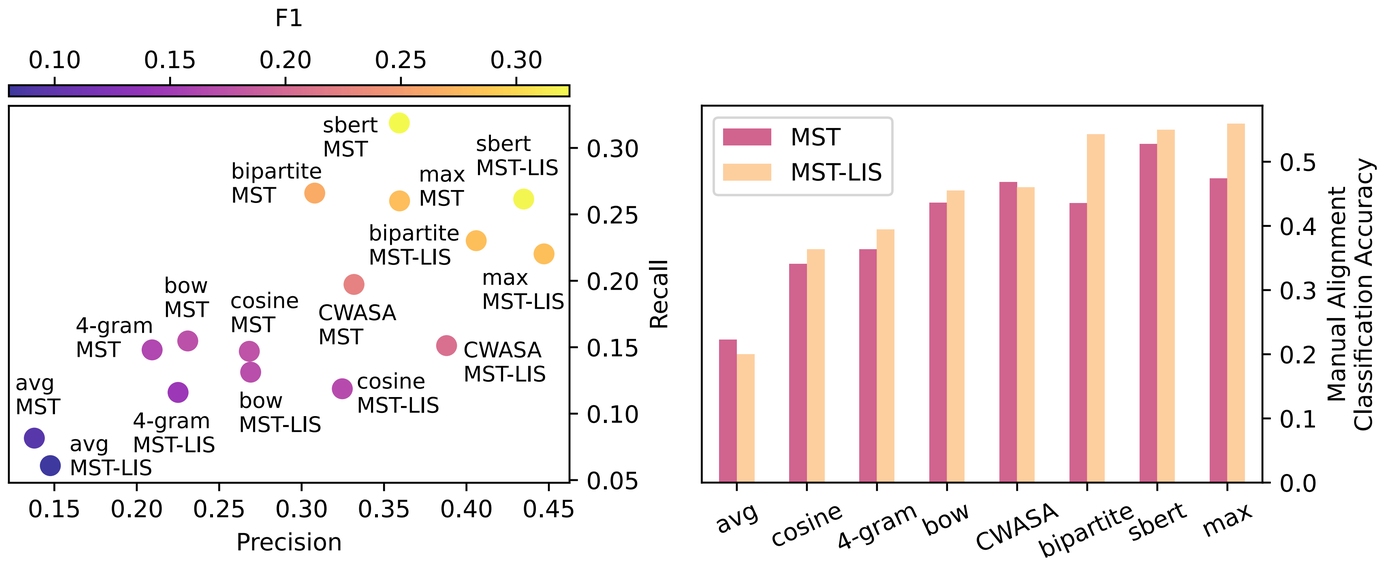
- Around 700 Simple German articles from eight different web sources with aligned German text
- Corpus spans different topics from general health information over administrative information to general news
- Seven sources in LS, one extensive source in ES
- Around 250,000 tokens in Simple German
- Article aligned corpus by design, which we extended by sentence alignments

		Simple German		German		
source		a	t	a	t	type
apotheke-umschau.de	General health information	168	94 808	166	187 427	ES
behindertenbeauftragter.de	Official office for disabled people	21	5 490	21	8 131	LS
brandeins.de	Various topics	47	9 634	47	9 728	LS
lebenshilfe-main-taunus.de	NPO for disabled people	45	6 946	45	9 023	LS
mdr.de	State-funded public broadcasting	322	53 277	322	126 191	LS
sozialpolitik.com	Explains social policy in Germany	15	5 122	15	11 790	LS
stadt-koeln.de	Administrative information	82	66 892	82	44 310	LS
taz.de	German Newspaper	8	7 924	14	8 171	LS
total		708	250 093	712	404 711	

Overview of the websites used for the corpus and some basic statistics. (a) articles and (t) tokens per source. The last column describes the type of Simple German used by the website „Einfache Sprache“ (ES) or „Leichte Sprache“ (LS).

Evaluation of Sentence Alignments

- Two fold evaluation of the combination of eight similarity measures and two matching methods
- (left) Manual creation of gold standard alignments allowed calculation of F1 score
  - sbert MST-LIS F1 score of 0.32
- (right) Manual evaluation of potential matches according to Xu et al. 2015 to calculate the Manual Alignment Classification Accuracy
  - max MST-LIS 55.94%
- We choose max MST-LIS for better precision



(left) Precision, recall and F1-score for all algorithm variants evaluated on the ground truth. (right) Manual alignment classification accuracy for the manually labelled matches.

Check out

- The biggest corpus in Simple German, completely publicly available
- Over 700 aligned documents from eight different web sources
- Over 10,000 matched sentences with a final F1 score of 0.28 compared to previously 0.085 (Klaper et al. 2013)
- Publicly available gold standard alignments
- Well preserved resources:
  - you can find the code on GitHub
  - the original sites are preserved at web.archive.org

Bibliography:
Coster, W., & Kauchak, D. (2011). Simple English Wikipedia: A New Text Simplification Task. ACL: Human Language Technologies.
Xu, W. et al. (2015). Problems in Current Text Simplification Research: New Data Can Help. ACL.
Klaper, D., Ebling, S., & Volk, M. (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. Second Workshop on Predicting and Improving Text Readability for Target Reader Populations.

UNIVERSITÄT BONN

LAMARR
Institute for Machine Learning and Artificial Intelligence