Language Level Classification on German Texts using a Neural Approach

Anonymous ACL submission

Abstract

Studies on language level classification (LLC) for German are scarce. Of the few existing, most use a feature-engineered approach. To the best of our knowledge, there is no deep learning approach on German texts yet. This paper shows that LLC can also be successfully applied to German texts by exploiting different pre-existing neural network architectures. Seven diverse corpora represent the data basis for training the networks: a web-scraped corpus, a corpus created from newspaper articles, three second language learner corpora, a corpus created by a company that translates complex texts into incremental simplified versions, and a corpus created from a collection of written examinations covering the whole CEFR spectrum (A1-C2). An approach based on the BERT architecture yielded the best results. The highest $F_1$ score achieved was 1.0 and 0.83 on a document and sentence level, respectively.

1 Introduction

Taking part in society requires access to textual information about culture, literature, politics, economics, etc. Simplified texts can be a support for people having difficulties receiving complex information. On the one hand, they can help people learning a second language, and on the other, first-language users of a language (Lotherington-Woloszyn, 1993; Yano et al., 1994; Long and Ross, 1993; Tweissi, 1998; Oh, 2001; Crossley et al., 2014). Indeed, simplified texts are mostly written for people with cognitive impairments. Article 21a of the UN Convention on the Rights of Persons with Disabilities states that state parties shall take measures for “providing information intended for the general public to persons with disabilities in accessible formats and technologies appropriate to different kinds of disabilities” (UN, 2006). Other target groups can benefit from simplified texts as well, such as persons with aphasia or dyslexia. Moreover, simplified texts can be an important resource for non-specialists in a certain domain.

LLC allows users to access texts on a language level adapted to their proficiency. In performing LLC for German, we achieve state-of-the-art results on the MERLIN corpus, a collection of written examinations by second language learners who had to write an e-mail, a letter or an essay according to their respective language level. We achieve a $F_1$ score increase of 0.18 in comparison to the previous state-of-the-art approach (Szügyi et al., 2019).

2 Related Work

So far, LLC has mainly been approached using a feature engineering approach (McCarthy, 2005; Vajjala and Meurers, 2012; Karpov et al., 2014; De Clercq and Hoste, 2016; Mesgar and Strube, 2018; Bestgen, 2020; Weiss et al., 2021; Imperial and Ong, 2021). In order to train a good machine learning algorithm, extensive feature engineering is required, which is costly and time consuming. Methods that use an artificial neural network have proven for years that they deliver state-of-the-art results in various natural language processing (NLP) tasks, such as part-of-speech (POS) tagging (Bohnet et al., 2018), named entity recognition (Yamada et al., 2020), sentiment analysis (Yang et al., 2019), machine translation (Edunov et al., 2018) or text simplification (Martin et al., 2020).

A neural approach to LLC was pioneered in 1994. McEneaney (1994) developed six back propagation networks. The networks either used a pre-existing readability formula of Fry (1968) or a sample of 50 words as a “visual pattern”, which in this case renders LLC a pattern recognition task. Considering the computing resources of the 90s, the author expressed doubts about the sufficient computing power needed to implement his approach and referred to future research.

24 years later the first follow-up article to neural
LLC was published in 2018. Nadeem and Osten-dorf (2018) applied two neural network architec-
tures on the WeeBit corpus (Vajjala and Meurers,
2012), firstly, a sequential recurrent neural network
(RNN), and secondly, a hierarchical one. It was
shown that the hierarchical outperformed the se-
quential RNN, achieving a correlation of 0.69 on
the WeeBit corpus. The authors also showed that
neural networks can be a good alternative to tradi-
tional feature-engineered models for texts shorter
than 100 words but do not perform adequately on
longer texts.

The WeeBit corpus was also used to test
the performance of different embedding models
(word2vec, GloVe, ELMo and BERT) on an LLC
task (Filighera et al., 2019). The embeddings
served as input to either an RNN or a convolu-
tional neural network (CNN). When combining all
models into an ensemble, the authors achieved an
accuracy of 0.813.

A multiattentive RNN architecture for automatic
multilingual readability assessment, Vec2Read,
was presented by Azpiazu and Pera (2019). A
multiattentive mechanism adapts and gives more
weight to specific data points depending on the
task. The authors observed for smaller datasets
that coarser information (e.g. POS tags) was used,
whereas for larger datasets more fine grained infor-
mation (e.g. word embeddings) was used by the
network. The authors reported a result of 0.527
accuracy on the Newsela corpus.

All aforementioned methodologies were applied
to English datasets. To the best of our knowledge,
a deep learning approach was never applied to Ger-
man texts as introduced here.

3 Experiments
In this section, we present the experimental setup
and results of applying a neural LLC approach on
German texts.

3.1 Data
Both learner corpora from the L2 domain and texts
human-translated into simplified language, sepa-
rated into one or more language levels, are utilised
in this study.

Presumably one of the first German learner cor-
pora is Falko, a corpus of argumentative texts writ-
ten by advanced learners of German (L2). The texts
in the corpus stem from two writing tasks: litera-
ture summaries and argumentative essays. For each
task, a control corpus of native speaker texts (L1)
has been compiled under the same conditions. The
two writing tasks resulted in two separate corpora,
the Falko Essays Corpus with 346 documents and
10,382 sentences and the Falko Summaries Corpus
with 164 documents and 3,294 sentences (Reznicek
et al., 2012).

The Corpus of LEarer German (CLEG), cre-
ated at Lancaster University (UK), consists of argu-
mentative writing of British students with Ger-
man as L2 (second language). All students had
English as L1 and had passed their A-Levels in
German. Free compositions, like critical commen-
taries, critical summaries and argumentative es-
says were collected. The CLEG contains 731 texts
with 18,619 sentences in total (Maden-Weinberger,
2013, 2015).

Boyd et al. (2014) introduced the MERLIN cor-
pus containing 2,286 written documents of language
learners in Czech, Italian and German. The
corpus covers the whole CEFR spectrum from A1
to C2. The sub-corpus of German includes 1,033
texts with 11,169 sentences. It was compiled from
standardised CEFR-related exams of L2 learners at
a language institute in Germany.

The simplified German Web corpus is a col-
lection of texts extracted from web sources in
Germany, Austria and Switzerland (Battisti et al.,
2020). Access to simplified information has re-
cently been introduced into legislation in those
countries. Acting as a role model, civil institutions
provide the public with texts in simplified language
on their websites. The Web corpus is separated
into a parallel corpus and a monolingual corpus.
The parallel data consists of 756 documents and
39,822 sentences. The monolingual data consists
of 1,916,045 tokens.

The Austrian Press Agency (APA) is the national
news agency and the leading information provider
in Austria. Since 2017, APA has published a sum-
mary of the four to six most important news of the
day in two language levels, B1 and A2. The APA
corpus was built by Säuberli et al. (2020); Spring
et al. (2021) and in its most recent version consists
of 6,012 documents and 79,085 sentences.

The capito corpus is a compilation of docu-
ments human-translated into simplified language
(levels A1, A2 and B1) at the Austrian company
CFS/capito (Spring et al., 2021). The company of-
ers specialises products and services for persons
with disabilities. The whole capito corpus includes
1,963 documents and 132,958 sentences.

### 3.2 Methods

Martinc et al. (2021) proposed a new approach to LLC using different deep learning techniques and applying them on English and Slovenian texts. The authors utilised three pre-existing neural network architectures: The first is a Hierarchical Attention Network (HAN) proposed by Yang et al. (2016), in which the authors made two assumptions: firstly, that documents have a hierarchical structure, and secondly, that there is more and less important content in the text when constructing an overall document embedding. The model proposed is a hierarchical approach with the aggregation of important words into sentence vectors constituting the lower level and the aggregation of important sentence vectors into document vectors, the higher level. The authors showed that the attention layers are effectively picking out semantically important words and sentences. In experimental results their model outperformed those of previous studies in six different classification tasks.

The second neural network architecture is the Bidirectional Long Short-term memory (BiLSTM) network proposed by Conneau et al. (2017). For a sequence of $W$ words, a bidirectional LSTM computes a set of $W$ vectors $V_i$, whereby $V_i$ is the concatenation of a forward LSTM and a backward LSTM that read the sentences in two directions. To combine the varying number of $V_i$, the authors experimented with different approaches: firstly, taking the average of the vectors, which is referred to as mean pooling, and secondly, taking the maximum value over each dimension of the hidden units, known as max pooling (Collobert and Weston, 2008). The BiLSTM with max pooling outperformed previous models in four out of six classification tasks.

The third neural network architecture applied was BERT (Devlin et al., 2018). The architecture consists of 12 layers of size 768 and 12 self-attention heads. For the sake of LLC, a linear classification head was added on top of the pre-trained language model (Huggingface, 2019). The model can be fine-tuned in different ways: e.g. Martinc et al. (2021) suggested a training of 3 epochs (which also showed best results in this study). The pre-trained German language model used was the bert-base-german-cased, open sourced by the German company deepset (deepset, 2019).

We applied the approach of Martinc et al. on a new language and on new data. Additionally, an alternative LLC method was set in comparison in order to evaluate our results, the Language Level Evaluator (LLE), which was developed by the German company L-Pub GmbH. LLE is hosted on a website that contains an input mask for sentences or documents (Steel, 2021). Classification within LLE is based on three different word lists.

### 3.3 Results

For each corpus shown in Table 1 the result of the best performing architecture is marked in bold: BERT outperformed the other architectures on all seven corpora. Deep learning approaches based on the Transformer architecture (Vaswani et al., 2017) have shown to deliver state-of-the-art achievements in NLP (Edunov et al., 2018; Yang et al., 2019; Yamada et al., 2020; Martin et al., 2020). The good performance of BERT applied to German texts (Table 1) substantiates the efficiency of this neural network architecture. All language levels were classified 100% correctly by applying BERT on the MERLIN dataset; the source code underlying this experiment has been published to allow for replication. Since BERT clearly outperformed the other two neural network architectures, only the results of BERT are depicted in Table 1 for the sentence level. Considering the small amount of textual

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HAN</th>
<th>BiLSTM</th>
<th>BERT</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falko Essays</td>
<td>0.852</td>
<td>0.830</td>
<td>0.882</td>
<td>0.830</td>
</tr>
<tr>
<td>Falko Summaries</td>
<td>0.167</td>
<td>0.577</td>
<td>0.931</td>
<td>0.626</td>
</tr>
<tr>
<td>CLEG</td>
<td>0.890</td>
<td>0.785</td>
<td>0.986</td>
<td>0.817</td>
</tr>
<tr>
<td>Web</td>
<td>0.861</td>
<td>0.806</td>
<td>0.894</td>
<td>0.732</td>
</tr>
<tr>
<td>capito</td>
<td>0.667</td>
<td>0.673</td>
<td>0.765</td>
<td>0.653</td>
</tr>
<tr>
<td>APA</td>
<td>0.821</td>
<td>0.755</td>
<td>0.867</td>
<td>0.777</td>
</tr>
<tr>
<td>MERLIN</td>
<td>0.941</td>
<td>0.969</td>
<td>1.0</td>
<td>0.822</td>
</tr>
</tbody>
</table>

Table 1: $F_1$ scores for distinct German corpora on a document and sentence level.
material per sentence during training and the high
total number of units during testing, the results were
still satisfactory, with Falko Essays performing best
(0.830) and Falko Summaries worst (0.626). Gener-
ally, BERT achieved better results on a document
than on a sentence level.

In order to compare BERT to the lexical lan-
guage level classification methodology of LLE, two
confusion matrices are depicted. Figure 1 shows
the confusion matrix of BERT applied on 99 ran-
dom documents of the capito corpus that were split
equally into three parts of complex, B1 and A2.
BERT achieved an overall accuracy of 74.75%.
LLE, in comparison, achieved an overall accuracy
of 37.37% (Figure 2). Hence, BERT yielded twice
the accuracy of LLE.

Recurrent models such as RNNs and LSTMs
process the input sequentially (right-to-left or left-
to-right). BERT’s mask technique allows the model
to read the entire sequence of words once at a time
to learn the context of a word based on all of its sur-
rroundings. Furthermore, BERT uses the Attention-
mechanism introduced by Vaswani et al.. With the
help of this mechanism the model is able to achieve
advanced mappings of relationships between indi-
vidual words. These are two probable explanations
why BERT is performing so well compared to the
other methodologies introduced.

4 Conclusion

To the best of our knowledge, our contribution is
the first to use a neural approach to language level
classification for German. Out of three neural net-
work architectures, BERT showed the best results
both on a document and on a sentence level. When

compared to an alternative lexical-based method-
ology, BERT was able to correctly classify the lan-
guage levels of twice the number of documents.

Acknowledgements

We would like to thank the Austrian Press Agency
and the CFS/capito company for their continuous
support and for providing us with their data for this
study.

Source Code

https://github.com/kinimod23/

GRANT

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019.
Multiattentive recurrent neural network architecture
for multilingual readability assessment. Transactions
of the Association for Computational Linguistics,
7:421–436.

A Battisti, S Ebling, D Pfütze, A Saeuberli, and
M Kostrzewa. 2020. A corpus for automatic readabil-
ity assessment and text simplification of German. In
Proceedings of the 12th International Conference on
Language Resources and Evaluation (LREC), Mar-
seille, France 2020.

Yves Bestgen. 2020. Reproducing monolingual, multi-
lingual and cross-lingual CEFR predictions. In Pro-
cceedings of The 12th Language Resources and Evalua-
tion Conference, pages 5595–5602.

Bernd Bohnet, Ryan McDonald, Goncalo Simoes,
Daniel Andor, Emily Pitler, and Joshua Maynez.
2018. Morphosyntactic tagging with a meta-
BiLSTM model over context sensitive token encod-


Heather Lotherington-Woloszyn. 1993. Do simplified texts simplify language comprehension for ESL learners?


Adel I Tweissi. 1998. The effects of the amount and type of simplification on foreign language reading comprehension.


