Analysing Direct Speech in German Novels

Fotis Jannidis, Albin Zehe, Leonard Konle, Andreas Hotho, Markus Krug

Universität Würzburg

DHd-Tagung 2018, Köln

Introduction

Detecting direct speech in fiction allows gaining insight into an important element of its narrative structure. In literary studies, there are assumptions on the factors influencing the distribution of direct speech, like genre, period and aesthetic complexity.

This paper aims to provide a detailed analysis of the use of direct speech across different time periods and domains. To create a reliable database for these analyses, we need to measure the usage of direct speech in a large and representative corpus. This task is more challenging than it may sound: While, nowadays, direct speech is often marked very explicitly by the use of quotes, this has not always been consistently the case. Many historical novels are not available in a well-edited form, meaning that there may be inconsistent use of quotation, or no quotation at all (Brunner, 2013). In this case, a more robust method for detecting direct speech is necessary.

Our first contribution is therefore a deep learning-based method to detect direct speech using large amounts of rule-based, but slightly flawed, labelled data extracted from raw text. This has multiple advantages over the use of manually annotated training data: First, manually annotating large amounts of text is very time-intensive and therefore costly. Furthermore, annotations for one type of texts may not be transferable to other types, leading to the necessity of new annotated data for new corpora. Being able to learn from the already existing weakly labelled

data is therefore desirable, as this data can automatically be extracted for a new corpus.

Our second contribution is the application of this approach on curated texts to gain insight in trends of direct speech distribution. On one hand we try to look for development of direct speech over time, analysing a large dataset of novels from the nineteenth century, on the other hand we focus on differences in genre comparing contemporary high and low brow literature.

Related Work and Task Description

There have been several previous approaches to direct speech detection applying machine learning methods.

For example, Brunner (2013) tests rule-based and machine learning driven classification, as well as combinations of both, on German novels. She recommends using a pure machine learning approach (Random Forest), reaching an F1 score of 0.87.

Scheible et al. (2016) employ a simple greedy algorithm and a semi-Markov model, showing that the latter outperforms the previous state-of-the-art by achieving a precision of 0.88.

Although the results seem quite satisfying, these systems require a relatively large amount of labelled data for training. As stated above, this is problematic because of the need for expensive annotation and lack of transferability to other domains. Thus, our goal in this paper differs from that in previous work. We do not aim to set a new state-of-the-art in direct speech detection, but instead:

a) present a method that can leverage large amounts of weakly labelled data extracted from raw text, and

b) use this model for the analysis of different distributions of direct speech across genres or time-periods.

To the best of our knowledge, the second task has never been done on a large collection of texts.

Corpus and Resources

The following experiments are based on three German corpora. The first one is a large corpus containing 4600+ public domain novels including texts from the TextGrid digital library[1] and Project Gutenberg[2]. We will refer to this as the Corpus *Public Domain*, PD. The second one contains 800+ texts of current popular genres like romance, crime or science-fiction (Corpus *Low Brow*, LB). Finally, we use a corpus with 200 novels nominated for the *German Book Prize* or the *Georg Büchner Prize* (Corpus *High Brow*, HB).

In order to train and evaluate our classifiers, we need to obtain labels specifying which parts of the texts contain direct speeches. To this end, we chose two strategies:

For training our classifiers, we decided to extract weak labels using a simple rule based on quotation, implying everything written between quotation marks is direct speech. To yield high accuracy for this approach, it is necessary to use a well-edited collection of texts. Our PD corpus contains such a subset, which we refer to as our *Kerncorpus*. This *Kerncorpus* consists of 250 high and middle brow texts (those from the TextGrid digital library), has been manually edited and is assumed to have a mostly consistent use of quotation.

Using our quotation rule on the *Kerncorpus* resulted in a dataset where about 36% of tokens were marked as direct speech. In order to assess the quality of these weak labels, we gave 500 of the sentences to domain experts for manual correction. We found that there was an error-rate of about 3% in those sentences, mostly caused by nested direct speech or inscriptions being enclosed by quotation marks.

For further evaluation, we chose to annotate a smaller subset of the corpus *LB* by hand. We selected 50 snippets from texts of low brow literature. This dataset, referred to as *ALB*, is relatively skewed towards text outside direct speech, with only about 18% of tokens in a direct speech.

[1] https://textgrid.de/digitale-bibliothek
[2] http://gutenberg.spiegel.de/

Experiments

The following experiments use both labelled subsets described above, the large *Kerncorpus* and the smaller *ALB*.[3] For all experiments, quotation marks are removed from the texts. This is done to avoid training models that rely only on the formal style of qualifying direct speech, but also consider implicit signs like the use of first person verbs or speech words.

We conducted experiments on two different levels, starting with a sentence classification task, which is then refined to detect direct speech on word-level.

Sentence-Level Classification

In our first classification task, documents are split into sentences and vectorised by storing each sentence in a bag-of-words representation. To create a baseline for measuring the advantage using deep learning for direct speech recognition, we compared the performance of traditional machine learning algorithms on our labelled datasets. Training and testing some of the most common machine learning classifiers to detect sentences containing at least one word of direct speech leads to an accuracy of **0.85** using Logistic Regression; for more results see Table 1.

Table 1: Results of traditional machine learning algorithms for direct speech detection.

Algorithm	Multinomial Naive Bayes	Random forest	SVM linear kernel	Logistic Regression	K-nearest neighbours	Passive Aggressive	Perceptron
Accuracy	0.84	0.78	0.80	**0.85**	0.67	0.78	0.76

Using the same setting and replacing machine learning with a combination of recurrent and convolutional neural networks (see Chollet 2017 and Goodfellow 2017) ended up with an accuracy of **0.84**.

[3] We cannot use the remaining texts for either training or evaluation, as we do not have any reliable source of labels for these texts.

Since we noticed that three of our classifiers all ended up with about the same score, we decided to give the task to two human annotators to establish an upper bound. We selected 250 sentences for manual annotation and again removed all quotation marks. Both annotators ended up with an accuracy comparable to that of the best machine learning methods, 84% and 82.8% respectively. From this result we concluded that it is not expedient to further optimise the sentence classification task, as we had already reached human-level accuracy.

Word-Level Classification

Because of the results from the previous section, we decided to modify our task to a word-level prediction, which enables us to include more context by ignoring sentence boundaries and at the same time make more fine-grained predictions. In this second classification task, each word is to be classified separately as inside or outside a direct speech. As baseline for this task, we trained a Linear Chain Conditional Random Field (CRF) that was only given the word itself and its part-of-speech tag. This CRF stagnated at a comparably low accuracy of **0.71** using cross-validation on the *Kerncorpus*.

Since our goal was to provide the classifier with more context, we chose to use an architecture based on recurrent neural networks, which are able to deal with relatively large contexts. Our assumption here is that, for a good classification, we need context from both before and after the target word itself, as markers for direct speech can be found at the beginning or the end of the direct speech. We thus designed a two-branch network, visualised in Figure 1. This network receives as input a text-segment, specifically the target word in its context. The words of the input are then passed through an embedding layer and split into two parts, where the first part contains the context up to the target word and the second part contains the context following the target word. The target word itself is contained in both parts. Each part is passed through three separate LSTM-layers. In the future-branch, the context is passed through the layers in reverse, so that the target word is the last word to be read in both branches. The LSTM-layers in the past-branch are stateful and can therefore theoretically retain the entire context of the novel up to the target

word. The outputs of the final LSTM-layer of both branches are concatenated. The final prediction is made based on this concatenation by a fully connected layer.

In our best setup, we used 60 words before and after the target word as context. Training on one half of the *Kerncorpus* and evaluating on the other half, this setup yielded an accuracy of **0.83**. Training on the full *Kerncorpus* and evaluating on the manually annotated *ALB* reached an even better accuracy of **0.90**.

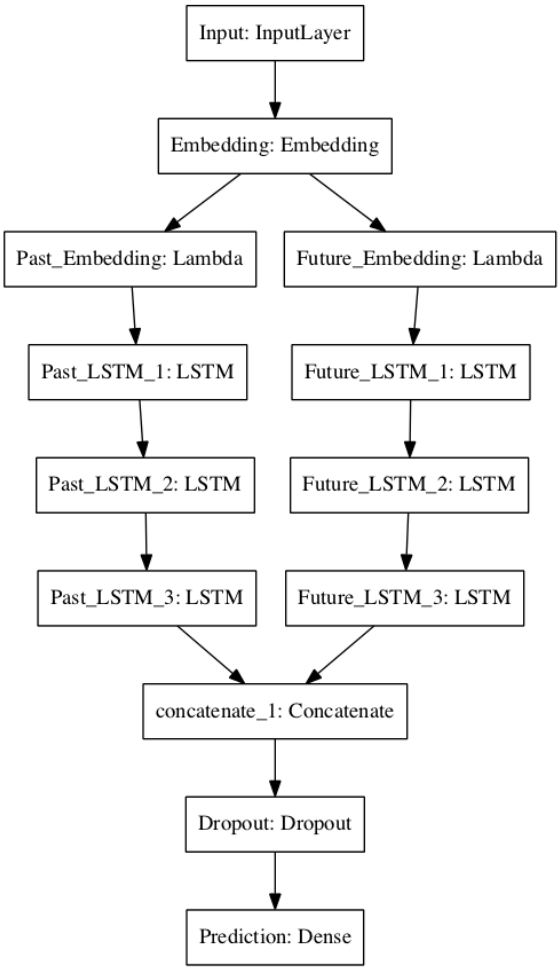

Figure 1: Architecture of recurrent network to detect direct speech

Distribution of direct speech

In the following experiments, we used the model based on the architecture described above. We trained this model on the *Kerncorpus* and used it to detect direct speech

in the complete corpora *PD*, *LB* and *HB*. Here, we describe our findings on these corpora.

Direct speech in 19th Century Fiction

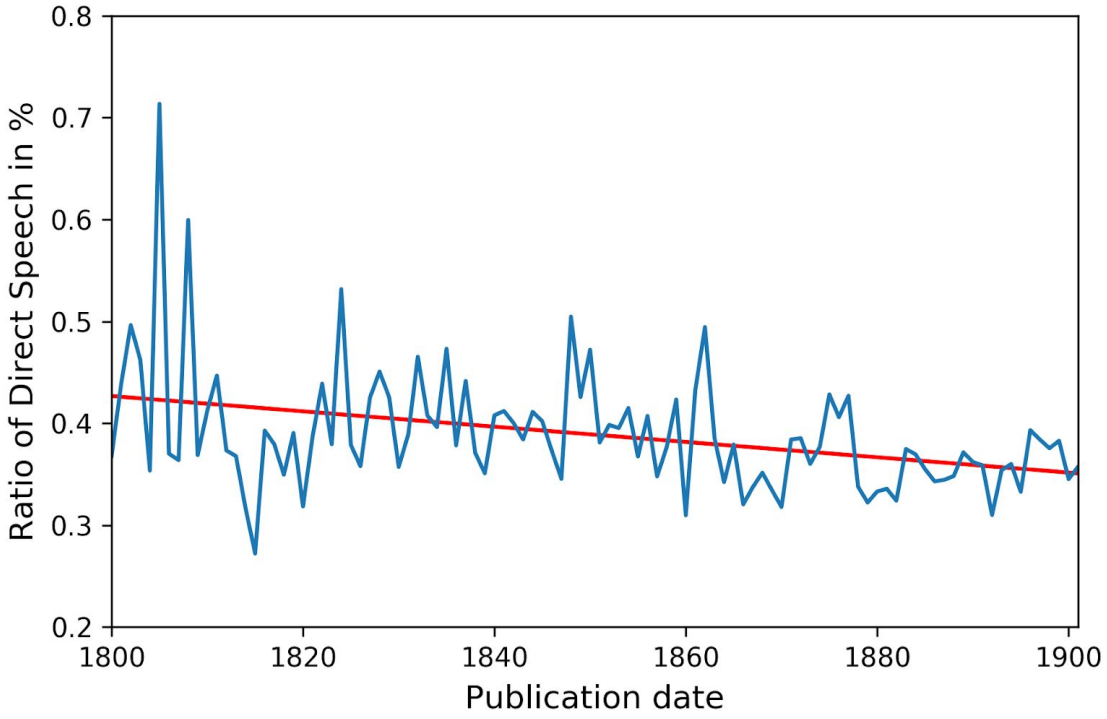

Figure 2: Ratio of direct speech in German novels from 1800 - 1900

Figure 2 shows the ratio of direct speech in German novels from 1800 till 1900 based on the texts from Corpus *PD*. The regression line indicates a decline of direct speech over time; at the same time, we can observe a decrease of variance. The strong variations between certain years, especially in the early 19th century, are caused by low numbers of provided texts (see Fig. 3). For instance, the peak in 1805 can be explained by the first publication of Denis Diderots ”Herrn Rameaus Neffe”, a philosophical dialogue-based novel.

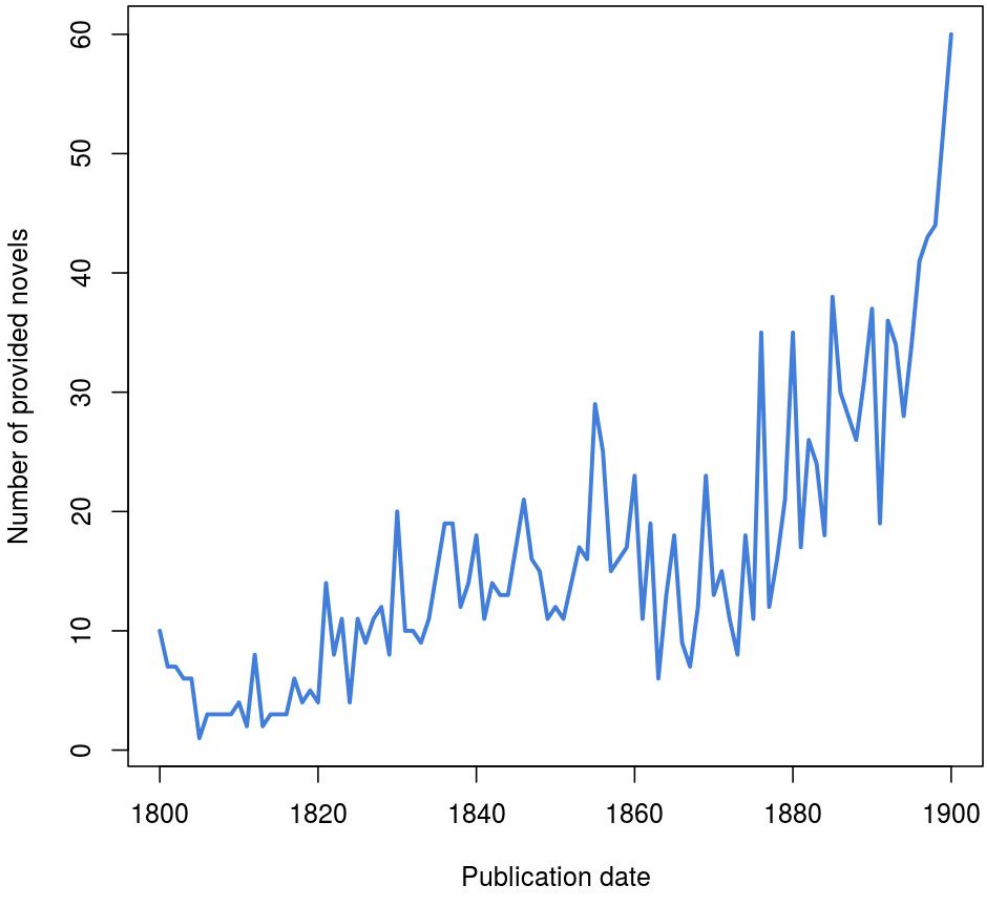

Figure 3: Number of provided novels per year

Distribution of direct speech in low and high brow literature

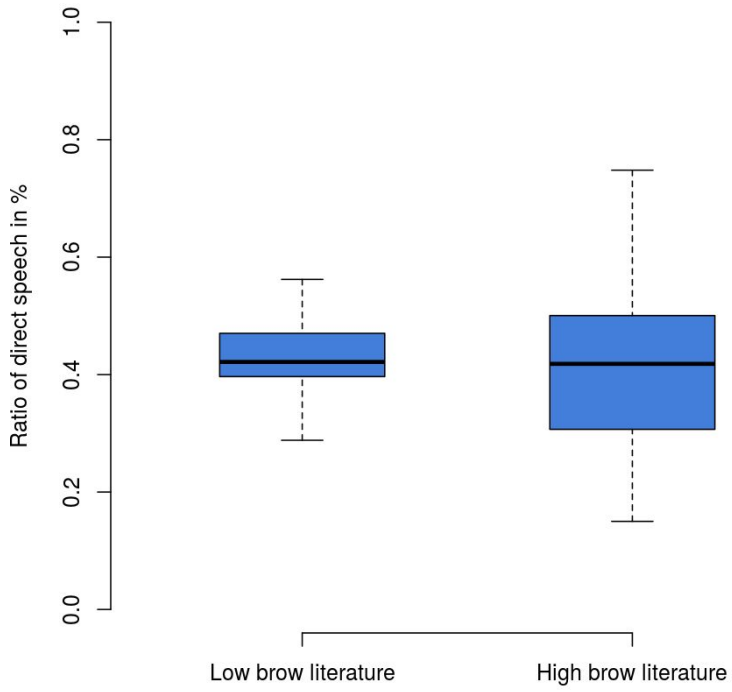

Figure 4: Ratio of direct speech in German low and high brow novels after 1945

There is an assumption in literary studies that a huge amount of direct speech is an indicator of low brow fiction. Figure 4 shows the ratio of direct speech between Corpus *LB* and *HB*. While the mean usage of direct speech is nearly equal in both groups, the high brow literature is far more variable.

This finding is contrary to the assumption mentioned above. We propose that, while there is no clear difference in the average use of direct speech between high and low brow literature, authors in high brow literature are far more flexible in choosing how much direct speech they use in their novels. Low brow literature, on the other hand, is expected to have a rather constant amount of dialogue.

Conclusion and Future Work

In this paper, we introduced a neural network architecture that is able to learn the classification of direct speech by training on weakly labelled data. This network works purely on the raw text of a novel by taking into account a relatively large context. We also demonstrate that training on weakly labelled data leads to satisfying results.

While an accuracy of 0.9 is remarkable, there is still need for optimisation. Recent developments in the performance of neural networks by adding an attention mechanism (see Rush 2015) could improve the results.

We used our neural network to analyse the distribution of direct speech over time and genres. Besides algorithmic refinements, there is a lot of potential in adding more text to our corpus and refining metadata to allow more sophisticated research questions like differences between or development of direct speech in certain genres.

References

Brunner, Annelen (2013): “Automatic recognition of speech, thought, and writing representation in German narrative texts”, in *Literary and Linguistic Computing. Vol. 28 (2013).*

Chollet, Francois (2017): “Deep Learning with Python”. Manning Publications. New York. (Preprint: https://www.manning.com/books/deep-learning-with-python)

Goodfellow, Ian / Bengio Yoshua / Courville, Aaron (2016): “Deep Learning”. MIT Press.(URL: http://www.deeplearningbook.org)

Rush, Alexander M. / Chopra, Sumit / Weston, Jason (2015): “A Neural Attention Model for Abstractive Sentence Summarization”. *arXiv preprint arXiv:1509.00685.*

Scheible, C., Klinger, R. & Padó, S. (2016): “Model Architectures for Quotation Detection”, in *Proceedings of ACL (p./pp. 1736--1745)*