Originally published in: Ebling, Sarah/Tuggener, Don/Hürlimann, Manuela/Cieliebak, Mark/Volk, Martin (Eds.):
Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS). Zurich, 23-25 June 2020.
– Aachen: CEUR Workshop Proceedings, 2020.

To BERT or not to BERT – Comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation

Annelen Brunner, Ngoc Duyen Tanja Tu
Leibniz-Institut für Deutsche Sprache
R5 6-13
D-68161 Mannheim
brunner|tu
@ids-mannheim.de

Lukas Weimer, Fotis Jannidis
Universität Würzburg
Am Hubland
D-97074 Würzburg
lukas.weimer|fotis.jannidis
@uni-wuerzburg.de

Abstract

We present recognizers for four very different types of speech, thought and writing representation (STWR) for German texts. The implementation is based on deep learning with two different customized contextual embeddings, namely FLAIR embeddings and BERT embeddings. This paper gives an evaluation of our recognizers with a particular focus on the differences in performance we observed between those two embeddings. FLAIR performed best for direct STWR (F1=0.85), BERT for indirect (F1=0.76) and free indirect (F1=0.59) STWR. For reported STWR, the comparison was inconclusive, but BERT gave the best average results and best individual model (F1=0.60). Our best recognizers, our customized language embeddings and most of our test and training data are freely available and can be found via *www.redewiedergabe.de* or at *github.com/redewiedergabe*.

1 Introduction

Speech, thought and writing representation (STWR) is an interesting phenomenon both from a narratological and a linguistic point of view. The manner in which a character's voice is incorporated into the narrative is strongly linked to narrative techniques as well as to the construction of the narrative world and is therefore a standard topic in narratology (e.g. McHale (2014); Genette (2010); Leech and Short (2013)). For some phenomena, such as free indirect discourse and stream of consciousness, there is a large amount of research (e.g. Banfield (1982); Fludernik (1993); Pascal (1977)). In linguistics, the grammatical, lexical and functional characteristics of STWR have also been of interest (e.g. Weinrich (2007); Zifonun et al. (1997); Hauser (2008); Fabricius-Hansen et al. (2018)).

To conduct either narratological or linguistic studies on STWR based on big data, being able to automatically detect different types of STWR would be of great benefit. This was our motivation to develop recognizers for the following four forms of STWR, which have been distinguished in literary and linguistic theory.[1]

Direct STWR is a quotation of a character's speech, thought or writing. It is frequently – though not always – enclosed by quotation marks and/or introduced by a framing clause.

> Dann sagte er: *"Ich habe Hunger."*
> (Then he said: *"I'm hungry."*)

Free indirect STWR, also known as "erlebte Rede" in German, is mainly used in literary texts to represent a character's thoughts while still maintaining characteristics of the narrator's voice (e.g. past tense and third person pronouns).

> Er war ratlos. *Woher sollte er denn hier bloß ein Mittagessen bekommen?* (He was at a loss. *Where should he ever find lunch here?*)

Indirect STWR is a paraphrase of the character's speech, thought or writing, composed of a framing clause (not counted as part of the STWR) with a dependent subordinate clause (often using subjunctive mode) or an infinitive phrase.

> Er fragte, *wo das Essen sei.* (He asked *where the food was.*)

[1] The stretch of STWR is printed in italics in the following examples.

Reported STWR is defined as a mention of a speech, thought or writing act that may or may not specify the topic and does not take the form of indirect STWR.

> *Er sprach über das Mittagessen.* (*He talked about lunch.*)

In the following, we will describe our approach in developing recognizers for these four STWR types and evaluate our results with a particular focus on the differences between the two contextual embeddings that proved most successful for our task, BERT and FLAIR.

2 Related work

2.1 STWR recognizers

Automatic STWR recognition focuses mainly on the forms direct STWR (e.g. Schricker et al. (2019); Jannidis et al. (2018); Tu et al. (2019); Brunner (2015); Brooke et al. (2015) for German texts; Schöch et al. (2016) for French texts) and indirect STWR (e.g. Schricker et al. (2019); Brunner (2015) for German texts; Lazaridou et al. (2017); Scheible et al. (2016) for English texts; Freitas et al. (2016) for Portugese texts). For free indirect and reported STWR, recognizers were implemented by Brunner (2015) and Schricker et al. (2019), the latter builds upon work by the former. In addition to that, Papay and Padó (2019) propose a corpus-agnostic neural model for quotation detection.

Since we developed recognizers for German, we will only take a closer look at recognizers trained and tested on German texts. Jannidis et al. (2018) developed a deep-learning based recognizer for direct speech in German novels, which works without quotation marks and achieves an accuracy of 0.84 in sentence-wise evaluation. The algorithm by Brooke et al. (2015) is a simple rule-based algorithm, which matches quotation marks. They do not report any scores. Like Jannidis et al. (2018), Tu et al. (2019) focus on developing a recognizer for direct speech which works without quotation marks, but they used a rule-based approach, achieving a sentence level accuracy between 80.5 to 85.4% for fictional and 60.8% for non-fictional data. Brunner (2015) uses a corpus of 13 short German narratives. For each type of STWR, she implements a rule-based model and trains a RandomForest model, evaluated in ten-fold cross validation. The best F1 scores, evaluated on sentence level, were achieved by the rule-based approach for indirect (F1=0.71) and reported (F1=0.57) STWR and by the RandomForest model for direct (F1=0.87) and free indirect (F1=0.40) STWR. Schricker et al. (2019) use the same corpus, but split the data into a stratified training and test set. They use different features than Brunner and train three different machine learning algorithms, RandomForest, Support Vector Machine and Multilayer Perceptron. RandomForest was most successful and gave sentence-wise F1 scores of 0.95 for direct, 0.79 for indirect, 0.70 for free indirect and 0.49 for reported STWR. Papay and Padó (2019) test their corpus-agnostic quotation detection model on Brunner's corpus and approximate her RandomForest results.

Our recognizers fill a need regarding the recognition of STWR in German texts, as they deal with all four forms of STWR and are, at the same time, trained and tested on a much larger data base than Brunner's corpus, making our results much more reliable. In addition to that, our data not only comprises fictional, but also non-fictional texts. An earlier version of our recognizer for free indirect STWR was discussed in Brunner et al. (2019). We improved on this version by adding more training data and achieving higher scores with the BERT based model.

2.2 Language embeddings

As the testing of different language embeddings was a central component in the development of our recognizers, we will briefly outline characteristics and research concerning the two most successful ones that will be in focus in the rest of this paper: FLAIR embeddings (Akbik et al., 2018) and BERT embeddings (Devlin et al., 2019).

Both have in common that they produce context-dependent embeddings as opposed to static word embeddings, such as fastText (Bojanowski et al., 2016), GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013). That means they assign an embedding to a word based on its context and are therefore able to capture polysemy. Contextual embeddings have been shown to be of great benefit in several NLP tasks, e.g. predicting the topic of a tweet (Joshi et al., 2019), part-of-speech-tagging, lemmatization and dependency parsing (Straka et al., 2019). Though FLAIR and BERT both produce contextual embeddings, they differ in several features.

FLAIR produces character-level embeddings. A sentence is passed to the algorithm as a sequence of characters and the task is to predict the next character based on the previous characters. When using FLAIR embeddings, it is recommended to combine two independently trained models: i) a forward model and ii) a backward model. The forward model reads every character in a sentence from left-to-right, the backward model from the opposite direction. This character-based embedding architecture gives advantages in capturing morphological and semantic structures (cf. Akbik et al. (2018)).

While FLAIR embeddings only know their previous context, BERT embeddings know the previous as well as succeeding context at the same time. The training of BERT embeddings consists of two tasks: 1) given a sequence of tokens, where 15% of tokens are masked by the masked language model, predict the masked token based on its context, 2) given two pair of sentences, predict if the second sentence is the subsequent sentence of the first one. The advantage of this model is that it learns associations between tokens as well as sentences. This is important for token-level tasks, such as question answering (cf. Devlin et al. (2019)).

There are no systematic analyses comparing the performance of embeddings for a sequence labeling task in German, like we will do in in this paper. However, there is work focusing on the use of different embeddings in NLP tasks in English texts: e.g. Wiedemann et al. (2019) compare BERT, ELMo and FLAIR embeddings in a word sense disambiguation task, where BERT performed best. Sharma and Daniel (2019) compare BioBERT (Lee et al., 2019) and FLAIR embeddings, more precisely the pubmed-x model, in a Biomedical Named Entity Recognition task. They find that stacking FLAIR embeddings with BioELMo yields better results than using only FLAIR. Compared to BioBERT the results of FLAIR are very close, although the FLAIR embeddings are pretrained on a much smaller dataset. There is also a comparison between BERT (bert-base-uncased, bert-base-chinese, bert-base-multilingual-uncased), FLAIR (bg, cs, de, en, fr, nl, pl, pt, sl, sv) and ELMo (english) in a part-of-speech-tagging, lemmatization and dependency parsing task on different languages: Straka et al. (2019) showed, that BERT outperforms ELMo as well as FLAIR embeddings in dependency parsing. As opposed to that, ELMo performs best in part-of-speech-tagging and lemmatization, followed by FLAIR. Therefore Straka et al. (2019) conclude that ELMo is best and FLAIR embeddings are second best in capturing morphological and orthographic information while BERT is best in capturing syntactic information.

3 Method

We defined the recognition of STWR as a sequence labeling task on token level. For each of the four types of STWR, a separate model was trained on binary labels ("token is part of this type of STWR: yes/no"). The input data consists of chunks of up to 100 tokens, which may span several sentences. The chunks may never cross borders between different texts or cut sentences (except when a sentence exceeds 100 tokens) and can therefore also be shorter than the maximum.

To train our tagging model, we used the SequenceTagger class of the FLAIR framework (Akbik et al., 2019) which implements a BiLSTM-CRF architecture on top of a language embedding (as proposed by Huang et al. (2015)). We use two BiLSTM layers with a hidden size of 256 each and one CRF layer. This setting was decided after running tests with only one BiLSTM layer, which gave considerably worse results, and with three BiLSTM layers, which led to no significant improvements.

We tested many different configurations for the language embeddings in this setup. Initial tests were done with just fastText embeddings. The results were much worse than the two configurations that became our main focus: a) a fastText model stacked with a FLAIR forwards and a FLAIR backwards model (as recommended in Akbik et al. (2018)) and b) a BERT model.

Except for free indirect, all of our recognizers were trained and tested on historical German. Using out-of-the-box embeddings, which are trained on modern texts like German Wikipedia dump, Open legal data dump, Open subtitles or the EU bookshop corpus (Tiedemann, 2012), is therefore problematic. So we custom-trained our own fastText and FLAIR embeddings and fine-tuned the BERT embeddings. The following settings were used:

Skip-Gram fastText models: We used the default setting as recommend by the fastText tutorial, i.e. we trained for five epochs, set the learning rate

to 0.05, adjusted the minimum of the character n-gram-size to 3 and the maximum to 6. We varied the model dimensions as well as the training material: fastTextTrain_clean is a smaller, cleaner corpus, fastTextTrain contains additional material with OCR errors (cf. section 4.1). On each training set, one model with 300 and one model with 500 dimensions was trained.

FLAIR: A forward and a backward FLAIR embedding with a hidden size of 1024 were trained. All settings were chosen according to the recommendation of the FLAIR tutorial, i.e. the sequence length was set to 250, the mini-batch size to 100, the learning rate to 0.20, the annealing factor to 0.4 and the patience value to 25. The model stopped training after 10 epochs due to low loss.

BERT: We used the PyTorch script fine-tune_on_pregenerated.py to fine-tune the pre-trained bert-base-german-cased-model with the recommended default configuration: epochs: 3, gradient_accumulation_steps: 1, train_batch_size: 32, learning rate: 0.00003 and max_seq_len: 128.

4 Data

4.1 Training data for the embeddings

For the training/fine-tuning of the embeddings 9,577 fictional and non-fictional German texts from the 19th and early 20th century were selected.

For the fine-tuning of the BERT embeddings, we fed all data – split into sentences – into the script pregenerate_training_data.py from PyTorch, which transforms it to BERT embedding compatible input data. The BERT fine-tuning tutorial recommends to create an epoch of input data for each training epoch, so BERT will not be trained on the same random splits in each epoch. We fine-tuned BERT for 3 epochs, so we generated 3 epochs of data.

For the FLAIR embeddings, 70%, i.e. 4,508,960 sentences, of the 6,441,372 sentences from the data were randomly drawn to form the training corpus. For the validation corpus 15%, i.e. 966,206 sentences, were randomly drawn. The rest was used for testing purposes.

For training the fastText embedding, we used two different inputs. FastTextTrain contained all of the 137,093,995 tokens of our data. From fastTextTrain_clean we removed all texts that were recognized with OCR and thus contained typical OCR errors. This resulted in a smaller input set of 131,360,863 tokens.

4.2 Training data for the recognizers

The recognizers for direct, indirect and reported STWR were trained on historical German texts – excerpts as well as full texts – that were published from the middle of the 19th to the early 20th century. It comprises fiction as well as non-fiction (newspaper and journal articles) in near equal proportion; fiction is somewhat more dominant. Roughly half of the data was manually labeled by two human annotators independently of one another. Then a third person compared the annotations, adjudicated discrepancies and created a consensus annotation. The rest was labeled by a single annotator.

For indirect STWR, the training data was supplemented with 16 additional historical full texts (9 fictional and 7 non-fictional) to increase the number of instances. To speed up the annotation process, these texts were automatically annotated by one of our earlier recognizer models and then manually checked. The annotators looked at the whole texts, so false negatives were corrected as well.

All the historical data is published as corpus REDEWIEDERGABE (Brunner et al., 2020) and freely available.

As the historical data contained much too few instances of free indirect STWR, we had to create a separate training corpus for this STWR type. The basis were 150 instances of free indirect STWR with little to no context, manually extracted from 20th century novels. In addition to that, full texts and excerpts from modern popular crime novels as well as dime novels were automatically annotated with a basic rule-based recognizer that used typical surface indicators. Those annotations were then verified by human annotators. On this data, we trained an early recognizer (Brunner et al., 2019) which was then used to annotate additional historical fictional texts. These annotations were again verified by human annotators before they were added to the training material as well. It should be noted that in this semi-automated annotation process, instances that were not detected by the early recognizers had no chance of being annotated. Because of this, the data most likely contains false negatives.

For model training, our data was split into a training corpus (648,338 tokens for direct and re-

	Training corpus			Validation corpus			Test corpus		
	Tokens	Percent	Instances	Tokens	Percent	Instances	Tokens	Percent	Instances
Direct	212,467	32.77	6,293	24,321	24.99	878	18,307	18.71	605
Indirect	49,222	7.03	3,505	8,502	8.74	571	8,664	8.86	545
Reported	66,817	10.31	7,522	11,404	7.73	1,219	10,696	10.93	976
Free Ind	236,011	6.30	6,887	7,005	3.85	205	3,002	13,09	98

Table 1: The occurrences of each form of STWR in the training, validation and test corpora given in tokens, percentage of tokens in the respective corpus, and instances.

ported; 700,202 tokens for indirect; 3,804,226 tokens for free indirect) and a validation corpus (97,316 tokens for direct, reported and indirect; 181.942 tokens for free indirect). Table 1 shows the occurrences of each form of STWR in its training and validation corpus, given in tokens, percentage of tokens in its corpus and instances[2].

4.3 Test data for the recognizers

Our test data for the direct, indirect and reported STWR recognizers has 97,863 tokens and comprises excerpts from historical fictional and nonfictional texts in equal proportions. They were labeled with a consensus annotation as described in section 4.2. The test data for the free indirect STWR recognizer has 22,935 tokens and comprises 22 excerpts from dime novels, which were manually labeled by one human annotator. Table 1 shows the occurrences of each form of STWR in its test corpus, given in tokens, percentage of test corpus tokens and instances.

5 Results

We report the scores of our most successful language embedding configurations, i.e. fine-tuned BERT and fastText stacked with FLAIR forwards and backwards.

Notably, our fine-tuned BERT model performed better than the regular BERT model for all STWR types, even free indirect, where the STWR recognizers were tested on modern German fiction (as opposed to historical German fiction and nonfiction for the other models). The same is true for the custom-trained fastText + FLAIR models which outperformed models pretrainend on modern German for all STWR types as well.

We speculate that this is because the customization made the models better suited for literary texts in general, even though it was done on historical German.

The most successful configuration of fastText + FLAIR varies slightly between the different forms of STWR with respect to the fastText model that gave the best results. The fastText specifications for the four types of STWR are detailed in table 2.

	Dimensions	Training data
Direct	500	fastTextTrain_clean
Indirect	300	fastTextTrain
Reported	500	fastTextTrain_clean
Free ind	300	fastTextTrain

Table 2: Dimensions and training data for the fastText models used by the different FLAIR based recognizers.

We trained each model with the same configuration for three times to correct for random variation in the deep learning results. Table 3 reports the average value of each score and the standard deviation, calculated on token level.

On average, the recognizers using BERT embeddings scored better for all types of STWR except direct, for which the recognizers of the stacked fastText and FLAIR embeddings proved consistently more successful. Most striking was BERT's advantage for free indirect, where especially the recall improved. It should be noted though, that the FLAIR-based freeIndirect model consistently gave better precision.

However, when looking at the standard deviation over the three runs and the range of results, we see that the F1 score ranges of the FLAIR and BERT recognizers overlap for reported, so the results of the comparison are not conclusive for this STWR type. For the other three STWR types, the F1 score ranges are clearly distinct, even though the free indirect models show a high variance.

Table 4 lists the scores for the individual recognizers from the three training runs that produced

[2] An instance is defined here as an uninterrupted sequence of tokens annotated as the same type of STWR, which can be longer than one sentence. This is of course a simplification, as two conceptually separate stretches, such as lines of dialogue by two different people, will be counted as one instance if they follow directly after each other, but can serve as a rough guideline. On average, a direct instance is 46 tokens long, an indirect instance 15 tokens, a reported instance 10 tokens and a free indirect instance 34 tokens.

	fastText + FLAIR						BERT					
	F1		Prec		Rec		F1		Prec		Rec	
Dir	**0.84**	(0.0047)	**0.90**	(0.0245)	**0.79**	(0.0094)	0.80	(0.0047)	0.86	(0.017)	0.74	(0.0082)
Ind	0.73	(0.0082)	0.78	(0.0082)	0.68	(0.0205)	**0.76**	(0.0)	**0.79**	(0.0236)	**0.73**	(0.017)
Rep	0.56	(0.0125)	0.68	(0.0094)	0.48	(0.0141)	**0.58**	(0.017)	**0.69**	(0.0163)	**0.51**	(0.034)
Fr ind	0.49	(0.017)	**0.86**	(0.0094)	0.35	(0.0125)	**0.57**	(0.0216)	0.80	(0.017)	**0.44**	(0.0309)

Table 3: Average scores over three runs for each form of STWR, standard deviation given in brackets. Best average scores are bolded.

the best results.

	F1	Prec	Rec	Embedding
Direct	0.85	0.93	0.78	cust. fastText+FLAIR
Indirect	0.76	0.81	0.71	BERT fine-tuned
Reported	0.60	0.67	0.54	BERT fine-tuned
Free ind	0.59	0.78	0.47	BERT fine-tuned

Table 4: Scores of our top models

To give an impression how difficult it is for humans to annotate these forms, table 5 presents the agreement scores between human annotators. The scores for direct, indirect and reported STWR are based on corpus REDEWIEDERGABE, the corpus of fictional and non-fictional historical texts our test data was drawn from. The score for free indirect was calculated directly on the free indirect test corpus.

	F1	Prec	Rec	Fleiss' Kappa
Direct	0.94	0.94	0.94	0.92
Indirect	0.75	0.77	0.74	0.73
Reported	0.56	0.56	0.56	0.49
Free ind	0.69	0.64	0.73	0.66

Table 5: Human annotator agreement for the STWR types.

We performed two types of error analysis: First, we looked at the first 10,000 tokens of our test data and categorized the types of errors made by our top recognizers (cf. table 4). This gives an impression of the types of challenges the four forms of STWR pose and how well our recognizers can deal with them which is important practical information for anyone using them.

Second, we also looked at the first 20 differences between the results of the best models trained with BERT vs. the best models trained with FLAIR. The goal was to find indicators which specific properties of the two different contextual embeddings made them better or worse suited to a particular task.

As the four types of STWR have very different characteristics, we will discuss each of them separately.

5.1 Direct STWR

Direct STWR has two main characteristics: First, being a quotation of the character's voice, it tends to use first and second person pronouns and present tense. Second, it is often marked with quotation marks, but the reliability of this particular indicator varies dramatically between different texts. Its instances can also be very long, spanning multiple sentences. We observed that about half of the false positives as well as the false negatives are partial matches, i.e. the recognizer did correctly identify a stretch of direct STWR, but either broke of too early or extended it too far.

A main cause for false negatives were missing quotation marks, i.e. unmarked stretches of direct STWR, especially if those occured in first person narration. In these cases, the recognizer is missing its two most reliable indicators to distinguish direct STWR from narrator text at the same time. Another source of false negatives are very long stretches of direct STWR, such as embedded narratives. The recognizer looses the wider context and tends to treat this STWR as narrator text, especially if it contains nested direct STWR and exhibits characteristics such as third person pronouns and past tense.

The main source of false positives is also related to narrative perspective: In a first person narration or a letter, the recognizer tends to annotate narrator text as direct STWR – the reverse problem to the one described above. Note that these cases are very hard for human annotators as well and can only be solved by knowing a wide context. The recognizer knows a context of 100 tokens maximum and we observed that wrong decisions often occur at the beginning of a context chunk and are then propagated to its end. Another source of false positives are – predictably – stretches of texts in quotation marks that are not direct STWR, though these are a relatively rare occurance. We also observed mix-ups with the forms indirect and free indirect STWR, especially if unsual punctuation was used, though this was rare as well.

In summary, we can say that for direct STWR narrative perspective is a major factor. The test material was deliberately designed to contain texts written both in first and third person perspective. If evaluated separately, we could observe a significantly better performance for third person perspective (see table 6).[3]

	F1	Prec	Rec
First person	0.80	0.86	0.75
Third person	0.87	0.97	0.79

Table 6: Evaluation for the direct recognizer (top model, FLAIR based) split into texts with first and third person perspective

Direct STWR is the only type of STWR where FLAIR embeddings performed better than BERT with a clear advantage. Looking at the first 20 differences between the recognizers, we found that BERT is more prone to annotate letters and first person perspective narratives as direct STWR. It also breaks off prematurely more often, indicating that FLAIR seems to be better in maintaining the context of the annotation. On the other hand, FLAIR tends to make more minor mistakes, such as not annotating a dash when it is used instead of a quotation mark to introduce direct STWR. This points to the more character-based behaviour of FLAIR which – in general – seems to serve well for direct STWR, maybe because of the prevalence of typographical indicators. The wider context of the BERT embeddings does not seem to help with the perspective problem, but instead introduces additional errors.

5.2 Indirect STWR

Indirect representation in our definition takes the form of a subordinate clause or an infinitive phrase, dependent on a framing clause which is not part of the STWR itself. Thus, instances of indirect STWR are always shorter than one sentence. Of the four STWR forms, it is the one that is most strongly defined by its syntactical form.

One difficulty are cases where the indirect STWR contains subclauses or, conversely, is followed by a subclause that is not part of the instance. In these structures, the recognizer tends to have trouble identifying the correct borders of the STWR. When looking at the error analysis, about one third of the errors for both false positives and false negatives are partial matches, mostly caused by this problem.

The biggest cause for errors are cases where the typical indirect structure – a subclause starting with *dass, ob* (*that, whether*) or an interrogative prounoun – is paired with an unusual frame. This leads to false positives, if the frame contains words that usually indicate STWR, such as *es scheint außer Frage, dass ...* (*it seems out of the question that ...*). Though this phrase does not introduce STWR, the word *Frage* (*question*) still triggers an annotation. On the flipside, cases of indirect STWR tend to be missed if they are introduced by phrases that have an unsual structure and don't contain words that are strongly associated with speech, thought or writing acts. We also observed that unusual punctuation, such as dashes, multiple dots and colon (used instead of comma at the border of an indirect STWR), have negative effects on recognition accuracy.

In a comparison between the indirect models using BERT and FLAIR embeddings, we observed that both models make errors of the types described above, though at different places. However, overall FLAIR seems more susceptible to interference in the form of unusual punctuation or framing phrases that are interjected in the middle of a stretch of indirect STWR. It is also less successful than BERT in recognizing STWR instances that are introduced with nouns instead of verbs.

5.3 Reported STWR

Reported STWR is a fairly difficult form even for human annotators, mainly because it is so similar to pure narration that it can be hard to distinguish. It should be noted that the gold standard annotation in this case contains a number of uncertain instances that could be debatable for humans as well. Reported instances tend to be rather short, varying from one token to one sentence at most, and syntactically diverse. The most reliable indicators are words referring to speech, thought and writing acts.

Only about a fifth of the false negatives and false positives observed for reported STWR were partial matches, a significantly lower percentage than for direct and indirect STWR. This indicates

[3] We experimented with training two specialized direct models, one only using texts with first person perspective and one only texts with third person perspective as training material, and evaluated them on the matching types of texts. Unfortunately the performance was worse than that of the model trained on the complete training corpus, probably because of the significant reduction of training material.

that for this form, finding the correct borders of the annotation is less of a problem than deciding whether STWR is present at all.

Most errors can be attributed to problems related to speech, thought or writing words, the main indicator for reported STWR. Such words can trigger a false annotation and are the main cause of false positives. The reverse problem is even more prominent: Instances that do not use lexical material commonly associated with speech, thought and writing tend to be overlooked. Missing such unusual instances is the main problem of the recognizer and though the direct and indirect recognizers also have better precision than recall scores, the difference for reported is clearly more pronounced. Another recurring error type is that the borders of the STWR were not detected correctly, missing modifiers or annotating part of the surrounding narration. We also observed some rare mixups with indirect STWR.

As noted above, the F1 score ranges of the FLAIR and BERT based recognizers are not distinct for reported, though BERT does perform somewhat better on average. Looking at the differences, we found that the recognizers make the same types of mistakes, but BERT is generally more open to unusal instances of reported STWR, leading to a better recall, which is the main reason for its better overall performance.

5.4 Free indirect STWR

Free indirect STWR is structurally similar to direct STWR in that it usually spans one or more consecutive sentences. It is very hard to identify using surface markers, as it is basically a shift to a characters internal thoughts, but still uses the same tense and pronouns as the surrounding narration. The best indicators are emphatic punctation such as *?, !,* -, words indicating a reference point in the present (such as *now, here*) and characteristics of informal speech such as dialect or modal particles.

The free indirect recognizers show the largest gap between precision and recall: nearly 0.4 points. Clearly, the problem here lies in undetected cases. Notably however, over 40% of the false negatives are partial matches, meaning that the recognizer at least correctly detected an instance of free indirect, though it failed to capture it completely.[4]

The main cause for false positives are cases in which some of the main indicators of free indirect (as described above) occur in narration. In addition to that, unmarked direct STWR is prone to be labeled as free indirect. As for the false negatives, about half of the missed instances contained at least one surface marker, but many are only recognizable via wider context clues or an understanding of the content.

Comparing BERT and FLAIR again, we find that BERT gives a much better recall – the same effect as with reported STWR, but more pronounced. BERT is clearly better in picking up subtle signals for free indirect STWR than FLAIR. The flip-side of this is that the BERT model also produces more false positives than the FLAIR model. An interesting observation is that it sometimes annotates sentences that are not part of the free indirect STWR itself, but introduce it. Though the borders of the STWR are not detected correctly in these cases, this might indicate that the model learned that these context clues are highly relevant to identify free indirect STWR which is indeed the case. An example for this scenario is the following passage:

Jetzt war er mit dem Lächeln an der Reihe. *Ihre Reaktionen kamen so spontan und waren so ungekünstelt und ehrlich. Hoffentlich würde sie das nie verlieren.* (Now it was his turn to smile. *Her reactions came so spontanous and were so genuine and honest. Hopefully she would never loose that.*)

BERT also marks the introductory sentence (underlined) that shifts the focus to the character to introduce the free indirect instance (in italics) that tells us his thoughts. The FLAIR model on the other hand has its strength in precision: The few false positives that it produced are often borderline cases that are attached to free indirect passages and could be read as plausible extensions.

6 Conclusion

We presented recognizers for four types of STWR which differ strongly in structure and difficulty. Our models for direct, indirect and reported were trained and tested on historical German fictional and non-fictional texts, the model for free indirect on modern German fiction. The success rates

[4]The recall problem might be exacerbated by the false negatives in the training data. We ran tests where we cut out the marked instances in the training data with some context (25 or 50 tokens) and used this as training input. A detailed evaluation is beyond the scope of this paper, but the resulting recognizers had better recall but worse precision, leading to similar F1 scores.

correspond closely to the reliability of humans: For indirect and reported, we even achieved similar scores to the human annotator agreement on a comparable corpus. For the types direct and free indirect, humans still clearly outperform our best models. In both cases, we believe that the need for wide contextual knowledge plays an important role to explain the gap: For direct, the models fail most often in distinguishing between a first person narrator and a character quote. Free indirect in general is a highly context dependent form that requires an understanding of the narrative structure.

We tested a variety of different language embeddings for our task and provided a comparison of the most promising: FLAIR and BERT embeddings. For both, we also trained/fine-tuned models on historical texts. FLAIR gave the best scores for direct, BERT for indirect and free indirect. For reported, the results were not conclusive: Though BERT performed better on average, we observed an overlap in F1 score range of the BERT and FLAIR models over multiple runs.

Most striking was the improvement achieved with BERT for free indirect STWR. In particular, BERT improved recall for the most difficult forms, free indirect and – to a lesser degree – reported, showing a greater ability to detect unusual instances.

Direct STWR was the only form where FLAIR clearly outperformed BERT. It seems like the higher sensitivity of BERT is more of a disadvantage here, as it tended to misclassify even more instances of first person narration than FLAIR.

To further improve performance, one idea is modifying our input strategy: instead of consecutive chunks of up to 100 tokens, overlapping chunks could be used as input. This might prevent the recognizers from loosing context at the beginning of a chunk, which would be especially relevant for the direct and free indirect recognizer.

The top models and customized embeddings described in this paper are freely available via our homepage *www.redewiedergabe.de* and via GitHub. In detail, our customized BERT embeddings can be found at *huggingface.co/redewiedergabe/bert-base-historical-german-rw-cased*, the custom-trained FLAIR embeddings are integrated into the FLAIR framework as *de-historic-rw-forward* and *de-historic-rw-backward*. The top recognizer models are available at *github.com/redewiedergabe/tagger* along with the code used for training and execution.

In addition to that, all the material used for the direct, indirect and reported recognizers and part of the material used for the free indirect recognizer[5] is available as corpus REDEWIEDERGABE (Brunner et al., 2020) at *github.com/redewiedergabe/corpus*. The rich annotation of corpus REDEWIEDERGABE also offers opportunities to train more complex recognizers, e.g. by providing labels for the medium of the STWR (speech, thought or writing) as well as annotation for the framing phrase for direct and indirect STWR and the speaker for all four forms of STWR.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ann Banfield. 1982. *Unspeakable sentences. Narration and representation in the language of fiction*. Routledge & Kegan Paul, Boston u.a.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR*, abs/1607.04606.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: An NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. *North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 42–47.

Annelen Brunner. 2015. *Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie*. Number 47 in Narratologia. de Gruyter, Berlin u.a.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020.

[5] Unfortunately we can only publish the historical part of the free indirect material due to copyright restrictions on the modern texts.

Corpus REDEWIEDERGABE. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 796–805, Marseille, France. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2019. Deep learning for free indirect representation. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Short Papers*, pages 241–245, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Cathrine Fabricius-Hansen, Kåre Solfjeld, and Anneliese Pitz. 2018. *Der Konjunktiv: Formen und Spielräume*. Number 100 in Stauffenburg Linguistik. Stauffenburg, Tübingen.

Monika Fludernik. 1993. *The fictions of language and the languages of fiction. The linguisitic representation of speech and consciousness*. Routledge, London/New York.

Cláudia Freitas, Bianca Freitas, and Diana Santos. 2016. QUEMDISSE? Reported Speech in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) – Book of abstracts*, pages 4410–4416.

Gérard Genette. 2010. *Die Erzählung*, 3 edition. Number 8083 in UTB. Fink, Paderborn.

Stefan Hauser. 2008. Beobachtungen zur Redewiedergabe in der Tagespresse. Eine kontrastive Analyse. In Heinz-Helmut Lüger and Hartmut Lenk, editors, *Kontrastive Medienlinguistik*, pages 271–286. Verlag Empirische Pädagogik, Landau.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR*, abs/1508.01991.

Fotis Jannidis, Albin Zehe, Leonard Konle, Andreas Hotho, and Markus Krug. 2018. Analysing Direct Speech in German Novels. In *Digital Humanities im deutschsprachigen Raum – Book of abstracts*.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. 2019. A Comparison of Word-based and Context-based Representations for Classification Problems in Health Informatics. In *Proceedings of the BioNLP 2019 workshop*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Konstantina Lazaridou, Ralf Krestel, and Felix Naumann. 2017. Identifying Media Bias by Analyzing Reported Speech. In *IEEE International Conference on Data Mining – Book of abstracts*, pages 943–948.

Jinhyuk Lee, Wonjin Yoon, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, Chan Ho So, and Jaewoo Kang. 2019. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*.

Geoffrey Leech and Mick Short. 2013. *Style in fiction. A linguistic introduction to English fictional prose*, 2 edition. Routledge, London u.a.

Brian McHale. 2014. Speech Representation. In Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert, editors, *The living handbook of narratology*. Hamburg University Press, Hamburg.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*.

Sean Papay and Sebastian Padó. 2019. Quotation detection and classification with a corpus-agnostic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 888–894, Varna, Bulgaria. INCOMA Ltd.

Roy Pascal. 1977. *The dual voice. Free indirect speech and its functioning in the nineteenth-century European novel*. Manchester University Press, Manchester.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Christof Schöch, Daniel Schlör, Stefanie Popp, Annelen Brunner, and José Calvo Tello. 2016. Straight talk! Automatic Recognition of Direct Speech in Nineteenth-century French Novels. In *Conference Abstracts*, pages 346–353, Jagiellonian University & Pedagogical University, Kraków.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model Architectures for Quotation Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) – Book of abstracts*, pages 1736–1745.

Luise Schricker, Manfred Stede, and Peer Trilcke. 2019. Extraction and Classification of Speech, Thought, and Writing in German Narrative Texts. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 183–192, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Shreyas Sharma and Ron Daniel. 2019. *BioFLAIR: Pretrained Pooled Contextualized Embeddings for Biomedical Sequence Labeling Tasks*.

Milan Straka, Jana Straková, and Jan Hajic. 2019. *Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ngoc Duyen Tanja Tu, Markus Krug, and Annelen Brunner. 2019. Automatic recognition of direct speech without quotation marks. A rule-based approach. In *Digital Humanities: multimedial & multimodal. Konferenzabstracts*, pages 87–89, Frankfurt am Main/Mainz.

Harald Weinrich. 2007. *Textgrammatik der deutschen Sprache*, 4., rev. aufl edition. Wiss. Buchges, Darmstadt.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Gisela Zifonun, Ludger Hoffmann, and Bruno Strecker. 1997. *Grammatik der deutschen Sprache*, volume 3 of *Schriften des Instituts für deutsche Sprache*. de Gruyter, Berlin/New York/Amsterdam.