Originally published in: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. Pp. 241-245

Deep learning for Free Indirect 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

In this paper, we present our work-in-progress to automatically identify free indirect representation (FI), a type of thought representation used in literary texts. With a deep learning approach using contextual string embeddings, we achieve f1 scores between 0.45 and 0.5 (sentence-based evaluation for the FI category) on two very different German corpora, a clear improvement on earlier attempts for this task. We show how consistently marked direct speech can help in this task. In our evaluation, we also consider human inter-annotator scores and thus address measures of certainty for this difficult phenomenon.

1 Introduction

In contrast to the well-known direct or indirect representation of speech, thought and writing, there have been hardly any attempts to tackle the automatic recognition of free indirect representation (FI) up until now. FI – in German also known as "Erlebte Rede" – is mainly used in literary texts to represent a character's thoughts while still maintaining characteristics of the narrator's voice. In the following example, the part in italics is FI:

> Er glaubte, sie zu kennen. *War das nicht die Grünkramfritzen von der Ecke?* [He thought he knew her. *Wasn't that the greengrocer gal from the corner?*]

While the third person pronouns and the past tense indicate the narrator's voice, the use of a question and the informal language makes the presentation similar to a direct quotation of the character's thoughts. FI has been a much discussed topic in literary theory since the early 20th century (overview in McHale (2014)). In our approach we follow the 'classical' definition of FI (e.g. Genette (2010), Leech and Short (2013)) that focusses on the representation of characters' thought processes. When a personal, character-focussed style gradually became mainstream over the last century, FI became a common narrative tool which today appears even in popular and children's literature. For quantitative studies of literary style, it would be highly useful to be able to detect the usage of FI automatically. However, FI is very context dependent, essentially a shift in perspective to the character, which can be hard to detect even for humans. In this brief presentation of our work-in-progress in this area, we will show preliminary results with a deep learning approach which we will contrast with a simple rule-based approach as well as a RandomForest approach from earlier research. We will also consider human inter-annotator agreement in our evaluation.

2 Related Work

The only attempt of automatic detection of FI in German texts known to us is the work by Brunner (2015). She implemented a simple rule-based algorithm and also trained a RandomForest model. On a corpus of 13 short German narratives (57,000 tokens) from the late 18th to early 20th century, she reports a sentence-based f1 score for the category FI (as opposed to non-FI) of 0.31 (rule-based) and 0.4 (RandomForest, 10-fold cross validation). We will compare our results to Brunner's in section 5.2.

With respect to automatization, we consider the detection of FI a sequence labelling task. In this area, much progress has been made recently employing deep learning and language embeddings. We use FLAIR (Akbik et al., 2019), a PyTorch-based framework that facilitates the use of language embeddings and model training for NLP tasks. The architecture of our deep learning model is adapted from Akbik et al. (2018). They propose 'contextual string embeddings', an approach which passes sen-

tences as sequences of characters into a character-level language model to form word-level embeddings. This approach achieves significant improvements for NER, especially for German, and state-of-the-art results for chunking and POS tagging and can be considered one of the leading architectures for sequence labelling tasks to date. We will detail the exact configuration of our model in section 4.2.

3 Training data

As mentioned above, FI became much more common in modern times. For this reason we decided to use mainly modern popular literature for our test and training data.

The bulk of our training data comprises dime novels as well as popular crime novels (full texts or excerpts). This data was preprocessed with a basic rule-based FI recognizer (description see section 4.1). The automatically detected instances were then presented to human annotators who could either dismiss or accept them. The annotators were also instructed to annotate any additional cases of FI in the direct vicinity of the automatically detected instances. This sped up the annotation process considerably, but of course also created a bias, as it is quite possible that valid instances of FI that were never detected by the rule-based recognizer have been missed. This material was supplemented by 150 instances[1] of FI with little to no context, manually extracted from 20th century novels. For model training, it was split into a training corpus (1,443,811 tokens with 5.46% FI, 2551 instances) and a validation corpus (181,916 tokens with 3.85% FI, 205 instances).

4 Automatic approaches

4.1 Rule-based recognizer

For the rule-based FI annotation, the text is preprocessed using OpenNLPSentenceDetector, OpenNLPTokenizer (https://opennlp.apache.org), MateLemmatize (Björkelund et al., 2010) and TreeTagger (Schmid, 1995) with the STTS tagset (Schiller et al., 1999). Sentences that contain direct speech (as identified by a simple approach matching quotation marks) are skipped, as it is relatively unlikely for them to contain FI and they exhibit many similarities to FI at the same time. The remaining sentences are categorized as FI if they contain any typical FI indicators, e.g. typographical markers like *!* ... *?* –, temporal markers indicating the present as a reference point (*gestern [yesterday], heute [today], morgen [tomorrow], jetzt [now]*), forms of *würde [would]*) which are commonly used to refer to the future in FI, or the STTS tags ITJ (interjection) or PTKANT (modal particles). This basic recognizer was mainly used to aid in the generation of training material, but its results will serve as a baseline for our evaluation.

4.2 Deep learning model

For language embedding, we used pre-trained models provided by the FLAIR framework, combining word embeddings with contextual string embeddings as recommended in Akbik et al. (2018) in the following combination: 'de' (fastText word embedding (Bojanowski et al., 2016) with 300 dimensions, trained over Wikipedia), 'german_forward', 'german_backward' (two contextual string embeddings trained with a mixed corpus of web texts, Wikipedia and subtitles).

To train our tagging model, we used FLAIR's SequenceTagger class which implements a BiLSTM-CRF architecture on top of the language embedding (as proposed by Huang et al. (2015)). After initial tests with one bidirectional LSTM layer with hidden size 256 and one CRF layer, we decided to add a second BiLSTM layer (hidden size also 256) on top of the first to account for the complexity of our task. This led to visible improvements in both precision and recall. The latter model was used to create the results presented below.

Some consideration was given to the format in which we present the data to our model. FI has a tendency to appear in blocks of several consecutive sentences and, as explained above, constitutes a shift in narrative perspective. Because of this, it is extremely difficult to identify a single sentence as FI without its context, even for humans. On the other hand, though FI most often comprises at least one sentence, it can also be shorter, if the perspective shift occurs within a sentence. We therefore opted to model the sequence labelling task on token level, but input the data as rather large chunks of up to 100 tokens, which may span several sentences. Note that the chunks can be shorter than this maximum, as they may never cross borders between different texts or cut sentences (unless a sentence is longer than 100 tokens).

[1] An instance is defined as a continuous passage of FI tokens and may span several sentences.

5 Results and discussion

5.1 Evaluation on the dime novel corpus

The test data comprises 22 excerpts from dime novels (romance and horror), each about 1,000 tokens long. These were manually annotated in full by humans. To give justice to the difficult nature of FI, we present two competing annotations: **anno1** was done by a single person who annotated all excerpts; **anno2** was done by two different people, each annotating half of the excerpts.[2] All annotators were trained to recognize FI according to the definition used in our project, but worked independently and did not discuss this annotation. The differences between their results are mostly due to true borderline cases rather than clear mistakes. In table 1 we present the agreement scores between the two human-made annotations followed by an evaluation of our deep learning recognizer. Note that the recognizer scores are the results of one model (as described in section 4.2), compared to four different gold standards: the two different human annotations (anno1 and anno2), the set of cases agreed upon by both anno1 and anno2 (**anno_all**; i.e. cases which can be considered fairly obvious for humans) as well as the set of cases marked by either anno1 or anno2 (**anno_any**; i.e. cases that at least some humans would see as FI). This gives us a better understanding of the performance in relation to human certainty. According to anno_all, the test corpus contains 163 (9%) FI sentences, according to anno_any there are 304 (16%) FI sentences.

In addition to that, we tested for the influence of quotation marks on the results. This is relevant, because by definition, FI has many similarities to direct representation (character specific speech patterns, questions, exclamations etc.). The presence of quotation marks makes it much easier to distinguish between the two forms. We tested our recognizer on one version of our test corpus that lacked any quotation marks and one that used a consistent pattern of quotation marks. The training data marked direct speech in most but not all cases, using a varity of patterns.

Table 1 shows the agreement scores between humans, our rule-based baseline as well as the results of our deep learning model on texts with and without quotation marks. As FI is a mostly sentence-based phenomenon, we calculate the scores on sentences, though the recognition happened on tokens. The (very rare) cases when FI was partially recognized were counted as correct. F1, precision and recall scores are provided for the category FI.

	f1	**prec**	**rec**	**acc**
human agreement				
anno1 vs. 2	0.7	0.73	0.67	0.93
rule-based (baseline)				
anno1	0.37	0.57	0.27	0.88
anno2	0.31	0.45	0.24	0.88
anno_all	0.35	0.42	0.3	0.9
anno_any	0.34	0.61	0.23	0.85
deep learning (data without quotation marks)				
anno1	0.46	0.61	0.37	0.89
anno2	0.46	0.58	0.38	0.89
anno_all	0.46	0.48	0.44	0.91
anno_any	0.46	0.71	0.34	0.87
deep learning (data with consistent quotation marks)				
anno1	0.45	0.78	0.32	0.9
anno2	0.48	0.79	0.35	0.91
anno_all	0.49	0.65	0.39	0.93
anno_any	0.45	0.92	0.3	0.88

Table 1: F1 score, precision, recall (for category FI) and overall accuracy on the dime novel test corpus, calculated over sentences. Results reported with varying gold standards.

The agreement score beween human annotators, f1=0.7 (fleiss kappa=0.66), can serve as an indicator on how much certainty can be expected when identifying FI in general. We can see that our deep learning model clearly outperforms the rule-based baseline, regardless of quotation marks. When quotation marks are added, you can observe a strong increase in precision and some decrease in recall.

In our error analysis we focussed on cases which have not been identified as FI by the model even though both annotations agreed on them ('clear cases') and, in contrast, cases that were identified by the model even though none of the humans considered them FI ('unlikely cases').

Table 2 lists those cases and sorts them into rough categories. In general we see that the recognizer has its weakness in recall much more than in precision. This is especially true if quotation marks are present: Their absence causes the annotator to categorize direct representation as FI. The

[2] As the skill levels of all annotators were similar and we are not interested in the performance of any one annotator, we believe it is valid to treat this annotation as though it was done by one person as well.

missing clear cases (false negatives for anno_all)		
	no quotes	quotes
no indicators	57	50
known indicators	17	10
partial passage	18	39
total	92	99
finding unlikely cases (false positives for anno_any)		
	no quotes	quotes
direct speech	38	0
known indicators	2	6
overlong passage	3	2
total	43	8

Table 2: Error analysis for the two versions of the test data: 'no quotes' = without quotation marks; 'quotes' = with consistently used quotation marks

addition of quotation marks eliminates this problem completely and hardly any 'unlikely' cases remain.

We sorted the error cases into the following categories:

- **no indicators:** isolated sentence with no obvious FI indicators (recognizable only by context); example:[3]

  Sie war glücklich mit dem Resultat, so viel war deutlich. Aber bestimmt nicht halb so glücklich wie er. *Sein Instinkt hatte ihn also nicht getäuscht, sie war perfekt.* [She was happy with the result. But certainly not half as happy as he was. *His instinct had not deceived him, she was perfect.*] (Perspective shift into the head of the man in the last sentence.)

- **known indicators:** isolated sentence which contains known FI indicators; these can be specific surface markers like the ones used by the rule-based recognizer, but also softer indicators like informal speech patterns; example:

  »Auch das noch!« *Nicht, dass es ihn überraschte – er hatte sie von Anfang an gewarnt.* [»Oh no, not that!« *Not that it surprised him – he had warned her from the beginning.*] (Ellipsis in the first sentence part and dash, which is a known surface indicator of FI.)

- **partial passage / overlong passage:** sentence adjacent to a longer passage of FI that is either not annotated or added incorrectly; example:

  Er hatte sie geküsst! Und es war noch herrlicher gewesen, als sie sich erträumt hatte. [*He had kissed her! And it had been even more glorious than she had dreamed.*] (Both sentences are FI. The recognizer detected the first but not the second.)

The cases adjecent to an FI passage were categorized separately, as one can argue that these errors are less grave: The recognizer at least identified that FI is present in this part of the text, but detected the wrong borders for the passage.

It is also heartening that the recognizer only incorrectly labeled sentences as FI that had at least some known FI indicators; example:

> Der Wagen auf der Achterbahn fuhr weiter. Jetzt hatte er den höchsten Punkt der Steigung erreicht. [The car on the roller coaster went on. Now it had reached the highest point of the ascent.] (The second sentence was incorrectly labeled as FI. It contains the surface indicator *jetzt [now]*.)

The biggest issue, both in numbers as well as in gravity, are the missing isolated sentences, especially the cases of context-based FI without clear indicators within the sentence itself.

5.2 Evaluation on the Brunner corpus

We also tested our deep learning model on the corpus used by Brunner (2015).[4] Table 3 shows the evaluation and contrasts them with the results of Brunner's RandomForest model.[5] We also provide the scores of the rule-based recognizer which were extremely poor for this corpus due to a large number of false positives in a text with unmarked dialogue and many false negatives for FI sentences without explicit indicators.

We are happy to see that our model gives comparable scores to the ones for the dime novel corpus even though Brunner's corpus is very different:

[3] The italicized parts of the examples are FI, according to anno_all.

[4] Brunner's corpus and all her annotations are available for download at http://hdl.handle.net/10932/00-027B-9E8A-9300-0B01-E

[5] The scores reported for Brunner's RandomForest model differ slightly from the ones reported in Brunner (2015), as we used a different sentence splitting tool on her corpus for easier comparison.

	f1	prec	rec	acc
rule-based	0.04	0.06	0.03	0.94
Brunner's RF model	0.41	0.61	0.31	0.96
our model	0.52	0.65	0.43	0.96

Table 3: Scores on Brunner's corpus, in comparison to the scores of the rule-based recognizer and of Brunner's RandomForest model; gold standard is Brunner's manual annotation.

It contains historical texts (1787-1913) with only partly modernized spelling and a lot of stylistic variation, while the model was trained almost exclusively on modern popular literature and uses language embeddings generated from modern German. The percentage of FI is also much lower than in the dime novel corpus, only 4.5% (99 FI sentences), and highly skewed towards one text. Still, our model clearly outperforms Brunner's RandomForest model, which was trained on her own corpus (in 10-fold cross-validation). It looks as though the FI characteristics learned by our model are valid for more than one genre and time period. The error analysis for the Brunner corpus showed the same tendencies as for the dime novel corpus.

6 Conclusion and outlook

We presented our deep learning model for FI and evaluated it on two very different corpora with similar results. Though the f1 scores are only in the 0.45 to 0.5 range and there are problems, especially with respect to recall, they clearly outperform a rule-base detection of FI as well as a RandomForest approach. Considering that trained human annotators only achieved an f1 score of 0.7 (fleiss' kappa 0.66), the results are promising. We also showed that the presence of quotation marks for direct representation has a strong effect on precision.

We will continue trying to improve our model. One focus is on training data: Apart from simply adding more data, we plan to add specifically more FI cases without clear surface markers in order to fix our recall problem. We also consider removing long passages without detected FI from our current training data, as due to the semi-automated annotation process those could easily contain valid FI. The second focus is on testing other leading language embeddings for this task, such as BERT (Devlin et al., 2018).

References

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, August. Association for Computational Linguistics.

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, June. Association for Computational Linguistics.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

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

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.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset)*. Institut für Maschinelle Sprachverarbeitung (Universität Stuttgart) / Seminar für Sprachwissenschaft (Universität Tübingen), August.

Helmut Schmid. 1995. Improvements on Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.