Automatic recognition of direct speech without quotation marks. A rule-based approach

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Brunner, Annelen
brunner@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Motivation

Many texts incorporate multiple voices: the voice of the narrator and those of the characters or people who are quoted by the narrator. Separating these quotations from the surrounding text is relevant for many applications: In the field of literary studies it is a requirement for studies concerning character representation like sentiment analysis (e.g. Blessing et al. 2016; Schmidt / Burghardt / Dennerlein 2018) or character networks (e.g. Rydberg-Cox 2011; Dimpel 2018). For non-fictional texts, recognizing quotations is relevant for question answering and similar tasks. As those applications rely on processing a lot of textual material, having a way to automatically detect instances of direct speech (DS) is crucial.

As long as a specific pattern of quotation marks is used consistently this is a trivial task. Unfortunately, this is not necessarily the case: First, there is an astounding number of ways to encode quotation marks and incorrect or inconsistent usage is very common. Mistakes like missing closing quotation marks happen easily and can throw off a parser relying solely on those markers. (Brunner 2015: 180-182) The problem is worse for older texts, where the typographic rules are even less standardized and additional errors can occur in the digitization process. Finally, in literature it is not uncommon that authors deliberately choose not to use any markers for stylistic reasons. Those types of texts are especially common in the Digital Humanities and it is useful to have a tool that is not dependent on the use of quotation marks. In addition to that, the tool presented in this talk is rule-based and thus requires no training material.

Related Work

The detection of DS is usually a pre-processing step for another task so it rarely gets much focus in the respective papers. There are many applications for English that could be cited here, but we will deliberately focus on applications for German.

Pouliquen / Steinberger / Best 2007 develop a tool for automatic quotation detection and speaker attribution in newspaper articles for several languages. They look for the proper name of a public character, followed by a verb associated with speech representation, followed by quotation marks. Due to this strict pattern, their recall is relatively low (76%) but they identified 81.7% correct quotations.

The tool GutenTag, developed by Brooke et al. 2015, implements several NLP techniques to use on the texts of Gutenberg corpus. The GutenTag DS recognition relies solely on quotation marks. Before processing the text, the tool checks which types of quotation marks (single or double quotes) are used in it.

In her study about the automatic recognition of speech representation in German literary texts, Brunner 2015 implements two strategies for DS detection: Her rule-based approach uses quotation marks as well as pattern matching to identify frames (proper name - speech verb - colon/ comma). This leads to some success in texts with unmarked DS. On a corpus of 13 German narrative texts an f-score of 0.84 for the category *direct* is achieved in a sentence-wise evaluation. Brunner also implements a machine learning approach using random forests and features based on POS-tags and markings for typical speech/thought/writing words. In a ten-fold cross validation, this model achieves an f-score of 0.87 for *direct*. If quotation marks are ignored entirely in training and evaluation, the model still achieves an f-score of 0.81.

Jannidis et al. 2018 implemented a recognizer for DS based on deep learning, specifically trained to work without quotation marks. It is trained on a corpus of 300 German fictional texts in which the quotation marks were removed. This recognizer achieves an accuracy of 0.84 in sentence-wise evaluation and 0.90 in token-wise evaluation on the corpus that is called “Gutenberg” in our evaluation.

To our knowledge, there is at the moment no recognizer for DS in German texts that is rule-based and does not use quotation marks.

Data

We evaluated our algorithm on four different and distinct data sets.

1) Gutenberg[1]: 33 short stories, taken from Project Gutenberg

2) DROC_red[2] (Krug et al. 2018): 85 text samples of German novels (1800-1950). Each sample spans at least one chapter.

3) RW[3]: Text samples (1840-1920) with 200-1000 tokens, manually annotated in the project "Redewiedergabe"
3a) RW_fict: 222 fictional text samples
3b) RW_nonfict: 206 non-fictional text samples

Methods

The algorithm tries to detect whether a given token or a given sentence is part of a DS. Currently, it cannot reliably determine the exact borders of individual DS instances. The technique is purely rule-based and does not rely on machine learning or training data of any sort. The following pre-processing steps were performed: a) tokenization with the OpenNLP Tokenizer[4], b) sentence detection with the OpenNLP SentenceDetector[5], c) tagging with the RFTagger[6], d) parsing with the mate dependency parser[7] and e) named entity recognition[8].

The algorithm tries to solve the problem in the following steps:

1) Segment the text into narrative and non-narrative sections (use paragraphs, if available):
If meaningful paragraphs are present in the text, those paragraphs usually tend to represent either narrative sections or dialogue sections. As it is unlikely that a narrative section contains DS at all, this knowledge can be used to adapt the recognizer rules. However if no such paragraphs are available, the algorithm starts by reconstructing surrogate sections, using the concept of "coherence" between narrative perspective and tempus. It is determined whether the sentence contains first/second or third person pronouns and which tense is used. If consecutive sentences agree on both, they belong to the same segment. If the narrative perspective is in third person and the dominant tense is past, the segment is categorized as 'narrative'. Inside those sections a penalty is introduced, so that more than a single weak indicator has to be found to assume DS.

2) Determine sentences containing DS with high probability: After the sections have been introduced, each sentence is classified as either "directspeech" or "other". The algorithm utilizes a scoring mechanism with manually defined features (e.g. imperative mode, interjections, tempus shift). These features (as well as their weights) were created by inspecting parts of the DROC corpus. They are optimized to identify DS with high precision. The intuition of this pass is that all sentences which are rather "obvious" (at least to the human reader) are now correctly marked as DS.

3) Use the sentences from step 2) as anchors to expand the annotation: The next pass expands the instances within their previously detected coherent sections. It starts at an anchor sentence and adds the adjacent sentences to the "directspeech" annotation if tempus and narrative perspective still agree. The border of a coherent section serves as a definitive stopping point for this process.

The resulting sentences are the final result for an evaluation based on sentence borders.

4) If token-level accuracy is required: Remove sentence parts that are considered frames from the annotation: If the exact span on a token level is required, the sentences are split into sub-sentences (enclosed by commas) and whenever a frame of a DS is detected, this sub-sentence is removed, yielding the final result of the detection.

Evaluation

For evaluation, two performance metrics are applied: 1) Sentence level accuracy # a true positive is achieved by correctly predicting whether DS is contained in the sentence 2) Token level accuracy # each token is evaluated individually.

The results on the different corpora are depicted in the table below. We report micro accuracy values to not favor documents which are much shorter than others.

Corpus	Sentence level accuracy (in %)	Token level accuracy (in %)
DROC_red	80.5	80.4
Gutenberg	85.4	86.8
RW_fict	81.9	81.7
RW_nonfict	60.8	53.4

Both accuracy metrics show very similar results on three corpora. An interesting fact is that the corpus which was used to create the rules (DROC_red) shows the worst results of all three fictional datasets. Gutenberg contains rather schematic narratives which appear to be easier compared to the other data sets. A more fine grained analysis shows that the best document for DROC_red yields a token level accuracy of 99.7% and the worst document an accuracy of 21.5%. The largest gap can be found in the RW_fict dataset with 0% accuracy for the worst and 100% accuracy for the best document. This shows that while in average the results are promising, there are still phenomena that need to be addressed separately.

For the only non-fictional corpus, RW_nonfict, the scores drop by a large margin. This is because the algorithm finds anchors in sections without DS and propagates those incorrectly to the surrounding section. Those incorrect detected anchors are sentence written in first person or rhetorical questions, which are mistaken as DS. This resulted in about 1800 false positives while only 89 sentences of DS were not detected.

Conclusion

We proposed a rule-based approach to detect DS without the help of any quotation markers. Our approach

creates coherent sections which segment the documents. Specialized rules detect DS on the sentence level inside those segments. The annotation is then expanded from those anchor sentences. Post-processing removes frame sub-sentences to get an exact span for the utterance. Our evaluation shows that the results appear stable throughout different datasets in the fictional domain and are comparable to the results achieved in related work. The tool even achieves a higher score compared to Jannidis et al. 2018 on the sentence level. The current algorithm still has issues with non-fictional texts and some types of fictional texts (especially romantic letters and reports written in first person singular) which suggests that it should be extended to detect the type of document in advance in order to classify in a more robust approach across different domains.

Fußnoten

1. http://gutenberg.spiegel.de/
2. https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release
3. www.redewiedergabe.de
4. https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/opennlp/tools/tokenize/Tokenizer.html
5. https://opennlp.apache.org/docs/1.8.2/apidocs/opennlp-tools/opennlp/tools/sentdetect/package -summary.html
6. http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/
7. https://code.google.com/archive/p/mate-tools/downloads
8. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/14333/file/Jannidis_Figurenerkennung_Roman.pdf

Bibliographie

Blessing, Andre / Bockwinkel, Peggy / Reiter, Nils / Willand, Marcus (2016): *"Dramenwerkbank - Automatische Sprachverarbeitung zur Analyse von Figurenrede"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 281-284.

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

Brunner, Annelen (2015): *Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie* (= Narratologia 47). Berlin: De Gruyter.

Dimpel, Friedrich Michael (2018): *"Narratologische Textauszeichnung in Märe und Novelle"*, in: **Bernhart, Toni / Willand, Marcus / Albrecht, Andrea / Richter, Sandra (eds.)**: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: De Gruyter 121-148.

Jannidis, Fotis / Zehe, Albin / Konle, Leonard / Hotho, Andreas / Krug, Markus (2018): *"Analysing Direct Speech in German Novels"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 114-118.

Krug, Markus / Weimer, Lukas / Reger, Isabella / Macharowsky, Luisa / Feldhaus, Stephan / Puppe, Frank / Jannidis, Fotis (2018): *"Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]"*, in: DARIAH-DE Working Papers Nr. 27 http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-2-9 [letzter Zugriff 27. September 2018].

Pouliquen, Bruno / Steinberger Ralf / Best, Clive (2007): *"Automatic Detection of Quotations in Multilingual News"*, in: International Conference 'Recent Advances in Natural Language Processing' – Proceedings 487-492.

Rydberg-Cox, Jeff (2011): *"Social networks and the Language of Greek Tragedy"*, in: Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1 (3): 1-11.

Schmidt, Thomas / Burghardt, Manuel / Dennerlein, Katrin (2018): *""Kann man denn auch nicht lachend sehr ernsthaft sein?" - Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen"*, in: Digital Humanities im deutschsprachigen Raum – Konferenzabstracts 244-249.