Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 100–104
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Detecting Direct Speech in Multilingual Collection of 19th-century Novels

Joanna Byszuk[1], Michał Woźniak[1], Mike Kestemont[2], Albert Leśniak[1], Wojciech Łukasik[1], Artjoms Šeļa[1,3], Maciej Eder[1]
[1]Institute of Polish Language, Polish Academy of Sciences; [2]University of Antwerp; [3]University of Tartu
Mickiewicza 31, 31120 Kraków, Poland; Prinsstraat 13, 2000 Antwerpen, Belgium; Ülikooli 18, 50090 Tartu, Estonia
{joanna.byszuk, michal.wozniak, albert.lesniak, wojciech.lukasik, artjoms.sela, maciej.eder}@ijp.pan.pl
mike.kestemont@uantwerp.be

Abstract

Fictional prose can be broadly divided into narrative and discursive forms with direct speech being central to any discourse representation (alongside indirect reported speech and free indirect discourse). This distinction is crucial in digital literary studies and enables interesting forms of narratological or stylistic analysis. The difficulty of automatically detecting direct speech, however, is currently under-estimated. Rule-based systems that work reasonably well for modern languages struggle with (the lack of) typographical conventions in 19th-century literature. While machine learning approaches to sequence modeling can be applied to solve the task, they typically face a severed skewness in the availability of training material, especially for lesser resourced languages. In this paper, we report the result of a multilingual approach to direct speech detection in a diverse corpus of 19th-century fiction in 9 European languages. The proposed method fine-tunes a transformer architecture with multilingual sentence embedder on a minimal amount of annotated training in each language, and improves performance across languages with ambiguous direct speech marking, in comparison to a carefully constructed regular expression baseline.

Keywords: direct speech recognition, multilingual, 19th century novels, deep learning, transformer, BERT, ELTeC

1. Introduction

Fictional prose can be broadly divided into narrative and discursive forms with direct speech being central to any discourse representation (alongside indirect reported speech and free indirect discourse). This distinction is crucial in digital literary studies and drives various forms of narratological or stylistic analysis: direct, or "mimetic" speech and thought (Gennette, 1980) was used to understand voice of literary characters (Burrows, 1987; Hoover, 2014) and study narrative representations of speech (Conroy, 2014; Katsma, 2014). Distinction between "mimetic" speech and "narration" helped to formalize free indirect discourse, defined as a linguistic mixture of these two types (Brooke, Hammond and Hirst, 2017; Muzny, Algee-Hewitt and Jurafsky, 2017). Sequences of direct exchanges between characters were studied to understand the evolution of dialogue as a literary device (Sobchuk, 2016) and dynamics of "dialogism" over the course of novel's history (Muzny, Algee-Hewitt and Jurafsky, 2017). Direct speech recognition is also closely related to the problem of identification and modeling fictional characters (He, Barbosa and Kondrak, 2013; Bamman, Underwood and Smith, 2014; Vala et al., 2015).

The majority of approaches to direct speech recognition (DSR) in prose remain language-specific and heavily rely on deep morphological and syntactic annotation of texts and depend on typographic conventions of marking direct speech within a given tradition. Rule-based solutions variably use punctuation, contextual heuristics, and morpho-syntactic patterns within clauses to identify direct and indirect speech (Krestel, Bergler and Witte, 2008; Alrahabi, Desclés and Suh, 2010; Brunner, 2013; Brooke, Hammond and Hirst, 2015; Muzny, Algee-Hewitt and Jurafsky, 2017), sometimes relying on external dictionaries of proper names and reporting verbs (Pouliquen, Steinberger and Best, 2007; Nikishina et al., 2019). When DSR does not use quotation marks, it utilizes pre-determined linguistic features – tense, personal pronouns, imperative mode or interjections – to guess speech type (Tu, Krug and Brunner, 2019). Similar assembling of mixed features that might be relevant for direct speech is implemented in supervised machine learning approaches to DSR in two-class classification task (Brunner, 2013; Schöch et al., 2016). Jannidis et al. (2018) constructed a deep-learning pipeline for German that does not rely on manually defined features. It uses simple regular expressions for "weak" labeling of direct speech and then feeds marked text segments to the two-branch LSTM network (one for the "past" and one for the future context of a token) that assigns speech types on a word-to-word basis.

State-of-the-art DSR performance seems to be revolving around 0.9 F1-score with the highest (0.939) for French 19th-century fiction with Random Forests classification (Schöch et al., 2016), 0.87 (Brunner, 2013) or 0.9 (Jannidis et al., 2018) for German novels, 0.85 for Anglophone texts with noisy OCR (Muzny, Algee-Hewitt and Jurafsky, 2017). Despite relatively high performance, all implementations require either a general language-specific models (for tagging corpus and extracting features) or standardized typographic and orthographic conventions, which we cannot expect in historical texts across uneven literary and linguistic landscape. Few attempts to make multilingual DSR used highly conventional modern news texts and benefited from databases specific to the media; at their core these implementations remain a collection of rules adjusted to several selected languages (Pouliquen, Steinberger and Best, 2007; Alrahabi, Desclés and Suh, 2010).

In this paper we propose a multilingual solution for direct speech recognition in historic fictional prose that uses transformer architecture with multilingual sentence embedding and requires minimum amount of "golden standard" annotation.

2. Data

The project was born in relation to Distant Reading for European Literary History (COST Action CA16204) project, and one of its subtasks – direct speech markup. We have therefore focused on the problems as observed in

the corpus created within the project: European Literary Text Collection (ELTeC), which is aimed to consist of "around 2,500 full-text novels in at least 10 different languages" (https://www.distant-reading.net/). Spanning from 1840 to 1920, ELTeC provides a cross-view of literary traditions and typography conventions.

The collection presents a number of challenges due to its historic variation, from typographic and orthographic differences, to old vocabulary, to the status of given languages at the time, with some, most notably Norwegian, undergoing at the time the process of being established as a standardized written language. Another challenge results from the varying origin of the texts in the subcollections – some were contributed from existing open-source collections, while others, e.g. Romanian, due to lack of digitized collections in respective languages were scanned, OCR-ed and annotated by the Action members specifically for ELTeC. Detailed information on the process and rules guiding the creation of the corpus can be found on the dedicated website https://distantreading.github.io/sampling_proposal.html.

We use ELTeC as in its first official release in Level 1 encoding (basic XML-TEI compliant annotation of the texts' division into chapters and paragraphs), covering the following languages: English, German, Italian, French, Romanian, Slovene, Norwegian, Portuguese, Serbian. We do not introduce changes in the original texts and select five samples per language of around 10,000 words each, with every sample drawn from a different novel. We use random sampling and preserve information about paragraphs and sentences.

The samples were manually annotated by JB, WŁ and AŠ, with two-fold purpose in mind: 1) they were used to train the model, 2) they were "the golden standard" to compare baseline performance to. At this early stage of the project we did not calculate inter-annotator agreement as in the case of some languages with which only one of us would be familiar the texts were annotated twice by the same person. In the next stage of the project we plan to involve the Action members in providing and verifying annotations, which will allow us to examine the quality of the annotations better.

Language	Paragraphs	Script	Direct speech ratio
English	989	Latin	0.684
French	1394	Latin	0.450
German	987	Latin	0.756
Italian	662	Latin	0.308
Norwegian	979	Latin	0.334
Portuguese	1573	Latin	0.583
Romanian	1522	Latin	0.597
Serbian	1278	Cyrillic	0.572
Slovene	1809	Latin	0.392

Table 1: Sample summaries and direct speech ratio (word level).

3. Method

3.1 Rule-based Approach and Baseline to Evaluate Model

Typographic conventions such as various quotation marks or dashes (see Table 2 below) are strong indicators of the direct speech. Based on them, we have constructed a baseline that relies on regular expressions to extract occurrences of unambiguously marked direct speech. In the languages that use dashes to mark dialogue, the challenge was to separate reporting clauses embedded in a sentence. The results obtained using this baseline were compared with those of manual annotation to assess its performance.

Language	Direct speech conventions
English	" … "
French	— … ; « … » ; « … ; » …
German	» ... «
Italian	— ... ; — ..., —; « ... » ; " ... "
Norwegian	— ... ; « ... »
Portuguese	— ... ; — ..., —
Romanian	— ... ; „ ... "
Serbian	— ... ; — ... —
Slovene	" ... " ; „ ... "

Table 2: Conventions of marking direct speech across languages, as accounted for in the baseline (the above conventions apply to non-normalized ELTeC corpus, but not necessarily to the 19th-century typographic traditions in general).

For many European languages with a high degree of standardization of typographic conventions this approach is extremely effective. For example, in English where the words spoken are enclosed in double quotation marks, narrator's inclusions are easy to identify, therefore the example sentence: *"I see," said Rachel; "it is the same figure, but not the same shaped picture."* may be captured using simple regular expression: (".+?"). Other languages, like French, not only use different symbols for quotations («…»), but also tend to omit them in dialogues for the initial dashes. Despite this, the performance of the rules-based approach decreases only slightly.

Language	Precision	Recall	Accuracy	F1-score
English	0.98	0.99	0.99	0.98
Slovene	0.99	0.97	0.99	0.98
Portuguese	0.95	0.94	0.96	0.94
Romanian	0.90	0.94	0.94	0.92
German	0.99	0.86	0.94	0.92
French	0.92	0.92	0.95	0.92
Italian	0.87	0.88	0.94	0.88
Serbian	0.90	0.85	0.93	0.87
Norwegian	0.72	0.59	0.84	0.65

Table 3: Performance of regular expression baseline in direct speech detection on manually annotated samples.

However, frequently the formal structure of a compound sentence delimited by commas does not allow distinguishing the narration from the direct speech for the baseline. As, for instance, in the sentences *—Et la bonne Rosalie,*

la gouvernante de Frédéric, l'accompagne sans doute! and *—Je ne demanderais pas mieux, dit Albert, en regardant madame Mansley*. With the lack of clear separation of the direct speech, which is often the case for the early 19th-century editions, baseline performance drops substantially: for the German sample without proper marks it achieves 0.68 accuracy and only 0.18 recall (F1 = 0.04).

Other common problems include no clear mark at the end of an utterance, no difference in marking direct speech and proper names, irony, or other pragmatic shifts that introduce subjective perspective, such as characters using metaphorical phrases, e.g. "*little man*" indicating not that the person addressed this way is short, but is treated with less respect by the speaker. These irregularities are the reason behind the decrease in baseline performance, with the worst results for Norwegian.

Deep learning solution that has distributed understanding of the direct speech features in multilingual environment may provide a way to get beyond typographic conventions or language-specific models.

3.2 Adopted Deep Learning Solution

While new developments in deep learning have had a significant impact on numerous natural language processing (NLP) tasks, one solution that has gained increased attention in recent months is BERT (Devlin et al., 2018), or Bidirectional Encoder Representations from Transformers. This new representation model holds a promise of greater efficiency of solving NLP problems where the availability of training data is scarce. Inspired by its developers' proposed examples of studies done on Named Entity Recognition (https://huggingface.co/transformers/examples.html), we adjusted discussed classifying method to work on the data annotated for direct speech utterances.

BERT is based on Transformer architecture, "an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms – an encoder that reads the text input and a decoder that produces a prediction for the task." (Horev, 2018). As learning in BERT happens both in left-to-right and right-to-left contexts, it manages to detect semantic and syntactic relations with greater accuracy than previous approaches. The model is trained on the entire Wikipedia and Book Corpus (a total of ~3,300 million tokens), currently covering 70 languages. The last part was specifically important for our purposes, given that we aimed to provide a solution that could work well across all languages in ELTeC corpus.

Our solution consisted of several steps. First, we sampled five 10,000 word samples per language collection of ELTeC and manually annotated it for direct speech. We followed TEI guidelines annotating spoken and marked thought-out utterances into `<said> </said>` tags. Based on that, we converted our datasets into BERT-accepted column format of token and label (I for direct, O for indirect speech), with spaces marking the end of a paragraph (in alteration to NER solution that divided the text into sentences). Our sample paragraph `<said>»Ich bin derselben Meinung«</said>, rief Benno Tönnchen eifrig.</p>` would thus be turned into:

```
Ich I
bin I
derselben I
Meinung I
, O
rief O
Benno O
Tönnchen O
eifrig O
. O
```

In the next step, we collated our samples together and divided our dataset into train, test, and dev text files, following proportion of 0.8, 0.1, 0.1, ending with ~40,000 tokens per language, and 360,000 or 320,000 tokens total in training data, depending on the test conducted. The number depended on whether we included all languages or conducted a leave-one-out test. To ensure that the model learned a multilingual perspective, we introduced paragraph mixing, so a paragraph in a given language would occur every 8 or 9 paragraphs.

We trained our model with similar parameters as the NER solution we followed, that is with 3 or 2 epochs and batch size of 32. We found that decreasing the number of epochs to 2 improved model performance by 1–2%. We also increased the maximal length of a sequence, due to errors coming from longer sentences in some of the languages.

While we attempted increasing the number of epochs in the training, we realized the model performance was reaching its plateau at 3, pointing to the need to adopt other solutions to further boost its efficiency. We have also tried training on 1/2 and 3/4 of the training dataset, noting that performance drop would only occur when going to half of the training set, again indicating the possibility of having reached plateau, or a need for introducing more variance of conventions when increasing the amount of training data.

4. Results

General model performance is presented in Table 4. Aligning with our intuition, the overall behavior of the multi-language model performs slightly worse than the rule-based approach applied individually to each language.

Loss	Precision	Recall	F1-score
0.306	0.873	0.874	0.873

Table 4: General model performance.

To scrutinize the above intuition, we performed a series of leave-one-out tests, recording the performance of each model with one of the languages being excluded. The results are shown in Table 5. The scores obtained while excluding Norwegian and Italian suggest that in our composite model, some of the less-standardized languages might distort the final results. While this in itself might speak against choosing a multi-language approach, the fact that inclusion of the more-standardized languages in the model improves direct speech recognition for all languages indicates the usefulness of such model for auto-

matic tagging of these parts of multilingual corpora for which regular expression based solutions are not good enough. The difference between the general model and the set of its leave-one-out variants turned out to be minor, leading to a conclusion that the general model exhibits some potential to extract direct speech despite local differences between the languages – suffice to say that the dispersion between the languages in the rule-based approach was much more noticeable.

Excluded language	Loss	Precision	Recall	F1-score
German	0.29	0.89	0.89	0.89
English	0.35	0.87	0.86	0.86
French	0.31	0.87	0.89	0.88
Italian	0.32	0.86	0.90	0.88
Norwegian	0.30	0.89	0.91	0.90
Portuguese	0.33	0.88	0.88	0.88
Romanian	0.30	0.89	0.89	0.89
Slovene	0.34	0.86	0.86	0.86
Serbian	0.40	0.87	0.88	0.89

Table 5: Leave-one-out performance.

Examination of the misclassifications of the model reveal three major sources of errors: narrative structures, size-related uncertainty and noise in pattern-learning. First person narration is often labeled as "direct speech" and linguistically these cases may appear inseparable. This applies not only to a general narrative mode of a novel, but also to the pseudo-documental entries (like letters, diaries) and other "intradiagetic" shifts, with characters becoming narrators. This points to the possible need of using separate DSR models for different narrative modes.

Size of the paragraph seems to influence model's judgement substantially: in longer paragraphs the model expects a mix of direct and indirect clauses (even if the text is homogenous), while one-sentence paragraphs tend to be marked as direct speech. This is in line with findings of Kovaleva et al. (2019) and Clark et al. (2019), showing that attention of BERT is strongly connected to delimiters between BERT input chunks and token alignment within them, as well as sentences across the training data that share similar syntax structure but not semantics. We also observed that many cases that would be easily detected by a rule-based approach are recognized wrongly by BERT-based model: this suggests a certain level of noise in model's decisions (e.g., quotation marks are used for different purposes within the corpus). Abundance of the `[reported clause] -> [reporting clause] -> [reported clause]` pattern also blurs the model and forces it to anticipate this structure.

It is unclear how important are linguistic features of direct and non-direct speech for the model, but errors suggest it pays some attention to imperative mode, personal pronouns, proper names, interjections and verb forms, while heavily relying on punctuation. The last one seems particularly important for misclassifications originating from the expectation that a sentence preceded by a colon or ending with a question or exclamation mark should be classified as direct speech. In a few cases we do not know if the model is wrong or right, because a context of one paragraph could be not enough for a human reader to make a correct judgement.

5. Conclusions

Our project gave us a number of findings in regard to the possibility of developing a uniform solution for direct speech annotation. First of all, we observe that inclusion of languages marking direct speech in more standardized conventions in the model boosts its general performance, improving classification also for literary traditions (or languages) with less regularities in spelling and typography. This is particularly important in the context of corpora such as ELTeC, which gather texts from several languages, including ones that are given relatively little attention in terms of the development of suitable NLP solutions, and present historical variants of the languages, often not well covered in contemporary language representations. It is also important for annotation of texts that feature extensive interjections from other languages, e.g. French dialogue in Polish and Russian novels, a phenomenon common in 19th-century literature involving gentry and bourgeoise characters.

The performance of the model also hints at possible latent imbalances in the corpus which may introduce additional noise and structural problems. In future tests it will be necessary to control the effects of texts coming from first editions (historical language and typographic conventions) and modern reprints (used in some of the ELTeC subcollections); and, while we have not observed significant correlated impact on the results, perhaps also account for language families (Germanic vs. Romance vs. Slavic) and scripts (Cyrillic vs. Latin). The impact of first-person narratives on the instability of the performance also seems to be a factor. Finally, imbalance of "quote"-based and "dash"-based conventions of marking direct speech in the corpus may have introduced additional punctuation-driven noise. Given the above, it is reasonable to attempt conducting experiments with removed direct speech marks altogether, examining the possibility of guiding a model away from the surface-level punctuation features.

Since the transformers-based solution performs better than the baseline in the situations of increased uncertainty and lack of orthographical marks, it is feasible to expect its stable performance also in texts with poor OCR or in historic texts in European languages unseen by the model. These conditions are easily testable in the future.

6. Acknowledgements

The project was launched as a part of a three-year collaborative research project "Deep Learning in Computational Stylistics" between the University of Antwerp and the Institute of Polish Language (Polish Academy of Sciences), supported by Research Foundation of Flanders (FWO) and the Polish Academy of Sciences. JB, ME, AL, AŠ and MW were funded by "Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics" (NCN 2017/26/E/ HS2/01019) project supported by Polish National Science Centre.

7. Supplementary materials

Model and detailed results available at https://gitlab.ijp.pan.pl:11431/public-focs/detecting-direct-speech

8. Bibliographical References

Alrahabi, M., Desclés, J.-P. & Suh J. (2010). Direct Reported Speech in Multilingual Texts: Automatic Annotation and Semantic Categorization. In *Twenty-Third International FLAIRS Conference*. Menlo Park: AAAI Press, pp. 162–167.

Bamman, D., Underwood, T., & Smith N.A. (2014). A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 370–379.

Brooke, J., Hammond, A., & Hirst G. (2015). GutenTag: An NLP-Driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver: Association for Computational Linguistics, pp. 42–47.

Brooke, J., Hammond, A. & Hirst G. (2017). Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. *Digital Scholarship in the Humanities*, 32(2), pp. 234–250.

Brunner, A. (2013). Automatic Recognition of Speech, Thought, and Writing Representation in German Narrative Texts. *Literary and Linguistic Computing*, 28(4), 563-575.

Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels*. Oxford: Clarendon Press.

Conroy, M. (2014). Before the 'Inward Turn': Tracing Represented Thought in the French Novel (1800–1929). *Poetics Today*, 35(1–2), pp. 117–171.

Devlin, J., Chang, M.-W., Lee K., and Toutanova K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805, (http://arxiv.org/abs/1810.04805).

Genette, G. (1980). *Narrative Discourse: An Essay in Method.* Ithaca, NY: Cornell University Press.

He, H., Barbosa, D., & Kondrak, G. (2013). Identification of Speakers in Novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1312–1320.

Hoover, D.L. (2014). The Moonstone and The Coquette : Narrative and Epistolary Style. In D.L. Hoover, J. Culpeper, K. O'Halloran. *Digital Literary Studies*: *Corpus Approaches to Poetry, Prose and Drama*. NY: Routledge, pp. 64–89.

Horev, R. (2018). BERT Explained: State of the art language model for NLP. *Medium*, 17.11.2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Jannidis, F., Zehe, A., Konle, L., Hotho, A., & Krug M. (2018). Analysing Direct Speech in German Novels. In *DHd 2018: Digital Humanities. Konferenzabstracts*, pp. 114–118.

Katsma, Holst. (2014). Loudness in the Novel. *Stanford Literary Lab Pamphlets*, 7.

Krestel, R., Bergler, S., & Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association, pp. 2823–2828.

Muzny, G., Algee-Hewitt M., & Jurafsky D. (2017). Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations. *Digital Scholarship in the Humanities*, 32(suppl. 2), pp. 1131–1152.

Nikishina, I.A., Sokolova I.S., Tikhomirov D.O., and Bonch-Osmolovskaya, A. (2019). Automatic Direct Speech Tagging in Russian prose markup and parser. In *Computational Linguistics and Intellectual Technologies*, 18.

Pouliquen, B., Steinberger R. & Best C. (2007). Automatic Detection of Quotations in Multilingual News. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007)*. Borovets, Bulgaria, pp. 487–492.

Schöch, C., Schlör D., Popp S., Brunner A., Henny U. & Calvo Tello J. (2016). Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. In *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 346–353.

Sobchuk, O. (2016). The Evolution of Dialogues: A Quantitative Study of Russian Novels (1830–1900). *Poetics Today*, 37(1), pp. 137–154.

Tu, N.D.T., Krug, M. & Brunner, A. (2019). Automatic Recognition of Direct Speech without Quotation Marks. A Rule-Based Approach. In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*. Frankfurt am Main, Mainz, pp. 87–89.

Vala, H., Jurgens D., Piper A., & Ruths, D. (2015). Mr. Bennet, His Coachman, and the Archbishop Walk into a Bar but Only One of Them Gets Recognized: On the Difficulty of Detecting Characters in Literary Texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 769–774.

9. Language Resource References

ELTeC (2019). European Literary Text Collection. Distant Reading for European Literary History (COST Action CA16204), https://github.com/COST-ELTeC.