Zero-shot cross-lingual identification of direct speech using distant supervision

Murathan Kurfalı and Mats Wirén
Department of Linguistics
Stockholm University
Stockholm, Sweden
{murathan.kurfali, mats.wiren}@ling.su.se

Abstract

Prose fiction typically consists of passages alternating between the narrator's telling of the story and the characters' direct speech in that story. Detecting direct speech is crucial for the downstream analysis of narrative structure, and may seem easy at first thanks to quotation marks. However, typographical conventions vary across languages, and as a result, almost all approaches to this problem have been monolingual. In contrast, the aim of this paper is to provide a multilingual method for identifying direct speech. To this end, we created a training corpus by using a set of heuristics to automatically find texts where quotation marks appear sufficiently consistently. We then removed the quotation marks and developed a sequence classifier based on multilingual-BERT which classifies each token as belonging to narration or speech. Crucially, by training the classifier with the quotation marks removed, it was forced to learn the linguistic characteristics of direct speech rather than the typography of quotation marks. The results in the zero-shot setting of the proposed model are comparable to the strong supervised baselines, indicating that this is a feasible approach.

1 Introduction

All narratives imply a speaker, that is, someone who tells the story (Bal, 2017). This is the basic narrative mode. But all stories are populated by characters who typically speak to each other, as opposed to telling the story. The characters' direct speech is therefore a distinct narrative mode, though it is embedded into the narration and mediated by the narrator (Koivisto and Nykänen, 2016). Typically, prose fiction consists of passages alternating between these two modes of narrative transmission. There are also hybrid forms, such as indirect discourse and free indirect discourse (Rimmon-Kenan, 2011, Chapter 8), but for the purpose of this work we are only interested in the binary distinction between narration (the telling of the story, in which we here include any hybrid forms) and the direct speech between characters.

Recently, there has been a growing interest in the detection of direct speech, with novel approaches to problems such as distinguishing narration and speech (Jannidis et al., 2018; Ek and Wirén, 2019; Brunner et al., 2020b), keeping track of speakers (He et al., 2013; Muzny et al., 2017; Ek et al., 2018), and keeping track of addressees (Ek et al., 2018). However, the lack of annotated resources has been a hindrance for progress in this field.

It might seem that an effective strategy would be to use quotation marks (single quotes, double quotes, dashes, etc.) for distant supervision, but this is also riddled with difficulties since typographical conventions vary a lot across languages and time periods (Byszuk et al., 2020). In addition, many types of quotation marks are overloaded in the sense of being simultaneously used for other purposes, such as pauses and contractions. Presumably as a result of all this, almost all approaches to identification of direct speech have been monolingual. For example, Quintão (2014) describes identification of direct speech for Portuguese, Pareti (2015) for English, Jannidis et al. (2018) and Brunner et al. (2020b) for German, and Ek and Wirén (2019) for Swedish.

Proceedings of LaTeCH-CLfL 2020, pages 105–111
Barcelona, Spain (Online), December 12, 2020.

In contrast, the only multilingual approach that we are aware of is Byszuk et al. (2020), who collected a corpus from nine languages which they annotated manually for direct speech. Based on this, they developed a classifier for token-level identification of direct speech using BERT (Devlin et al., 2018), reaching an F-score of up to 0.9 in their leave-one-out evaluation at the language level.

Our approach has similarities to that of Byszuk et al. (2020), but, to begin with, we have adopted a low-resource scenario in which we only assume the availability of a collection of raw texts. Using a set of heuristics to automatically detect texts where quotation marks appear sufficiently consistently (similar to Jannidis et al. (2018)), we created a training corpus to be exploited for the purpose of distant supervision. In addition, we fine-tuned multilingual contextual embeddings for the corpus. Furthermore, to prevent the system from relying on the typography of quotation marks, we removed them in the version that was used for fine-tuning, thereby forcing the system to instead learn the linguistic characteristics of direct speech, notably *speech-framing expressions*. These are the cues that narrators provide to allow readers to understand and assess what the characters are saying (and in our work, also who is saying it), consisting of concatenations of a referent, verb and possibly modifiers, such as "Zoe said" or "she inserted drily" (Caballero and Paradis, 2018, pages 45, 53). To retain the relevant information, we had instead labelled each token as belonging to narration, beginning of speech or inside speech in accordance with the IOB format. We then used this corpus to develop a token-level classifier for identification of direct speech in the absence of any quotation marks, using Multilingual-BERT (M-BERT), released by Devlin et al. (2018). The experimental results show that the performance of our classifier is comparable to supervised approaches even in the zero-shot scenario. To the best of our knowledge, this is the first study of zero-shot cross-lingual identification of direct speech.

2 Method

In this section, we describe our method for training a multilingual direct-speech classifier via distant supervision, i.e. without assuming any manually labeled dataset.

2.1 Automatic Extraction of Direct Speech

Relying on quotation marks to detect direct speech leads to a number of challenges due to their typographic variations across texts and their ambiguity between different functions (e.g. dash can be used to signal a dialogue line or to indicate a sharp pause, as in *"I know you are still there – somewhere in the sky."*). However, when used consistently, quotation marks can be exploited to generate large amounts of annotations without requiring any manual work. In the first step of our pipeline, among a large collection of texts, we try to find texts which we can use to extract direct speech with high confidence. That is, contrary to Jannidis et al. (2018), we do not a priori assume the texts in the collection to use quotation marks consistently.

We firstly compile a short list of the most frequent quotation marks which consists of the following markers: «», “ ” , "" , – , — . Then, the texts in the collection are filtered according to the following filtering steps:

- We firstly compute the frequency of each possible quotation mark in our list and filter out the book if one of the marks does not dominantly occur, i.e. does not account for ($> 90\%$) of all the occurrences. The rationale behind this step is to eliminate the cases where several markers may be used interchangeably, hence probably not consistently, which would lead to poor quality annotations.

- We further filter the books based on the ratio of the number of speech tokens to the whole book. Specifically, we count tokens which are between the quotation mark found in the first step and only accept a book if these speech tokens constitutes more than 30% and less than 50% of the whole book, aiming to have a balanced corpus

It is worthwhile to note that we did not include the single quotation mark ’ in our marker list as it is a very frequent punctuation mark used in different constructions, such as contractions. Moreover, as a quotation mark, in many languages it is only used to mark the quotations inside the direct speech (e.g.

	Training			Dev			Test		
	Speech	Narr	All	Speech	Narr	All	Speech	Narr	All
Riqua	81,614	166,684	248,298	7,662	19,389	27,051	26,125	36,691	62,816
SLäNDa	35,192	79,914	115,106	5,795	12,358	18,153	8,262	22,737	30,999
Redewiedergabe	202,499	435,465	637,964	22,990	72,775	95,765	17,084	79,423	96,507
QUAC	11,884	104,464	116,348	1,583	14,937	16,520	4,389	35,213	39,602
Silver data	271,240	403,695	674,935	75,078	85,945	161,023	-	-	-

Table 1: Statistics of how many speech and narration (narr) tokens are in the datasets, excluding the quotation marks.

Sam exclaimed "he said, 'I'll see you at the party'"). Since our annotation schema does not distinguish embedded levels of direct speech, we simply discarded single quotation mark as a direct speech signal.

As the final step, we annotate the remaining books by labeling each token within quotation marks as direct speech, and all the other tokens as narration (*O*). Following the IOB convention, we assign a special label to the first token of each direct speech chunk (*Speech-B*) whereas the remaining speech tokens are annotated with the same label (*Speech-I*). For example the sentence: *["Find any mineral?" asked Cameron, presently.]* is annotated as follows:

"	Find	any	mineral	?	"	asked	Cameron	,	presently	.
-	Speech-B	Speech-I	Speech-I	Speech-I	-	O	O	O	O	O

In line with our aim to train a general classifier which can identify direct speech without relying on any explicit typographic signal, all quotation marks are removed in the final data (thus the label – in the above example).

The collection of texts comes from Project Gutenberg. The collected books are cleaned using the *Gutenberg, dammit*[1] library which is partly based on the GutenTag project (Brooke et al., 2015). Based on the available metadata, we only considered the books which are written in English and listed as a fictional work, excluding the *plays*.

2.2 Classifier

We approach identification of direct speech as a sequence labeling problem. To this end, we use multilingual contextual embeddings, multilingual-BERT (mBERT), and apply the token classification procedure explained in the original paper by Devlin et al. (2018). The model is fine-tuned only on the automatically generated training data, as explained in the previous section.

The way data is presented to the classifier is of special significance. Direct speech and their signals (e.g., *. . . he, then, exclaimed . . .*) tend to span over multiple sentences; hence, accurate identification of direct speech usually requires a larger context than a single sentence. Therefore, we closely follow the configuration of Brunner et al. (2020b) and the text is divided into chunks of sentences provided that there is a maximum of 100 tokens in each chunk and no sentence is cut in half. Only those sentences which are longer than 100 tokens are split into sub-sentences. The sentence boundaries are detected using NLTK's sentence tokenizer (Bird et al., 2009) for all datasets except Redewiedergabe and QUAC which were already sentence tokenized.

3 Experimental Setting

The classifiers are implemented using the Transformers library (Wolf et al., 2019). All classifiers employ mBERT and were fine-tuned for 3 epochs, with a batch size of 32 and a learning rate of 3e-5. In all experiments, the quotation marks were removed from the datasets, thus forcing the model to identify the direct speech from other linguistic cues. We evaluated our classifier on the following manually annotated datasets, each representing a different language:

[1] https://github.com/aparrish/gutenberg-dammit

	RiQuA			SLäNDa			Redewiedergabe			QUAC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Brunner et al. (2020b)	-	-	-	-	-	-	0.87	0.74	0.80	-	-	-
Supervised Baseline	0.87	0.91	0.89	0.79	0.83	0.81	0.70	0.73	0.72	0.73	0.54	0.63
Our System	0.82	0.88	0.85	0.75	0.70	0.73	0.60	0.68	0.64	0.33	0.32	0.33

Table 2: Results on the evaluated datasets in terms of (P)recision, (R)ecall and F-score. The supervised baseline was trained on the training set of each respective dataset after all quotation marks had been removed (uses mBERT). The figures for Brunner et al. (2020b) are obtained using the German BERT model, trained on the whole dataset (including the quotation marks).

- **RiQuA (English):** RiQuA is built on 11 literary 19^{th} century texts and thoroughly annotated for direct and indirect quotation spans, cues, speakers and addressees (Papay and Padó, 2020). Since there is no official training/dev/test split for this dataset, we allocated *The Boscombe Valley Mystery* and the second passage of *Tom Sawyer* for development; *A Christmas Carol, The Lady with the Dog, The Red Headed League* and the second passage of *Emma* for test, and the remaining nine texts for training.
- **SLäNDa (Swedish):** SLäNDa is a recent annotation effort on eight Swedish novels written 1879–1940, including the annotations of speech segments, speech-framing expressions, and speakers (Stymne and Östman, 2020). We used the suggested training/test split and randomly selected four texts from the training data as the development set.
- **Redewiedergabe (German):** Redewiedergabe is the largest avaliable resource for speech, thought and writing representation (ST&WR), with annotation of four main categories: direct, indirect, free indirect and reported speech (Brunner et al., 2020a). In the experiments, we used the version of the corpus described by Brunner et al. (2020b)[2].
- **QUAC (Portuguese):** QUAC is the only corpus in the our experiments which is completely built on non-fiction text collection (Quintão, 2014). Although QUAC covers 403 news in total, only 212 of them are tagged as including direct speech. Hence, we only considered those documents in the experiments, and split the data into documents 0–150, 150–170 and 170–212 as training, development and test sets.

The detailed statistics about these datasets are provided in Table 1. As for our classifier, we prepared a training dataset with 680K words using the method described in Section 2.1, on par with the size of Redewiedergabe which is the largest corpus in our experiments. Due to lack of previous work, we compare our results against the supervised baseline which is obtained by training a separate classifier on the training set of each dataset. It must be noted that this is a very strong baseline as the classifiers are trained on the gold annotations and are monolingual in the sense that they are trained and tested on the same language.

4 Results and Discussion

The experimental results are provided in Table 2. Our model achieves performance competitive with the strong supervised baseline in all datasets except for QUAC by yielding an average F-score of 0.74 against the average F-score of 0.81 of the supervised baselines on these three datasets. Moreover, the performance is stable across different languages although relatively better for the RiQuA corpus which is also in English.

QUAC is the most challenging dataset because it is not only annotated in a zero-shot language, but also represents a non-fiction genre, namely, news. Therefore, unlike the other datasets, QUAC further tests the generalization capability of our classifier to other domains which unfortunately is not sufficient.

[2]https://github.com/redewiedergabe/corpus/blob/master/resources/docs/data_konvens-paper-2020.md

However, we believe that the performance is not surprising given that news tend to adopt very different linguistic cues to signal direct speech (e.g. *according to the chairman, "..."*) than that of fiction texts. Furthermore, QUAC is challenging even in the supervised setting suggested by the relatively lower performance of the supervised baseline.

A qualitative error analysis of 30 random paragraphs from each of the English and Swedish datasets indicated that the most salient characteristic of false negatives was absence of speech-framing expressions. Thus, it seems that narrators' cues to their readers were indeed useful also for the classifier. False positives were more difficult to categorize, but sometimes non-speech verbs generated false alarms, as in Example 1.[3]

(1) **It's time for me to go north,** thought Gurov as he left the platform. **High time!**

Likewise, the classifier sometimes struggled when the narration interrupting the speech extended over several sentences, as in Example 2. On the other hand, the model sometimes recognized very long narrations correctly, as in Example 3.

(2) But he made a dash, and did it: **Is your master at home, my dear?** said Scrooge to the girl. **Nice girl! Very.**

(3) **In everything that made my love of any worth or value in your sight. If this had never been between us,** said the girl, looking mildly, but with steadiness, upon him; **tell me, would you seek me out and try to win me now? Ah, no!**

A common problem was that the system repeatedly switched its binary decisions in the middle of clauses, resulting in spurious fragments, as in Example 4.

(4) **I** recognised as Peter **Jones, the official police agent, while the other was** a long, **thin, sad-faced man, with a very shiny hat and oppressively respectable frock-coat. Ha! Our party is complete,** said Holmes, buttoning up his pea-jacket and taking his heavy hunting crop from the rack.

It should be possible to handle errors of this kind by introducing a syntactic post-processing step which prevents sudden shifts between narration and speech within clauses.

Finally, as a further experiment, we evaluated our model on the indirect speech annotations of Redewiedergabe (Brunner et al., 2020a) to see if training without any quotation marks could help the model to generalize over these implicit cases, as well. However, with a recall of 0.20, it completely failed on this type of speech. Training the classifier on the same language does not help either; the classifier trained on the direct speech annotations of the same data also failed, with a recall of 0.16. Hence, this low performance is not due to the zero-shot configuration, but must be the case because direct-speech classifiers are unable to generalize to other (hybrid) forms of speech, highlighting the need for more studies on the simultaneous identification of these.

5 Conclusion

In the current study, we show that it is possible to perform zero-shot cross-lingual identification of direct speech with a performance comparable to that of supervised baselines. Since quotation marks are not in general reliable indicators of direct speech, we instead devise a set of heuristics for collecting data where they are used sufficiently consistently, and use this to elicit enough data to fine-tune contextual embeddings. As future work, we consider focusing on the cases where our model falls short, namely, generalization to (i) other domains and (ii) other speech forms. Finally, we hope that our study will also be useful in establishing a benchmark to evaluate multilingual direct-speech identification.

Acknowledgement

This work has been partly funded by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, 2019–24; contract no. 2017-00626).

[3] In the examples, the system's predictions are displayed in boldface and the gold standards are underlined.

References

Mieke Bal. 2017. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, Toronto, 4th edition.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Julian Brooke, Adam Hammond, and Graeme Hirst. 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*, pages 42–47.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020a. Corpus REDEWIEDERGABE. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 803–812.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020b. 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. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Joanna Byszuk, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 100–104, Marseille, France. European Language Resources Association (ELRA).

Rosario Caballero and Carita Paradis. 2018. Verbs in speech framing expressions: Comparing English and Spanish. *Journal of Linguistics*, 54(1):45–84.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Adam Ek and Mats Wirén. 2019. Distinguishing narration and speech in prose fiction dialogues. In *Digital Humanities in the Nordic Countries 4th Conference (DHN), Copenhagen, Denmark, March 5-8, 2019*, pages 124–132. CEUR-WS. org.

Adam Ek, Mats Wirén, Robert Östling, Kristina Nilsson Björkenstam, Gintarė Grigonytė, and Sofia Gustafson-Capková. 2018. Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320.

Fotis Jannidis, Leonard Konle, Albin Zehe, Andreas Hotho, and Markus Krug. 2018. Analysing direct speech in German novels. *Konferenzabstracts der DHd 2018. Kritik der Digitalen Vernunft*, pages 114–118.

Aino Koivisto and Elise Nykänen. 2016. Introduction: Approaches to fictional dialogue. *International Journal of Literary Linguistics*, 5.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Silvia Pareti. 2015. Attribution: A Computational Approach. Ph.D. thesis, School of Informatics, University of Edinburgh.

Marta E. Quintão. 2014. Quotation Attribution for Portuguese News Corpora. Master's thesis, Técnico Lisboa/UTL.

Shlomith Rimmon-Kenan. 2011. *Narrative Fiction: Contemporary Poetics*. Routledge, Taylor & Francis Group, London.

Sara Stymne and Carin Östman. 2020. SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 826–834.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.