Distinguishing Narration and Speech in Prose Fiction Dialogues

Adam Ek[†] and Mats Wirén

Stockholm University, SE-106 91 Stockholm, Sweden
{adam.ek, mats.wiren}@ling.su.se

Abstract. This paper presents a supervised method for a novel task, namely, detecting elements of narration in passages of dialogue in prose fiction. The method achieves an F_1-score of 80.8%, exceeding the best baseline by almost 33 percentage points. The purpose of the method is to enable a more fine-grained analysis of fictional dialogue than has previously been possible, and to provide a component for the further analysis of narrative structure in general.

Keywords: Prose fiction · Literary dialogue · Characters' discourse · Narrative structure

1 Introduction

Prose fiction typically consists of passages alternating between two levels of narrative transmission: the narrator's telling of the story to a narratee, and the characters' speaking to each other in that story (mediated by the narrator). As stated in [Dolezel, 1973], quoted in [Jahn, 2017, Section N8.1]: "Every narrative text T is a concatenation and alternation of ND [narrator's discourse] and CD [characters' discourse]". An example of this alternation can be found in August Strindberg's *The Red Room* (1879), with our annotation added to it:

(1) <NARRATOR>
Olle very skilfully made a bag of one of the sheets and stuffed everything into it, while Lundell went on eagerly protesting.
When the parcel was made, Olle took it under his arm, buttoned his ragged coat so as to hide the absence of a waistcoat, and set out on his way to the town.
</NARRATOR>
<CHARACTERS>
– He looks like a thief, said Sellén, watching him from the window with a sly smile. – I hope the police won't interfere with him! – Hurry up, Olle! he shouted after the retreating figure. Buy six French rolls

[†] Currently working at: Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg (adam.ek@gu.se).

and two half-pints of beer if there's anything left after you've bought the paint.
</CHARACTERS>

The work described here is part of a larger effort to develop methods for analysis and annotation of narrative structure in prose fiction. To this end, the two discourse levels require different types of analysis: In narrator's discourse, we are primarily interested in a sequence of events, the ordering of these, and how they form a plot. Although this is true also for characters' discourse, the fact that the narration is expressed through the speech of the characters makes the problem different.[1] In characters' speech, we are interested in what is said (the information conveyed. attitudes, beliefs, sentiments, etc.) and who is speaking to whom. Fortunately, distinguishing narrator's and characters' discourses in prose fiction is relatively straightforward, although the conventions vary between authors, works and printed editions. Typically, devices such as dashes, quotation marks and/or paragraph breaks are used for indicating alternations between the two types of discourse. Furthermore, the problem of identifying speakers and addressees in prose fiction has been dealt with by [Ek et al., 2018], [Muzny et al., 2017] and [He et al., 2013].

There is an additional problem, however: The alternation between narrator's and characters' discourses is not as clearcut as the quotation by Dolozel above may imply, since elements of narration can occur inside characters' discourses as well, interspersed with lines. One case of this is when the narrator indicates who is speaking, and possibly who is being addressed. In the example above, the speaker is indicated in the first and third lines by the speech-framing expressions (utterances that introduce or accompany instances of direct speech in narratives) "said Sellén" and "he shouted", respectively. Furthermore, in the third line the addressee is indicated by a description, "he shouted after the retreating figure". This addressee is also indicated by a vocative, "[Hurry up,] Olle", but that is part of the speech and hence does not belong to the narration.

Another case is when the narrator describes how something is being said or what is happening during the speech: "watching him from the window with a sly smile" in the first line, and "he shouted after the retreating figure" in the second line. In the latter example, "after the retreating figure" also illustrates that these elements of narration may serve more than one function, in this case both indicating who is the addressee and describing his activity.

Given that we want to analyse narrative structure, it is important to distinguish these elements of narration inside characters' discourses for several reasons: First, we need to recognize speech-framing expressions for the purpose of identifying speakers, and vocatives and other constructions for the purpose of identifying addressees. Furthermore, we need to extract other elements of narration related to the quality of the speech or the situation to determine their contributions to the overall plot.

[1] We refer to this interchangeably as "characters' discourse" or "dialogue". For the purpose of this work, we are only concerned with direct speech as opposed to indirect speech, the latter of which we here regard as part of the narration.

To the best of our knowledge, this paper provides the first approach to the problem of distinguishing narration and speech within passages of dialogue in prose fiction.

2 Data

The data consists of excerpts from four novels by Swedish authors (in Swedish): August Strindberg's *The Red Room* (1879), Hjalmar Söderberg's *The Serious Game* (1912), Birger Sjöberg's *The Quartet That Split Up*, part I (1924), and Karin Boye's *Kallocain* (1940). The number of lines and tokens are shown in Table 1. In total, 52.7% of the lines contain narration and 15.6% of all the tokens in the lines belong to narration.

Table 1. Dataset of lines and tokens.

Data	Total	Narration
Lines	1,620	855
Tokens	40,263	6,306

We began by extracting all the passages of dialogue (characters' discourse) in the works, each consisting of one or several lines. By "line" we here mean both the direct speech of the characters and any narration interspersed with this, in such a way that a dialogue passage is completely divided into lines, as exemplified in (3). Each dialogue was annotated by the authors using the opening and closing tags (<NC>, </NC> for "narrative constructions") to demarcate the narration in a line, for example:

(2) – He looks like a thief, <NC> said Sellén, watching him from the window with a sly smile. </NC>

– I hope the police won't interfere with him!

– Hurry up, Olle! <NC> he shouted after the retreating figure. </NC> Buy six French rolls and two half-pints of beer if there's anything left after you've bought the paint.

3 Method

3.1 Task

The problem was regarded as binary classification, where the task was to classify each token in a line as to whether it was an element of narration or not. Put differently, the task can be regarded as narration detection in passages of dialogue. To solve this problem, a logistic regression model was used.

3.2 Features

The data was preprocessed using the Swedish annotation pipeline of `efselab`.[2] The pipeline uses Stagger 3 [Östling, 2013] for part-of-speech tagging and tokenization, and MaltParser [Nivre et al., 2006] for dependency parsing. The final output is a CoNLL-U file[3]. A list of speech verbs was compiled manually by the authors from the corpus.

For each token in a line, a set of features were extracted to predict whether the token was part of narration or not. In addition to capturing the features of the current token, features were extracted from the four succeeding and preceding tokens. The features used are described below and summarized in Table 2.

- **Speech verb:** The token is a speech verb.
- **Lemma:** The lemma form of the token.
- **Part-of-Speech:** Part-of-speech tag of the token.
- **Punctuation:** Token is any punctuation mark.
- **Exclamation/question mark:** Token is an exclamation or question mark.
- **Grammatical features:** The grammatical features of the token.[4]
- **Dependency relation:** Dependency tag of the token.
- **Dependency root:** The token is the head (root) word of the sentence.
- **Sentence ID:** Each sentence in a line is numbered. Based on this, the ID of the sentence that the token belongs to (first, second, etc.).
- **Sentence termination:** The punctuation symbol used to terminate the sentence that the token is part of.
- **Unit ID:** Each line is segmented into numbered units delimited by punctuation marks. For example, the line:

  [– He looks like a thief, said Sellén, watching him from the window with a sly smile.]

  is segmented as follows:

  [– (He looks like a thief,)$_0$ (said Sellén,)$_1$ (watching him from the window with a sly smile.)$_2$]

[2] https://github.com/robertostling/efselab.
[3] https://universaldependencies.org/format.html
[4] For further information, see: https://universaldependencies.org/u/feat/index.html.

Table 2. Summary of the features extracted from the target token and the four preceding and succeeding context tokens. A star (*) indicates that the feature is used only for the target token and not for the context tokens.

ID	Feature	Description
1	Speech verb	The token is a speech verb
2	Lemma	Lemma form of the token
3	Part-of-speech	Part-of-speech tag of the token
4	Punctuation	Part-of-speech is punctuation
5	Exclamation/Question mark	Token is an exclamation or question mark
6	Grammatical features	The grammatical features of the token
7	Dependency relation	Dependency relation of the token
8	Dependency root	Token is the root of the dependency graph
9	Sentence ID*	Sentence id of the token
10	Sentence termination*	Punctuation token used to end the sentence
11	Unit ID*	Unit ID of the token

Punctuation tokens were excluded from the classification, since we did not consider it meaningful to predict whether they belonged to narration or not. They were still used as features for other tokens, however.

3.3 Experimental setup

The logistic regression model was implemented using Python 3 and the `sklearn` package [Pedregosa et al., 2011].[5] The data was split into 5% development and 95% train/test. The model was evaluated using 10-fold cross-validation with 10% of the data used for testing in each fold. To estimate the performance of the logistic regression model, it was compared against two baselines based on speech verbs, described below.[6]

– *SV* → *End*: The first baseline looks for lines which begin with direct speech and end with narration, for example (where boldface indicates narration):

[– He looks like a thief, **said Sellén, watching him from the window with a sly smile.**]

This baseline is found by identifying the first speech verb in the line and then labelling the speech verb and all subsequent tokens as narration.

– *SV* → *Punctuation*: The second baseline is a more specific extension of the first, which additionally captures lines where narration is surrounded by speech, for example:

[5] All resources and code used in this paper are available at https://github.com/adamlek/sv_narration.

[6] As with the model, none of the baselines label punctuation tokens.

[– Hurry up, Olle! **he shouted after the retreating figure.** Buy six French rolls and two half-pints of beer if there's anything left after you've bought the paint.]

This baseline is found by identifying the first speech verb in the line and then labelling the speech verb and all subsequent tokens as narration until a sentence-terminating punctuation mark (`. ! ?`) is encountered.

4 Results

The baselines and the model's performance were measured by calculating the precision, recall, and F_1-score from the token predictions of each line. The model's performance is reported as the average precision, recall and F_1-score from the cross-validation. The results from the baselines and the model are shown in Table 3.

Table 3. Evaluation of precision, recall and F_1-score for the model and baselines with respect to narration detection.

System	Precision	Recall	F_1-score
$SV \rightarrow$ *End*	22.0	56.7	31.9
$SV \rightarrow$ *Punctuation*	50.1	46.0	47.9
Logistic regression	84.2	78.0	80.8

As expected, the $SV \rightarrow$ *End* baseline has the lowest F_1 performance, but a higher recall. $SV \rightarrow$ *Punctuation* performs better in terms of precision and F_1-score. The logistic regression model shows great improvements on all metrics: a gain of 34.1 percentage points in precision, 21.3 percentage points in recall and 32.9 percentage points in F_1-score when compared against the best baselines.

To investigate the influence of the context window size, the model was tested with a window of 0–9 tokens. Figure 1 shows the results of this, which indicate that a context window of 0 to 1 performs poorly in comparison to larger windows. The model's performance stabilizes when the context window is 4 tokens or larger, only showing minor fluctuations thereafter.

In addition to evaluating precision, recall and F_1-score, the logistic regression model and the baselines were evaluated on their ability to find full and partial solutions to complete lines (see Table 4). In a Partial solution, at least one of the tokens belonging to narration in the line are found. Alternatively, if all these tokens are found, but in addition some direct speech tokens are classified as belonging to narration, the line is also regarded as partially correct. In a None solution, the model does not find any narration tokens in a line that contains narration tokens. The columns Full, Partial and None represent all lines that contain narration.

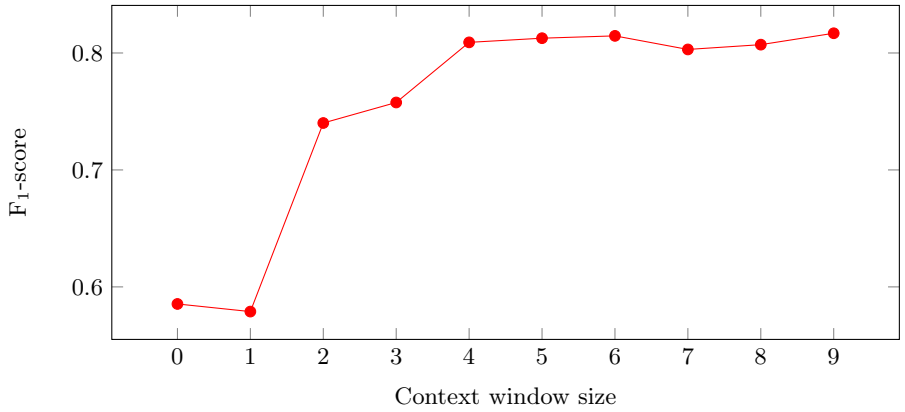

Fig. 1. Influence of context window size on F_1-score.

FP-ERROR is the proportion of lines predicted to have narration tokens but which do not have any narration tokens in the annotated data. The results, shown in Table 4, are based on the average number of solutions obtained using cross-validation.

Table 4. Evaluation of the model's ability to find full, partial and no solutions to dialogue lines (results in percentages). For NONE and FP-ERROR, lower numbers are better.

SYSTEM	FULL	PARTIAL	NONE	FP-ERROR
SV → *End*	23.3	37.7	38.8	13.8
SV → *Punctuation*	42.4	18.1	39.4	13.8
Logistic regression	62.0	34.9	3.0	13.3

Table 4 shows that both baselines have a large number of NONE errors but the *SV* → *End* baseline finds fewer FULL solutions than the *SV* → *Punctuation* baseline. FP-ERROR is the same for both baselines. The model shows major improvements compared to the baselines by finding 19.6 percentage points more FULL solutions, and reducing the number of NONE errors by 36.4 percentage points. However, the model only reduces FP-ERROR by 0.5 percentage points.

5 Discussion

5.1 Model evaluation

The problem appears to be easy at first glance, but the baselines show that the simple rule-based heuristics are unable to capture most cases of narration dialogue. A simple logistic regression model is able to overcome many of the

weaknesses of the baselines and performs well on the task. One of the main strengths of the model is that it is able to detect narration in lines solely based on the tokens in the line, with no other contextual information available. In Figure 1 we examined the influence of context window size on the performance and showed that using more than four context tokens only has minor effects. This indicates that while there may be long-range dependencies, the most important features are captured in a context window of four tokens.

5.2 Speech verbs

Table 4 shows that the model finds solutions to 96.9% (FULL + PARTIAL solutions) of the lines with narration, which is an increase of 36.4 percentage points compared to the $SV \rightarrow$ *Punctuation* baseline. This increase amounts to the number of additional solutions the model found compared to the baselines e.g. the reduction of NONE errors. A weakness of both baselines is their reliance on pre-defined speech verbs. Having a pre-defined list of speech verbs is not realistic when generalizing to unseen books. The results from Table 4 indicate that the model is able to recognize narration beyond the pre-defined list of speech verbs by instead learning these indicators from the data.

5.3 Error analysis

Looking closer at the errors the model makes on partial solutions, one problem is that the model predicts short segments of speech within narrations. For example, in the sentence below (bold indicates tokens tagged as narration):

(3) [– You spoke about an introduction, **said Rissen to the woman without attaching** himself to **me**. How do you get an introduction?]

Most of the narration tokens are correctly classified except the tokens [himself to] which have been classified as speech rather than narration. The model predicts that narration may be interrupted by speech and then continued. This has not been observed in the data and the error arises because the features used capture lexical properties but do not explicitly capture structural information about narration length. This problem could be avoided by including information about the typical length of narration within dialogues and by restricting the number of narration sequences to one per line. In other words, for any line, there would be at most one continuous sequence of narration tokens.

Related to the above problem, the model makes mistakes when a character uses a speech verb in direct speech. This problem could also be avoided by restricting the number of narration sequences to one. However, speech verbs appear to be important for the model and such a solution may prefer tagging only the speech verb as narration rather than the actual sequence of narration tokens. To avoid this, the restriction to one narration sequence should be combined with (a) narration length (e.g., prefer a longer sequence to a shorter one) and/or (b) assigning a lower weight to the speech verb feature.

6 Conclusion and future work

This paper introduces the novel task of narration detection in passages of dialogue in prose fiction, and reports the first results on this using a logistic regression model and data from four Swedish novels. Due to lack of previous research, the model is compared against two baselines and is shown to achieve significant gains over both of them. Most of our features are lexical in nature, but our error analysis indicates that more structural features would be needed in order to further improve the model. In future work, we hope to remedy this and plan to generalize the task to multiple languages. We expect this kind of method to be a valuable component for the purpose of a more fine-grained analysis of fictional dialogue than has previously been possible, and for the further analysis of narrative structure in general.

Acknowledgements

This work has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003). We want to thank Murathan Kurfalı for very valuable comments.

References

Dolezel, 1973. Dolezel, L. (1973). *Narrative Modes in Czech Literature*. University of Toronto Press.

Ek et al., 2018. Ek, A., Wirén, M., Östling, R., Nilsson Björkenstam, K., Grigonytė, G., and Gustafson Capková, S. (2018). Identifying speakers and addressees in dialogues extracted from literary fiction. In *Language Resources and Evaluation Conference, Miyazaki, Japan, 7–12 May 2018*. European Language Resources Association.

He et al., 2013. He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *ACL (1)*, pages 1312–1320.

Jahn, 2017. Jahn, M. (2017). Narratology: A guide to the theory of narrative.

Muzny et al., 2017. Muzny, G., Fang, M., Chang, A., and Jurafsky, D. (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*, volume 1, pages 460–470.

Nivre et al., 2006. Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Östling, 2013. Östling, R. (2013). Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Pedregosa et al., 2011. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.