Borrowing Words: Transfer Learning for Reported Speech Detection in Slovenian News Texts

Zoran Fijavž
Jožef Stefan Postgraduate International School
Peace Institute
Slovenia, Ljubljana
zoran.fijavz@mirovni-institut.si

Abstract

This paper describes the development of a reported speech classifier for Slovenian news texts using transfer learning. Due to a lack of Slovenian training data, multilingual models were trained on English and German reported speech datasets, reaching an F-score of 66.8 on a small manually annotated Slovenian news dataset and a manual error analysis was performed. While the developed model captures many aspects of reported speech, further refinement and annotated data would be needed to reliably predict less frequent instances, such as indirect speech and nominalizations.

Keywords

reported speech, natural language processing, transfer learning, news analysis

1 Introduction

Reported speech allows the rendering of speaker's statements in literary and news texts in which it is used ubiquitously. In the context of computational analysis of said genres, it raises an interesting question about which discourse is reflected in the analysis, as statements by sources may differ significantly from background information provided by writers and journalists. Reported speech uses explicit lexical and syntactic patterns to establish an intertextual link to a source text which should enable precise and robust modeling using natural language processing (NLP) methods. This paper adds to the existing literature by providing a provisional reported speech classifier for Slovenian by using transfer learning and a manual error analysis of its errors.

2 Related Work

2.1 Role of Reported Speech

Reported speech is a ubiquitous language feature in news texts. Broadly, reported speech refers to a reporting some prior utterance for the general purpose of informing the listener what was said in the past and generally understood in a binary between direct and indirect reported speech, with the original utterance repeated verbatim after a reporting verb in the first case, and the embedding of the utterance in a that-clause after a reporting verb in the second case [20] (e.g. *Jimmy said: "Another systematic review would be great!"* versus *Jimmy said that another systematic review would be great.*). However, more complex structures are possible including mixed reported speech (*City officials rebuffed the accusations as "groundless and blatantly false".* and linguistic structures not considered reported speech, which nevertheless fulfill a similar function, such as reportative nominalization (*The speaker particularly emphasized the pressures on the media and the illegal withdrawal of funds.*) [8].

Reported speech features prominently in news reporting: approximately 50% of sentences in newspaper corpora are sourced, primarily through direct and indirect speech, and the share of citations is consistent across texts of different lengths [19]. Corpora data suggests reported speech is most commonly cued by verbs, followed by prepositional phrases and other parts of speech in 96%, 3% and 1% of instances, respectively [14].

Reported speech is more broadly used to lend objectivity to statements [10] and in the context of media reporting to summarize and recreate the source statements [18]. As it signals an explicit intertextual link between a news piece and its sources, it has been productively used in fields, such as critical discourse analysis and communication studies. It may be used to analyze the representation of speakers in a discourse on characteristics, such as gender [1], institutional affiliations of speakers [9] and stances taken within a particular topic [17], or to serve as a variable in broader research topics for distinguishing between the voices of journalists and their sources [12].

2.2 Existing Datasets and Modelling Approaches

Specific annotated datasets exist for different aspects of reported speech. They are primarily based on either literary or news texts with some overlaps. A non-exhaustive list of corpora with some annotation identified in the literature include RiQuA [13], SLäNDa 2.0 [21], Redewiedergabe [3], QUAC [15], PolNeAR [11], Quotebank [23] STOP [24], an annotated Croatian news corpus [16] and a multilingual collection of annotated direct speech [4]. RiQuA and Redewiedergabe (RWG) are the largest manually annotated corpora with both direct and indirect speech annotated. Redewiedergabe consists of a mix of excerpts from German newspapers and fictional works, balanced across the period 1840–1919, while RiQuA is based on sections from 11 19th century English novels. QUAC contains 212 articles published in the 1990s by Público, a Portuguese newspaper with annotations for speakers and direct speech. Quotebank is a corpus comprising 162 million news articles covering the period 2008–2020 with automatic annotations for speakers and direct speech, while also including a manually annotated section for testing. PolNeAR is a corpus containing 1,028 news articles manually labeled with attributions, which are a superset of reported speech and include all quoting, paraphrasing or describing the statements and private states of a third party and primarily used for other tasks, such as event modeling. A summary of the datasets used in this paper can be found in Table 1. As no extensive Slovenian corpus of reported speech exists as of the writing of this paper, we used cross-lingual

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

Table 1: Summary of Datasets' Characteristics.

Corpus	Type	Annotations	Language	Sentence No.	Role	Positive Class
RiQua	fiction	direct and indirect speech, cues, speakers, addressees	English	38,610	72% train, 18% development, 10% test	48%
Redewiedergabe	fiction, news	direct, indirect, free indirect and reported speech, speaker, cues	German	24,033	76% train, 16% development, 9% test	33%
Quotebank (manual)	news	speaker, direct speech	English	9,071	test	30%
QUAC	news	speaker, direct speech	Portuguese	11,007	test	11%
PolNeAR	news	speaker, cues, attributions	English	34,153	test	59%
Slovenian parliamentary news	news	sentence-level binary labels	Slovenian	744	test	43%

transfer learning, which leverages multilingual models, such as mBERT [7] and XLM-R[5], which to embed the training data into overlapping vector spaces for several languages. Models with a narrow selection of languages, such as CroSloEngual BERT tend to outperform model trained on a broad selection of languages [22]. Existing papers approach reported speech by either analyzing particular constituent parts or using end-to-end models. The former include speaker detection [25], quotation detection [19], and speaker resolution across different texts (an entity linking process), or a combination thereof [1]. Reported speech may serve as an informative variable in discourse analysis and its lexical and syntactic regularities make it a promising feature to model. While hardly a novel task, the use of NLP methods in social science research has come under scrutiny for failing to theoretically ground the modeled phenomena or to reliably predict multiple text features, required to draw rich conclusions from texts on par with established methods, such as manual content analysis [2].

3 Experimental Setting

3.1 Task Definition

In this paper we approached reported speech detection as a sentence classification task. While this does not allow for the separate extraction of reported and reporting clauses, it offers advantages. It simplified the data annotation of Slovenian texts for which no extensive annotations are available. Past studies indicate sentence-level classification is warranted by the fact that voices of sources and journalists are mostly clearly delineated in news texts and has been demonstrated to achieve high classification reliability [19]. The latter aspect is particularly relevant for the potential use of models for pre-annotation or downstream research, where predicting a limited set of features reliably may be preferred over end-to-end reported speech classification.

3.2 Training and Test Data

Our experiments were based on existing annotated reported speech datasets and a small annotated Slovenian dataset. The training data consisted of a section from RiQuA and Redewiedergabe, as they are relatively large and include labels for both direct and indirect reported speech. For the training with CroSloEngual Bert, the Redewiedergabe training set was machine translated into English. Testing was done on the test sections of RiQuA, Redewiedergabe, as well as the entirety of the Portuguese newspaper corpus QUAC and the manually annotated portion of the English newspaper corpus Quotebank. Additionally, we manually annotated a small Slovenian dataset of 10 online newspaper articles from the national broadcaster RTV Slovenia. The datasets are summarized in Table 1.

The Slovenian dataset was a selection of 10 news texts of parliamentary sittings, which cover a variety of strategies for reporting utterances. The retrieved articles were split into sentences which were annotated. A sentence was considered reported speech if it included direct and indirect reported speech cued by a reporting clause or prepositional phrase. We excluded nominalizations and shorter quoted text fragments (e.g. *He also particularly emphasized the pressures on the media, currently on RTV and the "illegal non-funding of the Slovenian Press Agency."*) as implied quotes (e.g. *There will be more than 300,000 recipients, he emphasized. For this purpose, 169 million euros will have to be paid out.*). To provide a reference for model performance, we included a heuristic of assigning a positive label to all examples. The models' results on the test datasets were with a Friedman's test as suggested in the literature [6]. Predictions from the best performing classifier on the Slovenian data were manually analyzed further.

3.3 Training Settings

XLM-R and mBERT were used as base models with the default training settings from the *transformers* library with the exception of using 16 gradient accumulation steps and freezing the bottom 8 layers of all models. The latter reduces the training time without significant performance drops (Kovaleva idr., 2019; Merchant idr., 2020). Additionally, a Slovenian-Croatian-English BERT model was trained on English machine-translated data from Redewiedergabe.

4 Results

4.1 Model Results

The model performance differs significantly based on the training and testing data, based on the congruence between the language used and precise task definitions in each dataset. The differences between model predictions were not statistically significant when tested with a Friedman's test (χ^2_F = 9.66; df = 5; n = 8; p = 0.14) so post-hoc tests were not perform. As Table 2 demonstrates, XLM-R model trained on both RiQuA and Redewiedergabe performed well across the datasets with an F-score of 80.5 and 77.6 on the Redewiedergabe and RiQuA test set, respectively. Training on RiQuA improved performance on the Redewiedergabe test set and vice versa, the performance on both datasets was highest when training on combined data. This suggests that the two expansive datasets may still benefit from additional or complementary data, at least in a setting using cross-lingual transfer learning. The most successful strategy for Slovenian data was using the CroSloEngual BERT model in conjunction with machine-translated German training data (into English), reaching a F-score of 66.8. We did not evaluate the impact of translating training data with mBERT and XLM-R.

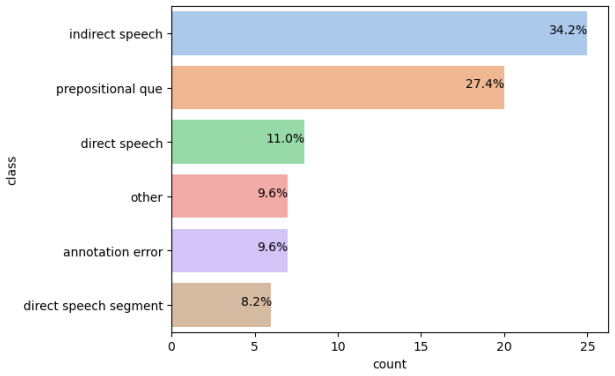

Figure 1: False Negatives from the CroSloEngual BERT Classifier.

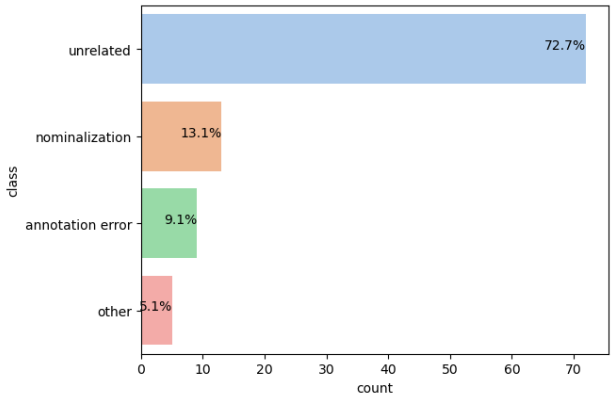

Figure 2: False Positives from the CroSloEngual BERT Classifier.

4.2 Error Analysis Results

The predictions of CroSloEngual BERT classifier on the Slovenian data were analyzed in detail. False positives were more common than false negatives, representing 23.4% and 9.8% of all examples (n = 744), respectively. Close reading of a sample of 100 false positives did not show a definite pattern for most (72.9%) of them, as the examples were clearly not related to reported speech, although some did include words lexically related to reporting verbs (e.g. *The proposed law is still under discussion*). The second category were nominalizations of reported statements (13.1%) not included in our annotation schema. The final source of false positives were annotation errors consisting of wrongly unmarked examples of direct or indirect speech (9.1%). The distribution of categories identified in the sample of false positives are illustrated in Figure 2. The most common errors identified among the 73 false negative examples were instances of indirect speech (34.2% of false negatives) or of prepositional queing of statements (27.4%). Instances of direct speech, direct speech fragments and annotation errors represented 11%, 8.2% and 9.6% of the false negatives, respectively. The annotation errors included nominalizations and statements reported as adjective complements (*The speaker was happy that the provisions were accepted)* not included in our annotation schema. A summary of the identified false negative categories can be found in Figure 1.

5 Discussion

This paper presents the development of a reported speech classifier, tested through a small annotated Slovenian dataset and manual error analysis.

Cross-lingual transfer learning from annotated datasets such as RiQuA and Redewiedergabe achieved an F-score of 66.8 on a small manually annotated dataset of Slovenian news of parliamentary sessions using the base CroSloEngual model with RiQuA and English machine-translated Redewiedergabe training data. This these results are in line with observations that language model trained on a limited number of languages may outperform less specialized ones such as mBERT and XLM-R [22]. The major source of errors were false positives (23.4% of all sentences) for which no systematic pattern was discernible in the majority (72.9%) of examples. The error analysis demonstrated a performance difference across sub-types of reported speech, as 61.6% of false negatives are instances of indirect speech and prepositional queing of statements. Although rare, nominalizations were present in both false positives and false negatives and should be considered in future annotation guidelines.

6 Conclusion

This study developed a sentence-level reported speech classifier for Slovenian news texts using cross-lingual transfer learning. By leveraging existing multilingual models (mBERT, XLM-R, and CroSloEngual BERT) with the English and German datasets RiQuA and Redewiedergabe, we demonstrated that sentence-level classification can detect some aspects of reported speech in Slovenian. However, the performance estimates are limited due to the small size of the Slovenian testing set and the limited definition used for the annotations. Future research should focus on developing a Slovenian annotated dataset, refining the annotation schema for multiple use cases, and exploring additional modeling features such as encoding broader sentence contexts. This work

Table 2: Model Performances across Datasets (F-scores).

	Redewiedergabe	RiQuA	PolNeAR	QUAC	Quotebank	Slovenian dataset
Positive by default	52.1	60.6	74.2	19.5	45.8	60.3
mBERT+Both	77.5	77.4	73.1	40.5	53.5	63.2
mBERT+RiQuA	68.2	76.9	72.6	31.1	52.6	39.1
mBERT+RWG	78.4	70.4	65.5	43.4	49.1	63.2
XLM-R+Both	80.5	77.6	70	38.8	57.7	63.2
XLM-R+RiQuA	66.6	76.7	73.6	25.5	53.7	60.3
XLM-R+RWG	80.9	70.7	66.4	43.9	50	63.2
CroSloEnBERT+Both+MT	54	76.6	73	24	52.5	66.8

contributes a provisional tool for computational discourse analysis of Slovenian media text, but further development is necessary for its application in more nuanced tasks.

Acknowledgements

This work was supported by the Slovenian Research Agency grants via the core research programs Equality and Human Rights in the Times of Global Governance (P5-0413) and Hate Speech in Contemporary Conceptualizations of Nationalism, Racism, Gender and Migration (J5-3102)

References

[1] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE*, 16, 1, (Jan. 29, 2021), e0245533. DOI: 10.1371/journal.pone.0245533.

[2] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16, 1, (Jan. 2, 2022), 1–18. DOI: 10.1080/19312458.2021.2015574.

[3] Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, (May 2020), 803–812. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.100.

[4] 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 LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*. LT4HALA 2020. Rachele Sprugnoli and Marco Passarotti, editors. European Language Resources Association (ELRA), (May 2020), 100–104. ISBN: 979-10-95546-53-5. https://aclanthology.org/2020.lt4hala-1.15.

[5] Alexis Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors. Association for Computational Linguistics, 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[6] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, 7, (Dec. 1, 2006), 1–30.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. [n. d.] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, 4171–4186. DOI: 10.18653/v1/N19-1423.

[8] Gabriel Dvoskin. 2020. Reported speech and ideological positions: the social distribution of knowledge and power in media discourse. *Bakhtiniana: Revista de Estudos do Discurso*, 15, 193–213.

[9] Zoran Fijavž and Darja Fišer. 2021. Citatnost in reprezentacija v spletnem migracijskem diskurzu. In *Sociolingvistično iskrenje*. Maja Bitenc, Marko Stabej, and Žejn Andrejka, editors. Založba Univerze v Ljubljani. Retrieved Apr. 3, 2024 from https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/259/370/6011.

[10] Elizabeth Holt. 1996. Reporting on Talk: The Use of Direct Reported Speech in Conversation. *Research on Language and Social Interaction*, 29, 3, (July 1, 1996), 219–245. DOI: 10.1207/s15327973rlsi2903_2.

[11] Edward Newell, Drew Margolin, and Derek Ruths. [n. d.] An Attribution Relations Corpus for Political News. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Nicoletta Calzolari et al., editors. European Language Resources Association (ELRA). Retrieved Apr. 10, 2024 from https://aclanthology.org/L18-1524.

[12] Mojca Pajnik and Marko Ribać. 2021. Medijski populizem in afektivno novinarstvo: časopisni komentar o »begunski krizi«. *Javnost - The Public*, (Dec. 14, 2021). Retrieved Apr. 24, 2024 from https://www.tandfonline.com/doi/abs/10.1080/13183222.2021.2012943.

[13] Sean Papay and Sebastian Padó. [n. d.] RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, 835–841. ISBN: 979-10-95546-34-4. Retrieved Apr. 21, 2024 from https://aclanthology.org/2020.lrec-1.104.

[14] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. [n. d.] Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors. Association for Computational Linguistics, 989–999. Retrieved Apr. 17, 2024 from https://aclanthology.org/D13-1101.

[15] Marta Ercília Mota Pereira Quintão. 2014. Quotation A ttribution for Portuguese News Corpora. In Retrieved Apr. 21, 2024 from https://www.semanticscholar.org/paper/Quotation-A-ttribution-for-Portuguese-News-Corpora-Quint%C3%A3o/69fea7d030d5e71b973ec67aa897a7c9aadadac2.

[16] Jelena Sarajlić, Gaurish Thakkar, Diego Alves, and Nives Mikelic Preradović. 2022. Quotations, Coreference Resolution, and Sentiment Annotations in Croatian News Articles: An Exploratory Study. Version 1. DOI: 10.48550/ARXIV.2212.07172.

[17] Masaki Shibata. 2023. Dialogic Positioning on Pro-Whaling Stance: A Case Study of Reported Speech in Japanese Whaling News. *Japanese Studies*, 43, 1, (Jan. 2, 2023), 71–90. DOI: 10.1080/10371397.2023.2191839.

[18] Michael Short. 1988. Speech presentation, the novel and the press. In *The Taming of the Text*. Willie Van Peer, editor. Routledge. ISBN: 978-1-315-54452-6.

[19] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying Informational Sources in News Articles. Version 1. DOI: 10.48550/ARXIV.2305.14904.

[20] Stef Spronck and Daniela Casartelli. 2021. In a manner of speaking: how reported speech may have shaped grammar. *Frontiers in Communication*, 6, 624486.

[21] Sara Stymne and Carin Östman. [n. d.] SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. LREC 2022. Nicoletta Calzolari et al., editors. European Language Resources Association, 5324–5333. Retrieved Apr. 21, 2024 from https://aclanthology.org/2022.lrec-1.570.

[22] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue* (Lecture Notes in Computer Science). Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors. Springer International Publishing, Cham, 104–111. ISBN: 978-3-030-58323-1. DOI: 10.1007/978-3-030-58323-1_11.

[23] Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. [n. d.] Quotebank: A Corpus of Quotations from a Decade of News. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining. ACM, 328–336. ISBN: 978-1-4503-8297-7. DOI: 10.1145/3437963.3441760.

[24] M. Wynne. 1996. Speech, Thought and Writing Presentation Corpus. Retrieved Apr. 21, 2024 from https://ora.ox.ac.uk/objects/uuid:6caa73c1-d283-4d51-a78f-55df69bae986.

[25] Dian Yu, Ben Zhou, and Dong Yu. [n. d.] End-to-End Chinese Speaker Identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors. Association for Computational Linguistics, 2274–2285. DOI: 10.18653/v1/2022.naacl-main.165.