SouDeC: Source Detection and Classification in Czech

Jiří Mírovský, Barbora Hladká
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
mirovsky|hladka@ufal.mff.cuni.cz

Abstract

We present a method of attribution source detection and classification in Czech. A plain text (typically, a newspaper article) enters the SouDec system, gets parsed with the external tool UDPipe into Universal-Dependencies style of sentence representation, and then is analyzed for occurrences of attribution signals and sources. The list of attribution signals has been extracted from a corpus of Czech newspaper articles annotated with interlinked attribution signals and sources, and has been complemented with context and syntax information to help distinguish relevant occurrences of the signals. The SouDec system further classifies the attribution sources in one of five classes: *anonymous*, *partially anonymous*, *unofficial*, *official non-political* and *official political*, using information from another external tool, a recognizer and classifier of named entities, NameTag 3. While our source detection method gets results comparable to existing systems for other languages, further improvements can be achieved by incorporating fully-fledged automatic coreference resolution into the classification method. In a focused case study, we test a possible usage of SouDeC for distinguishing domain-specific texts of less vs. more reputable origin.

Keywords: attribution signals, source detection, source classification

1. Introduction

In the age of information overload and increasing concern about misinformation, the credibility and transparency of journalism have become central issues. News articles often rely on various sources, including individuals, organizations or documents, to support claims or provide evidence. Identifying and analyzing these sources is crucial for assessing the reliability of the media, understanding journalistic practices and enabling automated fact checking systems. The automatic detection of information sources in news texts aims at recognizing and categorizing entities that provide information within an article. This task combines elements of natural language processing, information extraction and computational journalism.

We dedicate our effort to a specific task of automatically detecting and classifying sources that journalists credit in newspaper articles. We work with the definition of attribution formulated in Hladká et al. (2022): $attribution = source + information + signal$ where $source$ originally provided $information$ and $signal$ is a textual marker that identifies the source of the information.[1]

This paper introduces and describes in detail SouDeC, a tool for automatic detection and classification of attribution sources in news articles. In a case study, we demonstrate its application by exploring its potential in distinguishing sets of domain-specific texts of less and more reputable origin, namely a collection of chain mail texts from articles originating from a reputable online news outlet.

In Section 2, we review related work on attribution extraction. Section 3 briefly describes the SiR 1.0 corpus used in the development and evaluation of the tool. Section 4 is dedicated to a detailed description of the architecture and methodology of SouDeC. Section 5 presents a case study on distinguishing a collection of chain mails from articles originating from a reputable online news outlet. We conclude in Section 6 with a discussion of the results and future directions.

2. Related work

Most of the works on extraction of attribution focus on the detection of quotations. The first work on detection of direct quotations was published by Bruno et al. (2007). The authors used the Joint Research Centre's Europe Media Monitor system as a data source and they implemented a rule-based system for detection of reporting verbs (= signals), person names (= sources) and direct quotations (= information). O'Keefe et al. (2012) reformulated the quote attribution task as a sequence labeling machine learning task. Further efforts were focused on the development of corpora with manually annotated attribution, including the PARC3 corpus (Pareti, 2012). Janicki et al. (2023) developed a system for attribution information detection for Finish with F1 of 0.82 for a rule-based model and 0.91 for a BERT-based model. For Czech, Poláková et al. (2015) utilized features from deep-syntax annotations of the texts from the Prague

[1] We use mathematical notation intentionally, namely to emphasize that the order of $source$, $information$ and $signal$ in the sentence does not play a role, which is analogous to the commutative property of addition.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 685–693
11–16 May 2026.

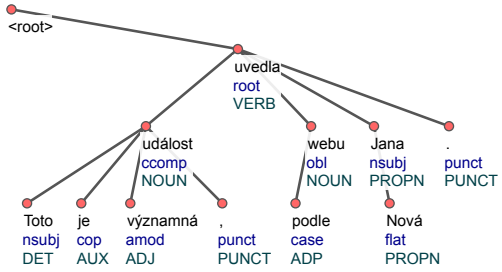

Figure 1: Two linearly embedded attributions in one sentence

Dependency Treebank (Bejček et al., 2013), reaching F1 measure of 0.85 on detecting attribution signals. Petersen-Frey and Biemann (2024) address the task of identifying and attributing quotations in German news texts and they propose a sequence-to-sequence transformer model with constrained decoding to tackle this task.

3. Data

In the development and evaluation of SouDeC, we have used data from a freely available corpus SiR 1.0 (Hladká et al., 2022),[2] an annotated corpus of Czech articles from a news server of a Czech public radio iRozhlas.[3] It is a collection of 1,718 articles (42,890 sentences, 614,995 words) with manually annotated attribution. Specifically, sources of attribution (such as *ministr zahraničních věcí* [*Minister of Foreign Affairs*]) along with attribution signals (such as *prohlásil* [*stated*]) are marked (and linked) in the data. The sources are classified into five classes:

1. *anonymous*: a completely anonymous source, for example *kdokoli* [*anyone*], *anonymní zdroj* [*an anonymous source*],
2. *partially anonymous*: a partially specified source, for example *většina lékařů* [*most doctors*],
3. *unofficial*: a specific person/entity without a mandate to speak for anyone else, e.g., *obyvatel Prahy Václav Novák* [*a Prague citizen Václav Novák*], *bývalý prezident Václav Klaus* [*former president Václav Klaus*], *New York Times*,
4. *official non-political*: a representative of a non-political entity, e.g. *ředitel Škoda Auto, a.s.* [*the director of Škoda Auto, Inc.*], *mluvčí fotbalového klubu* [*the spokesman of the football team*], and

[2] Published as a language resource (Hladká et al., 2022) and available from http://hdl.handle.net/11234/1-4840

[3] https://www.irozhlas.cz/

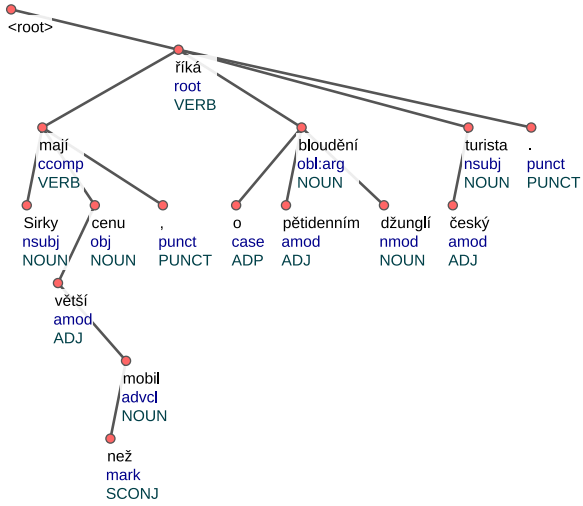

Figure 2: Dependency tree of the sentence in Example 2

5. *official political*: a political representative, e.g., *předseda vlády* [*the prime minister*], *Ministerstvo obrany* [*Ministry of defence*].

In the SouDeC tool, as well as in the experiments described in this paper, we have adopted the same taxonomy.

The annotations in SiR 1.0 come as a result of a crowdsourcing task and the corpus consists of three parts, depending on the quality of the annotations:

1. triple-annotated articles: 46 articles (933 sentences, 13,242 tokens) annotated independently by three annotators and subsequently curated by an arbiter,
2. double-annotated articles: 543 articles (12,347 sentences, 180,622 tokens) annotated independently by two annotators and automatically unified, and
3. single-annotated articles: 1,129 articles (29,610 sentences, 421,131 words) annotated only by a single annotator each.

As described later in Section 4.4, the triple-annotated articles formed a basis for our development and evaluation test data. Additionally, we have manually checked and corrected the automatic unification in 80 articles from the double-annotated section, using half of them as further development test data and the other half as additional evaluation test data.

4. Method

The source detection and classification method in *SouDeC* is implemented as a dependency tree matching method where input sentences are

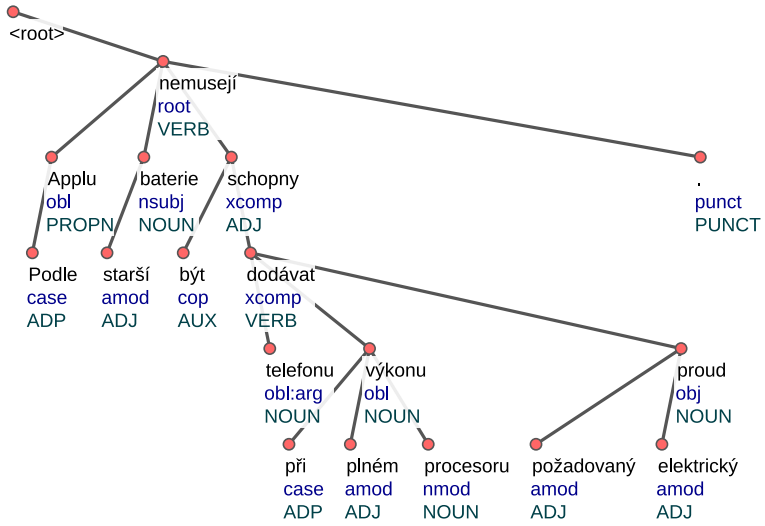

Figure 3: Dependency tree of the sentence in Example 3

first processed with two external tools: a syntactic parser and a named-entity recognizer. This method enables the detection of more intricately structured data compared to traditional regular expression queries, particularly in handling variations in Czech free word order. Consider the sentence in Example 1:

(1) Toto je významná událost, uvedla podle webu Jana Nová.
[*lit.* This is an important event, stated according to the website Jana Nová.]

The sentence contains two sources, *web*[4] [*the website*] signalled by the preposition *podle* [*according to*], and *Jana Nová* signalled by the verb *uvedla* [*stated*]. These attributions are linearly embedded and difficult to process, while in the dependency tree structure, they form two distinct subtrees (conf. Figure 1).

Figure 4 shows the architecture of the system. The system starts with an input in plain text and proceeds in several steps. Let us document the method on two example sentences (Examples 2 and 3) from the double-annotated section of SiR 1.0. The attribution signals are typeset in bold, the sources are underlined.

(2) Sirky mají větší cenu než mobil, **říká** o pětidenním bloudění džunglí český turista. (SiR 1.0, doc-5950761)
[Matches are worth more than a mobile phone, **says** a Czech tourist about his five-day jungle wandering.]

(3) **Podle** Applu starší baterie nemusejí být schopny dodávat telefonu při plném výkonu procesoru požadovaný elektrický proud. (SiR 1.0, doc-6596981)
[**According to** Apple, older batteries may not be able to supply the energy needed to power the phone at full processor power.]

4.1. Source Detection

To detect attribution sources, we first detect attribution signals. Once a signal is found, the source is identified by traversing the dependency tree. Its position may differ according to the type of attribution signal.

4.1.1. UD Parsing

In the first step, the text is parsed using an external web service UDPipe version 2 (its older version was described in Straka, 2018).[5] UDPipe is a pipeline tool that performs sentence segmentation, tokenization, lemmatization, morphological tagging, and dependency parsing in the Universal Dependencies framework (De Marneffe et al., 2021).[6] Figures 2 and 3 show the resulting representations of the two example sentences as dependency trees produced by UDPipe.

4.1.2. Named Entities

As the second step, another external web service, NameTag version 3, is called for named entity recognition (Straková et al., 2019).[7] NameTag identifies proper names in the text and classifies

[4] *webu* in genitive

[5] https://lindat.cz/services/udpipe/

[6] UDPipe has been trained for almost all UD treebanks, incl. Czech, with state-of-the-art results.

[7] NameTag 3 achieves state-of-the-art performance on 15 languages, incl. Czech.

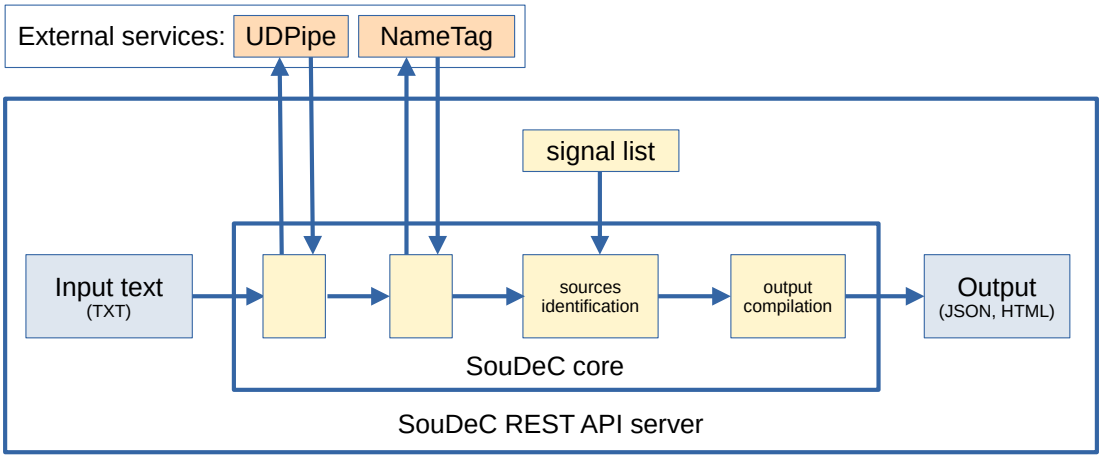

Figure 4: The architecture of the SouDeC REST API server

them into predefined categories, such as names of persons, locations, organizations, etc. It recognizes nested entities (embedded entities) of arbitrary depth.[8]

4.1.3. Signal Lemmas

The output of the two external tools serves as an input for further processing. We traverse the dependency trees of individual sentences and search for attribution signals, using a list of approx. 200 lemmas of potential signals. The list was extracted from the training data (see below in Section 4.4) and manually checked and cleaned. Some of the lemmas were complemented with hand-made constraints that help determine in what context and syntactic structure they represent an attribution signal. For example, the entry:

```
podělit se|názor-o
[share (refl.)|opinion-(acc.-prep)]
```

defines a constraint for lemma *podělit* [*share*] that in the dependency representation of the sentence requires two dependant sons of the node *podělit*, namely a node representing a reflexive particle *se* and another one representing the word *názor* [*opinion*], which in turn is required to have a son *o* (an accusative preposition). That is, this entry represents the attribution signal *podělit se o názor* [*share (one's) opinion*].

If a lemma from the list is found in the dependency tree, the properties defined by the constraint are checked and the lemma is considered an attribution signal only if the constraint is met. The same lemma can appear in the list with several different constraints, which are then checked independently. Each item in the list is complemented with statistics from the train data, namely the total number of occurrences of the lemma in the syntactic structure defined by the constraint vs. the number of its occurrences (with the constraint) as an attribution signal. The ratio of these two values represents a reliability of the lemma (and the constraint) and a minimal required reliability can be used as a parameter of the system.

4.1.4. Claim Identification

Although finding the claimed information is not part of our task, its existence or absence in the tree helps distinguish cases where the lemma serves as an attribution signal from cases where it does not. The constraints in the list of attribution signal lemmas may also indicate where in the tree the claimed information should be searched for (if not in the standard position among the signal lemma's sons). The non-standard position of the claim is indicated by an exclamation mark at one of the nodes of the constraint, for example for expression *přijít s tvrzením* [*come with a claim*], the constraint:

```
přijít tvrzením!-s
[come claim!-with]
```

specifies that the claimed information should be searched among the sons of the word (node) *tvrzením* [*claim*].

For more complex cases, there are several special constraint values. In fact, the whole constraint for lemma *podělit* [*share*] is:

```
podělit se|názor-o|POSTPOS
[share (refl.)|opinion-(acc.-prep)|
POSTPOS]
```

where the POSTPOS feature indicates that the attribution signal appears in the postposition relative to the claimed information, confr. Example 4,

(4) Jsem zásadně proti tomu, **podělil se o názor** trenér družstva.
[*lit.* I am strongly against it, **shared his**

[8] https://lindat.cz/services/nametag/

set	docs	sentences	tokens	sources
dtest	56	1,052	20,015	264
etest	70	1,414	24,225	297

Table 1: Size of the development (dtest) and evaluation test (etest) data and number of attribution sources marked in the data by annotators

opinion the team coach.]

while for the different syntactic structure from Example 5, there is a second entry for lemma *podělit* [*share*]:

```
podělit se|názor!-o
[share (refl.)|opinion!-(acc.-prep)
```

(5) Trenér družstva **se podělil o názor**, že je zásadně proti tomu.
[The team coach **shared his opinion** that he is strongly against it.]

Other possible values for defining non-standard positions of the claimed information are ANTEPOS for anteposition, and PREP for prepositional expressions such as *podle* [*according to*]:

```
podle PREP
[according_to PREP]
```

The PREP constraint for lemma *podle* [*according to*] helps find the claimed information in Example 3 at the grandfather of the node *podle* (see Figure 3), i.e., at the node *nemusejí* [*may not*]. Lemma *říkat* [*say*] from Example 2 does not have any constraint, i.e., the claimed information is to be searched among the sons of node *říká* [*says*] (see Figure 2).

4.2. Source Identification

Based on the position of the attribution signal in the dependency tree, finding the source is a straightforward step: in most cases, it is the (sometimes partial) subtree of the subject of the attribution signal lemma node, i.e., the son of the attribution signal with dependency relation *nsubj*, confr. node *turista* [*tourist*] and its son *český* [*Czech*] with dependency relation *amod* (adjectival modifier) in Figure 2.

For lemmas with the PREP constraint, the attribution source root node is the parent of the PREP lemma (e.g., the parent of the preposition *podle* [*according to*]), confr. node *Apple*[9] in Figure 3.

[9] *Applu* in genitive

development test data (dtest)				
reliability	F1	P	R	acc
10%	0.84	0.90	0.81	67%
20%	0.84	0.90	0.81	67%
30%	0.84	0.90	**0.81**	67%
40%	0.84	0.91	0.80	67%
50%	**0.84**	0.91	0.80	67%
60%	0.84	0.91	0.79	67%
70%	0.84	0.91	0.79	67%
80%	0.64	0.95	0.49	73%
90%	0.58	**0.98**	0.42	72%
100%	0.12	0.95	0.07	100%

Table 2: F1-measure on recognizing the source (along with the Precison and Recall values) and the accuracy of classifying the source on development test data for various thresholds of reliability of attribution signals

4.3. Source Classification

After its detection, the source is classified into one of five classes (the same classes that recognizes the SiR 1.0 corpus, see Section 3): *anonymous*, *partially anonymous*, *unofficial*, *official non-political* and *official political*. The solution of this task is based on combination of two sources: (i) information from named entities recognition provided by the external tool NameTag, and (ii) lexical rules based on lists devised by analyzing the development data.

evaluation test data (etest)				
reliability	F1	P	R	acc
30%	0.87	0.90	**0.85**	65%
50%	**0.87**	0.90	0.85	65%
90%	0.61	**0.92**	0.47	62%

Table 3: F1-measure on recognizing the source (along with the Precison and Recall values) and the accuracy of classifying the source on evaluation test data for selected thresholds of reliability of attribution signals

These hand-made lists comprise of keywords or phrases not recognized by NameTag but often associated with certain classes. For example, words such as *zdroj* [*source*], *pozorovatel* [*observer*]) are often connected with the *anonymous* class, words such as (*část* [*part of*], *některý* [*some*], *většina* [*most of*], etc.) indicate a *partially anonymous* source. Words *bývalý* [*former*] or *někdejší* [*erstwhile*] often indicate a source that may have been official in the past but now is *unofficial*. *Majitel* [*owner*] may signal an *official-non-political* source, while words such as *premiér*

development test data (dtest)					
Actual vs. Predicted	anonymous	anon.-partial	unofficial	official-non-political	official-political
anonymous	0	0	0	0	0
anonymous-partial	1	20	0	0	1
unofficial	0	10	50	25	2
official-non-political	0	10	7	51	6
official-political	0	1	2	4	23

Table 4: Confusion matrix on dtest, reliability threshold 50%

[*prime minister*], *poslanec* [*member of parliament*], *senátor* [*senator*] or *předseda vlády* [*prime minister*], etc. strongly indicate an *official-political* source.

In Example 2, NameTag does not find any named entity in the source *český turista* [*Czech tourist*], neither any of the words is found in our hand-made lists, therefore we mark (correctly) the source as *partially anonymous*, which is our default class for sources with no recognized named entity. In Example 3, the source *Apple* is marked by NameTag as an institution of type "companies, concerns", therefore we mark (also correctly) the source as *official-non-political*.

4.4. Evaluation

We have tested SouDeC on data from the SiR 1.0 corpus described in Section 3. 16 articles from the triple-annotated section plus 40 newly revised articles from the double-annotated section were used as development test data (dtest), the remaining 30 articles from the triple-annotated section plus 40 (also newly revised) articles from the double-annotated section were used as evaluation test data (etest), see Table 1. Further approx. 460 articles (without corrections) in the double-annotated section were used to extract a list of attribution signals.

Tables 2 and 3 show the results of the SouDeC system applied on the dtest and etest data. The first table shows F1 measure on recognizing the attribution sources (along with the respective Precision and Recall values) and the accuracy of classifying the sources[10] for various thresholds of reliability of attribution signals.[11]

Based on the results, we have selected three values of the threshold – 30% for best Recall with minimal impact on F1 measure, 50% for best F1 measure and 90% for best Precision with moderate impact on F1 measure. For these (based on the dtest data) best settings for Recall preference, F1 preference and Precision preference, respectively, we have tested the system on evaluation test data, the results are given in Table 3. Table 4 gives a confusion matrix between the correct and predicted attribution source classes, measured on the dtest data.

[10] Please note that the accuracy of source classification does not directly depend on the reliability threshold for the attribution signals.

[11] We do not start with threshold 0, as all attribution signals in our list have the reliability higher than 10%.

5. A Case Study

Counts of various classes of attribution sources detected in a text by SouDeC can be used to compile an overall distribution of attribution sources types in an entire set of documents. In a focused case study, we employ this method to try and distinguish two sets of texts: (i) a set of chain mails, as a representative of texts of a "less reliable origin", and (ii) articles from a reputable internet server of the Czech public radio iRozhlas, as a representative of texts of a "more reliable origin". We hypothesize that the public radio articles would contain (i) more attributions than the chain mails, and (ii) fewer anonymous or partially anonymous sources than the chain mails.

5.1. Data

The collection of chain mail data has come from the Czech civic movement *Čeští elfové* that "monitors, analyzes, and actively combats foreign disinformation campaigns on the Czech Internet. They drew loose inspiration from a similar movement of Elves in the Baltic states, who have been fighting Russian hybrid and disinformation operations for years."[12] The chain mails collection comprises of 417 articles with 28,732 sentences.

The data collection of the iRozhlas articles comprises of all articles from the server archive[13] between the years 2017 and 2022 (incl.). It contains 104,666 articles with 3,191,936 sentences.

Both these collections were processed with SouDeC (reliablity threshold set to 30%). In the chain mails collection, 1,524 attributions were detected, with 3.7 attributions per article and 0.05 at-

[12] https://cesti-elfove.cz/

[13] https://www.irozhlas.cz/zpravy-archiv

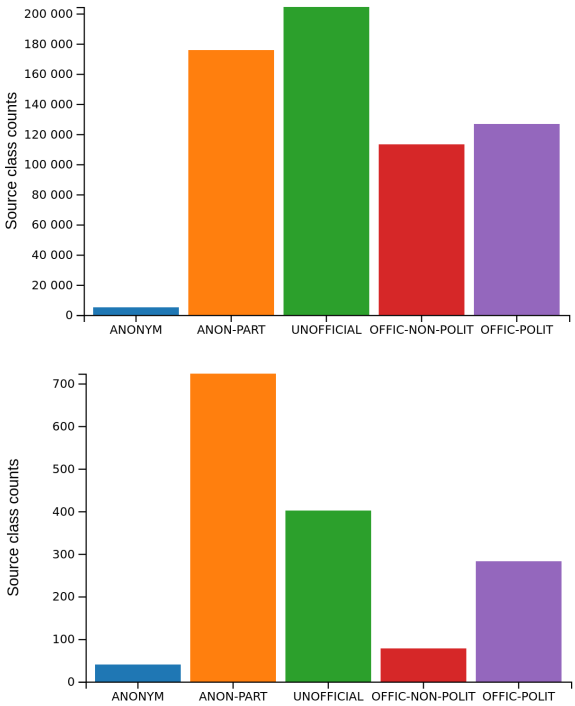

Figure 5: All iRozhlas articles (top) vs. chain mails (bottom)

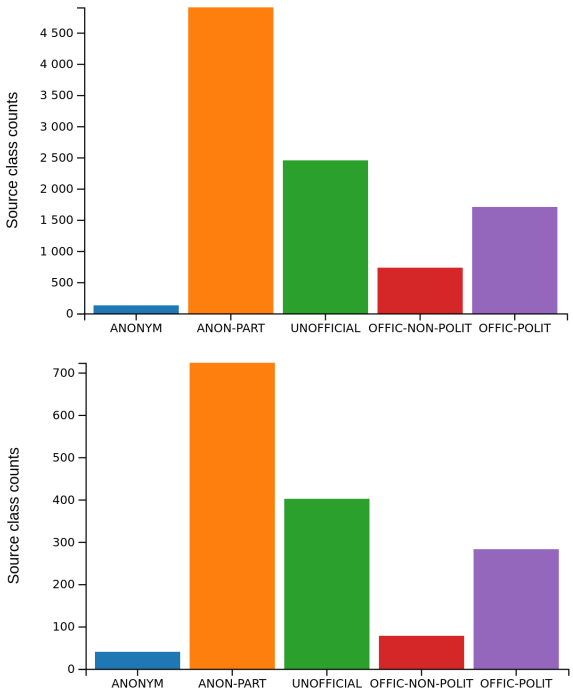

Figure 6: Commentaries section from iRozhlas (top) vs. chain mails (bottom)

tributions per sentence. In the iRozhlas articles, SouDeC recognized 624,882 attributions, with 6.0 attributions per article and 0.20 attributions per sentence. Figure 5 shows distributions of individual source classes in the two collections. Both the overall attribution counts normalized per article and sentence, and the distributions of source classes seem to confirm our hypothesis: texts of less reputable origin contain in total less attributions, and relatively more attributions are of *partially anonymous* class.

However, we should also take into account that the iRozhlas collection contains articles of various kinds; they are in different sections of the archive, such as *Travelling*, *Theatre*, *Sport*, *Society*, *Culture*, *Economics*, and many others (in total 45 sections). The opinion types of texts in the chain mails collection best fit the section *Commentaries*. If the SouDeC analysis of the iRozhlas texts is only restricted to this section (4,162 articles with 134,144 sentences), the overall SouDeC numbers are 9,927 attributions, 2.4 attributions per article, 0.07 attributions per sentence, the distribution of attribution classes is given in Figure 6. For the *Commentaries* section, both the overall number of attributions normalized per article and sentence, and the attribution classes distribution are strikingly close to the chain mails collection.

6. Conclusion

SouDeC demonstrates how far we can get in the attribution source detection and classification task in Czech if we employ the language analysis approach using state-of-the-art tools for syntactic parsing and named entities recognition. Our results on attribution source detection (F1: 0.87) compare fairly with similar rule-based systems, e.g. for English (0.75–0.89 for various settings on claimed information identification reported in Pareti et al., 2013), for Finish (0.82 on claimed information identification reported in Janicki et al., 2023),[14] and for Czech (0.85 on signal recognition reported in Poláková et al., 2015, they however relied on information from manually annotated deep-syntax trees).

Analysis of the cases of misclassification of source classes (in the development test data) reveals several principal causes of classification errors:

(i) SouDeC does not analyze the structure of the source. Thus, for example, the source *lidé, které Radiožurnál oslovil v centru Prahy* [*people interviewed by Radiožurnál in the Prague centre*] gets classified as *unofficial*, because NameTag marks *Radiožurnál* as a media name of type “periodical”. However, the correct class is *partially anonymous*.

(ii) The confusion matrix in Table 4 shows that

[14] Their BERT-based model with F1 of 0.91 outperforms our system.

a large portion of the misclassified cases occurs between the *unofficial* and the *official-non-political* classes. The boundary between these two types of sources is indeed difficult to find in many real world cases even for human annotators. If it is not specified in the context, it is not trivial to decide whether *pan Pavel Procházka z Ústavu jaderné fyziky* [*Mr. Procházka from the Institute of Nuclear Physics*] is speaking (and is entitled to speak) on behalf of the institute. This raises a related open question: SouDeC does not take the attributed content in consideration. Should we classify the source in connection with the attributed claim? I.e., for example, can political representatives or company directors speak in some cases for themselves (e.g., speaking about their vacations) and thus being in some context considered *unofficial* sources?

(iii) SouDeC in its present version does not employ a full-fledged coreference resolution (unlike, for example, Almeida et al., 2014). While several heuristic rules have been implemented to solve the most frequent and typical anaphora references, still many errors in attribution source classification might be solved by a correct anaphora resolution. This is one of the near-future features planned for SouDeC, as a coreference resolver is expected to become available soon as an addition to the UDPipe parser.

In a focused case study, we investigated a research question whether a reputability of collections of journalistic texts can be determined based on the relative frequency of attributions in the documents and on the distribution of attribution source classes. While the experiments with SouDeC showed clear differences in the observed phenomena between a collection of chain mails and the whole archive of articles from a reputable online news outlet, no difference was detected if the articles from the news outlet were restricted to the section “Commentaries”.

The SouDeC system is available as a command-line program, web service and a REST API server[15] and the source code is downloadable under the Mozilla Public License 2.[16]

Acknowledgements

The authors gratefully acknowledge support from the TAČR project TL05000057 and the National Recovery Plan projekt MPO 60273/24/21300/21000 NPO. The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

[15] https://quest.ms.mff.cuni.cz/soudec/

[16] https://github.com/ufal/soudec/

7. Bibliographical References

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

Pouliquen Bruno, Steinberger Ralf, and Best Clive. 2007. Automatic detection of quotations in multilingual news.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Barbora Hladká, Jiří Mírovský, Matyáš Kopp, and Václav Moravec. 2022. Annotating attribution in Czech news server articles. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1817–1823, Marseille, France. European Language Resources Association.

Maciej Janicki, Antti Kanner, and Eetu Mäkelä. 2023. Detection and attribution of quotes in finnish news media: Bert vs. rule-based approach. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, number 52 in NEALT Proceedings Series, pages 52–59, Estonia. University of Tartu Library. Nordic Conference on Computational Linguistics , NoDaLiDa 2023 ; Conference date: 22-05-2023 Through 24-05-2023.

Timothy O’Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, Jeju Island, Korea. Association for Computational Linguistics.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3213–3217, Istanbul, Turkey. European Language Resources Association (ELRA).

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Fynn Petersen-Frey and Chris Biemann. 2024. Fine-grained quotation detection and attribution in German news articles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 196–208, Vienna, Austria. Association for Computational Linguistics.

Lucie Poláková, Pavlína Jínová, and Jiří Mírovský. 2015. Signals of attribution in the prague dependency treebank. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 292–299, Warszawa, Poland. IPI-PAN, IPIPAN.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2019. Neural architectures for nested ner through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Stroudsburg, PA, USA. Association for Computational Linguistics.

8. Language Resource References

Eduard Bejček and Eva Hajičová and Jan Hajič and Pavlína Jínová and Václava Kettnerová and Veronika Kolářová and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Jarmila Panevová and Lucie Poláková and Magda Ševčíková and Jan Štěpánek and Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. Univerzita Karlova v Praze, MFF, ÚFAL. Univerzita Karlova v Praze, MFF, ÚFAL.

Barbora Hladká and Jiří Mírovský and Matyáš Kopp and Václav Moravec. 2022. *SiR 1.0*. Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, Lindat/CLARIAH, http://hdl.handle.net/11234/1-4840.