

Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT

Hanh Thi Hong Tran^{1,2,3}, Vid Podpečan¹,
Mateja Jemec Tomazin⁴, Senja Pollak¹

¹Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

²Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana,
Slovenia

³University of La Rochelle, 23 Av. Albert Einstein, La Rochelle, France

⁴ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenia

E-mail: tran.hanh@ijs.si

Abstract

Definition Extraction is a Natural Language Processing task that automatically identifies definitions from unstructured text sequences. In our research, we frame this problem as a binary classification task, aiming to detect whether a given sentence is a definition or not, using text sequences in Slovene. The main contributions of our work are two-fold. First, we introduce a novel Slovene corpus for the evaluation of Definition Extraction named *RSDO-def*. The dataset contains labeled sentences from specialized corpora using two different extraction processes: random sampling and pattern-based extraction. Both sets contain manual annotations by linguists with three labels: Definition, Weak definition, and Non-definition. Second, we propose the benchmarks for Slovene Definition Extraction systems that use (1) rule-based techniques; (2) Transformers-based models as binary classifiers; (3) ChatGPT prompting, and evaluate them on both sets of *RSDO-def* corpus. When only the small sample *RSDO-def-random* is considered, the pattern-based rules surpassed the performance of language models classifiers or ChatGPT in terms of F1 on definition class in the strict evaluation setting (considering Weak definition as Non-definition). Meanwhile, language models (classifiers and ChatGPT) outperformed rule-based approaches when applied to the data with a higher number of definitions and more relaxed evaluation scenarios (considering Weak definition as Definition). Comparing ChatGPT and language models classifiers on the definition class of *RSDO-def-random* and *RSDO-def-large*, we observe that higher precision was obtained with classifiers, but higher Recall with ChatGPT.

Keywords: Definition Extraction; RSDO-DEFT; Rule-based; Transformers; ChatGPT

1. Introduction

Definition Extraction is a Natural Language Processing (NLP) task that extracts textual definitions from naturally occurring texts (Navigli & Velardi, 2010). While extracting the definitions of words or phrases from general language corpora is needed for the creation of general dictionaries and lexical databases, extracting definitions of terms from specialized domain corpora can serve for creating specialized dictionaries and glossaries. In our paper, we are interested in the latter.

Definition Extraction is closely tied to the needs of lexicography and terminography. For example, in our recently developed terminological portal¹ supporting also Term Extraction, the Definition Extraction module is used to select selecting examples of use and support the user in the manual definition construction process (currently, the pattern-based extractor is implemented in the portal). Definition Extraction is applied in many other NLP downstream tasks as well, including glossary creation (Klavans & Muresan, 2002; Park et al., 2002), question answering (Cui et al., 2005; Saggion & Gaizauskas, 2004), taxonomy learning (Velardi et al., 2013; Espinosa-Anke et al., 2016), knowledge base generation (Bovi et al., 2015), to cite a few.

SemEval-2020 Task 6: DeftEval: Extracting Term-Definition Pairs in the Free Text (Spala et al., 2020) recently introduced the novel human-annotated English dataset, namely Definition Extraction from Texts (DEFT) corpus and formulated the task as three consecutive subtasks: (1) classification of sentences as definition or non-definition, which is also the task addressed in our work (2) labeling of definitional sentences, and (3) relation classification. However, when it comes to lesser-resourced languages like Slovene, there is no dedicated publicly available annotated collection designed for the development and evaluation of Definition Extraction.

The contribution of this paper is threefold: (1) The creation of a novel Slovene corpus for Definition Extraction evaluation (consisting of *RSDO-def-random* and *RSDO-def-larger*) with three labels: not a definition, weak definition, and strong definition; (2) Filling the research gap in the Definition Extraction for Slovene by experimenting with different neural approaches (3) An empirical evaluation of rule-based, language model based binary classifier, and ChatGPT prompting benchmarks for Definition Extraction task.

This paper is organized as follows: Section 2 presents the related work in Definition Extraction and addresses also the gap between the approaches developed for rich-resourced (e.g., English) and less-resourced (e.g., Slovene) languages. In Section 3, we briefly introduce the novel Slovene corpus for Definition Extraction with two different sample acquisition strategies: random sampling and pattern-based methods. Next, we present the methods and implementation details (Section 4), followed by the description of experimental results (Section 5) and error analysis (Section 6). Section 7 presents the conclusion and our plans for further work.

2. Related Work

Most of the early approaches to Definition Extraction were inspired by the work of Hearst (1992) on lexico-syntactic patterns. The cues of definition sentences include lexical cues (e.g., “*is*”, “*means*”, “*are*”, “*is defined as*”), which are often combined with grammatical rules and syntactic patterns (Klavans & Muresan, 2001; Cui et al., 2004, 2005; Sarmiento et al., 2006; Storrer & Wellinghoff, 2006). As these approaches are only able to detect explicit and structured definitions, they fail to detect sentences containing definitions without predefined linguistic clues, and thus, often suffer from low Recall.

With the advent of machine learning (ML) methods, several supervised and semi-supervised models have been proposed (Fahmi & Bouma, 2006; Westerhout, 2009; Reiplinger et al., 2012; Jin et al., 2013; Espinosa-Anke & Saggion, 2014; Espinosa-Anke et al., 2015). The

¹ <https://terminoloski.slovenscina.eu>

task was then considered as a binary sentence classifier (Fahmi & Bouma, 2006) where they took advantage of features based on bag-of-words (BoW), n-grams, and syntactic information, to mention a few. A hybrid approach (Westerhout, 2009) that combined the rule-based and ML-based classifier was then introduced with further exploration and exploitation in both linguistic and structural information, while Borg et al. (2009) proposed the use of genetic algorithms to learn distinguishing features of definitions and non-definitions and weight the individual features. Different bootstrapping algorithms (Reiplinger et al., 2012; De Benedictis et al., 2013) were also experimented with to boost the performance of the extraction. However, the mentioned methods often depend heavily on manual rules and handcrafted features, which are time- and effort-consuming to design as well as domain-specific. As a result, they are hard to adapt and generalize to a new domain or another specific task. An interesting approach by Navigli & Velardi (2010) proposed automatically learned Word Class Lattices (WCLs), a generalization of word lattices, to model textual definitions, where lattices are learned from an annotated dataset of definitions from Wikipedia.

Recent years have witnessed a shift toward neural network-oriented solutions to prevent the issues from rule-based and traditional ML approaches and better capture a large variety of possible definition realizations. Li et al. (2016) proposed a Long Short-Term Memory (LSTM) classifier, where the features were automatically generated from the raw input sentences and part-of-speech (PoS) sequences. Meanwhile, various neural hybrid methods have been released regarding either the combination between two models (e.g., Anke & Schockaert (2018) combined Convolutional Neural Network (CNN) and bi-LSTM) or between two different representations as an input for the neural model (e.g., Veyseh et al. (2020) leveraged Graph Convolutional Neural (GCN) by concatenating both syntactic and semantic information). Furthermore, Kannan & Ponnusamy (2020) presented a combination of both hybrid strategies by concatenating GloVe and on-the-fly PoS information as an input and feeding them to the bi-LSTM with an additional 1-dimensional Convolution and MaxPool layer on top of that. Meanwhile, Kaparina & Soboleva (2020) made use of both global and contextual information by ensembling FastText and ELMo word embeddings to a Recurrent Neural Network (RNN) architecture.

However, until recently, the existing methods have not yet benefited from the large pretrained language models and the transfer learning paradigm (Kenton & Toutanova, 2019), which is today a standard for developing state-of-the-art (SOTA) solutions to a large variety of NLP downstream tasks. This has changed with the introduction of *SemEval-2020 Shared Task DeftEval: Extracting Term-Definition Pairs in the Free Text* (Spala et al., 2020) and its novel human-annotated English dataset, namely Definition Extraction from Texts (DEFT) corpus. There, several solutions based on transfer learning and Transformer architecture, have been proposed. While several participants opted to simply fine-tune BERT (Davletov et al., 2020; Jeawak et al., 2020; Singh et al., 2020), RoBERTa (Avram et al., 2020) or XLNet (Ranasinghe et al., 2020), others opted for more specific approaches. For example, Caspani et al. (2020) captured contextual information from the input sentence using RoBERTa and applied the Stochastic Weight Averaging (Izmailov et al., 2018) to combine weights of the same network at different stages of training, whereas Zhang & Ren (2020) incorporated several LSTM layers into different Transformer architectures to boost the performance of their definition extractor.

Whilst recent Definition Extraction methods are leveraging Transformers-based language models (see e.g., winning approaches in the *SemEval-2020 Task 6* (Spala et al., 2020)), these methods were not yet sufficiently applied to lesser-resourced language such as Slovene. Here, the related work is limited to rule-based approaches (Pollak, 2014b,a; Pollak et al., 2012) or feature-based ML methods (Fišer et al., 2010). The rule-based approaches have been applied to the specialized corpora in various domains, including karstology (Pollak et al., 2019; Vintar & Martinc, 2022). However, no language-model-based approaches have been tested on the task of Definition Extraction for Slovene.

3. RSDO-def Datasets

One of the contributions of this paper is the creation of the Slovene Definition Extraction evaluation datasets, *RSDO-def* (Jemec Tomazin et al., 2023), publicly available through the CLARIN.SI data sharing repository: <http://hdl.handle.net/11356/1841>. The dataset was annotated in the scope of the project Development of Slovene in Digital Environment (RSDO). The sentences were extracted from the Slovene domain-specific corpora (Jemec Tomazin et al., 2021), collected in the scope of the project Development of Slovene in a Digital Environment, containing texts with annotated terms from four different domains: biomechanics, linguistics, chemistry, and veterinary science. Two different sampling strategies were used to create two different sub-corpora: random sampling (*RSDO-def-random*) and pattern-based selection (*RSDO-def-larger*). While random sampling represents the most realistic evaluation scenario and allows for assessment of Recall of various methods, the number of definitions is very small and therefore represents a too small sample to support a reliable quantitative evaluation of methods. Therefore, in order to increase the number of definitions, we added pattern-based sampling, where a pattern-based Definition Extraction approach (Pollak, 2014a) was used for sentence selection. This approach results in a larger sample of definitions, but on the other hand, as the pattern-based method was used in the data collection process, we had to exclude the method from the evaluation phase.

Both sets were manually annotated by five terminographers, where after individual assessments in case of discrepancies between annotators, a consensus was reached and the final label was confirmed by all five annotators. In the resulting dataset, the sentences were annotated with one of the three labels: Definition, Weak definition, and Non-definition. The criteria for annotation are based on the standard *ISO 1087-1:2000 (E/F) Terminology Work - Vocabulary, Part 1, Theory and Application*, which explains the *definition* as follows: “Representation of a concept by a descriptive statement which serves to differentiate it from related concepts”. The most common are intensional definitions, which state the superordinate concept and the delimiting characteristics. Such definitions are often provided in student handbooks and specialized manuals. *Weak definition* labels were assigned if the extracted sentences contained a term and at least one delimiting feature without a superordinate concept, or sentences consisting of superordinate concepts without delimiting features but with some typical examples. Instances were labeled as *Non-definition* if the sentence with the extracted concept did not contain any information about the concept or its delimiting features. Such sentences are also more common in scientific texts, so the imbalance is not unexpected.

The label distribution statistics (after removing a small number of duplicates) are presented in Figure 1. For evaluation (see Section 5), we consider two scenarios: Weak definition is

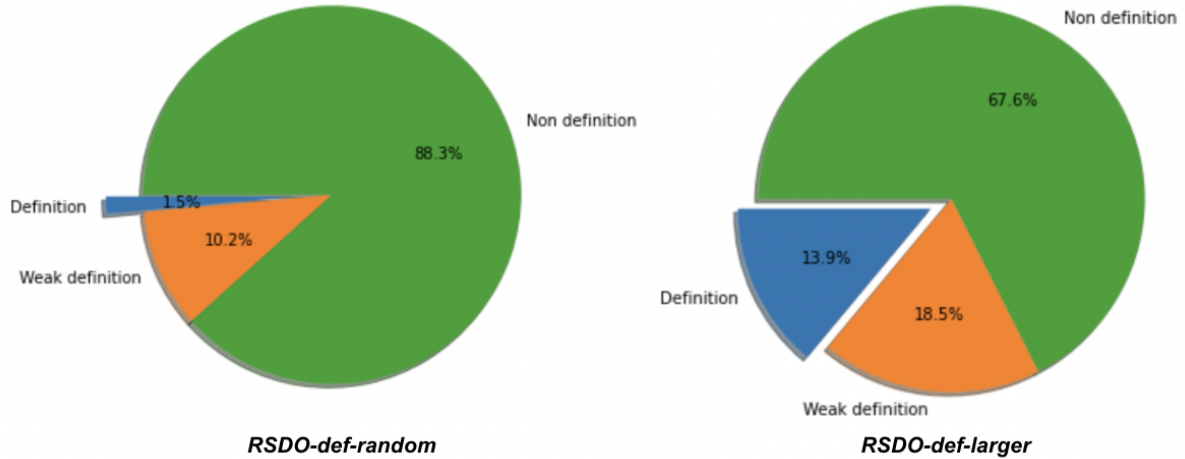


Figure 1: The distribution of Definition, Weak definition, and Non-definition category in each corpus.

considered as (1) Non-definition (Strict evaluation scenario); and (2) Definition (Relaxed evaluation scenario). In both scenarios, there is an imbalance between the number of samples belonging to the positive and negative classes, which reflects also the real-life Definition Extraction settings. Note that these samples are useful for the evaluation of the Definition Extraction methods, but we consider them too small for use as datasets for training definition extractor systems.

3.1 RSDO-def-random

In *RSDO-def-random* corpus, where sentences were selected by random sampling techniques, 961 sentences were manually annotated, out of which 14 examples were assigned the Definition class, 849 examples were assigned the Non-definition class, and 98 examples were assigned the Weak definition class. For the strict evaluation scenario (Weak definitions considered as Non-Definitions) the class distribution is 1.5% (Definitions) vs. 98.5% (Non-definitions), while for the relaxed evaluation scenario (Weak definitions considered as Definitions), the class distribution is 11.7% (Definitions) vs. 88.3% (Non-definitions).

3.2 RSDO-def-larger

In order to increase the number of definitions that represent only a small fraction of the randomly sampled *RSDO-def-random* set, we extended this initial corpus using pattern-based extraction methods. The pattern-based extractor was run on the whole corpus and the results were manually labeled. The resulting *RSDO-def-larger* dataset contains 175 Definitions and 848 Non-definitions while the rest 232 examples are Weak definitions. For the strict evaluation scenario (Weak definitions considered as Non-Definitions) the class distribution is 13.9% (Definitions) vs. 86.1% (Non-definitions), while for the relaxed evaluation scenario (Weak definitions considered as Definitions), the class distribution is 32.4% (Definitions) vs. 67.6% (Non-definitions).

4. Methodology

The main goal of the paper is to evaluate different Definition Extraction approaches. We compare the baseline pattern-based approach (Pollak, 2014a) (with two variants, *is a* pattern type and extended pattern list), with four newly implemented deep-learning Transformer classifiers, and a ChatGPT-based solution. For training the Transformer models, we used Wikipedia as a training set, as in Fišer et al. (2010). The experimental workflow is presented in Figure 1. In this section, we present the methods for the three approaches—Rule-based ones in Section 4.1, Transformer classifiers in Section 4.2, and the ChatGPT-based ones in Section 4.3—followed by the evaluation metrics used in for experiments.

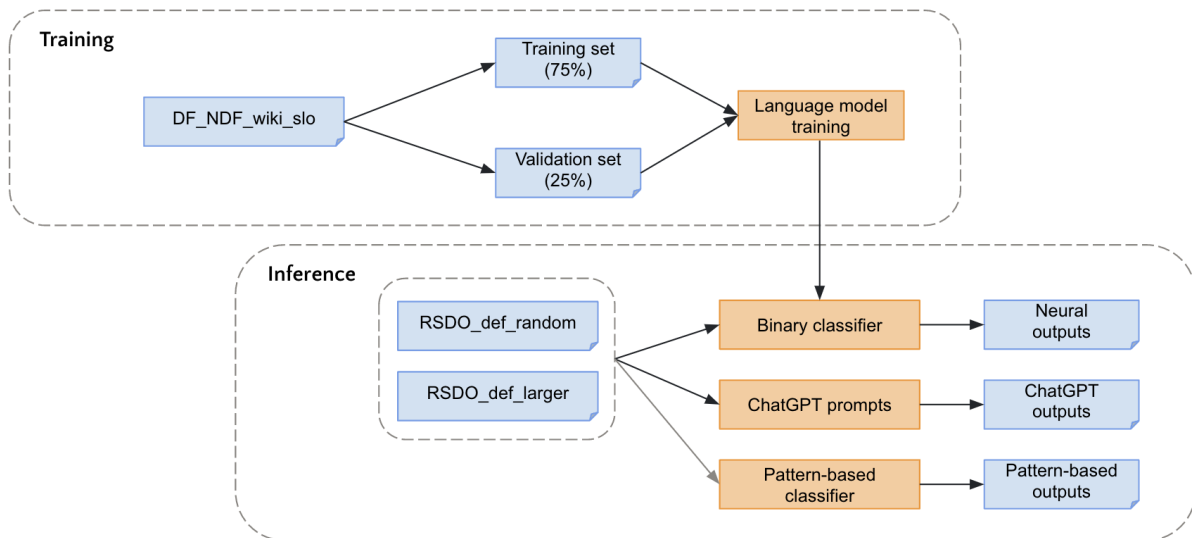


Figure 2: The experimental workflow.

4.1 Rule-based approach

We apply the pattern-based approach, based on a set of soft lexico-syntactic patterns, which was developed by Pollak (2014a,b). The patterns contain a combination of lexical cues (either on the lemma or token level) and information from morphosyntactic descriptions, more specifically part-of-speech and case information (e.g., to detect nominative case forms). In total, 12 soft patterns are defined to extract sentences of type (e.g., *NP-nom je* [Eng. *is*] *NP-nom*, *NP-nom se nanaša na* [Eng. *refers to*] *NP-nom*, *NP-nom pomeni* [Eng. *denotes*] *NP-nom*).

The simplest pattern is “*X je Y*” [“*X is Y*”], where X is the term (noun phrase) to be defined and Y is another noun phrase (usually its hypernym), followed by the differentia (part listing the differences from other types belonging to this class of entities). This corresponds to the genus and differentia definition type, meaning that if we have the term X to be defined, we define it by using its hypernym (Y) and by listing the differences from other types belonging to this class of entities (“X is Y that...”). Since Slovene is a highly inflected language, we can add the condition that the noun phrases should agree in

case and that the case should be nominative (i.e., “*NP-nom is NP-nom*”), where “*NP*” means noun phrase and “*NP-nom*” stands for noun phrase in the nominative case. There can be several variations, for example, including an English translation of a Slovene term. While the majority of the patterns are designed to extract genus and differentia type of definitions, other patterns are targeting Weak definitions, where patterns are designed to find paraphrases and synonym relations or cues of specific functional definitions (e.g., *Naloga [The task of] NP-gen is*). For a detailed description of the pattern-based approach, see Pollak (2014b).

In the evaluation, we consider two different pattern lists:

- *PatternAll* is the list of all 12 patterns;
- *JeStaSoPatterns* is a subset of the entire pattern set, containing only the patterns of “*X is a Y*” type and its variations.

4.2 Transformer classifiers

For the experiments described in this section, we compare different pretrained language models, one monolingual (SloBERTa) and three multilingual ones (mBERT, mDistilBERT, and XLMR)², which are fine-tuned for the definition classification task. Given that the size of our RSDO-def datasets is too small to use them for training a classifier, we use a dataset created from Wikipedia as training data, based on the work by Fišer et al. (2010).

4.2.1 Training datasets

For training our Transformers-based classifiers, we use the dataset *DF_NDF_wiki_slo* created from Wikipedia by Fišer et al. (2010). In the data construction process, the authors considered the first sentence of a Wikipedia article containing an entry term as a definition (Y) and other sentences with the same entry term as non-definitions (N). While this is not the ground truth, it can be considered a silver standard to be used as training data for ML approaches. We prepared the dataset for public release (Podpečan et al., 2023), available via CLARIN.SI: <http://hdl.handle.net/11356/1840>.

For our experiments, we compare two different samples for the negative class. In the dataset, all the sentences with the article key term but the first one, the non-definitions are labeled as N1, and in the version where the key term is not at the beginning of a sentence, these are labeled as N. The rationale of testing also the second approach, is that it is not impossible that non-first sentences in the Wikipedia articles are also definitions, and terms at the beginning of a sentence could indicate such examples, which would introduce noise when treated as a negative class. In total, 34,084 examples were collected, out of which 3,251 belong to the positive (definition) class Y, and 20,684 to the N1 class (non-definitions). In the second scenario, excluding the terms at the beginning of non-initial sentences, the distribution is 3,251 vs. 14,678, for definitions and non-definitions (N), respectively. When training a classifier, we compare both scenarios, one with the negative class of all non-first sentences with the term (N1) and one with only those that do not contain terms at the beginning of the sentence (N). The labeling ratio is presented in Figure 3 with free-text examples.

² Available in the HuggingFace library: <https://huggingface.co/models>

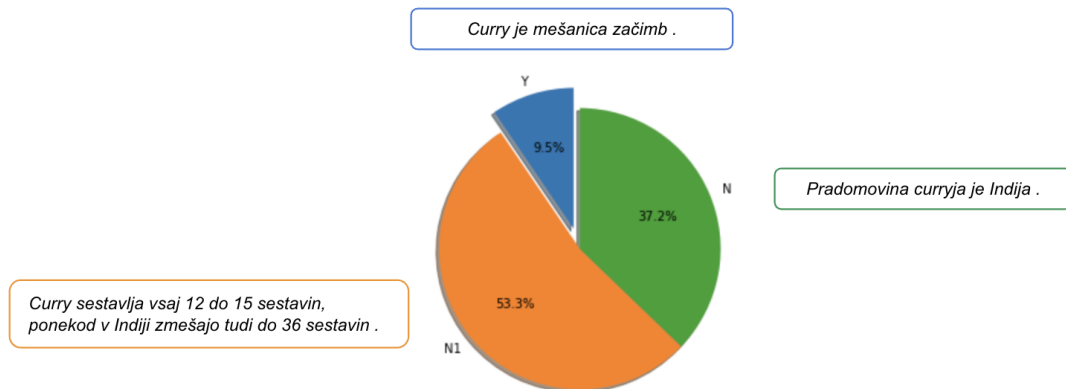


Figure 3: The distribution of each label in the DF_NDF_wiki_slo dataset.

4.2.2 Models

We consider one monolingual and three multilingual pretrained models that we fine-tune for the definition classification task using the training data described in Section 4.2.1.

SloBERTa³ (Ulčar & Robnik-Šikonja, 2021) is a cutting-edge Slovene language model trained as a masked language model, using *fairseq*⁴ toolkit. The corpora used for training the model have 3.47 billion tokens in total with a subword vocabulary of 32,000 tokens, making it a comprehensive resource for Slovene NLP research and development. The model’s performance on benchmark tests highlights its effectiveness for a wide range of NLP downstream tasks, especially with Slovene data.

mBERT⁵ (Kenton & Toutanova, 2019) is a multilingual Transformer-based model pretrained in a self-supervised regime on a massive corpus consisting of 104 languages using two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). While MLM randomly masks 15% of the words in the input sentences and then fed the entire masked sentence to the model to predict the masked words so that the model can learn the bidirectional representation of the given input sentences, NSP concatenates two masked sentences as input during pretraining and predicts if two sentences were following each other or not so that the model can an inner representation of the languages in the training set that can then be used to extract features useful for downstream tasks.

mDistilBERT⁶ (Sanh et al., 2019) is a distilled (smaller and faster) version of BERT, which uses BERT base model as a teacher and was pretrained on the same corpus as BERT in a self-supervised regime with three objectives: MLM as BERT, Distillation loss where the model was trained to return the same probabilities as BERT, and Cosine embedding loss where the model was also trained to generate hidden states as close as possible as the BERT. Similar to mBERT, the model was trained on a concatenated corpus of 104 different languages from Wikipedia. With only six layers, 768 dimensions, and 12 heads, mDistilBERT used 134M parameters in total, which is significantly fewer than the 177M

³ <https://huggingface.co/EMBEDDIA/sloberta>

⁴ <https://github.com/facebookresearch/fairseq>

⁵ <https://huggingface.co/bert-base-multilingual-uncased>

⁶ <https://huggingface.co/distilbert-base-uncased>

parameters used in mBERT-base. As a result, mDistilBERT is twice as fast as mBERT while maintaining a certain level of performance on various benchmark tests

XLMR⁷ (Conneau et al., 2019) is a multilingual RoBERTa-based version (Liu et al., 2019), which was pretrained on a vast corpus of 2.5TB of filtered CommonCrawl data, spanning 100 different languages. It uses the same training procedure as RoBERTa (Liu et al., 2019) which used only the Masked Language Model (MLM) technique without using Next Sentence Prediction (NSP) technique. The model was demonstrated to outperform other pretrained models on a variety of natural language understanding tasks (e.g., question answering, and natural language inference). This open-sourced model is also designed to be fine-tuned on specific tasks (e.g., NER, term extraction), making it a versatile tool for a wide range of NLP applications.

4.2.3 Implementation Details

We divide *DF_NDF_wiki_slo* into two parts: a training set and a test set with a ratio of 0.75: 0.25 in a stratified fashion, respectively. Class weighting is applied for each class in order that the classifier is aware of how to treat each class well in the cost function to improve the performance of the target.

As three labels are proposed in the training and test corpus, we build two separate classifiers: (1) one that predicts definitional sentences (Y) and non-definitional sentences which may also contain the term at the beginning of the sentence (N1); and (2) one that predicts between definitional sentences (Y) and non-definitional sentences that do not start with the key term (N). For each classifier, we evaluate the performance of our two evaluation datasets, *RSDO-def-random* and *RSDO-def-larger* with two distinctive settings: (1) in the strict evaluation scenario, considering Weak definitions as Non-definition (negative class); and (2) in the relaxed scenario, considering Weak definitions as Definitions (positive class).

The training and validation samples were binarized to the desired format. We fine-tuned weight decay and dropout coefficients due to high-performance costs. The learning rate was set equal to 1e-05. All models were trained for 5 epochs with a batch size of 8 and validation occurred at the end of each epoch.

4.3 ChatGPT

Introduced by OpenAI⁸ at The 36th Conference on Neural Information Processing Systems (NeurIPS⁹ 2022), ChatGPT quickly gained immense popularity with more than 1 million users in less than a week due to its ability to generate human-like and convincing responses. The underlying architecture of this conversational agent is GPT-3.5, a large generative pretrained Transformer model containing over 175 billion parameters. Figure 4 demonstrates how ChatGPT describes its capability in Definition Extraction.

We followed a straightforward zero-shot mechanism to classify a given Slovene sequence as a definition or not by accessing the ChatGPT via the official web interface¹⁰ with *ChatGPT*

⁷ <https://huggingface.co/xlm-roberta-base>

⁸ <https://openai.com/>

⁹ <https://nips.cc/>

¹⁰ <https://chat.openai.com>

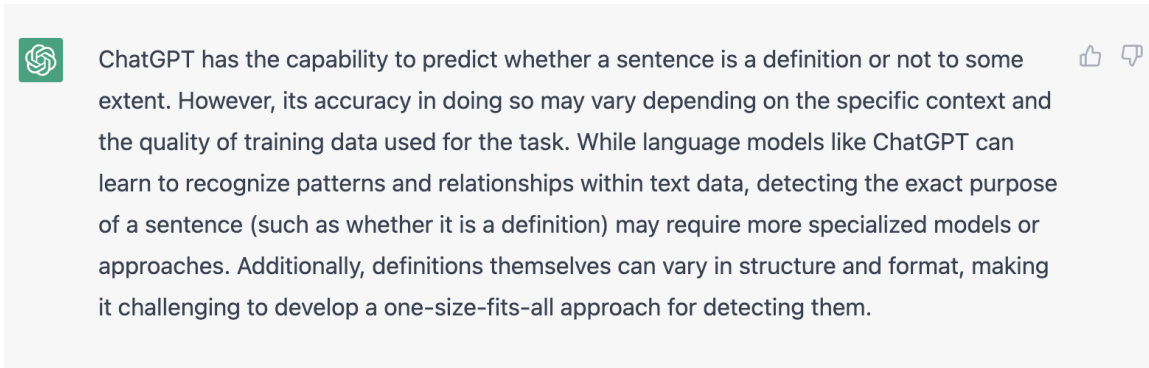


Figure 4: ChatGPT’s capability in Definition Extraction defined by ChatGPT.

Mar 23 Version¹¹ between 5th and 10th April 2023. We defined the vanilla prompt for both of our Slovene subsets as demonstrated in Figure 5.

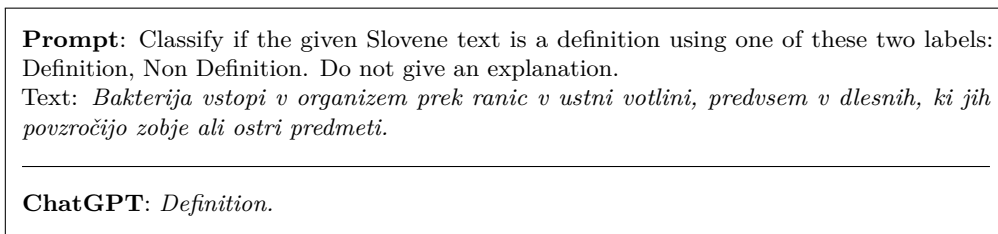


Figure 5: Slovene dataset prompt scenario

Besides the clear instruction on the specific task in our specific language, we also require ChatGPT not to explain in detail the reason why ChatGPT gives the labels for the given sequence so that we can capture only the necessary answer for the output format consistency.

4.4 Evaluation settings

Different approaches are evaluated on the two *RSDO-def* corpora (*RSDO-def-random* and *RSDO-def-larger*). We use two different evaluation scenarios. In the first, stricter setting, we consider Weak definitions as a negative class (Non-definitions), while in the second, relaxed setting, we consider Weak definitions as a positive class (Definitions), as they still produce very relevant content, but, for example, are not formulated as an intensional definition with the superordinate concept. For *RSDO-def-random*, we compare the performance among all the approaches—pattern-based, Transformer classifiers, and Chat-GPT—, whereas for *RSDO-def-larger*, we only consider Transformer classifiers and Chat-GPT results, as the pattern-based approach was used in the dataset construction phase.

As both corpora are imbalanced, we evaluate the performance of the classifier separately for each class by classifying all examples and comparing the predictions with the groundtruth using Precision (P), Recall (R), and F1-score (F1) for both the minority and majority

¹¹ <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

classes in each setting. Note that the minority (definition) class is the most important for our task. In addition to evaluating each class separately, we also calculate the macro-averaging of these three mentioned evaluation metrics, where no weights are applied for aggregation. Therefore, the evaluation metrics will have a bigger penalization when our classifier does not perform well with the minority classes.

From the terminographical perspective, we consider Precision more important than Recall. With possible integration into the Terminology portal in mind, where different features serve as a tool for compiling new terminological resources for human users, better Precision provides users with sentences that can be turned almost directly into definitions. Our quick survey among professional users revealed that they would only include definitions if they were informative enough, otherwise, they would settle for terms in the source language and terms in the target language without definitions, and not consider automated retrieval of sentences as a useful feature. Higher Precision of extracted sentences facilitates and simplifies the automatization of terminology management, as only with sufficient quality these tools are to be adopted by the users.

5. Results

We report the performance of different setups on *RSDO-def-random* and *RSDO-def-larger* dataset using P, R, and F1 for each class and macro-average —with best results in bold for each of them— in Tables 1 and 2, for the strict and the relaxed evaluation scenarios with regard to the weak definitions, respectively. We highlight the definition class results in each table, as this is the category of our main interest.

The results demonstrate that when the number of definitional samples is higher, as in the relaxed evaluation scenario, the language model tends to capture definitional sequences better than rule-based ones, while if the amount of definitions is very small (strict evaluation scenario), well-structured linguistic patterns used in the rule-based approaches have the advantage if we consider F1-score on the definition class.

Table 1: Comparative evaluation in Precision, Recall, and F1-score in the *strict* evaluation scenario, where we consider Weak definition as Non-definition.

Methods		RSDO-def-random									RSDO-def-large								
		Definition			Non-definition			Macro avg.			Definition			Non-definition			Macro avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pattern-based	JeStaSo	0.11	0.31	0.17	0.99	0.97	0.98	0.55	0.64	0.57	-	-	-	-	-	-	-	-	-
	Patterns all	0.11	0.31	0.16	0.99	0.96	0.98	0.55	0.64	0.57	-	-	-	-	-	-	-	-	-
Transformers Y/N classifier	SloBERTa	0.09	0.15	0.11	0.99	0.98	0.98	0.54	0.57	0.55	0.64	0.40	0.49	0.91	0.96	0.93	0.77	0.68	0.71
	mBERT	0.12	0.15	0.13	0.99	0.98	0.99	0.55	0.57	0.56	0.67	0.37	0.48	0.90	0.97	0.94	0.79	0.67	0.71
	DistilBERT	0.08	0.15	0.11	0.99	0.97	0.98	0.53	0.56	0.54	0.60	0.39	0.47	0.90	0.96	0.93	0.75	0.67	0.70
	XLM-R	0.09	0.15	0.11	0.99	0.98	0.98	0.54	0.57	0.55	0.65	0.41	0.50	0.91	0.96	0.94	0.78	0.69	0.72
Transformers Y/N1 classifier	SloBERTa*	0.14	0.15	0.15	0.99	0.99	0.99	0.57	0.57	0.57	0.68	0.27	0.39	0.89	0.98	0.93	0.79	0.63	0.66
	mBERT*	0.00	0.00	0.00	0.99	0.98	0.98	0.49	0.49	0.49	0.66	0.23	0.34	0.89	0.98	0.93	0.77	0.60	0.64
	DistilBERT*	0.00	0.00	0.00	0.99	0.98	0.98	0.49	0.49	0.49	0.64	0.25	0.35	0.89	0.98	0.93	0.76	0.61	0.64
	XLMR*	0.10	0.15	0.12	0.99	0.98	0.98	0.54	0.57	0.55	0.64	0.26	0.37	0.89	0.98	0.93	0.77	0.62	0.65
ChatGPT		0.03	0.93	0.06	1.00	0.54	0.70	0.51	0.73	0.38	0.22	0.78	0.34	0.94	0.54	0.68	0.58	0.66	0.51

In the strict evaluation scenario, where we consider Weak definition as Non-definition, the pattern-based JeStaso approach (Precision: 0.11, Recall: 0.31, F1-score: 0.17) surpasses the performance of Transformer classifiers (Y/N and Y/N1) that we proposed in Definition class for *RSDO-def-random* dataset. Despite the lower Recall and F1-score, the Precision of the SloBERTa model is higher, which is also the most important metric in the opinion of the terminographers involved in the terminological portal development. One should notice that this test set suffers from a significant lack of definitions (only 14 instances), which makes the quantitative results non-reliable. Meanwhile, in *RSDO-def-large*, pattern-based approaches were used for data preparation in the annotation process. Therefore, we do not consider the evaluation of the pattern-based methods. The best single neural classifier is XLMR with a Precision of 0.65, Recall of 0.41, and F1-score of 0.50 in predicting the Definition class. Despite lower Precision and F1-score in comparison with language models, ChatGPT dominates Recall with nine times higher in *RSDO-def-random* and three times higher in *RSDO-def-large*, but it strongly underperforms in terms of Precision.

In the relaxed evaluation scenario (see Table 2, where we consider Weak definition as definition, the mBERT Y/N classifier presents the best performance if we consider Precision as the most important metrics for the terminographers (Precision: 0.47, Recall: 0.08, F1-score: 0.13 in *RSDO-def-random*; Precision: 0.89, Recall: 0.22, F1-score: 0.35 in *RSDO-def-large*). Despite mBERT being the classifier with the best Precision, ChatGPT provides the best Recall and F1-score on the definition class, which is twice as high as other classifiers in both *RSDO-def-random* and *RSDO-def-large*. Note that both corpora, regardless of sharing the same characteristics of imbalance, have different proportions of Definitions, and the results are therefore not expected to be comparable.

Table 2: Comparative evaluation in Precision, Recall, and F1-score in the *relaxed* evaluation scenario, where we consider Weak definition as Definition.

Methods		RSDO-def-random									RSDO-def-large								
		Definition			Non-definition			Macro avg.			Definition			Non-definition			Macro avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pattern-based	JeStaSo	0.23	0.08	0.12	0.89	0.97	0.93	0.56	0.52	0.52	-	-	-	-	-	-	-	-	-
	Patterns all	0.21	0.08	0.11	0.89	0.96	0.93	0.55	0.52	0.52	-	-	-	-	-	-	-	-	-
Transformers	SloBERTa	0.36	0.08	0.13	0.90	0.98	0.94	0.63	0.53	0.53	0.86	0.24	0.38	0.74	0.98	0.84	0.80	0.61	0.61
	mBERT	0.47	0.08	0.13	0.90	0.99	0.94	0.68	0.53	0.54	0.89	0.22	0.35	0.73	0.99	0.84	0.81	0.60	0.60
Y/N classifier	DistilBERT	0.40	0.10	0.16	0.90	0.98	0.94	0.65	0.54	0.55	0.85	0.24	0.38	0.74	0.98	0.84	0.79	0.61	0.61
	XLM-R	0.36	0.08	0.13	0.90	0.98	0.94	0.63	0.53	0.53	0.87	0.24	0.38	0.73	0.98	0.84	0.80	0.61	0.61
Transformers	SloBERTa*	0.43	0.06	0.10	0.89	0.99	0.94	0.66	0.52	0.52	0.86	0.15	0.26	0.71	0.99	0.83	0.79	0.57	0.54
	mBERT*	0.19	0.03	0.05	0.89	0.98	0.93	0.54	0.51	0.49	0.79	0.12	0.21	0.71	0.99	0.82	0.75	0.55	0.52
Y/N1 classifier	DistilBERT*	0.00	0.00	0.00	0.89	0.99	0.93	0.44	0.49	0.47	0.83	0.14	0.24	0.71	0.99	0.83	0.77	0.56	0.53
	XLMR*	0.30	0.06	0.10	0.89	0.98	0.94	0.60	0.52	0.52	0.78	0.14	0.24	0.71	0.98	0.82	0.75	0.56	0.53
ChatGPT		0.17	0.68	0.27	0.93	0.56	0.70	0.55	0.62	0.49	0.43	0.68	0.53	0.79	0.57	0.67	0.61	0.63	0.60

6. Error Analysis

6.1 Pattern-based classifier

In the analysis, we observed several sources of errors in the pattern-based approaches.

First, when sequences are too long, contain lists of items, or have wrong sentence segmentation, the segments rarely contain definitions. Multiple sentences instead of one pose also a problem with regard to evaluation.

Second, there are several examples formulated as questions, but still matching the patterns (e.g., “*Kolikšen delež besed je enopojavnic?*”). These typical syntax errors contained no definitions at all and were also the most common error type. These kinds of errors could be easily removed in the future adaptation of a pattern-based approach. There are examples, where insufficient context is provided, or when a definition is in fact correct, but not related to the term of our interest.

Other sources of errors related to insufficient context for the extracted term, too general phrases, and errors in automated morpho-syntactic annotations as a source of wrongly extracted sentences.

6.2 XLMR classifier

In this section, we discuss the behavior of our best single classifier (XLMR) trained on *DF_NDF_wiki_slo* where Weak definition is considered as Non-definition, regarding F1-score of Definition label.

Figure 6 presents the Kernel Distribution Estimation (KDE) plot, which is often used for depicting the probability density function of the continuous or non-parametric data variables. Here, we plot the density distribution of each sequence regarding the sequence length where the left side of the figure refers to wrong predictions and the right side to the true prediction of the XLMR classifier. The horizontal or x-axis presents the range of values for sequence length in the data set while the vertical or y-axis in a density plot is the probability density function for the kernel density estimation.

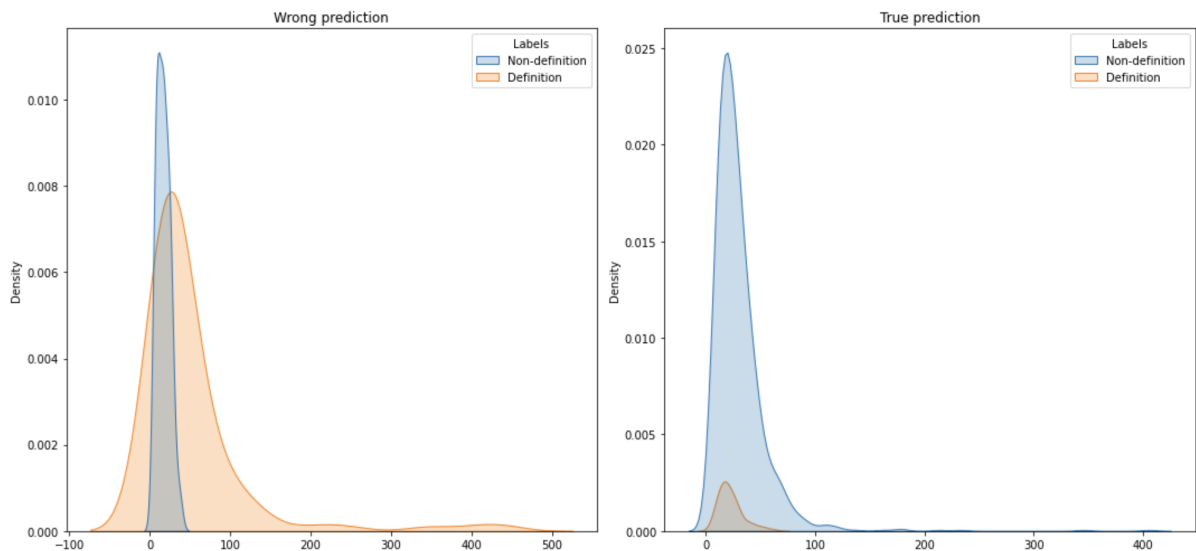


Figure 6: The distribution of sequence length each label wrongly predicted when using XLMR classifier trained on *DF_NDF_wiki_slo* in relaxed definition evaluation scenario.

Despite the infrequency of definitions being present in lengthy sequences, the classifier’s performance was impeded by data poisoning, resulting in the misclassification of instances in both classes. This anomaly is particularly evident in the misidentification of definition classes, where the model erroneously assigned examples with a long right tail to the definition class.

Table 3 lists randomly selected examples of the incorrectly classified sentences from *RSDO-def-large* when we use XLMR classifier trained on *DF_NDF_wiki_slo* where Weak definition is considered as non-definition. Looking at these sequences, we can see that some gold standard definitions are either incorrectly labeled or very difficult to classify even for a human.

For instance, the sentence “*Steklina je zoonoza , prisotna na vseh celinah in ugotovljena že v najmanj 150 državah sveta .*” (translated version: *Rabies is a zoonosis, present on all continents and found in at least 150 countries of the world.*) contains a term and its hypernym, but is not in the definition class, as the differentia part is not well expressed.

Table 3: Examples of the incorrect classified short sentence from *RSDO-def-large* using XLMR classifier trained on *DF_NDF_wiki_slo* where Weak definition is considered as non-definition.

Sentences	Labels	Predictions
Bruceloza je nalezljiva bolezen , ki se pogosto pojavlja pri kozah , ovce so nekoliko manj dovzetne .	0	1
Terminologija in sodobna terminografija (ur . Nina Ledinek , Mojca Žagar Karer in Marjeta Humar) .	0	1
Hitrost je fizikalna količina , ki nam pove , koliko se telo premakne na časovno enoto , in ima enoto [m/s] .	0	1
Prenašalci so okuženi konji in žrebci .	0	1
Steklina je zoonoza , prisotna na vseh celinah in ugotovljena že v najmanj 150 državah sveta .	0	1
Bakterijska bolezen , ki se pojavlja v obliki septikemije , hemoglobinurije , ikterusa in abortusa .	0	1
Slovenija je država , uradno prosta bruceloze .	0	1
...
4 . Amonijak je plin značilnega neprijetnega vonja , ki je dobro topen v vodi .	1	0
V glavi je onaglašena iztočnica z morebitnim izgovorom in slovničnimi podatki .	1	0
Terminov je v praktičnostrokovnih besedilih manj kot v drugih skupinah besedil .	1	0
Sila podlage Fp je reakcija , ki jo povzroči podlaga na opazovano telo .	1	0
Na primer , pojmovni sistem elektrotehnike je množica pojmov , ki določajo področje .	1	0
Izjema je → Toplarna Ljubljana , ki je namenjena tudi oskrbovanju mesta s toplo vodo .	1	0
Glavni predstavnik didaktičnostrokovnih besedil je učbenik (osnovnošolski , srednješolski , univerzitetni) .	1	0
Ponazorimo jih lahko s stavkom a je vrsta b . Gre za nadredne , podredne in priredne odnose .	1	0
3 . Najbolj pričakovano okolje terminov so strokovna besedila .	1	0
Izobraževalni proces je osebni odnos , didaktična in socialna interakcija .	1	0
Obe spojini sta klorida – soli klorovodikove kisline HCl .	1	0
Teoretična in empirična podlaga vprašalnika so sodobne motivacijske teorije (Juriševič , 2006) .	1	0
Izkustveno učenje je proces , ki poteka celo življenje .	1	0
Predilecijska mesta za razmnoževanje so možgani , stena prebavil in uterus .	1	0
Klinični znaki so anoreksija , depresija , povišana telesna temperatura .	1	0
Značilen znak je anemija , smrtnost pa je nizka .	1	0
Najpogostejša metoda za omamljanje kopitarjev je uporaba strelne naprave s penetrirnim klinom .	1	0
Najprimernejše mesto strela je vertikalna sredinska linija čela od 1 do 2 cm nad očmi .	1	0
...

Beside, above all the misclassified instances where the model failed to predict the true labels for definition, 24.5 % of the false negative belongs to the instances which constitutes from multiple sentences. As an example, this is a multiple-sentence instance: “*Imenujemo ga udarni zvok. « (52) ; » Kontaktor je mehanski stikalni aparat , ki ima samo en*

*mirovni položaj , ki ga ne upravljamo ročno , in je sposoben vklapljeti , prevajati in izklapljeti tok v normalnih pogojih obratovanja , upoštevajoč tudi preobremenitve. « (53) ; » Rele je električna naprava , ki povzroči v odvisnosti od spremenljive električne ali druge neelektrične veličine določeno spremembo v istem ali v drugih električnih tokokrogih. « (54) ; » Talilna varovalka je v bistvu namenoma ustvarjena šibka točka (močno zmanjšan prerez vodnika , poskus 4.5.2) na dostopovnem mestu električnega kroga , njeno delovanje pa temelji na odvisnosti toplotnega učinka od gostote toka. « (55) ; » Žirator je vezje , ki omogoča pretvorbo poljubne znane impedance v njeno dualno obliko ali inverzno vrednost. « (56) ... » (translated version: *We call it a percussive sound. « (52) ; » A contactor is a mechanical switching device that has only one rest position, which is not operated manually, and is capable of switching on, transferring and switching off the current under normal operating conditions, taking into account overloads as well. « (53) ; » A relay is an electrical device that causes, depending on a variable electrical or other non-electric quantity, a certain change in the same or other electrical circuits. « (54) ; » The fuse is basically a purposely created weak point (highly reduced conductor cross-section, experiment 4.5.2) at the access point of the electric circuit, and its operation is based on the dependence of the thermal effect on the current density. « (55) ; » A gyrator is a circuit that allows the conversion of any known impedance into its dual form or inverse value. « (56) ... »).**

This type of error could be removed by post-processing rules, based on the sequence length.

6.3 ChatGPT's prompting

Similar to the problems faced by sequence classifier, there exist multiple instances of ambiguity, which affect the model performance. For example, in Figure 7, the sentence *“Danes uporabljamo pretežno tripolne (ali kot jih običajno imenujemo, dvopolne z zascitnim kontaktom) . « (53) ; » Kontaktor je stikalni aparat, ki ga vklapljammo daljinsko (zato ga pogosto imenujemo tudi daljinsko stikalo) . « (53) ; » Oporniki ali izolatorniki so jekleni nosilci okrogle ali pravokotne oblike , vroce pocinkani, s pomocjo katerih pritrjujemo izolatorje na konzole , lesene drogove ali jeklene konstrukcije. « (54) ; » Mocnostni ali smerni releii”* (translated: *“Today we mostly use three-pole (or as they are usually called, two-pole with a protective contact). ” (53) ; » A contactor is a switching device that can be turned on remotely (that’s why it is often called a remote switch). ”(53) ; » Supports or insulators are steel supports of round or rectangular shape, hot-dip galvanized, with the help of which insulators are attached to consoles, wooden poles or steel structures. ” (54) ; » Power or direction relays”*) was detected by ChatGPT as a group of different sentences separated the full stop, and therefore, was tagged with the predicted labels separately. Additional major voting is applied for this scenario to create a comparative result with pattern-based and token classification approaches.

Although we do not require ChatGPT to explain in detail the reason behind the label choice, the chatbot still gave some explanation as demonstrated in Figure 8.

7. Conclusion and Further Work

We presented a novel evaluation dataset for Slovene Definition Extraction and proposed the benchmarks given three mechanisms: rule-based extractor, Transformer-based classifier, and prompt-based extractor with ChatGPT. Regarding the newly introduced corpus,

An example of ChatGPT’s response for the ambiguous cases.

Prompt: Classify if the given Slovene text is a definition using one of these two labels: Definition, Non Definition. Do not give an explanation. Text: *Danes uporabljamo pretežno tripolne (ali kot jih obicajno imenujemo, dvopolne z zascitnim kontaktom) . « (53) ; » Kontaktor je stikalni aparat, ki ga vklapljammo daljinsko (zato ga pogosto imenujemo tudi daljinsko stikalo) . «(53) ; » Oporniki ali izolatorniki so jekleni nosilci okrogle ali pravokotne oblike , vroce pocinkani, s pomocjo katerih pritrjujemo izolatorje na konzole , lesene drogove ali jeklene konstrukcije. « (54) ; » Mocnostni ali smerni releii*

ChatGPT: *Definition, Definition, Definition, Non Definition.*

Groundtruth: *Definition.*

Figure 7: Slovene dataset prompt gave multiple labels for a given text.

An example of ChatGPT’s response for the fail cases with explanation.

Prompt: Classify if the given Slovene text is a definition using one of these two labels: Definition, Non Definition. Do not give an explanation. Text: *Prav tako je za tipizacijo virusov stekline z monoklonalnimi protitelesi pomembna sestava aminokislinskega zaporedja nukleoproteina, saj imajo razlike v aminokislinskem zaporedju lahko za posledico tudi razlike v epitopih, kamor se vezejo monoklonalna protitelesa (Flamand in sod., 1980a, 1980b; Dietzschold in sod., 1987; Smith, 1989).*

ChatGPT: *Definition. The text explains the importance of the amino acid sequence composition of the nucleoprotein for the typing of rabies viruses using monoclonal antibodies, as differences in the amino acid sequence can result in differences in the epitopes to which monoclonal antibodies bind.*

Groundtruth: *Non Definition.*

Figure 8: Slovene dataset prompt answers with explanation.

RSDO-def-random and *RSDO-def-larger* were collected using random sampling techniques for the first, and including also the pattern-based sentence selection method for the second. Regarding the methods, if there are only a few well-structured instances of definitions that have clear linguistic characteristics (e.g., in the strict evaluation scenario, where Weak definitions are considered as non-definitions), a rule-based technique performed better in terms of F1-score (on the Definition class) than language models or prompting. However, for less structured examples (relaxed evaluation scenario with Weak definitions considered as definitions), ChatGPT prompting and language models were more effective than classical rule-based approaches. When comparing prompting and language model classifiers, for the Definition class, classifiers lead to higher Precision, while in terms of Recall, ChatGPT has better results.

The usefulness of the Definition Extraction process in practical applications is severalfold. In our work, we focused more specifically on terminological applications, as the definitions are extracted from the domain-specific corpora. The pattern-based approach is already implemented as part of the Slovene terminological portal as a tool for providing good examples and inspiring the users in their manual Definition Extraction process. In this paper, we also show the potential of language-model-based classifiers, that will be considered for inclusion after additional analysis.

There are also several points that have to be considered in using automated processes in terminographical work. First, for obtaining usable and applicable results, the user must

prepare a specialized corpus for the given domain. If the texts are not representative and include noisy material, also the Definition Extraction process will likely be not relevant. However, approaches that are not considering the task as extraction but as generation could be considered to overcome this limitation.

Next, we should also note that although Definition Extraction can be very useful for the manual definition process, as the Precision of the systems is still far from perfect, the users might be tempted to use the unmodified examples as final definitions and not use the tool as support. Therefore, in our project, we consider the output of Definition Extraction systems as good examples and do not call them definitions.

Our research provides the first study in neural Definition Extraction for Slovene and a multi-domain dataset for Definition Extraction evaluation. In future work, we plan to investigate several directions. First, increasing the size and diversity of the evaluation dataset to improve the reliability of the quantitative results is of crucial importance. Second, one of our goals is to develop a larger collection of labeled sentences that could be used not only for evaluation but also for training Definition Extraction systems. Next, as the annotation is often time- and effort-consuming, active learning can be considered to more efficiently use labeled examples and reduce human efforts. Furthermore, investigating the use of ensemble methods regarding the combination of different meaning information (e.g., local, global, contextual) and/or the combination of multiple models may help improve the overall performance. Last but not least, while in our study, we used ChatGPT for classifying sentences, we plan to leverage the large generative language models also for definition generation.

The definition evaluation gold standard dataset [Jemec Tomazin et al. \(2023\)](#) is publicly available at <http://hdl.handle.net/11356/1841>, the pattern-based approach from [Pollak \(2014a\)](#) is available at https://github.com/vpodpecan/definition_extraction and the code for the language models classifiers at https://github.com/honghanhh/definition_extraction. We also released the silver standard training data ([Podpečan et al., 2023](#)) by [Fišer et al. \(2010\)](#), which is now available via CLARIN.SI: <http://hdl.handle.net/11356/1840>.

8. Acknowledgements

The work was partially supported by the Slovene Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103), and the Ministry of Culture of the Republic of Slovenia through the project Development of Slovene in Digital Environment (RSDO), as well as the project Formant combinatorics in Slovenian (J6-3131). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

9. References

Anke, L.E. & Schockaert, S. (2018). Syntactically Aware Neural Architectures for Definition Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. pp. 378–385.

- Avram, A.M., Cercel, D.C. & Chiru, C. (2020). UPB at SemEval-2020 Task 6: Pretrained Language Models for Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 737–745.
- Borg, C., Rosner, M. & Pace, G. (2009). Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop on Definition Extraction*. pp. 26–32.
- Bovi, C.D., Telesca, L. & Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3, pp. 529–543.
- Caspani, F., Ratnamogan, P., Linger, M. & Hajaiej, M. (2020). ACNLP at SemEval-2020 task 6: A supervised approach for definition extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 479–486.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Cui, H., Kan, M.Y. & Chua, T.S. (2004). Unsupervised learning of soft patterns for generating definitions from online news. In *Proceedings of the 13th international conference on World Wide Web*. pp. 90–99.
- Cui, H., Kan, M.Y. & Chua, T.S. (2005). Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 384–391.
- Davletov, A., Arefyev, N., Shatilov, A., Gordeev, D. & Rey, A. (2020). Gorynych Transformer at SemEval-2020 Task 6: Multi-task Learning for Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 487–493.
- De Benedictis, F., Faralli, S. & Navigli, R. (2013). Glossboot: Bootstrapping multilingual domain glossaries from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 528–538.
- Espinosa-Anke, L. & Saggion, H. (2014). Applying dependency relations to definition extraction. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*. Springer, pp. 63–74.
- Espinosa-Anke, L., Saggion, H. & Delli Bovi, C. (2015). Definition extraction using sense-based embeddings. In *Gupta P, Banchs RE, Rosso P, editors. International Workshop on Embeddings and Semantics (IWES'15); 2015 Sept 15; Alicante, Spain.[Place unknown]:[CEUR]; 2015.[6 p.]*. CEUR.
- Espinosa-Anke, L., Saggion, H., Ronzano, F. & Navigli, R. (2016). Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Fahmi, I. & Bouma, G. (2006). Learning to identify definitions using syntactic features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*.
- Fišer, D., Pollak, S. & Vintar, S. (2010). Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. In *LREC*. Citeseer.
- Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.
- Izmailov, P., Wilson, A., Podoprikin, D., Vetrov, D. & Garipov, T. (2018). Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. pp. 876–885.
- Jeawak, S., Anke, L.E. & Schockaert, S. (2020). Cardiff university at semeval-2020 task 6: Fine-tuning bert for domain-specific definition classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 361–366.

- Jemec Tomazin, M., Podpečan, V., Pollak, S., Thi Hong Tran, H., Fajfar, T., Atelšek, S., Sitar, J. & Žagar Karer, M. (2023). Slovenian Definition Extraction evaluation datasets RSDO-def 1.0. URL <http://hdl.handle.net/11356/1841>. Slovenian language resource repository CLARIN.SI.
- Jemec Tomazin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T. & Erjavec, T. (2021). Corpus of term-annotated texts RSDO5 1.0.
- Jin, Y., Kan, M.Y., Ng, J.P. & He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pp. 780–790.
- Kannan, M. & Ponnusamy, H.S. (2020). TüKaPo at SemEval-2020 Task 6: Def (n) tly Not BERT: Definition Extraction Using pre-BERT Methods in a post-BERT World. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 724–729.
- Kaparina, J. & Soboleva, A. (2020). DeftPunk at SemEval-2020 Task 6: Using RNN-ensemble for the Sentence Classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 697–703.
- Kenton, J.D.M.W.C. & Toutanova, L.K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. pp. 4171–4186.
- Klavans, J.L. & Muresan, S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 324.
- Klavans, J.L. & Muresan, S. (2002). A method for automatically building and evaluating dictionary resources.
- Li, S., Xu, B. & Chung, T.L. (2016). Definition extraction with lstm recurrent neural networks. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, pp. 177–189.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Navigli, R. & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. pp. 1318–1327.
- Park, Y., Byrd, R.J. & Boguraev, B. (2002). Automatic Glossary Extraction: Beyond Terminology Identification. In *COLING*, volume 10. pp. 1072228–1072370.
- Podpečan, V., Pollak, S., Fišer, D., Vintar, Š. & Tran, T.H.H. (2023). Slovenian Definition Extraction training dataset DF_NDF_wiki_slo 1.0. URL <http://hdl.handle.net/11356/1840>. Slovenian language resource repository CLARIN.SI.
- Pollak, S. (2014a). Extracting definition candidates from specialized corpora. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 2(1), pp. 1–40.
- Pollak, S. (2014b). Semi-automatic domain modeling from multilingual corpora. *PhD thesis*. Ljubljana: Department of Translation Studies, Faculty of Arts.
- Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex*, 2019, pp. 934–956.
- Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, N. & Vintar, S. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In J. Jancsary (ed.) *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*. ÖGAI, Wien, Österreich, pp. 53–60. URL http://www.oegai.at/konvens2012/proceedings/10_pollak12o/.

- Ranasinghe, T., Plum, A., Orašan, C. & Mitkov, R. (2020). RGCL at SemEval-2020 Task 6: Neural Approaches to Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 717–723.
- Reiplinger, M., Schäfer, U. & Wolska, M. (2012). Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 special workshop on rediscovering 50 years of discoveries*. pp. 55–65.
- Saggion, H. & Gaizauskas, R.J. (2004). Mining On-line Sources for Definition Knowledge. In *FLAIRS Conference*. pp. 61–66.
- Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Sarmiento, L., Maia, B., Santos, D., Pinto, A. & Cabral, L. (2006). Corpógrafo v3: From terminological aid to semi-automatic knowledge engine. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.
- Singh, A., Kumar, P. & Sinha, A. (2020). DSC IIT-ISM at SemEval-2020 Task 6: Boosting BERT with Dependencies for Definition Extraction. In *SemEval@ COLING*. pp. 710–716.
- Spala, S., Miller, N.A., Dernoncourt, F. & Dockhorn, C. (2020). Semeval-2020 task 6: Definition extraction from free text with the deft corpus. *arXiv preprint arXiv:2008.13694*.
- Storrer, A. & Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.
- Ulčar, M. & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.
- Velardi, P., Faralli, S. & Navigli, R. (2013). Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3), pp. 665–707.
- Veyseh, A., Dernoncourt, F., Dou, D. & Nguyen, T. (2020). A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. pp. 9098–9105.
- Vintar, Š. & Martinc, M. (2022). Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28.
- Westerhout, E. (2009). Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*. pp. 61–67.
- Zhang, H. & Ren, F. (2020). Bertatde at semeval-2020 task 6: Extracting term-definition pairs in free text using pre-trained model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 690–696.