SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation

Matej Klemen, Aleš Žagar, Jaka Čibej, Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
{matej.klemen, ales.zagar, jaka.cibej, marko.robnik}@fri.uni-lj.si

Abstract

Natural language inference (NLI) is an important language understanding benchmark. Two deficiencies of this benchmark are: i) most existing NLI datasets exist for English and a few other well-resourced languages, and ii) most NLI datasets are formed with a narrow set of annotators' instructions, allowing the prediction models to capture linguistic clues instead of measuring true reasoning capability. We address both issues and introduce SI-NLI, the first dataset for Slovene natural language inference. The dataset is constructed from scratch using knowledgeable annotators with carefully crafted guidelines aiming to avoid commonly encountered problems in existing NLI datasets. We also manually translate the SI-NLI to English to enable cross-lingual model training and evaluation. Using the newly created dataset and its translation, we train and evaluate a variety of large transformer language models in a monolingual and cross-lingual setting. The results indicate that larger models, in general, achieve better performance. The qualitative analysis shows that the SI-NLI dataset is diverse and that there remains plenty of room for improvement even for the largest models.

Keywords: natural language inference, Slovene, cross-lingual, transformers

1. Introduction

Natural language processing (NLP) is a rapidly developing area, with many new ever-more capable language models (LMs) being regularly introduced. While initial development focused primarily on processing a small pool of broadly spoken languages such as English and Chinese, later development started investing effort in a broader set of languages, as well as support for multilinguality and cross-linguality. This enables a more complete evaluation of the models as different languages contain different linguistic phenomena that may have a notable effect on the difficulty of tasks and consequent models' performance. As the datasets in less-resourced languages are commonly smaller, multilingual evaluation may also reveal performance discrepancies due to the data requirements of the models.

The language coverage varies significantly between NLP tasks: while certain tasks such as POS-tagging and dependency parsing have a wide coverage due to extensive international projects such as Universal Dependencies (Nivre et al., 2020), most tasks cover significantly fewer languages. In our work, we focus on extending the resources for natural language inference (NLI), an important semantic task with low coverage in languages other than English (see Section 2 for examples). NLI is an extension of the textual entailment recognition task (Dagan et al., 2006): given a premise and a hypothesis text, the goal is to determine whether the hypothesis is definitely true given the premise (entailment, E), definitely wrong given the premise (contradiction, C), or the relation is not decisive (neutral, N).

To extend the coverage of resources to less-resourced languages, authors use techniques such as translation of existing datasets for the same task into a new language (Obadić et al., 2023), recasting existing datasets (Uppal et al., 2020) for a different task by transforming the target class via a known relation, and construction of a new resource from scratch (Hu et al., 2020). While the translation approach is convenient as it can be done (semi-) automatically, the constructed dataset may not fully represent the complexity of the target language, especially if machine translation is in question. For example, a translation of an English dataset into Slovene would not produce texts involving the dual grammatical number as the English language does not feature this phenomenon.

Data recasting does not suffer from this issue, but it may produce datasets that do not fully represent the complexity of the task. For example, Uppal et al. (2020) convert a sentiment classification dataset into a natural language inference dataset by converting each sentiment class into a hypothesis via a template, e.g., "This product got positive reviews". As a result, the hypothesis in the produced dataset can only be one of a handful of texts, while the hypothesis, in general, can be an arbitrary text.

In our work, we opt for the third option, i.e., we construct a new resource from scratch. We build it using the hypothesis editing protocol described by Bowman et al. (2020) that has shown promise in reducing the amount of unwanted statistical cues enabling models to learn shortcuts

LREC-COLING 2024, pages 14859–14870

(Geirhos et al., 2020). More specifically, we gather candidate premises and hypotheses from publicly available Slovene reference corpus ccKres (Logar et al., 2013), and ask linguist students to edit or rewrite the hypothesis for each of the three NLI relations, following customized annotation guidelines designed to warn against pitfalls of existing NLI datasets.

In summary, we make the following contributions:

1. We introduce SI-NLI (Klemen et al., 2022), the first dataset for Slovene natural language inference. The dataset is constructed from scratch and aims to avoid pitfalls commonly encountered in existing NLI datasets for other languages. In addition, we manually translate the dataset to English to enable cross-lingual model training and evaluation.
2. Using the SI-NLI dataset, we train and evaluate a variety of language models in a monolingual, as well as cross-lingual (from Slovene to English) setting, setting baseline results and analyzing the model performance.

The remainder of this paper is structured as follows. In Section 2, we overview existing NLI resources. In Section 3, we describe the construction of the SI-NLI dataset and its essential statistics. In Section 4, we analyze the accuracy of language models using our resource. In Section 5, we summarize the findings and suggest possible directions for further work.

2. Related Work

Designing systems that can capture the meaning of the text is at the core of artificial intelligence and natural language processing. Motivated by the importance of this aspect, Dagan et al. (2006) proposed textual entailment recognition as an abstract task to compare how well different systems capture the meaning, and released an English dataset. Given two texts, the task is to determine if the meaning of the second text can be inferred from the first text (entailment) or not (non-entailment). In the following years, shared tasks such as the PASCAL Recognising Textual Entailment Challenge (Dagan et al., 2006) and SemEval-2014 Task 1 (Marelli et al., 2014) have contributed to the rapid development of semantic systems and the popularity of the task. Motivated in part by the introduction of data-hungry neural models, Bowman et al. (2015) released the large-scale Stanford NLI dataset. To account for cases where the relation cannot be determined certainly, the authors used a three-class annotation scheme by dividing non-entailment into a contradiction and neutral class, and started the popularization of the modified task of natural language inference. As the dataset only covers one genre of texts, the Multi-Genre NLI (MNLI) dataset (Williams et al., 2018) was released as a generalized option for NLI, covering ten genres instead. Using the same procedure as in MNLI, Conneau et al. (2018) collected additional validation and test examples, and manually translated them into 15 languages, enabling the evaluation of multilingual and cross-lingual systems. NLI datasets were introduced for several other languages such as Turkish (Budur et al., 2020), Korean (Ham et al., 2020), Persian (Khashabi et al., 2021), Croatian (Obadić et al., 2023), and Chinese (Hu et al., 2020). The datasets are typically constructed either by translating existing datasets such as XNLI (Obadić et al., 2023), recasting data primarily used for other tasks such as sentiment classification (Uppal et al., 2020), or from scratch using a custom annotation procedure (Hu et al., 2020).

Despite becoming popular evaluation benchmarks, many NLI datasets suffer from annotation artifacts, which enable performing NLI using shortcuts. For example, Gururangan et al. (2018) mention the issue of keywords that are very indicative of a class, such as negation words for contradiction, and the hypothesis-only bias, due to which models are able to correctly classify text pair relations using only one of the texts. Constructing NLI datasets from scratch, authors have tried to mitigate the number of artifacts, e.g., by using professional annotators, modified annotation guidelines, or an alternative dataset construction protocol (Parrish et al., 2021; Hu et al., 2020; Bowman et al., 2020).

In our work, we use all three options to create a quality resource: (1) we use a hypothesis editing dataset construction protocol, which has previously shown promise in reducing some artifacts (Bowman et al., 2020); (2) we use skilled linguist students instead of crowdsourcing; (3) we design annotation guidelines that warn against common artifacts in existing datasets. Our main focus in this work is not the evaluation of our annotation procedure, but an introduction of a NLI dataset for a previously unsupported morphologically rich language (Slovene), as well as the demonstration of the usability of the dataset.

3. The SI-NLI Dataset

An outline of the dataset construction process is shown in Figure 1: we construct it by sourcing pairs of sentences with similar meaning from a publicly available Slovene reference corpus ccKres (Logar et al., 2013) as the seed premises and hypotheses, and asking skilled annotators to edit the hypotheses three times, once for each of the three NLI relations, following customized annotation guidelines in the process. By sourcing semantically similar pairs from a reference corpus and asking multiple native

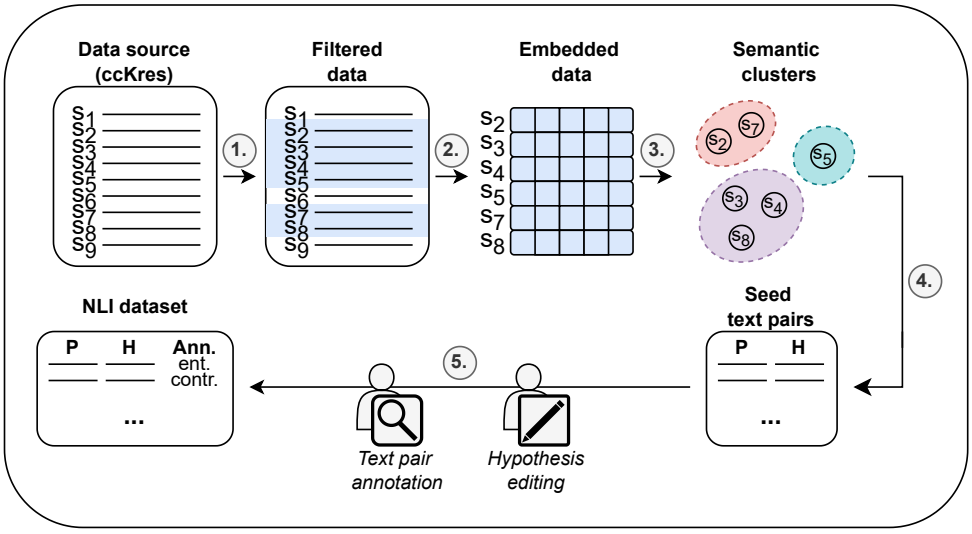

Figure 1: An overview of the SI-NLI construction process. We start with an open data source, from which we filter out (1) irrelevant sentences with inadequate structures, embed the remaining sentences (2), and cluster them (3). We obtain groups of semantically similar sentences, from which we sample pairs (4). These are handled by human annotators through hypothesis editing and text pair annotation (5) to produce SI-NLI.

speakers to construct examples, we strive towards capturing the realistic complexity of the Slovene language which might not be present when translating an existing resource in another language. We describe the steps in more detail in Sections 3.1, 3.2, and 3.3.

3.1. Sourcing the pairs

To allow the public distribution of our dataset, we use the 10-million token Slovene reference corpus ccKres 1.0 (Logar et al., 2013) whose license allows free non-commercial use of the data. As the corpus also contains sentences that are not suitable for our task, such as partial sentences (with no verbs) or sentences consisting of only several tokens (signatures, dates, etc.), we keep only sentences with between 10 and 40 tokens that contain a verb and at least one noun, determiner, proper noun, or pronoun.

At this stage, we could sample premises from the filtered data and ask the annotators to construct the hypotheses. However, this unconstrained setting did not work well in previous dataset construction attempts, as annotators resort to a few standard patterns of entailment and contradiction. Therefore, we decided to further guide and diversify the annotation procedure by providing annotators seed premises as well as hypotheses, and asking them to modify the hypotheses according to the desired class.

To construct hypotheses, we embedded the sentences using the Language agnostic BERT Sentence Embeddings (Feng et al., 2022), projected the embeddings to a lower dimensionality (100 dimensions) using PCA (Pearson, 1901), and clustered the embeddings using DBSCAN (Ester et al., 1996) clustering algorithm[1]. Then, we sampled non-overlapping sentence pairs from the obtained clusters. For example, if a cluster contained five sentences, we constructed two pairs, while one remained unused. We obtained a large pool of sentence pairs, from which we drew data for annotation according to our budget.

3.2. Annotation Process

The annotation process was divided into several steps: (1) preliminary annotation, which we performed ourselves to produce and refine a set of reliable guidelines that the annotators could follow; (2) introductory and training sessions for annotators; (3) hypothesis editing/formation by annotators; and (4) cross-checking of hypotheses by annotators. We summarize them in the following subsections.

3.2.1. Preliminary Annotation and Guidelines

The review of related work and instructions for similar tasks has shown that the guidelines for NLI categorization are somewhat unsatisfactory, with a noticeable lack of examples beyond the basic ones to illustrate, e.g., the use of negation to form a contradictory hypothesis ("John has a book." $\rightarrow$ "John has

[1]We used the scikit-learn (Pedregosa et al., 2011) implementation of DBSCAN, using parameters $eps = 0.3$, $min\_samples = 2$. The parameters were determined based on qualitative analysis.

no book."). We devised a set of detailed guidelines that contain a thorough list of strategies on how to form more complex hypotheses, with explanations. We provided general principles that advise against high overlap between premise and hypothesis, and described each of the three hypothesis classes with both adequate and inadequate examples in order to emphasize good and bad practices. For instance, the use of synonyms, acronyms, metaphorical expressions, active/passive voice conversion, and common-sense reasoning are all listed as adequate strategies, while simple negation, overlapping and shortening of original premises, or minimal substitutions (e.g., nouns to pronouns) are discouraged. An example from the guidelines is shown below where [P] is the premise, while [H-] and [H+] are examples of a bad and good hypothesis for entailment, respectively.

[P] *Vse vilice in nože, ki so jih pobrali iz lijaka, so zložili v predale v omari.* ("All the forks and knives that they took from the sink, they placed in the drawers in the cupboard")

[H-] *V predale v omari so zložili vse vilice in nože, ki so jih pobrali iz lijaka.* ("Into the drawers in the cupboard, they placed all the forks and knives they had taken from the sink.")

[H+] *Lijak je bil poln pribora, zato so ga pospravili v omaro.* ("The sink was full of cutlery, so they put it away in the cupboard.")

In [H-], the modification is insufficient as only the word order is changed in Slovene. The [H+] example, on the other hand, changes the word order and substitutes original expressions with a synonym (*zložili* "placed" - *pospravili* "put away") and a hypernym (*pribor* "cutlery").

The guidelines were designed based on a test annotation of approximately 50 premises and hypotheses; for each premise, at least three hypotheses were formed (one for each class) and then discussed to harmonize the decisions, particularly for borderline cases. The guidelines were later, during the annotation campaign, updated with additional examples and explanations. Although the guidelines contain some language-specific strategies that do not necessarily pertain to English (such as elaborations on whether a change in an inflected verb form affects the meaning of the hypothesis with respect to the premise), they can mostly be generalized to other languages. We translated them into English and made them publicly available[2].

3.2.2. Introductory and Training Sessions

For the annotation campaign, a total of 8 annotators were recruited, all of whom were students of translation and linguistics at the Faculty of Arts of the University of Ljubljana. The annotators were selected based on several factors, such as their educational background (field of study and year; with higher years (e.g., MA students) prioritized), availability (at least 8 hours per week), and previous experience with linguistic annotation tasks. During the introductory session (a meeting with the researchers to annotate examples and jointly discuss the guidelines and basic concepts of entailment, contradiction, and neutrality), the annotators were trained for the task, and initial misconceptions were resolved. We first demonstrated the workflow by forming three hypotheses for several premises, and continued with a joint annotation session, during which each annotator handled hypotheses construction (entailment, contradiction, neutrality) for additional premises, and had an opportunity to discuss adequate and inadequate strategies. We emphasized that the guidelines should not be taken as strict instructions or a checklist of strategies that need to be implemented in every single hypothesis, but as suggestions to avoid forming completely inadequate examples. This demands a great deal of creativity with paraphrasing and accuracy in conveying the correct meaning, which is why students of translation were very suitable for this task.

After the introductory session, each annotator received a separate online spreadsheet with a batch of approximately 30 premises for individual hypothesis formation as part of the training session. Once the first batch was done, we manually checked the formed hypotheses and provided feedback: first to individual annotators (who received examples of good and bad hypotheses along with our suggestions for better examples and the rationale behind them), as well as to the group as a whole - especially in the case of frequent errors and bad practices. For instance, neutral hypotheses proved to be the most difficult to form, and at the beginning, many annotators resorted to forming them by simply adding additional information not present in the premise, which was allowed in the guidelines, but not encouraged as a go-to strategy. A mailing list was created in order to allow annotators to post questions and dilemmas so we could resolve them with the whole group.

3.2.3. Hypothesis Formation and Cross-checking

Based on the results of the initial annotation, we estimated that 80 premises (which result in 240 hypotheses) take approximately 8 hours of work, and set this as the minimum weekly quota for annotators. From the beginning of the main phase of the annotation campaign, the annotators received an additional 80 premises in their online spreadsheet every week until approximately $6,000$ hypotheses

[2]SI-NLI Guidelines: https://wiki.cjvt.si/books/si-nli/page/si-nli-guidelines

were formed, which was the maximum number according to our plan based on the annotation budget.

In the second part of the main phase of the campaign, the annotators were tasked with classifying their colleagues' hypotheses in a double-blind setup. They were provided with random premises and one of the hypotheses formed by other annotators and asked to categorize them as either entailment, contradiction, or neutral[3]. This allowed for some degree of quality control and calculation of inter-annotator agreement.

3.2.4. Inter-Annotator Agreement

Table 1 shows the formed hypotheses by the annotators' agreement: the first categorization indicates the relation the original annotator needed to form, while the second is the one assigned by the cross-checker. In the majority of cases (79.26%), the cross-checker confirmed the hypothesis class, while 18.59% of examples show lesser disagreement, i.e., a tie between neutrality (N) and contradiction (C) or, much more frequently, between neutrality and entailment (E). Only 2% of cases show major disagreement (a tie between contradiction and entailment).

Table 1: Hypotheses by annotators' agreement. E stands for Entailment, C for Contradiction, and N for Neutrality.

Categorization	Number	%
C-C	1,699	29.17
E-E	1,705	29.28
N-N	1,212	20.81
C-N	159	2.73
N-C	82	1.41
E-N	200	3.43
N-E	642	11.02
C-E	76	1.30
E-C	41	0.70
Errors	8	0.15
Total	5,824	100.00

To check the inter-annotator agreement, we calculate Cohen's kappa (Table 2) for all annotator pairs that had more than one annotation in common (minimum 6, maximum 644). The average coefficient value is approximately 0.74, which indicates a reasonably high agreement. The lowest agreement values (between 0.21 and 0.60) can all be attributed to a single annotator who left the campaign after finishing the training session. The results show that in general, the annotators were consistent both in forming hypotheses as well as classifying them.

Table 2: Inter-annotator agreement statistics.

κ	Value
Average	0.739
Median	0.739
Minimum	0.213
Standard Deviation	0.198

The first round of agreement analysis was followed by an additional round of cross-checking. 583 examples of minor disagreement (N-E and E-N) were assigned a third annotation from the more reliable annotators. As before, the annotators only checked examples formed and classified by others. The results are shown in Table 3. Almost 70% of ambiguous examples were classified as neutral, while 30% leaned more towards entailment. An additional five examples were annotated as inadequate.

Table 3: Additional cross-checking statistics of annotators' agreement. E stands for Entailment, C for Contradiction, and N for Neutrality.

Categorization	Number	%
E-N-E	74	12.69
E-N-N	56	9.61
N-E-E	103	17.67
N-E-N	345	59.18
Errors	5	0.85
Total	583	100.00

Finally, 117 examples with major disagreement (C-E and E-C) were resolved by ourselves by either assigning the final annotation if the hypothesis was adequate, or by adapting the hypothesis to make it in line with the first annotation (according to which it was formed).

In the dataset, the hypotheses that exhibited complete agreement (e.g., C-C) and those disambiguated in the last phase (e.g., N-E-N) were assigned a final annotation based on the majority vote. However, all annotations and the IDs of their annotators are listed separately to allow for effective filtering and provide more transparency.

3.3. The SI-NLI 1.0 Dataset

The constructed dataset consists of 5937 examples, split into 4392 training, 547 validation, and 998 test examples. To construct the split, we selected all examples where the first and second annotation disagreed, as well as all examples with the same premise as those examples, and placed them into

[3]An additional category *X* was used for completely inadequate examples, such as hypotheses that annotators accidentally left unfinished. As these examples were rare, we resolved them manually at a later stage.

the training set. We grouped the remaining examples based on their premise, and split the groups between the validation and test set so that the test set contains approximately 1000 examples and the validation set contains approximately 500 examples. The remaining examples were put into the training set. We publish the dataset for non-commercial use (Klemen et al., 2022). The public test set does not contain class annotations in order to reduce potential issues with overfitting the test set, and to encourage submissions on the SloBench evaluation portal for Slovene natural language processing tasks[4].

In total, the construction of the dataset (including cross-checking) required approximately 200 hours of work and cost around €$2,000$. Half of the allocated time and budget were used for cross-checking.

To enable cross-lingual analysis, we translated the dataset into English by first automatically translating the examples into English with the DeepL machine translation tool and then manually correcting them. For this paper, 514 pairs were manually checked to get the quality evaluation statistics. 127 (70%) premises and 374 (73%) hypotheses required no further editing. Only 4 premises and 11 hypotheses contained major semantic errors that completely changed their meaning. Other errors included erroneously translated named entities, minor omissions (e.g., omissions of adverbs), and minor grammatical errors (most frequently mistranslations of pronouns, e.g., "she" instead of "it"). Overall, the manual analysis showed that the machine translations did not require much editing. We release the translated data publicly (Klemen et al., 2024).

4. Model Evaluation

In this section, we present the results of our NLI experiments. We start by describing the tested language models (Section 4.1) and the experimental settings (Section 4.2). Then, we present the results of two experiments. First, we present the experiments in Slovene using monolingual and multilingual models (Section 4.3); then, we perform cross-lingual experiments using a subset of the Slovene test set translated into English (Section 4.4). For reproducibility of our work, we publish the source code online[5].

4.1. Tested Language Models

Using the created SI-NLI dataset, we trained several classifiers using monolingual, few-lingual, and massively multilingual pretrained transformer language models of three types: encoder-based, decoder-based, and encoder-decoder models. The three types differ based on using only encoder, only decoder, or encoder and decoder transformer layers (Vaswani et al., 2017). The initial motivation for different types of models was their usability for different tasks. For example, decoder-based models are mostly used for (autoregressive) text generation tasks, encoder-based models for text representation and classification tasks, and encoder-decoder models for sequence-to-sequence transformation tasks. However, the distinction in their common use is blurred as tasks can be converted into a common text-to-text format (Raffel et al., 2020) or approached via instruction-augmented text generation (Ouyang et al., 2022), a conversion we also utilize in our experiments.

We next describe the used pretrained models and summarize them in Table 4.

Encoder-based models. We use the monolingual Slovenian SloBERTa (Ulčar and Robnik Šikonja, 2021), the trilingual Croatian-Slovene-English CroSloEngual BERT (CSE-BERT) (Ulčar and Robnik-Šikonja, 2020), the cased base multilingual BERT model (mBERTc) (Devlin et al., 2019), and the base and large multilingual XLM-RoBERTa (XLM-R) (Conneau et al., 2020). We use the models in a discriminative setting, i.e. we extend the pretrained models with a linear layer, and fine-tune a three-class classifier.

Decoder-based models. We use the Slovenian gpt-sl-base model[6], and the multilingual GPT-3.5-turbo instruction-tuned model (OpenAI, 2023). We use gpt-sl-base in a discriminative setting, while we use GPT-3.5-turbo in a generative setting, i.e. we fine-tuning the model to produce the NLI class in its text form (e.g., "entailment").

Encoder-decoder models. We use the small and large Slovenian T5 (Ulčar and Robnik-Šikonja, 2023), and the small, base, and large multilingual T5 models (Xue et al., 2021). We use the models in a generative setting.

4.2. Experimental Settings

We performed the monolingual Slovene NLI experiments using the dataset splits described in Section 3.3, and the cross-lingual NLI experiments using the Slovene training and validation set together with the manually translated English test set.

To train the models, we used reasonable hyperparameter values instead of performing thorough hyperparameter tuning as we are interested in general baselines rather than optimal model performance. We trained the autoregressive and masked language model classifiers for up to 10 epochs, and

[4] https://slobench.cjvt.si/leaderboard/view/9

[5] https://github.com/matejklemen/si-nli

[6] https://huggingface.co/cjvt/gpt-sl-base

Table 4: Summary of the used language models. The size of the GPT3.5-turbo model is marked with ? as its size is undisclosed and estimated based on its predecessor.

Model	Languages	# param.
gpt-sl-base	Slovene	110M
SloBERTa	Slovene	110M
CSE-BERT	triling.	110M
mBERTc-base	multiling.	110M
XLM-R-base	multiling.	125M,
large	multiling.	355M
T5-sl-small	Slovene	60M
-large	Slovene	750M
mT5-small	multiling.	300M
-base	multiling.	580M
-large	multiling.	1.2B
GPT3.5-turbo	multiling.	175B?

the sequence-to-sequence model classifiers for up to 300 epochs, selecting the best model based on the validation set accuracy. We vary the training time as certain Slovene sequence-to-sequence models converged very slowly and required significantly more training, while the autoregressive and masked language models converged significantly faster than in 10 epochs. For GPT3.5-turbo, we used 3 epochs, a setting which was automatically suggested by the OpenAI training platform. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate $2 \cdot 10^{-5}$. To constrain memory usage, we used a maximum input sequence length equal to the 99th percentile of all training sequence lengths (between 100 and 150 tokens), and truncated sequences beyond this length.

We report results using the mean and standard deviation of the test set accuracy across five runs of the training procedure, except for GPT3.5-turbo, for which we report a single-run accuracy. We assume that the standard deviation for GPT3.5-turbo would be comparable to other models.

4.3. Slovene NLI Experiments

Table 5 shows the accuracy of the tested models on the Slovene NLI task. In an unconstrained comparison, the best accuracy is achieved by GPT3.5-turbo (0.857), followed by XLM-R-large (0.791). This shows the impressive ability of GPT3.5-turbo to adapt to the Slovene language despite Slovene texts likely not being largely present during its pretraining. In addition, the results suggest that improved accuracy can be obtained using larger models, although the improvement is diminishing: in comparison with the accuracy of the best base-sized model SloBERTa (0.744), the 3-times larger XLM-R-large achieves an absolute improvement of $+0.047$, while the >1500-times larger GPT3.5-turbo achieves an absolute improvement of $+0.113$. Surprisingly, the large Slovenian T5 model t5-sl-large achieves lower accuracy (0.590) than its smaller equivalent t5-sl-small (0.653). This confirms observations of Ulčar and Robnik-Šikonja (2023) of the inadequacy of t5-sl-large, likely due to insufficient training data involved in its pretraining.

Table 5: Classification accuracy of different models for Slovene NLI experiments.

Model	Accuracy
majority	0.344
gpt-sl-base	0.479 (0.019)
SloBERTa	0.744 (0.008)
CSEBERT	0.667 (0.005)
mBERTc-base	0.595 (0.011)
XLM-R-base	0.670 (0.011)
XLM-R-large	0.791 (0.014)
T5-sl-small	0.653 (0.004)
T5-sl-large	0.590 (0.007)
mT5-small	0.540 (0.012)
mT5-base	0.621 (0.007)
mT5-large	0.767 (0.005)
GPT3.5-turbo	**0.857**

Comparing models of the base size (i.e., 110M parameters) in isolation, we see that the models trained for a smaller set of languages regularly outperform the broadly-focused multilingual models, with the monolingual SloBERTa model achieving accuracy 0.744, followed by the trilingual CSEBERT, and massively multilingual mBERTc-base. A notable exception is the monolingual gpt-sl-base model with a low accuracy 0.479. We hypothesize this is due to suboptimal pre-training of gpt-sl-base and not the architectural differences. To validate this, we ran the same experiment with an English GPT-2 model (Radford et al., 2019) of similar size (124M) on an automatically translated English version of the SI-NLI dataset, and observed the model achieving a significantly higher accuracy (0.588) with a similar standard deviation (0.020).

Despite being primarily aimed at generative tasks, the T5 models are able to perform the NLI classification task relatively well, although they commonly lag behind the encoder-based models in terms of accuracy. The best performing T5 model mT5-large achieves accuracy 0.767 and comes close to the accuracy of the best encoder-based model XLM-R-large (0.791), although the significantly larger size of mT5-large likely plays an important role in this. A small drawback of the T5 models we have observed in our experiments and previously mentioned in Section 4.2 is their occasional slow convergence. We attribute this to the models' need to learn how to generate the class in

the text form in addition to discriminating between three class values.

Table 6: Four manually selected erroneously classified examples by GPT3.5-turbo on Slovene NLI.

Example 1:
[P] V njej je prebujal občutke, ki jih ni poznala. *(He made her feel feelings she did not know before.)*
[H] Ob njem je začutila doslej neznana občutja. *(By his side she felt feelings she did not feel thus far.)*
predicted: neutral, **correct:** entailment

Example 2:
[P] »Ja?« sem rekel, ko sem zaslišal trkanje po vratih. *("Yes?" I uttered once I heard the knocking on the door.)*
[H] Zaradi zanimanja sem se oglasil na trkanje po vratih. *(Due to my curiosity I answered the door knock.)*
predicted: neutral, **correct:** entailment

Example 3:
[P] S prijateljem še ne bova vrgla puške v koruzo. *(Me and my friend will not give up yet.)*
[H] "S prijateljem sva se naveličala neuspehov, zato bova odnehala." *(Me and my friend got tired of failure, so we will quit.)*
predicted: neutral, **correct:** contradiction

Example 4:
[P] Vabljeni ste na konferenco, ki se bo zgodila v četrtek, 11. 11. 2010 v Ljubljani. *(You are invited to the conference happening on Thursday 11. 11. 2010 in Ljubljana.)*
[H] Vabimo vas na slavni dogodek, ki bo potekal 11. novembra 2010 v prestolnici. *(We invite you to the gala event happening on November 11 in the capital.)*
predicted: neutral, **correct:** entailment

To observe if the misclassifications are primarily actual errors or a consequence of different annotation perspectives (Plank et al., 2014), we qualitatively observe the errors of the best-performing model GPT3.5-turbo. While it achieves high accuracy and certain misclassified examples could also be labeled differently depending on the assumed background context, the model nonetheless makes multiple unambiguous errors due to not possessing certain capabilities. This is likely the consequence of being pre-trained on a negligible amount of Slovene texts. We show a small manual selection of errors in Table 6.

In **Example 1**, the model seems to misclassify the example as the pronoun "he" is implicitly present in the verb; this is a special property of the Slovene language.
In **Example 2**, the model seems to misclassify the example due to the lack of common sense that the utterance "Ja?" ("Yes?") implies a person's curiosity.
In **Example 3**, the model seems to to misclassify the example because it does not understand the idiom "vreči puško v koruzo" ("to give up").
In **Example 4**, the model seems to misclassify the example as it does not possess background knowledge that Ljubljana is also referred to as "prestolnica" ("the capital").

In general, our qualitative error analysis suggests there is room for improvement despite the already high accuracy achieved by the best model.

4.4. Cross-lingual NLI Experiments

Table 7 shows the accuracy of models on the cross-lingual NLI task, where the training data is Slovene, and the test data is a test set translated into English. We only evaluate models capable of handling English and Slovene, i.e., only a subset of the models evaluated in Section 4.3.

Table 7: Results of cross-lingual experiments (SL $\rightarrow$ EN) on a manually translated subset of the test set.

Model	Accuracy
majority	0.344
CSE-BERT	0.623 (0.011)
mBERTc-base	0.489 (0.031)
XLM-R-base	0.620 (0.038)
XLM-R-large	0.765 (0.017)
mT5-small	0.428 (0.011)
mT5-base	0.579 (0.029)
mT5-large	0.746 (0.017)
GPT3.5-turbo	**0.837**

We observe similar trends in the model performance as on the Slovene test set: the best accuracy is achieved by GPT3.5-turbo (0.837) and XLM-R-large (0.765), indicating the influence of the model size. Similarly as before, more narrowly focused models beat massively multilingual ones when comparing similar sizes of models. Concretely, the trilingual CSE-BERT with accuracy 0.623 performs better than mBERTc-base (0.489) and equivalently to the XLM-R-base model (0.620) containing 15 million more parameters.

To assess the influence of manual translation on the model performance, we test a well-performing model (XLM-R-large) on the automatically translated test set, and compare its accuracy to the accuracy achieved on the manually translated test set. On the automatically translated set, the model achieves the classification accuracy of 0.768 with the standard deviation 0.014, indicating no significant difference. This strengthens the observation

of the high quality of automatic translations previously noted in Section 3.3. Although the manual corrections improved the text legibility, they do not seem to have an influence on the model's prediction accuracy.

5. Conclusion

We presented the Slovene NLI dataset SI-NLI and released it publicly for non-commercial use. We described its creation process and set the initial prediction baselines using a diverse LM selection. The quantitative results show that models (particularly the larger ones) can perform the task relatively well, while the qualitative analysis reveals that several deficiencies are still present in the models. Improving models by reducing these deficiencies is a promising direction for further work. Additionally, we translated the dataset into English and performed cross-lingual NLI experiments. These models perform slightly worse. The patterns where models struggle also present a sensible direction for further improving the created resource by introducing more challenging examples using the same underlying linguistic phenomenona (e.g., metaphors). This will ensure the relevance of the resource for evaluating the increasingly powerful LLMs.

In addition to improving the models, we see another promising direction in approaching NLI through the lens of data perspectivism (Plank et al., 2014). In our work, we have assumed the existence of a single ground truth label, while the true label might be fuzzy for legitimate reasons instead of annotation errors.

6. Limitations

Our work introduces SI-NLI, a NLI dataset for Slovene. Previous work has shown that NLI datasets can contain biases that enable models to learn shortcuts. While we designed the annotation guidelines to mitigate these issues, and performed annotator training and quality checks, we did not extensively evaluate the potential biases in the created dataset. Our goal was the introduction of a new resource for NLI in a previously unsupported language and the initial testing of models, while we see a more thorough bias analysis and use in downstream applications as logical next steps in future work.

7. Ethical statement

We aim to maintain the privacy of participants in the annotation process. However, further research on the annotation process could benefit from this metadata. Therefore, we release pseudonymized identifiers of the annotators while their true identity is not shared. We see no further ethical issues in the conducted research.

Acknowledgements

The dataset construction campaign was funded by CLARIN.SI. The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411, young researcher grant, as well as projects J7-3159, CRP V5-2297, and L2-50070.

8. Bibliographical References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and representation for Turkish natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

Leo Obadić, Andrej Jertec, Marko Rajnović, and Branimir Dropuljić. 2023. C-XNLI: Croatian extension of XNLI dataset. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2258–2267.

OpenAI. 2023. GPT-4 technical report. ArXiv preprint 2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901.

Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: Less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, page 104–111.

Matej Ulčar and Marko Robnik Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of SI-KDD within the 24th International Multiconference Information Society 2021*, page 17–20.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6.

Shagun Uppal, Vivek Gupta, Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. Two-step classification using recasted data for low resource settings. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 706–719.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

9. Language Resource References

Klemen, Matej and Žagar, Aleš and Čibej, Jaka and Robnik-Šikonja, Marko. 2022. *Slovene Natural Language Inference Dataset SI-NLI*. Slovenian language resource repository CLARIN.SI.

Klemen, Matej and Žagar, Aleš and Čibej, Jaka and Robnik-Šikonja, Marko. 2024. *English translation of the Slovene Natural Language Inference Dataset SI-NLI-en 1.0*. Slovenian language resource repository CLARIN.SI.

Logar, Nataša and Erjavec, Tomaž and Krek, Simon and Grčar, Miha and Holozan, Peter. 2013. *Written corpus ccKres 1.0*. Slovenian language resource repository CLARIN.SI.