# Simple corpus analysis

## SI-NLI

SI-NLI is a dataset that contains 5937 examples. It was sourced from a Slovene reference corpus ccKres and then formed by annotators, who followed given guidelines, and who were from a linguistic field. The dataset was also translated into English to enable cross-lingual model evaluation.

Data collection and the annotation process:

* Constructing premise-hypothesis pairs out of the reference corpus ccKres, which were then handed to the annotators (using the language agnostic BERT sentence embeddings),
* only sentences with between 10 and 40 tokens containing certain syntactical elements were kept (to exclude partial sentences etc.),
* annotators were asked to edit the hypothesis 3 times, each of the edits expressing one NLI relation,
* additional guidance has been offered to the annotators,
* various steps in the annotation process.

The resulting corpus:

* Consists of 5937 examples (4392 training, 547 validation and 998 test examples),
  + training set – examples with a disagreement between the first and the second annotation + all of the remaining ones with the same premise + all the ones that were remaining after collecting the validation and the test set,
  + test set – approximately 1000 examples, validation set – approximately 500 examples (both randomly chosen after the initial selection for the training set had been done).

Tested language models:

* Several classifiers based on monolingual, few-lingual and massively multilingual pretrained transformer language models of three types were trained:
  + on encoder-based models,
  + on decoder-based model,
  + on encoder-decoder models.
* reasonable hyper-parameter values were used to enhance the study of the general baselines

Dev set appearance:





## MultiNLI

MultiNLI is a collection of sentence pairs, **based on the SNLI** corpus (and distributed in the same format). The difference lies in that it is **more diverse**. The **433k sentence** pairs cover a **range of genres of spoken and written text**. Here are some of the examples listed:

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjen

As stated under the given example, the sentence pairs included vary in genre. Furthermore, we can see the annotations given by different human annotators, and the selected gold label, which is the chosen final annotation.

All of the genres are included in the test and development sets, but only five are included in the training set. That way, models can be evaluated either on **matched test examples**, or the **mismatched** examples.

Data collection:

* Premise from a preexisting source, hypothesis from a human annotator,
  + 10 genres, 10 sources (9 of them from the same corpus – OANC)
    - Face-to-face (in-person conversations),
    - government (public reports, speeches etc.),
    - letters ,
    - 9/11 (public report on the terrorist attack),
    - OUP (non-fiction works by the Oxford University Press),
    - SLATE (articles from the Slate magazine),
    - telephone (transcriptions of telephone conversations),
    - travel (travel guides),
    - VERBATIM (short posts about linguistics from the Verbatim archives),
    - fiction (compiling works from different fiction genres).
* minimal preprocessing of the premise sentences,
  + exclude sentences under 8 characters

Hypothesis collection:

* Asking a crowdworker to create three novel sentences out of the premise (one for entailment, one for neutral, one for contradiction),
* Slightly tailored prompts with example premise and hypothesis to make the instructions clearer (e.g. to fit the emotional content of 9/11),
* relabeling done by 4 workers to create the final label – “the gold label”.

The resulting corpus:

* 2 formats - .json and .tsv,
* Hypotheses tend to be fluent and correctly spelled, but not all
  + Punctuation often omitted,
  + not all are complete sentences,
  + often don’t correspond closely to the syntactic structure of the premises.

License:

* corpus freely available
* fiction pairs, for example, are under different licences

Partition:

* train/test/development split
* test and dev set 🡪 2000 randomly selected examples genres each from each of the genres
* no premise occurs more than once

Statistics:

* premise sentences max 401 words, mean 22.3 words; hypotheses shorter (max 70 words)
* parsed as complete sentences in 91% of cases
* trivial editing strategies not common when creating hypotheses
  + only 0.9% differ from their premises by deletion
  + only 1.2% by addition, substitution or shuffling a word
* 30% of hypothesis-premise pairs with high token overlap (>37%)

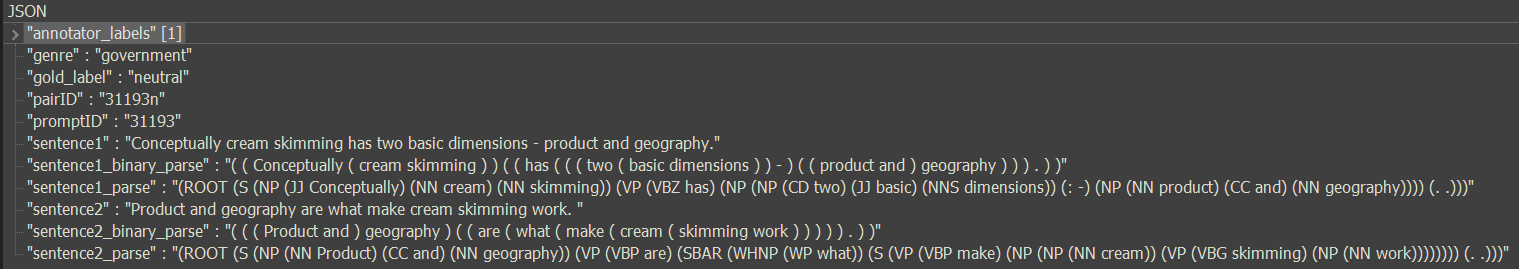
Baselines:

* Testing the difficulty of the corpus with three neural network models:
  + CBOW model,
  + LSTM RNN model,
  + ESIM model.
* MultiNLI dramatically more difficult than SNLI

Analysis:

* done by using a set of annotation tags that reflect linguistic phenomena which are known to be potentially difficult
* two methods to annotate
  + Penn Treebank
  + hand-chosen key words
* 🡪 “sentences containing negation are slightly more likely to be labelled ‘contradiction’”

Training set appearance



## e-SNLI

The e-SNLI dataset is also based on the SNLI (Stanford NLI dataset), to which an additional layer of human-annotated explanations for entailment relations is added. The goal is to enhance machine learning models’ interpretability and robustness by incorporating these explanations into the NLI task.

The resulting corpus:

* The explanations provide justifications for model decisions, making them more interpretable,
* models are trained with these explanations and can output them during testing.

Goals:

* Obtain full sentence justifications for model decisions,
* improve universal sentence representations,
* transfer knowledge to out-of-domain NLI datasets.

Experiments:

* Done by doing different combos of explanation, prediction and labelling, eg.:
  + “PREMISEAGNOSTIC: Generate an explanation given only the hypothesis”,
  + “PREDICTANDEXPLAIN: Jointly predict a label and generate an explanation for the predicted label”...

Format:

* .csv
* dev set
* test set
* 2 training sets

## DocNLI

The added value of the DocNLI dataset is its document granularity. So rather that being only sentence-based it extends the length of its premises to a paragraph or even document length. The goal was to create a dataset that would offer a higher capability of solving open-ended NLP problems, for example, to verify the factual correctness of a document summary.

Data collection:

* From diverse domains,
* without severe artifacts (eg. the ones that obviously indicate that a sentence is incorrect),
* multiple genres are covered.

Data preprocessing:

* ANLI to DocNLI,
* SQuAD to DocNLI,
* Summarization.

Format:

* .json
* dev, test, train sets

Dev set appearance:

