A deep Natural Language Inference predictor in Italian without Italian training data

Lorenzo Corradi[1,*], Alessandro Manenti[1], Francesca Del Bonifro[1] and Dario Del Sorbo[1]

[1]Lutech S.p.A., via Massimo Gorki, 30, Cinisello Balsamo (MI), 20092, Italy

Abstract

We present an application in the sphere of Natural Language Processing to tackle the problem of inference relation (NLI) between pairs of sentences in Italian without an Italian NLI training dataset. The model has been developed making use of the Knowledge Distillation technique. The model may be exploited in industrial scenarios to analyse any kind of user-generated content. Specifically, applications may include product or service review analysis, brand reputation analysis, or social media sentiment analysis. To support this claim, we evaluate the proposed architecture over the native Italian ABSITA dataset on different tasks, such as Aspect-Based Sentiment Analysis, Sentiment Analysis, and Topic Recognition. We balanced the dataset used for each task to obtain a 50/50 division, and we respectively achieve the performances in terms of accuracy of **94.03%**, **88.12%**, and **71.19%**. By applying the model over different scenarios, we empirically demonstrate the generality and exploitability of the NLI technique. While the developed model has still room for improvement, it is suitable to operate unsupervised free-text information extraction — mainly over reviews — in a production environment.

Keywords

Natural Language Inference, Knowledge Distillation, Information extraction

1. Motivation

1.1. Natural Language Inference

Natural Language Inference (NLI) is the task of determining the inference relation between two short and ordered texts, usually defined "premise" and "hypothesis". It is a challenging task that requires understanding the nuances of language and context, as well as the ability to reason and make logical implications.

One common approach to tackle NLI is to use Neural Networks, such as Recurrent Neural Networks or Transformer models [1], to learn to map the premise and hypothesis to a shared representation space and make a prediction.

Examples from the benchmark NLI dataset are shown in Table 1 — namely, Stanford NLI [2] — to detail the standard structure of a NLI dataset.

If the hypothesis may be inferred from the premise, the NLI task may be reinterpreted as a task of information extrapolation. We could query a dataset of reviews to extract any custom information — e.g. topic, or sentiment.

Premise	Hypothesis	Label
"A soccer game with multiple males playing"	"Some men are playing a sport"	Entailment
"An older and younger man smiling"	"Two men are smiling and laughing at the cats playing on the floor"	Neutral
"A man inspects the uniform of a figure in some East Asian country"	"The man is sleeping"	Contradiction

Table 1
NLI dataset. A label is produced based on logical interaction between sentences.

1.2. Goal

Let us have a large corpus of sentences in Italian, with the objective to extract information from each sentence.

The goal of this research is to build a model with the following traits:

- Ability to understand the inference relation between sentence pairs in a specific language. Informally, it means to train a model on a NLI dataset.
- Ability to understand Italian. The majority of the published models are in English. This reflects the fact that research is published in English, and relevant NLI datasets are all in English. To the best of our knowledge, we did not find any comprehensive NLI datasets in the Italian language to train a Deep Learning model.

The industrial applications include, among others:

- Product or service review analysis. We focused on this application to run our tests.
- Brand reputation analysis.
- Social media sentiment analysis.

The lack of a comprehensive NLI dataset in Italian has been the key driver to adopt the architecture described

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
l.corradi@lutech.it (L. Corradi)

in 2. This procedure can be generalised to any task involving a language-specific dataset. Another plus of this methodology is that we require no dataset labeling costs. This architecture has been demonstrated to have NLI capability in Italian without being exposed to a NLI training dataset in Italian.

2. Method

2.1. NLI training

The first step of our methodology is retrieving an encoder model, based on Transformers, available online. This encoder model is already fine-tuned for general purposes over different languages. Since Transformers are computationally expensive to train from scratch, we decided to test multilingual architectures of Transformers and fine-tune those that suited the needs the most (on Stanford NLI and Multi-Genre NLI [3]).

After a training session over a NLI English dataset, the result is a model, based on Transformers, that can proficiently address the NLI task — only in English though, despite being originally trained on multiple languages. More information about this work available at Ref. [4].

In our work, we want to enable a multilingual model, based on Transformers, previously fine-tuned for a specific task only in one specific language, to proficiently address that specific task in Italian.

2.2. Knowledge Distillation

The second step of our methodology is to employ a training without language-specific NLI training data. The selected approach is Knowledge Distillation [5].

We employed Knowledge Distillation to perform NLI in Italian, with the objective of forcing a translated sentence to have the same location mapping in the vector space as the original sentence. Compared to other approaches, this has the following advantages:

- Easy to extend models with few samples to new languages.
- Easy to ensure desired properties for the vector space.
- Low hardware requirements.

We require a teacher model (encoder) T, that maps sentences in the source language to a vectorial representation. Further, we need parallel (translated) sentences $D = ((source_1, target_1), ..., (source_n, target_n))$ with $source_j$ being a sentence in the source language and $target_j$ being a sentence in the target language. We train a student encoder model S such that $T(source_j) \approx S(target_j)$. For a given mini-batch B, we minimise the Mean Squared Error loss function:

$$MSE_{(S,T,D=(source_j,target_j))} = \frac{1}{|B|} \sum_{j \in |B|} (T(source_j) - S(target_j))^2 \quad (1)$$

Two instances of the encoder described in 2.1 have been taken for the experiment. One acts as teacher encoder model T, the other as a student encoder model S.

The application of Knowledge Distillation has the objective to share the domain knowledge of the teacher encoder model to the student encoder model, and at the same time learn a new vectorial representation for the target language. A schematic representation is provided in Figure 1.

The obtained NLI classifier, able to understand Italian, accepts a sentence pair to output a NLI label.

Execution-wise, the Knowledge Distillation task on a Tesla P100–PCIE–16GB GPU was completed in approximately five (5) hours on the TED2020 (English–Italian) dataset [6], consisting of more than 400k parallel sentences, with the following main parameters:

- $batch\_size = 24$
- $max\_sentence\_length = 256$
- $max\_tokens\_length = 128$
- $epochs = 6$
- $learning\_rate = 2e - 5$
- $epsilon = 1e - 6$
- $weight\_decay = 1e - 2$
- $accumulation\_step = 4$

The parameters were selected in an unstructured way; either by trial-and-error, or suggested online by informal sources. No cross-validation or grid-search analyses have been performed for computational constraints. Therefore, no guarantees on the optimality of the parameters can be made.

3. Results

3.1. NLI results

The discussed architecture has been tested over the standard NLI task in Italian. We made use of small Italian NLI datasets, available online, such as RTE3-ITA and RTE-2009. We also exploited an open-source machine translation model, developed by Facebook, named No Language Left Behind [7], to obtain a comprehensive Italian NLI dataset, suitable for testing.

The discussed architecture, based on Knowledge Distillation, demonstrated to perform better than another tested architecture that was directly trained over machine translated NLI datasets, despite having an objective disadvantage. We stress the fact that the proposed architecture was never directly trained over any kind of Italian NLI data.

3.2. ABSA results

Aspect-Based Sentiment Analysis at EVALITA (ABSITA) [8] is an Aspect-Based Sentiment Analysis dataset. Contains Italian hotel reviews that may touch different topics (such as price, location, cleanliness, etc.) and a sentiment associated to each topic (knowing that sentiments for different topics may be contrasting). By choosing arbitrary NLI hypotheses, this dataset may emulate a total of three (3) different tasks, namely Sentiment Analysis, Topic Recognition, and Aspect-Based Sentiment Analysis. The core idea behind this setting comes from the desire to query a text — in NLI, a set of premises (e.g. a set of reviews), in an unsupervised way, to receive specific answers from a predefined list of answers (e.g. the presence of a topic from a list of topics). In the case of open answers, a question-answer architecture would have been more suitable.

3.2.1. Sentiment Analysis

Sentiment Analysis is the task to recognise the overall sentiment of a sentence. As detailed above, we would like to exploit the models to apply Sentiment Analysis in an unsupervised manner — to do this, we fix a hypothesis arbitrarily. We assume that the hypothesis we have chosen captures the logical implication that is the core of NLI. Follow results for the ABSITA dataset, detailed in Table 2. Note that the hypothesis has been arbitrarily set to "Sono soddisfatto" ("I feel satisfied").

Dataset	Balancing	Task	Metric	Result
ABSITA	1:1	Sentiment Analysis	Accuracy	88.12%
ABSITA	1:1	Sentiment Analysis	Min F1-Score	86.89%
ABSITA	1:1	Sentiment Analysis	Macro-Avg F1-Score	88.02%

Table 2
ABSITA results, over the Sentiment Analysis task.

3.2.2. Topic Recognition

Topic Recognition is the task to recognise whether or not a sentence is about a topic. As detailed above, we would like to exploit the models to apply Topic Recognition in an unsupervised manner — to do this, we fix a hypothesis arbitrarily. We assume that the hypothesis we have chosen captures the logical implication that is the core of NLI. Follow results for the ABSITA dataset, detailed in Table 3. Note that the hypothesis has been arbitrarily set to "Parlo di pulizia" ("I'm talking about cleanliness").

Where the seven (7) in the "Balancing" column stands for the number of different topics in the dataset. The 1:1 balancing has been obtained by randomly sampling sentences from the seven (7) classes that do not compose the target. The two scenarios have been proposed to extensively test the generalisation capability of the models.

Dataset	Balancing	Task	Metric	Result
ABSITA	1:1	Topic Recognition	Accuracy	68.09%
ABSITA	1:1	Topic Recognition	Min F1-Score	65.75%
ABSITA	1:1	Topic Recognition	Macro-Avg F1-Score	67.97%
ABSITA	1:7	Topic Recognition	Accuracy	71.11%
ABSITA	1:7	Topic Recognition	Min F1-Score	37.94%
ABSITA	1:7	Topic Recognition	Macro-Avg F1-Score	59.56%

Table 3
ABSITA results, over the Topic Recognition task.

3.2.3. Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) is the task to recognise the sentiment about each sub-topic in a sentence. As detailed above, we would like to exploit the models to apply ABSA in an unsupervised manner — to do this, we fix a hypothesis arbitrarily. We assume that the hypothesis we have chosen captures the logical implication that is the core of NLI. Follow results for the ABSITA dataset, detailed in Table 4. Note that the hypothesis has been arbitrarily set to "La camera é pulita" ("The room is clean").

Dataset	Balancing	Task	Metric	Result
ABSITA	1:1	ABSA	Accuracy	94.03%
ABSITA	1:1	ABSA	Min F1-Score	93.90%
ABSITA	1:1	ABSA	Macro-Avg F1-Score	94.02%
ABSITA	1:15	ABSA	Accuracy	78.42%
ABSITA	1:15	ABSA	Min F1-Score	37.66%
ABSITA	1:15	ABSA	Macro-Avg F1-Score	62.30%

Table 4
ABSITA results, over the Aspect-Based Sentiment Analysis task.

Where the fifteen (15) in the "Balancing" column stands for the number of different tuples (topic, sentiment) in the dataset. The 1:1 balancing has been obtained by randomly sampling sentences from the fifteen (15) classes that do not compose the target. The two scenarios have been proposed to extensively test the generalisation capability of the models.

3.2.4. Summary

In this work we tested different architectures showing that it is possible to obtain reasonable accuracies over different Natural Language Processing tasks by fine-tuning a single architecture based on sentence embeddings over the NLI task.

We showed that various Natural Language Processing problems may be mapped into a NLI task — in this way, we empirically proved the generality of the NLI task. We would like to stress over the lack of need to retrain any models to obtain the results over each specific task.

The full, updated model is available at this link[1], along with all instructions for usage. We named it **I-SPIn**

[1] https://huggingface.co/Lutech-AI/I-SPIn

(**I**talian-**S**entence **P**air **In**ference).

To test our model perfomances over Sentiment Analysis, Topic Recognition, and Aspect-Based Sentiment Analysis, we employed arbitrary hypotheses. We tried our best to avoid any biases (e.g. hypotheses were chosen by colleagues that had never taken a look at the dataset), but we acknowledge that some bias may have been introduced. This is currently considered an open problem.

4. Figures

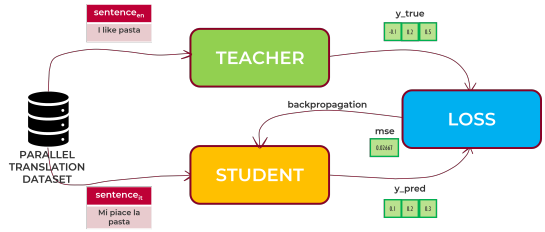

Figure 1: Knowledge Distillation. Teacher encoder model receives source sentences, student model receives target sentences. Student encoder model is updated with new information from the teacher.

5. Citations and Bibliographies

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2017). doi:https://doi.org/10.48550/arXiv.1706.03762.

[2] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference. (2015). doi:https://doi.org/10.48550/arXiv.1508.05326.

[3] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference (2017). doi:https://doi.org/10.48550/arXiv.1704.05426.

[4] A. Manenti, A. Braunstein, Deep learning techniques for natural language processing: A multilingual encoder model for nli task (2022). URL: https://webthesis.biblio.polito.it/24750/.

[5] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation (2020). doi:https://doi.org/10.48550/arXiv.2004.09813.

[6] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, G. Neubig, When and why are pre-trained word embeddings useful for neural machine translation? (2018). doi:10.18653/v1/N18-2084.

[7] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation (2022). doi:https://doi.org/10.48550/arXiv.2207.04672.

[8] P. Basile, D. Croce, V. Basile, M. Polignano, Overview of the evalita 2018 aspect-based sentiment analysis task (absita) (2018). URL: https://ceur-ws.org/Vol-2263/paper003.pdf.