Fantastic Labels and Where to Find Them: Attention-Based Label Selection for Text-to-Text Classification

Michele Papucci[1,2], Alessio Miaschi[1], Felice Dell'Orletta[1]

1 - ItaliaNLP Lab @ Institute for Computational Linguistics "Antonio Zampolli" (CNR-ILC), Pisa.

2 - Pisa Università di Pisa, Pisa.

Istituto di Linguistica Computazionale "Antonio Zampolli"

Consiglio Nazionale delle Ricerche

NL4AI @ AIxIA 2024, Bolzano, November 26-27 2024

Introduction

- Motivations:
  - **Large Language Models** (LLMs) demonstrated remarkable capabilities in solving a variety of tasks and shifted the focus towards the text-to-text framework (each task is verbalized in having a textual input and a textual output);

Introduction

- Motivations:
  - **Large Language Models** (LLMs) demonstrated remarkable capabilities in solving a variety of tasks and shifted the focus towards the text-to-text framework (each task is verbalized in having a textual input and a textual output);
  - This has resulted in a variety of studies on ***input representations*** (aka how to represent the input to the model and what effect does it have: *prompting, few-shot, in-context learning,* etc.);

Introduction

- Motivations:
  - **Large Language Models** (LLMs) demonstrated remarkable capabilities in solving a variety of tasks and shifted the focus towards the text-to-text framework (each task is verbalized in having a textual input and a textual output);
  - This has resulted in a variety of studies on ***input representations*** (aka how to represent the input to the model and what effect does it have: *prompting, few-shot, in-context learning,* etc.);
  - Few works have studied the effects of ***output representation***. (E.g. how to decide which textual output generation should be produced in a classification scenarios to represent the classes) and found out that **it matters** for some tasks' performances.

Introduction

- Motivations:
  - **Large Language Models** (LLMs) demonstrated remarkable capabilities in solving a variety of tasks and shifted the focus towards the text-to-text framework (each task is verbalized in having a textual input and a textual output);
  - This has resulted in a variety of studies on ***input representations*** (aka how to represent the input to the model and what effect does it have: *prompting, few-shot, in-context learning,* etc.);
  - Few works have studied the effects of ***output representation***. (E.g. how to decide which textual output generation should be produced in a classification scenarios to represent the classes) and found out that **it matters** for some tasks' performances.

(Topic Classification Task) Input: *"Partita dura ma dobbiamo dare di più ragazzi forzaaaaa"*
What should the model produce to classify the instance as the topic **Sport**?

Output: *Sport* - Output: *Gelato* - Output: *Calcio*

Introduction

- Motivations:
  - **Large Language Models** (LLMs) demonstrated remarkable capabilities in solving a variety of tasks and shifted the focus towards the text-to-text framework (each task is verbalized in having a textual input and a textual output);
  - This has resulted in a variety of studies on ***input representations*** (aka how to represent the input to the model and what effect does it have: *prompting, few-shot, in-context learning,* etc.);
  - Few works have studied the effects of ***output representation***. (E.g. how to decide which textual output generation should be produced in a classification scenarios to represent the classes) and found out that **it matters** for some tasks' performances.
  - Specifically, the impact of label selection in Classification tasks have been demonstrated, there are few studies on how to select good-performing **label representations** a-priori;

Our Approach to Select Label Representations

- We hypothesize that we can leverage the **attention mechanism** of the model to find in the **task's training set** suitable *label representations;*

Our Approach to Select Label Representations

- We hypothesize that we can leverage the **attention mechanism** of the model to find in the **task's training set** suitable *label representations;*
- To do so, we take each training instance and pass them in inference through the model and see how its **attention pattern behave** with respect to specific *important tokens* in the text.

Our Approach to Select Label Representations

- We hypothesize that we can leverage the **attention mechanism** of the model to find in the **task's training set** suitable *label representations;*
- To do so, we take each training instance and pass them in inference through the model and see how its **attention pattern behave** with respect to specific *important tokens* in the text.

    - Which *important token* should we look at?
        1. **</s> (End-of-Sentence)** - It'used to represent the end of the sentence during generation.
        2. **Appended Label** - We append the translated class name to the end of the sentence.
        3. **Appended Label with Prompt** - We append a prompt to give more context.

Our Approach to Select Label Representations

- We hypothesize that we can leverage the **attention mechanism** of the model to find in the **task's training set** suitable *label representations;*
- To do so, we take each training instance and pass them in inference through the model and see how its **attention pattern behave** with respect to specific *important tokens* in the text.

  - Which *important token* should we look at?
    1. **</s> (End-of-Sentence)** - It'used to represent the end of the sentence during generation.
    2. **Appended Label** - We append the translated class name to the end of the sentence.
    3. **Appended Label with Prompt** - We append a prompt to give more context.

  - How to evaluate which tokens are the most *salient*?
    - **Value Zeroing** - It's a technique used to determine how important is each token in the construction of the vectorial representation of each other token in the sentence inside the Self-Attention layer of a Transformer;

Dataset and Model

The Dataset is **Tag-IT** (Cimino et al., 2020), Topic Classification Task, 11 Classes.

Categories	# Data	# Training	# Test
Anime	3,972	2,894	1,078
Auto-Moto	3,783	2,798	985
Bikes	520	365	155
Celebrities	1,115	754	361
Entertainment	469	354	115
Medicine-Aesthetics	447	310	137
Metal-Detecting	1,382	1,034	348
Nature	516	394	122
Smoke	1,478	1,101	377
Sports	4,790	3,498	1,292
Technology	136	51	85
All	18,608	13,553	5,055

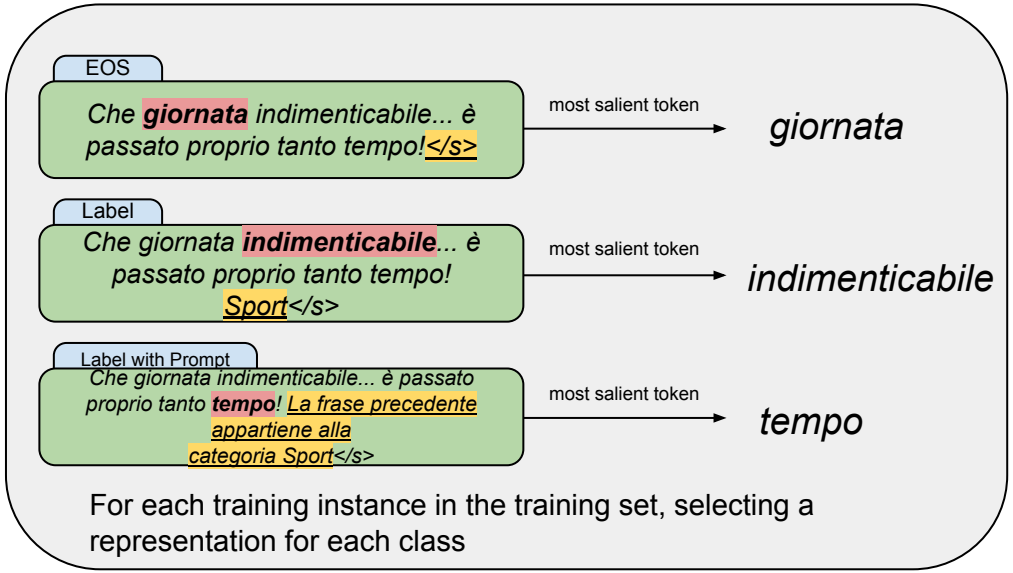
The model is **IT5** (Sarti & Nissim, 2024), a T5-base pre-trained on cleaned Italian Sentences from the mC4 Corpus.

To apply our technique the model was adapted to make its encoder calculate the **Value Zeroing matrix** for each input.

The classification capabilities were then evaluated using F-Score on the test set.

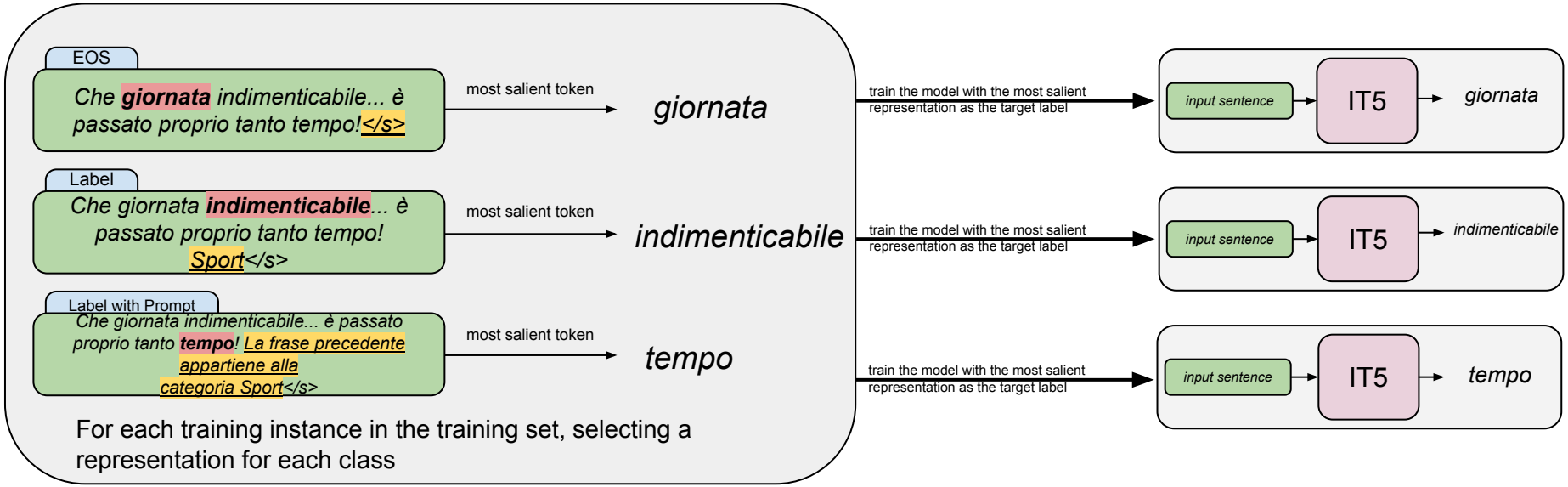
How we selected Label Representations

(1) Using each of the three strategy, we selected 10 representations for each class and **ranked** them on their Value Zeroing Score.



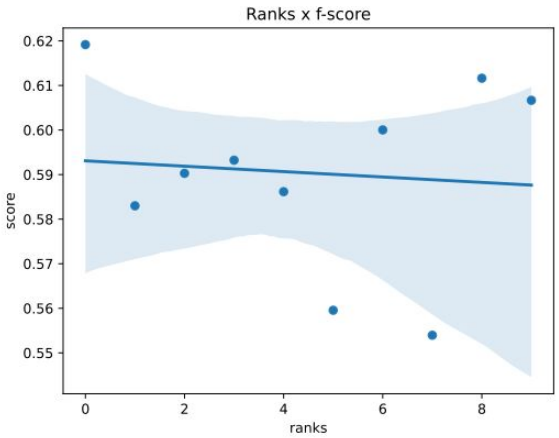
How we selected Label Representations

(2) Then, we trained **10 models for each strategy** using the Ranked Representation sets going from Rank 0 (the best) to rank 9 (the worst).

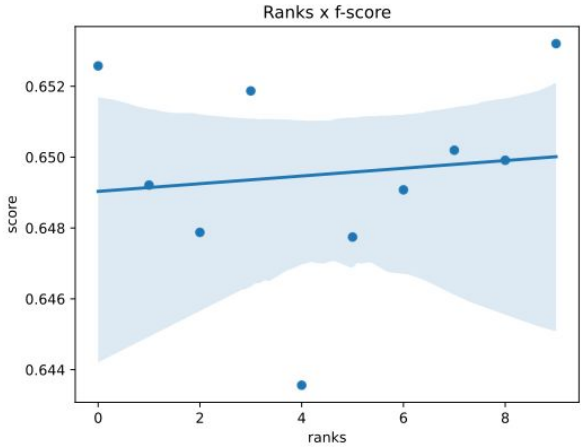


Preliminary Results

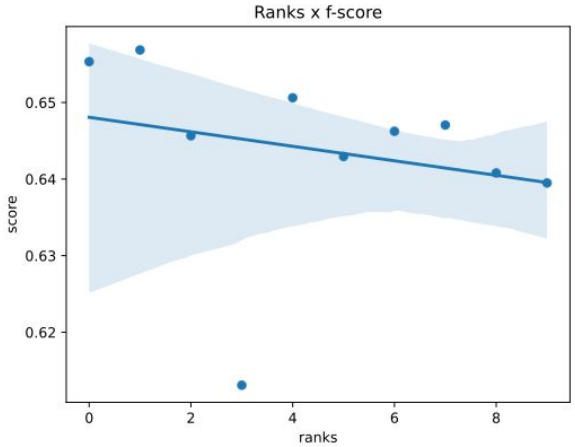
Label



Label with Prompt



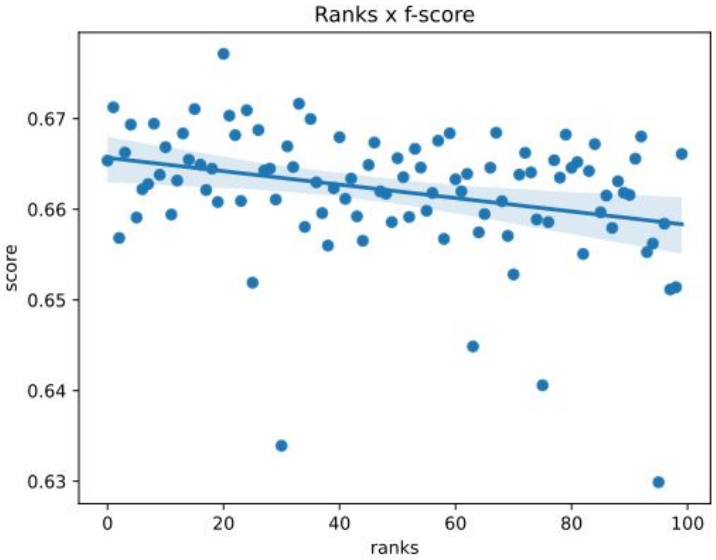
End of Sentence



To further test the **EOS Method**, we extracted 100 representation, ranked them and trained 100 models on it.

We prepared a baseline composed of other 100 models whose representations were: **the original class names** translated, 9 synonyms of the class names, 90 randomly chosen full words from the pre-training dataset of the model.

Results on EOS

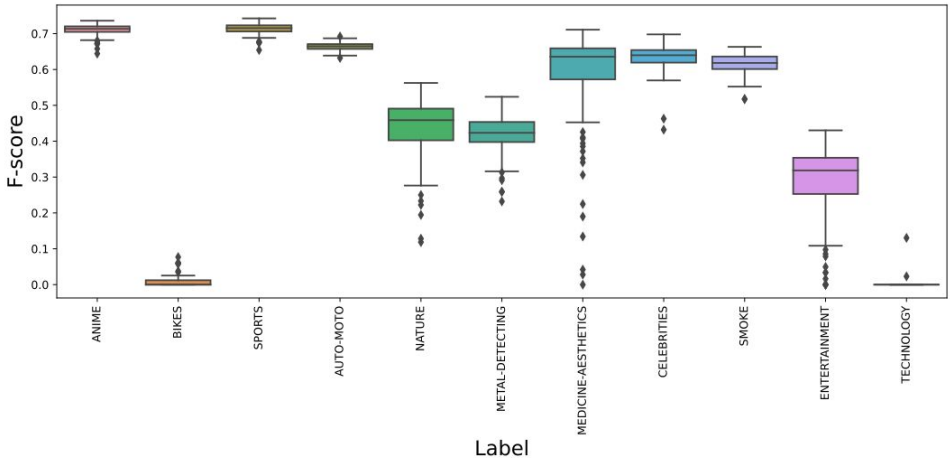


We obtained a statistically significant negative correlation between the rank of the Representations set and the obtained F-Score of **-0.314**.

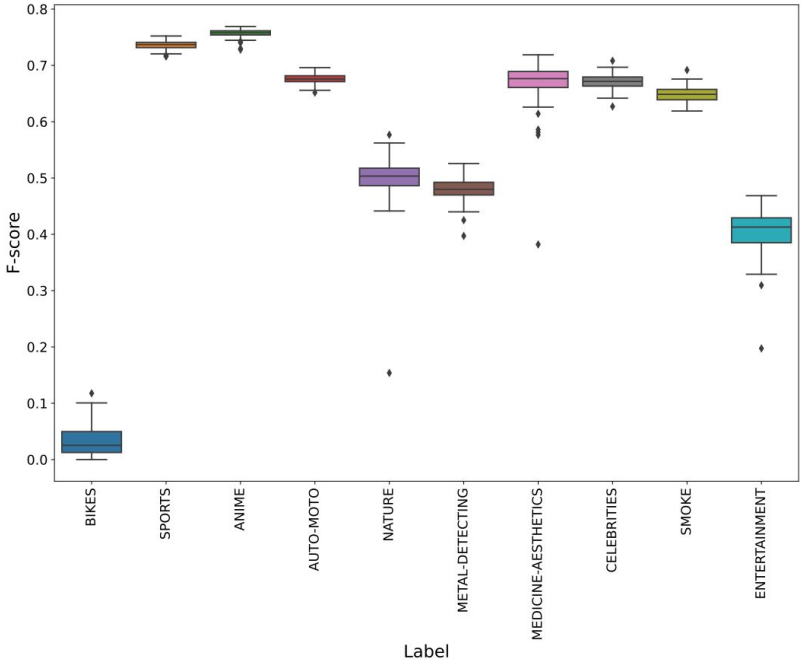
Representation Set	F-Score
0 (First Set)	0.66
20 (Best Performing Set)	0.68
95 (Worst Performing Set)	0.63
Baseline (Trained with original class names)	0.63

Results on EOS

10 synonyms + 90 random words (Baseline)



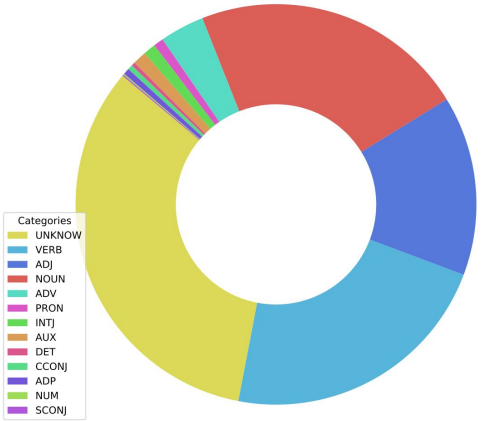
EOS Selection method



Qualitative Analysis on the Representations

Looking at the distribution of Parts-of-Speech in the selected representations, we found that the majority are **verbs**, **adjectives** and **nouns**.

In some categories like ANIME, SPORTS and CELEBRITIES we found also that most of the representations were **proper nouns** and **English words**.



We analyzed the following metrics to find correlations with the models performances:

TF-IDF of the representation: we found that ANIME and SPORTS had a low (0.20 Spearman rank) correlation with the model performances;

Representation Frequency in the training set: we found no correlation;

Sub-token length of the representation: we found no correlation;

Conclusions and Future Work

We performed an extensive evaluation of a novel **Label Representation selection strategy** that leverage the attention mechanism to extract representation from the training set.

We tried different technique on how to apply it, and we saw that the best performing one **outperforms the baseline strategies**: our method gives better performing Representation set for the classes with a **higher average and less variance in performance**.

Future works should assess the performance of this selection strategy on other classification tasks and in more languages.

Thanks for your attention!

Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

Briefly, on Value Zeroing

- Value Zeroing values the influence of a token, on the construction of another, by **zeroing** a certain token *j* when building the token *i*.

Briefly, on Value Zeroing

- Value Zeroing values the influence of a token, on the construction of another, by **zeroing** a certain token *j* when building the token *i*.

- Since removing a token from the text without changing its semantics (and therefore, its representations) is very hard, in Value Zeroing this is done during the **Self-Attention** calculation by **zeroing the Value vector** of *j* when building the representation of *i*.

Briefly, on Value Zeroing

- Value Zeroing values the influence of a token, on the construction of another, by **zeroing** a certain token *j* when building the token *i*.

- Since removing a token from the text without changing its semantics (and therefore, its representations) is very hard, in Value Zeroing this is done during the **Self-Attention** calculation by **zeroing the Value vector** of *j* when building the representation of *i*.

- We obtain the activation $x_i^{\neg j}$ and we compare to the activation that contains *j* by the means of cosine distance, obtaining a saliency metrics.

Briefly, on Value Zeroing

- Value Zeroing values the influence of a token, on the construction of another, by **zeroing** a certain token *j* when building the token *i*.
- Since removing a token from the text without changing its semantics (and therefore, its representations) is very hard, in Value Zeroing this is done during the **Self-Attention** calculation by **zeroing the Value vector** of *j* when building the representation of *i*.
- We obtain the activation $x_i^{\neg j}$ and we compare to the activation that contains *j* by the means of cosine distance, obtaining a saliency metrics.
- By zeroing each token in the sequence for each other tokens we obtain a matrix $n \times n$ where *n* is the sequence length. Each cell indicates how salient is each token in the construction of the representations of each other.