Vardial 2022 shared task: Italian dialect Identification (Project proposal)

Computational Semantics for NLP, Spring 2022

Quynh Anh Nguyen, Giacomo Camposampiero and **Francesco Di Stefano**
ETH Zurich - Swiss Federal Institute of Technology
{quynguyen, gcamposampie, fdistefano}@ethz.ch

1 Introduction

Dialect classification represents a key task in the improvement of many other downstream tasks such as opinion mining and machine translation, where the enrichment of text with geographical information can potentially result in improved performances for real-world applications (Zampieri et al., 2020).

As a result, the interest in the study of language variation has been steadily growing in the last few years, as highlighted by the increasing number of publications and events related to the topic. However, little has been done so far by researchers in the context of automatic dialect recognition for the Italian language.

In this context, the *Languages and Dialects of Italy* (ITDI) task of VarDial Evaluation Campaign 2022[1] aims to bridge this gap, facilitating the development of models capable of properly classifying 11 regional languages and dialects from both Italy's mainland and islands (Piedmontese, Venetian, Sicilian, Neapolitan, Emilian-Romagnol, Tarantino, Sardinian, Ligurian, Friulian, Ladin, Lombard).

The shared task organizers provide a dataset consisting of a large pool of Wikipedia articles written in one of these dialects, in the form of Wikipedia dump. The task is closed and, therefore, participants are not allowed to use external data to train their models (exception done for off-the-shelf pretrained language models from the HuggingFace model hub or similar, the use of which however has to be clearly stated).

The predictions are evaluated at sentence level using F_1 score.

2 Related works

In this section, we will briefly introduce several methods (Zampieri et al., 2020) which used to be applied on classification problems on similar languages or different dialects of a language.

[1]The website will be publicly available from end of April.

Italian dialect identification SoTa For what concerns Italian dialect identification, there are several researches working on analysing Italian dialects features (Zugarini et al., 2020) but no previous experiments deal with language identification task is found. Thus, our scope is to find a way to tackle this problem drawing inspiration from related works in dialect identification in different languages rather than overcome a particular state of the art model. In particular, we aim to obtain good results with the use of deep neural networks, more specifically adopting CNN and transformer architectures.

Machine learning models vs CNN Although deep learning models yield state of the art performances in many NLP tasks, an ensemble of SVM and Naive Bayes models was the best performing model in the Uralic Language Identification task in the VarDial Evaluation Campaign 2021 (Ceolin, 2021). *Linear SVM* classifier, *Naive Bayes* model, the combination of the two methods *Linear SVM + Naives Bayes* as well as *CNN* were implemented to classify target languages. Three machine learning approaches were all trained on TF-IDF character n-grams. The experiment shows that CNN did not always outperform the other machine learning approaches.

CNN Even if the state of the art in many dialect identification tasks has been reached through the application of transformer-based models, the use of CNNs is still high in this type of task. In particular, taking as an example the Romanian vs Moldavian dialect, CNN-based approaches achieved competitive results in both VarDial 2019 Evaluation Campaign (Tudoreanu, 2019) and VarDial 2020 Evaluation Campaign (Rebeja and Cristea, 2020).

Transformer The introduction of transformers (Vaswani et al., 2017) has revolutionized many tasks in NLP and the identification of dialects isn't

an exception. Models based on this architecture achieved state-of-the-art results in many application. A recent example is again VarDial 2020 Evaluation Campaign, where the use of a fine-tuned version of BERT previously trained on three publicly available Romanian corpora (Zaharia et al., 2020) reached a weighted F_1 score of 96.25% on the MOROCO dataset (Butnaru and Ionescu, 2019) in the Romanian vs Moldavian identification task.

3 Dataset

As already mentioned in the section 1, the training dataset is provided by the organizers and consists of 265 016 selected Wikipedia articles from 1st March 2022 dumps. As the training data is provided in the form of raw Wikipedia dumps, careful pre-processing of the data is part of the task. The development and test sets have not been disclosed yet, and will be made available to the participants according to the task milestones reported in Section 6.

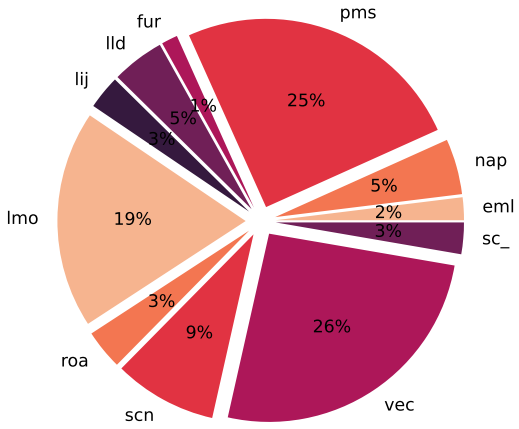
Since the data doesn't come from a well-know documented dataset, a preliminary exploration has been initially conducted to gain useful insight about it. This investigation highlighted a huge imbalance between classes as shown in Figure 1a, since the 3 most represented dialect (Venetian, Piedmontese and Lombard) account for almost three quarters of the entire articles in the training data. However, as shown in Figure 1b, the number of articles per dialect might not be fully representative on itself, since the size of each articles has also to be taken in account. This is the case of Lombard dialect, for example, which accounts for a large portion of training articles (19%), but more than a quarter of them are not longer than 2 sentences.

Nonetheless, imbalanced data seems to represent the main challenge posed by this dataset and should be addressed during the evaluation of the model. Possible solutions to the problem are data augmentation, weighted modelling and data sub-sampling.

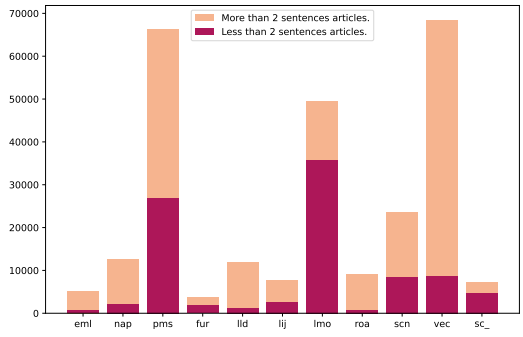
4 Approaches

In order to tackle the Italian dialect identification tasks, we plan to investigate different approaches, including Machine Learning methods and Deep learning methods.

ML approaches Linear SVM and Naive Bayes are the two methods which are still popularly used to handle this specific task. Thus, we would like



(a) Distribution of Wikipedia articles across dialects.



(b) Number of sentences of articles in the training set, grouped by dialect.

Figure 1: Preliminary data exploration on training set.

to implement them with tf−idf character n-grams varying from 3 to 5 grams as in the experiment of (Ceolin, 2021). These approaches could be also used as baseline models to compare to deep neural network approaches such as CNN or Transformers.

CNN Convolutional Neural Networks were employed with success by many teams in the previous editions of the dialect identification task. Therefore, we plan to experiment with a classifier that uses a CNN for feature extraction on top of an embedding layer - that could be a skip-gram model (Mikolov et al., 2013) as in the case of (Tudoreanu, 2019).

Transformer As said above, the use of transformer-based models yield state of the art results even in the task of dialect identification. In particular, the fine-tuning of pre-trained versions of BERT obtained good results in this field. Following this line, our approach with this type of model would be to adopt a version of BERT pre-trained with the Italian language (Polignano et al., 2019) and to fine-tune this latter on our task.

5 Expected Result

The main objective of this project is twofold.

- To develop a model based on the approaches described in Section 4; the resulting model should output sufficiently correct and consistent predictions to eventually allow our group to make a submission to the shared task.
- To gain useful insights about the resulting model, through ablation studies and error analysis.

Besides, since we participate to a shared task, we hope to achieve a good rank in the task leader board.

6 Milestones

The timeline outlined by the organizers is defined as follows.

1. *Now*: Training set and task description are already provided.
2. *End of April*: The shared task will be published in the Vardial 2022 website and will officially start. The development set will be disclosed.
3. *End of June*: The system should be finished.
4. *End of July*: deadline to submit the system description paper.

References

Andrei M. Butnaru and Radu Tudor Ionescu. 2019. Moroco: The moldavian and romanian dialectal corpus.

Andrea Ceolin. 2021. Comparing the performance of CNNs and shallow models for language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–112, Kiyv, Ukraine. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Petru Rebeja and Dan Cristea. 2020. A dual-encoding system for dialect classification. In *VARDIAL*.

Diana Tudoreanu. 2019. DTeam @ VarDial 2019: Ensemble based on skip-gram and triplet loss neural networks for Moldavian vs. Romanian cross-dialect topic identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 202–208, Ann Arbor, Michigan. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the power of Romanian BERT for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Andrea Zugarini, Matteo Tiezzi, and Marco Maggini. 2020. Vulgaris: Analysis of a corpus for middle-age varieties of italian language. *CoRR*, abs/2010.05993.