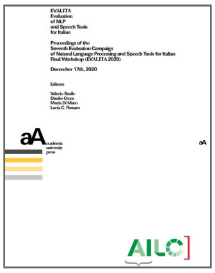

Valerio Basile, Danilo Croce, Maria Maro and Lucia C. Passaro (dir.)

EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020
Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop

Accademia University Press

B4DS @ PRELEARN: Ensemble Method for Prerequisite Learning

Giovanni Puccetti, Luis Bolanos, Filippo Chiarello and Gualtiero Fantoni

DOI: 10.4000/books.aaccademia.7528
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2020
Published on OpenEdition Books: 11 May 2021
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9791280136329

Electronic reference
PUCCETTI, Giovanni ; et al. *B4DS @ PRELEARN: Ensemble Method for Prerequisite Learning* In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop* [online]. Torino: Accademia University Press, 2020 (generated 17 mai 2021). Available on the Internet: <http://books.openedition.org/aaccademia/7528>. ISBN: 9791280136329. DOI: https://doi.org/10.4000/books.aaccademia.7528.

B4DS @ PRELEARN: Ensemble Method for Prerequisite Learning

Giovanni Puccetti
Scuola Normale Superiore
giovanni.puccetti@sns.it

Luis Bolanos
Texty S.r.l.
luis.bolanos@texty.biz

Filippo Chiarello
Universitá di Pisa
filippo.chiarello@unipi.it

Gualtiero Fantoni
Universitá di Pisa
g.fantoni@ing.unipi.it

Abstract

English. In this paper we describe the methodologies we proposed to tackle the EVALITA 2020 shared task PRELEARN. We propose both a methodology based on gated recurrent units as well as one using more classical word embeddings together with ensemble methods. Our goal in choosing these approaches, is twofold, on one side we wish to see how much of the prerequisite information is present within the pages themselves. On the other we would like to compare how much using the information from the rest of Wikipedia can help in identifying this type of relation. This second approach is particularly useful in terms of extension to new entities close to the one in the corpus provided for the task but not actually present in it. With this methodologies we reached second position in the challenge[1].

1 Introduction

The PRELEARN task consists in classifying pairs of concepts according to whether one is a prerequisite for the other or not. The concepts are presented as Wikipedia pages and they are divided into four different domains, physics, precalculus, data mining and geometry.

The task was organized in 4 subtasks: i) two of them concerned with the type of information that can be exploited by the submitted models, either solely textual or including metadata, e.g. Wikipedia hyperlinks; ii) the other two based on different classification scenarios, training and testing could happen either on the same domain or three domain could be used as training set and the fourth as testing. A more extensive description of the task together with all the results and more information is found in the report (Alzetta et al., 2020) which is part of the EVALITA 2020 (Basile et al., 2020). The concept of being a prerequisite is highly complex and can be misunderstood from humans as well. Indeed, this relation can be subtle and depending on the domain it may take a deep level of expertise to recognize. One of the reasons this challenge is very interesting, is the fact that several application can arise from this same setting. Regarding this, we point out how it could be interesting to apply the systems we develop for this task to evaluate teaching modules. Indeed, one could design a quality assessment for courses based on the level of agreement between subsequent chapters and sections and their prerequisite relations. A different application, could be the definition of a new way to move around Wikipedia itself, identifying which links move in the same direction as the prerequisite relation and which on the contrary move against it.

Let us now outline three main aspects common to different works tackling similar tasks. We will take into into account these specifics while developing our own models. The first is that hand crafted features are commonly used, in (Miaschi et al., 2019) they develop these features mostly analysing textual statistics, for example the occurrence of one concept in the page of another one. In (Liang et al., 2015) they also develop top down features, however the information they structure does not come from the body of the pages, instead they use the structure of Wikipedia as a graph with hyperlinks. Following this line, the second aspect is the use of graph structures. In most of the works predicting prerequisites, we see how they interpret pages as nodes and hyperlinks as edges. Both in (Talukdar and Cohen, 2012) and in (Liang et al., 2015) they use this feature, in some cases joining it with textual information, whereas in others as a stand alone one. On the contrary, in (Adorni et al., 2019) they use a bottom up graph structures created to help in the prediction. The third and last is the use of neural networks, as done in (Miaschi et al., 2019), where they are employed to create representations of text that can afterward be fed as features to simpler classifiers. We remark how structuring information into a graph is a practice used also in other tasks involving several documents. One example is topic modeling (Gerlach et al., 2018), it is interesting to notice how this task shares some of the steps needed for prerequisite learning. Indeed, in both cases one needs to crate a hierarchy of concepts which is then exploited in different ways. Since we wish to exploit textual knowledge, we can also employ word embeddings. For the Italian language they are developed in (Berardi et al., 2015). On top of them we will use ensemble methodologies since they can proficiently exploit information in these representations. Notice how in principle more modern techniques, such as transformer models (Devlin et al., 2019) could be used to help performance in this task, however as we will see we preferred not to do so. The main reason supporting this choice is the fact that the dataset provided for this task is not too big and thus we avoided too large models. The systems we developed try to enclose all these pieces of information we reported. Indeed, we try to exploit both knowledge strictly present within the Wikipedia pages provided for this task as well as information coming from the rest of the online encyclopedia.

2 Description of the System

In this report we describe the methodology we developed to tackle the PRELEARN task. We report the choices made and the steps that led us to them. In particular, We focused on the raw-text setting, for which we adopted two systems with the goal of prerequisite learning. Although both use the Wikipedia pages' texts, each one does it in different ways.

2.1 Model 1

This model exploits a combination of pretrained word embeddings, of GloVe type (Pennington et al., 2014), as trained for Italian in (Berardi et al., 2015) and handcrafted features, the latter inspired from (Miaschi et al., 2019). In particular, for each page title in a concept pair (A, B), we computed a 300-dimension vector by averaging the word embeddings of each word in the A/B title. These two resulting vectors were concatenated together with the following 14 handcrafted features.

- Is B(A) in A(B)'s text?
- Number of occurrences of B(A) in A(B)'s text
- Is B(A) in the first sentence of A(B)?
- Is B in A's title?
- Length of A(B)
- Jaccard similarity between the texts
- Jaccard similarity between nouns in the texts
- Difference in length between first paragraphs
- Difference in number of nouns in first paragraphs
- Jaccard similarity between nouns in first paragraphs

Then, for each pair (A,B) the final feature vector of 614 dimensions, was fed to a XGBoost classifier (Chen and Guestrin, 2016), whose model selection was performed via a nested cross validation with grid search.

2.2 Model 2

This model takes as information the first 400 words of each Wikipedia page, and for each pair (A,B) predicts if word B is a prerequisite for word A. It is composed of a Gated Recurrent Unit (Cho et al., 2014) with hidden size of 8 and encoding size 32, and a linear layer taking as input the concatenation of the two vectors representing the two Wikipedia pages to check and predict the prerequisite relation. This model, similar to model M1 in (Miaschi et al., 2019), though simpler, performs well enough and is fast to train. The parameters are chosen based on a grid search selecting the best results achieved on a validation set. The aforementioned values are the best performing choices for all settings and we keep them for the cross domain task as well. We tried different learning rates, though ultimately a constant one of 0.01 for the whole training was the best choice.

3 Discarded Models

We attempted to perform the structured data task as well, in particular adding the Wikipedia link

	Data-mining	Geometry	Physics	Precalculus
In-domain				
GRU + GCNConv[1]	0.74	0.74	0.85	0.84
Model 1	0.80	0.92	0.82	0.93
Model 2	0.81	0.91	0.81	0.89
Cross-domain				
Model 1	0.51	0.72	0.60	0.77
Model 2	0.48	0.71	0.61	0.77

Table 1: Accuracies obtained on the task test set. For the GCN see footnote.

structure to see if it would be useful. In order to exploit this knowledge we tried to use a Graph Convolutional Network (GCN) (Kipf and Welling, 2017). To do so we added the GCN between the Gated recurrent unit and the linear layer in Model 2 so as to perform the prediction based on the concatenation of the embedding of each node (Wikipedia page) in each pair. However this methodology resulted into lower scores in all dataset so we ended up not submitting it. We believe this is due to the fact that this is not the appropriate way to leverage the information present in the Wikipedia structure. Since we know from (Miaschi et al., 2019) that the information itself is relevant.

For Model 1 instead, a variation was tested with a multi-layer perceptron as well, but results were below those reported for the XGBoost ensemble.

An overall different approach we rejected is using transformer models. Indeed to obtain a representation of the text composing each page we could employ a representation extrapolated from BERT. However, after seeing how, much smaller models were overfitting the training set, we concluded that the amount of available textual data is not enough to exploit this model and avoided it.

4 Results

In Table 1 we report the achieved accuracy on the test set. As we can see, Model 1 outperformed Model 2. This is remarkable in the sense that the former is simpler than the one based on recurrent networks. The same can be said about the hand-crafted features, which are mostly statistics of each pair of pages based on occurrences. Indeed, as proven also in (Miaschi et al., 2019), this information does help the model. We believe Model 1 attained a higher score thanks to its pretrained word embeddings and the larger corpora they are trained upon. Indeed, the dataset used to create those vectors is composed of the whole Italian Wikipedia and of a large amount of novels. This encodes within these representations a wider knowledge than the one provided for this task only. Looking at the accuracy achieved with the GCN layer, we see how performances are systematically lower than the others, that is why we chose not to submit it.

After looking at the challenge results, we proceeded to explore more in general how well our models performed. In order to do so, for each one, we estimated precision, recall, accuracy and f1 score (reported in Table 2).

When comparing Model 1 and 2 between them, we noticed that the latter exhibited higher precision in 3 of the 4 areas, but also lower recall in 3 of them. As a result, there was a systematic difference in accuracy and f1-scores favouring Model 1 over Model 2. If we look closely at Model 1 scores in Table 2 we see how Physics and Precalculus show a broader difference between precision and recall. This underlines how in these two domains there are some concepts that despite being involved in several prerequisite relations are less represented in the general knowledge. Moreover, the same behavior is experienced for Model 2, indicating how the models started to miss some positive samples. The fact that it happens for this second setting makes us believe this phenomenon is also due to the presence of more spread information within the Wikipedia pages of the concepts enclosed in these domains. As we mentioned the second model has higher precision in three cases, whereas the first has higher recall, in two cases the

[1] Values from our own validation set split

	Precision	Recall	Accuracy	F1
Model 1				
data_mining	0.80	0.80	0.80	0.80
geometry	0.92	0.92	0.92	0.92
physics	0.84	0.82	0.82	0.81
precalculus	0.93	0.93	0.93	0.93
Model 2				
data_mining	0.82	0.80	0.81	0.81
geometry	0.90	0.91	0.91	0.91
physics	0.87	0.73	0.81	0.79
precalculus	0.95	0.82	0.89	0.88

Table 2: All scores obtained by Models 1 and 2.

difference in recall is much in favor of the latter and indeed it is the better performing one.

5 Discussion

Regarding the first model, we see how the vectorization obtained from the Wikipedia corpus performs well, particularly considering that it represents exclusively the pages' titles. We also notice that the comparison between the two models is not straighforward since the ensemble model we used was not tested on the vectors obtained from the recurrent neural networks. We did not experiment in this mixed setting, since we believe it would not make sense to deploy a methodology with the power of XGBoost on embeddings solely based on the information present in the pages provided for this task. Indeed, there are high chances that the results for such complex model would still be worse than the one with the pretrained embeddings, since, as we mentioned in Section 4, the knowledge available exclusively in the pages proposed for this task is limited.

The other remarkable aspect is that to surpass the performance of the GRU, handcrafted features were helpful, despite them being mostly word occurrences counts. This same information is available to the GRU models, which performs worse. This underlines how the recurrent architecture, though powerful and able to capture long distance relations, can not retain this type of substantial details. Regarding the second model introduced, we remark how the hidden units size and the encoding size are very small. This is coherent with the fact that the dataset is not large enough to exploit the scaling potential of a recurrent neural network with a larger size. However, with this small model the results are better than with a baseline and as we mentioned the training times are all quite small. Thus, the idea of performing more ablation studies where bag of words methodologies are used together with recurrent ones, could lead to further improvements still supporting a more bottom-up solution than hand crafted features.

Following the analysis of the models we used, we can conclude that the property of being a prerequisite is a complex characteristic and thus the use of large amounts of data can be useful. On the other hand, the fact that the model solely based on the data at hand performs only marginally worse than the other underlines how this information is present in the pages themselves. Possibly a mixed dataset contained between the one at hand and the whole Italian Wikipedia could be a solution to move further in prerequisites learning.

References

Giovanni Adorni, Chiara Alzetta, Frosina Koceva, Samuele Passalacqua, and Ilaria Torre. 2019. Towards the identification of propaedeutic relations in textbooks. In Seiji Isotani, Eva Millán, Amy Ogan, Peter M. Hastings, Bruce M. McLaren, and Rose Luckin, editors, *Artificial Intelligence in Education - 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I*, volume 11625 of *Lecture Notes in Computer Science*, pages 1–13. Springer.

Chiara Alzetta, Alessio Miaschi, Felice Dell'Orletta, Frosina Koceva, and Ilaria Torre. 2020. Prelearn @ evalita 2020: Overview of the prerequisite relation learning task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign* *of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. A network approach to topic models. *Science Advances*, 4(7).

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674, Lisbon, Portugal, September. Association for Computational Linguistics.

Alessio Miaschi, Chiara Alzetta, Franco Alberto Cardillo, and Felice Dell'Orletta. 2019. Linguistically-driven strategy for concept prerequisites learning on italian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 285–295.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Partha Talukdar and William Cohen. 2012. Crowdsourced comprehension: Predicting prerequisite structure in Wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 307–315, Montréal, Canada, June. Association for Computational Linguistics.