PRELEARN @ EVALITA 2020: Overview of the Prerequisite Relation Learning Task for Italian

Chiara Alzetta[•◇], Alessio Miaschi[⋆◇], Felice Dell'Orletta[◇],
Frosina Koceva[•], Ilaria Torre[•]

[•]DIBRIS, Università degli Studi di Genova, [⋆]Dipartimento di Informatica, Università di Pisa,
[◇]CNR, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa - ItaliaNLP Lab
{chiara.alzetta,frosina.koceva}@edu.unige.it, alessio.miaschi@phd.unipi.it,
ilaria.torreunige.it, felice.dellorletta@ilc.cnr.it

Abstract

The Prerequisite Relation Learning (PRELEARN) task is the EVALITA 2020 shared task on concept prerequisite learning, which consists of classifying prerequisite relations between pairs of concepts distinguishing between *prerequisite* pairs and *non-prerequisite* pairs. Four sub-tasks were defined: two of them define different types of features that participants are allowed to use when training their model, while the other two define the classification scenarios where the proposed models would be tested. In total, 14 runs were submitted by 3 teams comprising 9 total individual participants.

1 Introduction

The present paper provides an overview of the systems participating to PRELEARN, the first shared task on automatic prerequisite learning between educational concepts.

In the past decades we have witnessed a great revolution in the field of Education: advancement of technologies drastically transformed the teaching method and the setting of the learning process thanks to the raise of e-learning platforms and electronic educational materials. While so far they've been mainly used in lifelong learning, the current pandemic situation made very clear that distant learning is a valuable resource at all educational levels. This new era in education is commonly referred to as Education 4.0 (Saxena et al., 2017; Hussin, 2018; Salmon, 2019) and its main novelty is to put students at the core of every learning activity promoting the mission of fostering and improving personalisation techniques. While there is still much work to do to develop usable and scalable personalisation systems, much of the attention has been devoted to building and testing the building blocks of such applications.

The massive use of distance learning platforms has shed light on the need of developing intelligent agents able to support both students and teachers by, e.g., automatically identifying educational relations between learning concepts. Educational resources are designed to guide students through learning paths consisting of concepts related to each other. Among all pedagogical relations, prerequisite is the most fundamental since it establishes which sequence of concepts allows students to have a full understanding of the domain. In fact, the order in which concepts are presented to the learner plays a crucial role in avoiding student's frustration and misunderstandings while approaching a new topic, so teachers are very careful to organise the content of their learning materials accordingly and to highlight relevant connections to their students. Doing this automatically is still challenging from many perspectives.

The NLP community has tackled automatic prerequisite learning in the past with the goal of integrating prerequisite relations in systems for, e.g., curriculum planning (Agrawal et al., 2016), reading list generation (Gordon et al., 2017; Fabbri et al., 2018), automatic assessment (Wang and Liu, 2016) and automatic educational content creation (Lu et al., 2019). Wikipedia is rightfully considered a rich and freely available resource for training and testing educational applications, and this is also true in the case of prerequisite learning systems, which are often evaluated against manually annotated prerequisite relations between Wikipedia pages (Talukdar and Cohen, 2012; Gasparetti et al., 2018; Zhou and Xiao, 2019).

Based on the works available in the literature, we distinguish prerequisite learning systems in two main categories: 1) those based on relational metrics and 2) those on machine learning approaches. Relational metrics are designed to capture the strength of the relation between co-occurring concepts and identify pairs of concepts obtaining low values as non-prerequisites. The *RefD* metric (Liang et al., 2015) is possibly the most popular and measures how differently two concepts refer to each other considering the Wikipedia links of the pages associated with the concepts of the pair. Prerequisite concept learning from textbook concepts is addressed in Adorni et al. (2019), which presents a method based on burst analysis combined with temporal reasoning to identify possible propaedeutic relations and compare it with a concept co-occurrence metric. Among machine learning approaches, we distinguish between those that exploited link-based features (e.g. (Liang et al., 2015; Gasparetti et al., 2018)), text-based features only (e.g. (Miaschi et al., 2019; Alzetta et al., 2019)), or a combination of the two (Liang et al., 2018).

Unfortunately, the results obtained by those systems are not directly comparable: their approaches are based on different assumptions of what a concept is and which are the distinctive features for a prerequisite relation. Moreover, knowledge structures defined by domain experts are not always easily available or are missing for some domains. With PRELEARN, we are proposing the first shared task on automatic prerequisite learning, at least to the best of our knowledge. Located in the context of EVALITA 2020 evaluation campaign (Basile et al., 2020), the task challenges participants to develop prerequisite learning systems that can exploit either only information derived from textual educational resources or that can combine those information with structural properties of knowledge structure. We aim to compare the performances of systems based on these two different approaches and verify if they can obtain similar results or, conversely, one strategy is far better performing than the other. The goal of PRELEARN shared task is not only to offer a setting where different approaches and systems can be directly compared, but also to gather the research teams working on automatic prerequisite learning, which is distributed and doesn't have dedicated venues, and possibly fostering collaborations within the community. More broadly, we expect the outcomes of the task to be relevant to the wider information extraction and knowledge structure construction communities, as it offers the opportunity to test which information – either textual or extracted from a knowledge structure – are more effective for retrieving pedagogical relations in educational data.

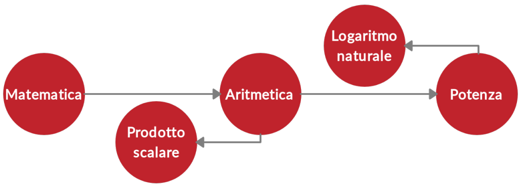


Figure 1: Example of prerequisite relations between concepts.

2 Task Description

PRELEARN (Prerequisite Relation Learning) is a shared task on concept prerequisite learning which consists of classifying prerequisite relations between pairs of concepts. This is the first time, to the best or our knowledge, that *automatic prerequisite learning* is addressed in a shared task. PRELEARN challenges participants to test their models for automatic prerequisite learning on four different domains and four training scenarios.

2.1 Problem Formulation

For the purposes of this task, prerequisite relations learning is proposed as a binary classification problem of concept pairs: given a pair of concepts (A, B), we ask to predict whether or not concept B is a prerequisite of concept A. We define a "*concept*" as single or multi word domain terms corresponding to the title of a page on the Italian Wikipedia: *Prodotto scalare* and *Aritmetica* are both concepts of the precalculus domain and are also the titles of two Italian Wikipedia pages. Prerequisite relations instead are dependency relations that naturally occur between educational concepts determining their learning precedence.

Consider the knowledge structure proposed as an example in Figure 1. Here, nodes represent concepts while links identify the prerequisite relations that connect them. According to the graph, "Aritmetica" is a prerequisite of "Potenza" since, if a student wants to understand what "Potenza" is, he/she has to know "Aritmetica" first. Hence, we formally define a *prerequisite relation* as a relation connecting a target and a prerequisite concept if the second has to be known in order to understand the first. In other words, the Wikipedia page of the prerequisite concept contains the prior knowledge required to understand the content of the Wikipedia page of the target concept.

2.2 Task Settings

We defined four sub-tasks for addressing automatic concepts prerequisite learning: two of them concern the model used by participants for tackling the task, the other two distinguish different classification scenarios where the proposed model can be tested. In order to make a valid submission, we asked participants to submit at least one model complying with at least one of these settings:
i) *Raw features setting* (RF): a model that acquires information only from raw text (e.g. textual content of the Wikipedia pages offered as training set, corpora for acquiring distributional representations, etc.);
ii) *Raw and structured features setting* (RnS): a model that can rely both on raw text and structured information (e.g. Wikipedia graph structure of a domain and metadata of a Wikipedia page, DBpedia, page hierarchical structure in terms of sections and paragraphs, etc.).

Each submitted model was tested in two evaluation scenarios, defined as follows:
i) *In-domain scenario*: the model(s) can be trained on data belonging to any domain, including the one appearing in the test set;
ii) *Cross-domain scenario*: the model(s) can be trained on data belonging to any domain but the domain of the test set.

Overall, we defined a total of four sub-tasks:
1) RF setting in an in–domain scenario;
2) RF setting in a cross–domain scenario;
3) RnS setting in an in–domain scenario;
4) RnS setting in an cross–domain scenario.

Only few work in the literature test their systems in a cross-domain scenario: our previous attempts in this direction (Miaschi et al., 2019) highlighted some issues in transferring the information acquired from one domain to an unknown one. At the same time, although the two proposed settings correspond to the most widely used approaches for automatic prerequisite learning, systems only rarely rely on textual information only, and when they do performances are generally worse than those obtained by exploiting structural information extracted from knowledge bases. This makes, in our view, the RF setting tested in the cross-domain scenario the most challenging sub-task.

2.3 Evaluation

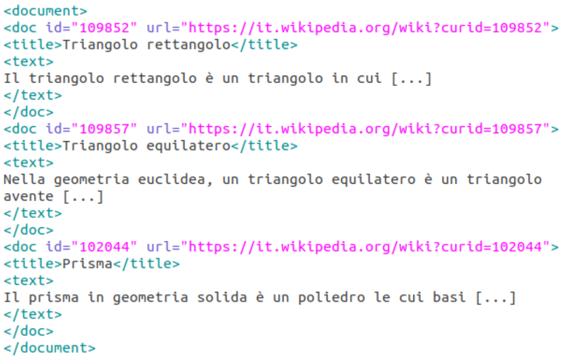
Metrics. Evaluation of participants' systems outputs was carried out on four balanced datasets, one for each domain, used for both in– and cross–domain evaluation. The size of the test sets is reported in Table 1. Each sub-task (i.e. each model on each scenario) was evaluated independently from the others by using standard metrics, such as Accuracy (A), Precision (P), Recall (R) and F$_1$-score (F_1). Since the test sets are balanced, we used Accuracy metric to rank participants' submitted runs.

Baseline. We used for all settings a linear SVM classifier trained using two binary features capturing the presence of a mention of concept *B*/*A* in the text of the Wikipedia page of concept *A*/*B*. Each feature returns 1 if the name of concept *B*/*A* is mentioned in the text of the Wikipedia page of concept *A*/*B*, while it returns 0 otherwise.

3 Data

We relied on ITA-PREREQ dataset (Miaschi et al., 2019), a dataset annotated with prerequisite relations between pairs of concepts in Italian. The dataset was built upon the AL-CPL dataset (Liang et al., 2018), a collection of binary-labelled concept pairs extracted from textbooks on four domains: data mining, geometry, physics and precalculus. In AL-CPL, for each domain, the authors extracted the relevant terms from the textbook: those appearing in the title of a English Wikipedia page were promoted as domain concepts and matched with their corresponding page. Finally, domain experts were asked to manually annotate the presence of absence o a prerequisite relation between all concept pairs. The final dataset consists of both positive and negative concept pairs that can be represented as a concept map, a specific type of knowledge graph where each node is a scientific concept and edges represent pedagogical relations.

The construction of ITA-PREREQ was carried out as follows, as described in (Miaschi et al., 2019). First, we took the Italian version of the Wikipedia pages considered for AL-CPL, excluding from the dataset those concepts (and the relations where they are involved) for which an Italian page was not available. Then, we mapped both positive and negative relations between pairs



```
<document>
<doc id="109852" url="https://it.wikipedia.org/wiki?curid=109852">
<title>Triangolo rettangolo</title>
<text>
Il triangolo rettangolo è un triangolo in cui [...]
</text>
</doc>
<doc id="109857" url="https://it.wikipedia.org/wiki?curid=109857">
<title>Triangolo equilatero</title>
<text>
Nella geometria euclidea, un triangolo equilatero è un triangolo
avente [...]
</text>
</doc>
<doc id="102044" url="https://it.wikipedia.org/wiki?curid=102044">
<title>Prisma</title>
<text>
Il prisma in geometria solida è un poliedro le cui basi [...]
</text>
</doc>
</document>
```

Figure 2: Example of Wikipedia pages (with cut off texts) from the "Wikipedia pages file".

of the remaining concepts from AL-CPL to ITA-PREREQ. As in AL-CPL, ITA-PREREQ dataset was expanded by creating irreflexive relations (add (*B*, *A*) as a negative sample if (*A*, *B*) is a positive sample) and transitive pairs (add (*A*, *C*) if both (*A*, *B*) and (*B*, *C*) are positive sample). In summary, ITA-PREREQ consists of pairs of concepts (*A*, *B*), labelled as follows: 1 if *B* is a prerequisite of *A* and 0 in all other cases. It was not allowed to use any sort of prerequisite-labelled data apart from ITA-PREREQ dataset provided by task organisers as official training set.

3.1 Format

PRELEARN participants were provided, upon request, with five files: a "concept pairs file" for each of the four domains containing the labelled concept pairs and one "Wikipedia pages file" containing the raw text and the link of the Wikipedia pages referring to the concepts appearing in the dataset. Here's an example of the pairs contained in the "concept pairs file":

```
Riflessione interna totale,Luce,1
Plasticita' (fisica),Durezza,0
...
Campo magnetico,Magnete,1
```

Figure 2 on the other hand shows an excerpt of the content of the "Wikipedia pages file". The content of the Italian Wikipedia pages was extracted using WikiExtractor[1] on a Wikipedia dump from January 2020.

3.2 Train and Test Sets

Table 1 provides a summary of the content of ITA-PREREQ, both for each domain covered by the dataset and overall. The number of concepts and pairs varies for each domain: while Geometry and Data Mining have a comparable amount of concepts, the latter shows a significantly smaller number of labelled pairs. It is interesting to note that, although not being the richer domain in terms of concepts, Physics shows the higher number of relations. As can be noted, regardless of the domain the dataset is strongly unbalanced since the majority of concept pairs do not show a prerequisite relation (*Non-PR Pairs*). For each domain we split the pairs into a portion of training and a portion of test data. For the test portion, we defined a fixed number of pairs to include (i.e. 200 pairs), with the exception of Data Mining where, given the limited number of total pairs, we included only 99 pairs. The distribution of prerequisite and non-prerequisite labels was balanced (50/50) for each domain only in the test datasets.

4 Participants

Following a call for interest, 16 teams registered for the task and thus obtained the training data. Eventually, three teams submitted their predictions, for a total of 14 runs, each executed on all four domains of the dataset. Two teams participated in all four sub-tasks while one team submitted results only for the two sub-tasks involving the RF setting. A summary of participants is provided in Table 2.

4.1 Submitted Systems

NLP-CIC (Angel et al., 2020) presented three different systems trained on both hand-crafted and embedding-based features. In particular, the team developed one model for the RF setting and two models for the RnS setting. Concerning the RF setting, the submitted model corresponds to a single layer Neural Network trained using concept pairs representations extracted from a BERT Italian model[2] fine-tuned on the training datasets. With respect to the RnS setting, the two submitted models are quite similar and differ only for one feature. The first model (Complex) is based on a tree-ensemble learner and trained it using a set of complexity-based features based on those defined by Aroyehun et al. (2018) combined with a feature capturing concept view frequency, i.e. the daily average of unique visits to the concept page by Wikipedia users (including editors, anonymous

[1] https://github.com/attardi/wikiextractor

[2] https://huggingface.co/dbmdz/bert-base-italian-cased

Domain	Concepts	Pairs	PR Pairs	non-PR Pairs	Pairs in Train set	Pairs in Test set
Data Mining	76	523	159 (30.40%)	364 (69.59%)	424	99
Geometry	74	1,748	432 (24.71%)	1,316 (75.28%)	1,548	200
Physics	130	2,420	415 (17.14%)	2,005 (82.85%)	2,220	200
Precalculus	177	1,916	508 (26.51%)	1,408 (73.48%)	1,716	200
Total	457	6,607	1,514 (22.91%)	5,093 (77.08%)	5,908	699

Table 1: Number of concepts, pairs, pairs showing a prerequisite [PR Pairs] (absolute and relative) or non-prerequisite relation [non-PR Pairs] (absolute and relative) for each domain of the ITA-PREREQ dataset. We also report the number of pairs (either prerequisite or not) released in the official training and test sets.

Team	Research Group	# Tasks	# Runs
NLP-CIC	Instituto Politécnico Nacional	4	6
B4DS	Università di Pisa	2	4
UNIGE_SE	Università degli Studi di Genova	4	4

Table 2: Teams participating in EVALITA 2020 PRELEARN shared task with number of sub-tasks they particpated in and number of submitted runs.

editors and readers) over the last year. The second model (Complex+wd) is an improved version of the first one: it takes as input the same set of features along with the Wiki-data embedding of each concept appearing in the concept pairs of ITA-PREREQ dataset.

B4DS (Puccetti et al., 2020) presented two different classification models, one based on XGBoost (Chen and Guestrin, 2016) classifier and one based on a Gated Recurrent Unit (GRU) model. The first classifier, Model 1, was trained using a combination of lexical and hand-crafted features. Specifically, lexical features were computed by averaging 300-dimensions pretrained word2vec embeddings (Berardi et al., 2015) of title *A* and *B* respectively, with *A* and *B* being the two concepts involved in a pair. The set of 14 hand-crafted text-based features, inspired by Miaschi et al. (2019), are extracted for each pair of the datasets and aim at capturing mentions and lexical similarity between the two pages associated with the concepts in the pair. The second classifier (Model 2) was trained with a GRU model (hidden size=8, encoding size=32, learning rate=0.01) that takes as input the first 400 words of each Wikipedia page of the (*A*, *B*) pair. The output was computed with a linear layer that takes the concatenation of the two learned vectors.

UNIGE_SE (Moggio and Parizzi, 2020) proposed a classifier based on a two-dense-layers Neural Network trained using a set of features automatically extracted from the Wikipedia pages associated with the concepts appearing in ITA-PREREQ dataset. In particular, the RF model was trained exploiting features that capture concepts co-occurrence and the lexical similarity between the pages referring to the concepts of a pair. On the other hand, the RnS model is trained combining the previous set of features with information based on the hyperlink and category structure of Wikipedia.

5 Results

In this section we provide both a discussion of the approaches and an analysis of the results reported in Tables 3 and 4.

Participants experimented with more classical machine learning algorithm as well as with Neural Networks (NN): we received results computed exploiting 7 different systems, 4 trained using only raw text features (RF setting) and 3 exploiting also structural information (RnS setting). Considering their average performances across all four domains, all systems outperformed the baseline. In this Section, we describe the results obtained by the submitted models and compare their performances on the official test set based on their average accuracy scores over the four domains (column *AVG* in the Tables).

5.1 Comparing Scenarios

In–Domain Scenario. As shown in Table 3, overall the model showing the best performances is Italian BERT, achieving an average accuracy score of 0.887 in the RF setting. Such result is not surprising if we consider the state-of-the-art performances obtained by recent Neural Language Models in the resolution of downstream NLP tasks. However, results obtained by BERT show only a small gap with respect to some of the other models. For instance, B4DS' Model

RF Setting

Place	Team	Model	Data mining	Geometry	Physics	Precalculus	AVG
1	NLP-CIC	BERT	0.838	0.925	0.855	0.930	0.887
2	B4DS	Model 1	0.797	0.920	0.815	0.930	0.866
3	B4DS	Model 2	0.808	0.905	0.810	0.890	0.853
4	UNIGE_SE	NeuralNet	0.595	0.620	0.530	0.675	0.605
5	Baseline	Occurrence	0.494	0.675	0.500	0.675	0.586

RnS Setting

Place	Team	Model	Data mining	Geometry	Physics	Precalculus	AVG
1	NLP-CIC	Complex+wd	0.808	0.905	0.795	0.915	0.856
2	NLP-CIC	Complex	0.828	0.895	0.785	0.885	0.848
3	UNIGE_SE	NeuralNet	0.565	0.755	0.725	0.755	0.700
4	Baseline	Occurrence	0.494	0.675	0.500	0.675	0.586

Table 3: Results in terms of Accuracy of the EVALITA 2020 PRELEARN RF and RnS models in the in–domain evaluation setting for each domain and on average.

RF Setting

Place	Team	Model	Data mining	Geometry	Physics	Precalculus	AVG
1	NLP-CIC	BERT	0.565	0.785	0.635	0.775	0.690
2	B4DS	Model 1	0.505	0.720	0.600	0.765	0.648
3	B4DS	Model 2	0.484	0.710	0.605	0.785	0.646
4	UNIGE_SE	NeuralNet	0.565	0.515	0.465	0.595	0.535
5	Baseline	Occurrence	0.494	0.500	0.605	0.500	0.525

RnS Setting

Place	Team	Model	Data mining	Geometry	Physics	Precalculus	AVG
1	NLP-CIC	Complex+wd	0.535	0.775	0.600	0.760	0.668
2	NLP-CIC	Complex	0.494	0.735	0.595	0.730	0.639
3	UNIGE_SE	NeuralNet	0.545	0.665	0.560	0.710	0.620
4	Baseline	Occurrence	0.494	0.500	0.605	0.500	0.525

Table 4: Results in terms of Accuracy of the EVALITA 2020 PRELEARN RF and RnS models in the cross–domain evaluation setting for each domain and on average.

1, exploiting a decision tree based on XGBoost framework and trained using both word embedding and handcrafted features, achieved 0.866 accuracy thus gaining the second place in the in–domain scenario. Similar competitive results are obtained by the Complex+wd model submitted by NLP-CIC team: this model combines Wiki-data embedding of each concept with a set of manually defined features that measure concept complexity and were designed to solve the task of complex word identification (Aroyehun et al., 2018). B4DS team submitted also a more sophisticated model (i.e. a GRU-based classifier) trained using only Word2vec embeddings with no other handcrafted features. Considering the results, combining lexical features, like word embeddings, with handcrafted features allows to achieve better performances regardless of the model employed for classification, while using these two types of features independently seems a worse strategy. As proof, B4DS' Model 2, despite being more sophisticated, achieved lower scores than Model 1. The fact that these models obtained similar results suggests that automatic prerequisite learning is more affected by predictors rather than the model used for classification.

Among submitted systems, only three didn't exploit word embeddings: NLP-CIC team submitted a tree-ensemble learner trained using only complexity features, and UNIGE_SE team used two versions of a two-layer NN trained with different sets of handcrafted features to comply with settings requirements. The results obtained by these models provide some interesting insights on the role of raw and structural features for solving the task. First, we observe that exploiting raw textual features based on lexical similarity and topic modelling (UNIGE_SE NN in the RF setting) only slightly outperforms the baseline, thus, when no lexical features are available, it seems more useful to rely on structural information. Anyways, complexity-based features exploited by NLP-CIC are more informative for prerequisite learning task than Wikipedia category and link structure. The intuition behind the NLP-CIC team approach is that less complex concepts are prerequisite for the more complex ones and, considering that the results are only slightly below those obtained using word embeddings, the intuition that complexity is involved in the process of defining prerequisite sequences seems confirmed.

Cross–Domain Scenario. Moving to the cross–domain evaluation scenario (see Table 4), we observe only small variations in the ranking of the submitted systems. In spite of this, we also observe a consistent drop of the accuracies obtained by the submitted systems.

Considering again the average accuracy scores, BERT model proved to be the best performing model also in this scenario. Interestingly, this time NLP_CIC's Complex+wd model outperforms B4DS's Model 1: both models are trained using both word embeddings and handcrafted features, with the latter being more useful possibly because capturing domain independent properties. The different performances of the two systems could be again due to the higher effectiveness of complexity-based features for identifying prerequisite relations. Consequently, these results suggest that, unlike the in-domain scenario, lexical information are not enough to identify prerequisite relations. Nevertheless, lexical features proved somehow useful since using handcrafted features only, as in the case of Complex NLP-CIC model and the NN models submitted by UNIGE_SE team, is outperformed by B4DS's Model 2 (based solely on word embeddings).

5.2 Domains Impact

Focusing on the differences between the four domains, we observe that for almost all submitted systems the results obtained on concept pairs belonging to the Data Mining domain are lower than the others. This is especially true for the cross–domain scenario and seems to corroborate what was already stated in Miaschi et al. (2019), namely that Data Mining is a relatively new and more specialised topic that presents shorter pages and, therefore, that contains less clear prerequisite relationships. Nevertheless, the model submitted by the UNIGE_SE team for the RF setting achieved the lowest results when tested on concept pairs belonging to the Physics domain.

With the exception of the UNIGE_SE's RF model in the cross–domain setting, all systems achieved best (and similar) results when classifying Geometry and Precalculus concepts pairs. This might be due to the fact that these two domains are more fundamental and broad subjects and, therefore, present more clear learning dependencies expressed through Wikipedia. Furthermore, since Geometry and Precalculus share more lexicon that the others, we believe that the models can take advantage of this overlap to better classify concept pairs, especially for the cross–domain evaluation setting.

6 Conclusion

Automatic prerequisite learning was for the first time the focus of a dedicated shared task. In particular, PRELEARN task was aimed at comparing the performances of different approaches and models tested within and across the four domains of ITA-PREREQ dataset. Although the results of 14 submitted runs were all above baseline, we observe several differences within the proposed settings and across domains. In particular, results suggests that automatic prerequisite learning is more affected by the predictors rather than by the classification models. Results also confirm that the RF cross–domain setting is the most challenging scenario. Nevertheless, BERT achieved best scores in both RF settings, also outperforming models trained with structural features extracted from the knowledge structure of Wikipedia.

For the future, it would be interesting to test the impact of hand-crafted features combined with a contextual language model such BERT and, considering the effectiveness of complexity–based features, explore the contribution of predictors encoding text readability properties in prerequisite learning systems.

References

Giovanni Adorni, Chiara Alzetta, Frosina Koceva, Samuele Passalacqua, and Ilaria Torre. 2019. Towards the identification of propaedeutic relations in textbooks. In *International Conference on Artificial Intelligence in Education (AIED)*. Springer.

Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. 2016. Toward data-driven design of educational courses: A feasibility study. *Journal of Educational Data Mining*, 8(1):1–21.

Chiara Alzetta, Alessio Miaschi, Giovanni Adorni, Felice Dell'Orletta, Frosina Koceva, Samuele Passalacqua, and Ilaria Torre. 2019. Prerequisite or not prerequisite? that's the problem! an nlp-based approach for concept prerequisites learning. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481. CEUR-WS.

Jason Angel, Segun Taofeek Aroyehun, and Alexander Gelbukh. 2020. Nlp-cic @ prelearn: Mastering prerequisites relations, from handcrafted features to embeddings. 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.

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 322–327.

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*, pages 785–794.

Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. TutorialBank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–620, Melbourne, Australia, July. Association for Computational Linguistics.

Fabio Gasparetti, Carlo De Medio, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. 2018. Prerequisites between learning objects: Automatic extraction based on a machine learning approach. *Telematics and Informatics*, 35(3):595–610.

Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully Burns. 2017. Structured generation of technical reading lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 261–270.

Anealka Aziz Hussin. 2018. Education 4.0 made simple: Ideas for teaching. *International Journal of Education and Literacy Studies*, 6(3):92–98.

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.

Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C Lee Giles. 2018. Investigating active learning for concept prerequisite learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Weiming Lu, Pengkun Ma, Jiale Yu, Yangfan Zhou, and Baogang Wei. 2019. Metro maps for efficient knowledge learning by summarizing massive electronic textbooks. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–13.

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.

Alessio Moggio and Andrea Parizzi. 2020. Unige_se @ prelearn: Utility for automatic prerequisite learning from italian wikipedia. 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.

Giovanni Puccetti, Luis Bolanos, Filippo Chiarello, and Gualtiero Fantoni. 2020. B4ds @ prelearn: Ensemble method for prerequisite learning. 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.

Gilly Salmon. 2019. May the fourth be with you: Creating education 4.0. *Journal of Learning for Development-JL4D*, 6(2).

Rajan Saxena, Vinod Bhat, and A Jhingan. 2017. Leapfrogging to education 4.0: Student at the core.

Partha Pratim Talukdar and William W Cohen. 2012. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 307–315. Association for Computational Linguistics.

Shuting Wang and Lei Liu. 2016. Prerequisite concept maps extraction for automatic assessment. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 519–521. International World Wide Web Conferences Steering Committee.

Yang Zhou and Kui Xiao. 2019. Extracting prerequisite relations among concepts in wikipedia. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.