

Linguistically-Driven Strategy for Concept Prerequisites Learning on Italian

Alessio Miaschi^{*,◊}, Chiara Alzetta^{*,◊}, Franco Alberto Cardillo[◊], Felice Dell’Orletta[◊]

^{*}Dipartimento di Informatica, Università di Pisa

[•]DIBRIS, Università degli Studi di Genova

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR), Pisa

ItaliaNLP Lab – www.italianlp.it

alessio.miaschi@phd.unipi.it,

chiara.alzetta@edu.unige.it,

{francoalberto.cardillo,felice.dellorletta}@ilc.cnr.it

Abstract

We present a new concept prerequisite learning method for Learning Object (LO) ordering that exploits only linguistic features extracted from textual educational resources. The method was tested in a cross- and in-domain scenario both for Italian and English. Additionally, we performed experiments based on an incremental training strategy to study the impact of the training set size on the classifier performances. The paper also introduces ITA-PREREQ, to the best of our knowledge the first Italian dataset annotated with prerequisite relations between pairs of educational concepts, and describe the automatic strategy devised to build it.

1 Introduction

Learning Objects (LO) are digital or non-digital educational resources deliverable over the Internet that can be employed in technology-supported learning (Wiley, 2000). According to the Learning Technology Standards Committee, being small and re-usable educational elements (e.g. lecture notes, multimedia content, presentations) is what mostly distinguishes LOs from other educational resources (IEEE, 2002). Recommendations for creating LOs in fact suggest that, although there is no standard LO structure, the content should be direct, succinct and homogeneous (Thompson and Yonekura, 2005). Grounded in the notion of object-oriented computing and programming, LO are designed according to the idea that combining small chunks of knowledge is what builds up an effective learning path. In order to promote sharing and re-usability, LO repositories were made available on the web, where LOs are stored, collected and can be searched by means of metadata provided by their authors (Tzikopoulos et al., 2009). Teachers and instructional designers can highly benefit from LO repositories since they can use

them to build educational materials such as textbooks, courses or, more in general, learning paths by combining various LOs of the same subject.

Being able to give a pedagogical meaning to the content of a set of LOs by ordering them respecting their pedagogical precedence is not trivial: uncovering educational relationship between LOs is a difficult and time consuming practice usually performed by domain experts (Gordon et al., 2017). Among all pedagogical relations, the most fundamental is the prerequisite relation, which best describes pedagogical precedence since it defines what one needs to know before approaching a new content.

Previous work in course and LO sequencing and knowledge tracing infers prerequisite relation between LOs based on their metadata and/or students’ preferences and competences (De-Marcos et al., 2009; Vuong et al., 2011; Piech et al., 2015; Méndez et al., 2016). Educational Data Mining methods usually rely also on graph information of ontologies, university programs or Wikipedia graph structure (Scheines et al., 2014; Chen et al., 2016).

In this paper we present a novel method based on deep learning applied to the task of automatic prerequisite relations identification between concepts to automatically create pedagogically motivated sequences of LOs. To the best of our knowledge, this is the first method that exploits exclusively linguistic feature extracted from textual resources. Considering only textual content is possibly the most complex condition to infer relationships between educational concepts since it cannot rely on any structured information. At the same time this is also the closest condition to a real world scenario, hence we aim to demonstrate that textual content can be sufficient to infer a pedagogically motivated ordering of LO pairs.

To verify the effectiveness of our strategy, we

performed experiments on the AL-CPL dataset (Liang et al., 2018b), an English dataset manually annotated with prerequisite relations between educational concepts, and on an Italian dataset we created. Hence, we introduce ITA-PREREQ¹, the first Italian dataset, to the best of our knowledge, annotated with prerequisite relations between pairs of concepts, built completely automatically.

Along the paper, we use the terms *Learning Object* (LO) and *Wikipedia page* interchangeably: in a broad sense, Wikipedia entries can be considered Ls (Nash, 2005), moreover previous work in related fields represent educational units as Wikipedia pages (Gasparetti et al., 2018). This fits our needs since a Wikipedia page consists of textual content pertaining to a single unit of learning. The term *concept* is also frequently used in the literature referring to educational units in general, and annotated dataset are usually described as identifying prerequisite relations between concepts. In this paper we use the term *concept* relying on the same sense of Liang et al. (2018b) as equivalent to the term *LO*.

The remaining part of the paper is organised as follows. First we present related work (Sec 2), then, after briefly presenting our approach (Sec. 3), we describe in more detail the data (Sec. 3.1), used features (3.2) and the classifier (Sec. 3.3). We also provide an insight of feature analysis in Sec 3.2.1. Experiments, results and incremental training tests are described in Section 4. In Section 5 we conclude the paper.

Our Contribution. In this paper, we present: (i) the first system based on neural network which exploits only linguistic features extracted from LO content and does not rely on Wikipedia graph or LO metadata information; (ii) the first Italian dataset annotated with prerequisite relations between pairs of concepts (ITA-PREREQ) and the automatic strategy devised to construct it; (iii) the first system for prerequisite relations extraction on Italian.

2 Related Work

Identifying prerequisite relations between educational materials is a task that has recently gained much attention both in the NLP community, aided by the fact that it is applicable to many contexts,

such as curriculum planning (Agrawal, 2016), course sequencing (Vuong et al., 2011), reading list generation (Gordon et al., 2017), automatic assessment (Wang and Liu, 2016) and domain ontology construction (Zouaq et al., 2007; Larranaga et al., 2014).

NLP techniques usually exploit structured information (e.g. hyperlinks, citations, DBpedia structure) combined with content-based information extracted from educational materials, like scientific literature (Gordon et al., 2016; Li et al., 2019), knowledge units in courses (Yang et al., 2015; Chaplot et al., 2016; Pan et al., 2017; Li et al., 2019) or Learning Objects (Gasparetti et al., 2018), often understood as Wikipedia pages (Gasparetti et al., 2015). Talukdar and Cohen (2012) presented the first work on predicting prerequisite structure of concepts using Wikipedia, which eventually became the most widely used resource for this task. They collected a manually annotated dataset of page pairs using crowd-sourcing and then trained a MaxEnt classifier using Wikipedia graph features, page content and edits to reproduce the prerequisite structure between pages. The classifier was tested both in and across domains, obtaining higher results in terms of accuracy if compared against a random baseline. The same dataset was used by (Liang et al., 2015) to test the RefD metric, that models the prerequisite relation by measuring how differently two concepts refer to each other using the tf-idf measure. Results are comparable with the MaxEnt classifier but the metric does not take into account all the available information in the resource so we argue that it could be improved further. The RefD metric was also used by (Wang et al., 2016) in a method that jointly extracts relevant concepts and prerequisite structure from textbooks exploiting also external knowledge from Wikipedia. Relying on textbooks but not on structured resources, Adorni et al. (2019) describe a method to infer prerequisite relations between concepts using burst analysis of concept occurrences in text and patterns based on temporal reasoning to identify possible propaedeutic relations.

Machine and deep learning techniques have been applied only recently to the prerequisite learning task. In (Liang et al., 2018b,a), the authors investigated the effects of integrating an active learning strategy in automatic extraction of prerequisites using a Random Forest classifier.

¹<http://www.italianlp.it/resources/>

Gasparetti et al. (2018) proposed a ML methods based on Multilayer Perceptron exploiting LOs, Wikipedia pages of concepts mentioned in the LOs and Wikipedia hierarchical category structure. Roy et al. (2018) presented a supervised learning method using a Siamese Network to predict prerequisite relations between University and MOOC courses.

The above methods strictly rely on Wikipedia graph information, which they report as highly informative, but that is not available when applying the method on different educational materials. We show how comparable results can be obtained considering only textual information.

Another acknowledged limit of the above methods is the need of large annotated datasets. Manual annotation by domain experts is the most commonly adopted strategy to build such resources, regardless the knowledge unit considered (Wang et al., 2016; Pan et al., 2017; Liang et al., 2017; Alzetta et al., 2018; Fabbri et al., 2018), with the notable exception of the crowd-sourcing strategy of Talukdar and Cohen (2012). The dataset we present in this paper is the first dataset annotated with prerequisite relations between concepts for Italian build completely automatically.

3 Our Approach

We tackle the problem of LO ordering as a task of automatic prerequisite relationship identification between LOs, here defined as follows: given a pair of LOs (A, B), we predict whether or not B is a prerequisite of A . As mentioned above, we define a LO as a concept corresponding to a Wikipedia page.

We trained deep learning models to predict whether or not two concepts are in a prerequisite relationship using a pre-trained word embedding lexicons and a set of linguistic features extracted from the pages of the concepts in the pair. The model was tested on two datasets: ITA-PREREQ, an Italian dataset annotated with prerequisite relations, and, to prove the effectiveness of the model, also on AL-CPL, an English dataset already used for the task of automatic prerequisite identification. In particular, the AL-CPL dataset was used both in its original and reduced version, as described in the next Section.

AL-CPL			
Domain	Concepts	Pairs	Prerequisites
Data Mining	120	826	292
Geometry	89	1,681	524
Physics	153	1,962	487
Precalculus	224	2,060	699
Total	586	6,529	2,002
ITA-PREREQ / English Reduced			
Domain	Concepts	Pairs	Prerequisites
Data Mining	75	429	154
Geometry	73	1,338	430
Physics	131	1,651	409
Precalculus	176	1,504	502
Total	455	4,922	1,495

Table 1: Number of concepts, pairs, pairs showing a prerequisite relation for each domain of each dataset and total values considering all domains for each dataset.

3.1 Dataset

For our experiments on the English language, we relied on the AL-CPL Dataset (Liang et al., 2018b), which is in turn based on the Wiki Concept Map dataset (Wang et al., 2016).

The Wiki Concept Map dataset is a manually constructed dataset consisting of binary-labelled concept pairs collected from textbooks on different educational domains: data mining, geometry, physics and precalculus. Concepts mentioned in the textbooks and appearing in the title of a Wikipedia page were considered domain concepts. Among them, key concepts and prerequisite relationships between them were annotated by experts for each domain, resulting in a concept map, a specific type of knowledge graph where each node is a scientific concept and edges represent pedagogical relations. Pairs not having prerequisite relation were also annotated, therefore the final dataset consists of both positive and negative pairs.

In Liang et al. (2018b) the dataset was expanded by adding (i) irreflexive and (ii) transitive relations: considering A, B and C as distinct concepts, (i) add (B, A) as a negative sample of (A, B) ; (ii) add (A, C) as positive sample if (A, B) and (B, C) are positive samples.

The AL-CPL dataset was also used by us to build ITA-PREREQ, the first Italian dataset annotated with prerequisite relation between pair of concepts, which we used to test our model on Italian. Considering the concepts of the AL-CPL dataset, we retrieved their Italian Wikipedia pages by matching the page title with the concept name. If the Italian page of a concept was not available,

the concept was excluded from the dataset. At the end of this process, we obtained an automatically constructed version of the AL-CPL dataset for Italian with a subset of 418 concepts (77.40% of the original dataset).

Note that the dataset only provides concept names (i.e. page titles), which means that downloading the pages from Wikipedia at different times might result in a slightly different corpus, since Wikipedia pages are frequently edited. In our case, we used the latest Wikipedia dump at the time of the experiments (February 2019).

Considering such Wikipedia impact factors (i.e. editing and differences between languages), we created a third dataset, again generated starting from AL-CPL. We call this version *English Reduced* since it is built excluding all those English Wikipedia pages that do not have a corresponding Italian one. Therefore, the size of English Reduced is the same of ITA-PREREQ. The aim of having this dataset is to check the real impact of languages differences by balancing the number of pages taken into account, as we will discuss further in the next Section.

Table 1 summarises the statistics of the three datasets. Although ITA-PREREQ and English Reduced resulted in smaller datasets in terms of both concepts and relations, their sizes are suited for training our systems.

3.2 Features

For each concept pair, we extracted two different sets of linguistic features: (i) *lexical features*, i.e. features that pertain to a single concept/page, and (ii) *global features*, i.e. features derived from the combination of concepts in pairs. All features were extracted for the AL-CPL, ITA-PREREQ and English Reduced datasets.

Hereafter, we denote by A and B the content of the Wikipedia page A or B ; (A, B) is how we refer to the concept pair, while A_t/B_t refers to the title of the corresponding page.

Lexical features. The first type of feature corresponds to pre-trained word embeddings (WE) computed for the first 400 words of each Wikipedia page. Specifically, we used a WE lexicon with 128 dimensions built with word2vec (Mikolov et al., 2013) both for Italian and English. We generated the two lexicons using the itWac and ukWac corpora, two collections of approximately 2 billion words constructed from web pages under,

respectively, the .it and .uk domains (Baroni et al. 2009, Ferraresi et al. 2008).

Global features. The second type of feature was devised to extract linguistic information from both A and B Wikipedia pages. Specifically, for each pair (A, B) , we extracted the following set of 16 text-based features:

- In text (#1, #2): if B_t/A_t appears in A/B .
- Count (#3, #4): how many times B_t/A_t is mentioned in A/B .
- In first line (#5, #6): if B_t/A_t appears in A/B 's first line, i.e. A/B definition.
- In title (#7): If B_t appears in A_t .
- Length (#8, #9): the number of words of A/B .
- Jaccard Sim. (#10): the Jaccard similarity between A and B .
- Jaccard Sim. (Nouns) (#11): the Jaccard similarity between nouns appearing in A and B .
- RefD (#12): the RefD metric between A and B (Liang et al., 2015).
- LDA Entropy (#13, #14): the Shannon entropy of the LDA vector of A/B . Note that we trained three different LDA (Deerwester et al., 1990) topic models, one for each dataset.
- LDA Cross Entropy (#15, #16): the cross entropy between the LDA vector of A/B and B/A .

Features from #1 to #6, #8, #9 and from #13 to #16 were used also in Liang et al. (2018a), but we expanded their set of features considering mentions in titles and Jaccard Similarities between both whole page contents and nouns only. Our set of linguistic features includes also the RefD metric (Liang et al., 2015), a feature usually considered a graph-based feature, that we adapted in order to be applicable also to those contexts where no structured information (i.e. hyperlinks) is provided. In fact, contrary to Liang et al. (2015) where the RefD value is computed considering hyperlinks between Wikipedia pages, we computed the metric using the mention of concept B/A in the page content of A/B , regardless the association of an hyperlink to that mention. Specifically, we im-

plemented RefD as follows:

$$RefD(A, B) = \frac{\sum_{i=1}^N r(c_i, B) \cdot w(c_i, A)}{\sum_{i=1}^N w(c_i, A)} - \frac{\sum_{i=1}^N r(c_i, A) \cdot w(c_i, B)}{\sum_{i=1}^N w(c_i, B)}$$

where c_i is a concept from our concept space C (all the Wikipedia articles in the domain); $r(c_i, B)$ is a binary indicator showing whether c_i is mentioned in the content of page B ; $w(c_i, A)$ is a weight indicator of the importance of c_i to page A (measured in terms of tf-idf).

3.2.1 Feature Analysis

In order to understand the relevance and behaviour of the *global features* in different domains and languages, we decided to perform feature analysis.

Following Liang et al. (2018a), we computed the feature importance by "mean decrease impurity" using an Extra-Trees Classifier, an implementation of a decision tree classifier. We decided also to perform the analysis using both the ITA-PREREQ and English Reduced dataset in order to compare the results in a cross-lingual scenario.

As we can see in Table 2, the results obtained from the English Reduced dataset show that, despite the ranking positions, there are many features that are common to all four domains. Specifically, the top features are RefD, LDA (entropy and cross-entropy), Length of B and B_t in first line of A. Moreover, we can notice that, except for graph features, our results are comparable to those obtained by Liang et al. (2018a). Nevertheless, the two rankings present some differences: e.g. Length of A/B tends to be more significant for our dataset. This could be to the fact that, as mentioned, the Wikipedia version used for our experiments could be different if compared to the one used by Liang et al. (2018b). It is also important to notice that the English Reduced dataset contains less concept pairs with respect to the original one.

Comparing the results according to the two languages, we can notice that the most important features tend to be quite similar. We can, however, identify some differences. For instance, we observe that Data Mining in ITA-PREREQ is the only domain for which RefD is not the most significant feature. Interestingly enough, the first 4 positions in the English Reduced dataset for Geometry, Physics and Precalculus seem to be more homogeneous when compared to the Italian ones.

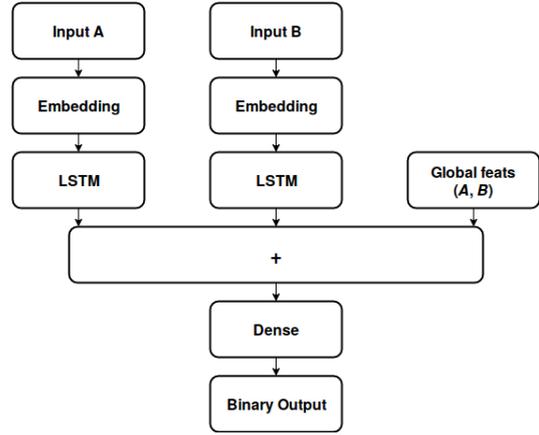


Figure 1: M3 architecture. M1 roughly corresponds to the left part of the architecture.

3.3 Classifier

For our LOs ordering experiments, we tested three different neural network models: (**M1**) one that learns to classify the binary labels using only pre-trained WE, (**M2**) one that learns using the global features automatically extracted and (**M3**) the last one which merges M1 with the input of M2.

M1 is composed of two identical LSTM-based sub-networks with 32 units, whose outputs are concatenated and classified by the outer Dense Layer. Each sub-network receive as input the WE of the first 400 words of the corresponding Wikipedia page of a given concept pair (A, B) . The two LSTM outputs are then concatenated ($V_A \oplus V_B$) and passed to a last Dense Layer.

M2 is based on a feedforward neural network that takes as input the global features of the pair (A, B) and passes them to a multilayer perceptron neural network (3 layers with ReLU activation).

M3 (represented in Figure 1) combines the previous two, joining the two sub-networks of M1 with the input of M2.

Each output layer of the three models consists of a single dense unit with sigmoid activation function.

The models are trained maximising the F-Score on the validation set, which corresponds to the 30% of the training data. The training stops after a certain number of epochs without improvement.

4 Experiments

We tested our approach predicting in-domain and cross-domain prerequisite relationships. Since the majority of (A, B) pairs do not present a prerequisite relation, we balanced the training and test sets

ITA-PREREQ			
Data Mining	Geometry	Physics	Precalculus
Length of B	RefD	RefD	RefD
Length of A	B_t in text of A	Length of B	Length of B
RefD	Length of B	LDA entropy of B	B_t in first line of A
Jaccard Sim.	LDA entropy of B	Length of A	LDA entropy of B
Jaccard Sim. (Nouns)	B_t in first line of A	LDA cross-entropy of B/A	Length of A
LDA entropy of B	A_t in text of B	LDA cross-entropy of A/B	Jaccard Sim.
LDA entropy of A	Length of A	LDA entropy of A	B_t in text of A
LDA cross-entropy of B/A	LDA entropy of A	Jaccard Sim.	Jaccard Sim. (Nouns)
LDA cross-entropy of A/B	LDA cross-entropy of A/B	Jaccard Sim. (Nouns)	LDA cross-entropy of A/B
B_t in first line of A	LDA cross-entropy of B/A	B_t in first line of A	LDA entropy of A
English Reduced			
Data Mining	Geometry	Physics	Precalculus
RefD	RefD	RefD	RefD
Length of B	B_t in first line of A	B_t in first line of A	B_t in first line of A
LDA entropy of B	LDA entropy of B	Length of B	Length of B
LDA entropy of A	Length of B	LDA entropy of B	LDA entropy of B
LDA cross-entropy of B/A	A_t in text of B	Length of A	A_t in first line of B
Length of A	LDA cross-entropy of B/A	LDA cross-entropy of A/B	Length of A
LDA cross-entropy of A/B	LDA entropy of A	Count of B_t in text of A	LDA entropy of A
Jaccard Sim.	B_t in text of A	LDA cross-entropy of B/A	Jaccard Sim. (Nouns)
Jaccard Sim. (Nouns)	Jaccard Sim.	Jaccard Sim. (Nouns)	Jaccard Sim.
B_t in first line of A	Jaccard Sim. (Nouns)	A_t in text of B	LDA cross-entropy of A/B

Table 2: Rankings of the first 10 features for each domain in the ITA-PREREQ and English Reduced datasets.

by oversampling the minority class.

All experiments were performed on AL-CPL, ITA-PREREQ and English Reduced datasets. As baseline, we used the Zero Rule algorithm, and F-Score as evaluation metric.

4.1 Experimental Settings

We run experiments using the three classifiers presented in Sec 3.3 on each dataset, considering each of the four domains independently (i.e Data Mining, Geometry, Physics and Precalculus). Each classifier was tested both in a in-domain and cross-domain scenario.

To perform in-domain experiments, we trained and tested the classifiers on concept pairs belonging to the same domain. The evaluation is performed using a 5-fold cross validation. Cross-domain experiments were performed in a leave-one-domain-out manner: classifiers were trained on three domains and tested on the fourth.

4.2 Results and Discussion

In-domain. As it can be noted in Table 3, our systems performs extremely well for the in-domain setting, achieving high scores for both English and Italian pages. Note that our results always outperform both the Zero Rule baseline and the results obtained by Liang et al. (2018a) for all domains. This confirms our hypothesis: it is possible to identify prerequisite relations between edu-

cational materials using linguistic information extracted from textual content alone.

Best results are obtained using M3, the classifier that exploits both lexical and global features. Interestingly, M1 model performs are constantly better than M2, especially for the Data Mining domain: this suggests that lexical information from the WE lexicon contributes significantly.

Although comparable, the AL-CPL dataset is the one obtaining best results, with an average F-Score of 92.21%. This is probably due to the fact that the other two datasets are smaller than AL-CPL in terms of number of Wikipedia pages. However, comparing the results obtained with ITA-PREREQ and English Reduced we notice that ITA-PREREQ is the one that achieves lower results. This could be due to differences in the composition of the two datasets. For example, we noticed that there is a high difference in the average page length (number of tokens) of the two languages: for English Wikipedia pages it is about twice the Italian ones (2,728 and 1,073 tokens respectively). The impact of this characteristic can be twofold, both on the lexical and global features. For what concerns the lexical features, since we considered WE of the first 400 tokens of each Wikipedia page this means that if a page is shorter then that our network can acquire less information. As proof, the number of pages shorter than 400 tokens is higher in ITA-PREREQ (138)

In-domain						
		Data Mining	Geometry	Physics	Precalculus	Avg.
ITA-PREREQ	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	72.45	86.89	79.28	90.53	82.28
	M2	64.25	85.27	76.26	89.02	78.7
	M3	77.91	90.01	85.08	93.91	86.72
English Reduced	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	85.36	92.03	84.4	90.84	88.15
	M2	70.78	89.05	78.52	89.62	81.99
	M3	85.6	94.1	88.49	95.22	90.85
AL-CPL	Baseline	66.66	68.82	75.17	66.66	69.32
	M1	88.81	92.43	83.49	92.48	89.30
	M2	73.29	89.66	80.72	90.9	83.64
	M3	89.66	95.69	88.54	94.95	92.21
Liang et al. (2018a)	RF	76.7	89.5	69.9	88.6	81.17
Cross-domain						
		Data Mining	Geometry	Physics	Precalculus	Avg.
ITA-PREREQ	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	28.07	62.99	45.34	59.88	49.07
	M2	37.09	79.53	71.56	83.66	67.96
	M3	30.36	76.33	69.6	83.4	64.92
English Reduced	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	47.83	69.17	28.97	69.18	53.78
	M2	59.91	75.8	75.05	85.81	74.14
	M3	41.9	80.24	58.33	79.52	64.99
AL-CPL	Baseline	66.66	68.82	75.17	66.66	69.32
	M1	37.89	70.04	39.31	71.98	54.80
	M2	50.89	80.41	74.74	87.14	73.29
	M3	38.78	82.53	63.67	84.41	67.34

Table 3: In- and cross-domain results in terms of F-Score obtained by the three models and the baseline on each domain for each dataset. The in-domain setting also shows results obtained by [Liang et al. \(2018a\)](#) using a Random Forest (RF) classifier.

than English Reduced (8). Additionally, global features could be affected by the fact that English Wikipedia pages tend to be linguistically richer than their Italian counterparts containing more information and mentions to related concepts.

Cross-domain. Observing the results obtained in the cross-domain setting, we notice a significant performance drop if compared to in-domain results (see Table 3). The reason might be due to differences in the topic coverage of some domains in Wikipedia. Following ([Wang et al., 2016](#)), we believe that fundamental and broad subjects, such as precalculus and geometry, have more clear learning dependencies expressed through Wikipedia, while Data Mining, which obtained the lowest scores with our models, being a specific and relatively newer topic presents shorter pages, which means less information. Another possible explanation could be that pages belonging to the same domain are more homogeneous internally in terms of content structure, so it is easier for the networks to identify regularities.

Interestingly, contrary to what happens for the in-domain setting, Table 3 shows that the main

contribution to the cross-domain results is given by the global features, most likely because they can detect domain-independent properties. Word embedding lexicon alone (M1 model) is not sufficient to exceed the baseline, with the only exception of Geometry and Precalculus. Since these two domains share more lexicon than the others, we assumed that our model could perform better if trained only on a single domain that is lexically close to the testing one.

To test this hypothesis, we computed the Jaccard similarity between pairs of domains using all their Wikipedia pages in order to define lexically close pairs and use them to perform cross-domain experiments, i.e. one domain for training and one for testing.

Despite low results still below the baseline, we identified a correlation between lexical similarity between domains and obtained scores. For instance, results achieved using two domains with high Jaccard similarity (0.35) such as Precalculus and Geometry are much higher than those obtained comparing two domains with low Jaccard similarity (0.28), such as Precalculus and Physics

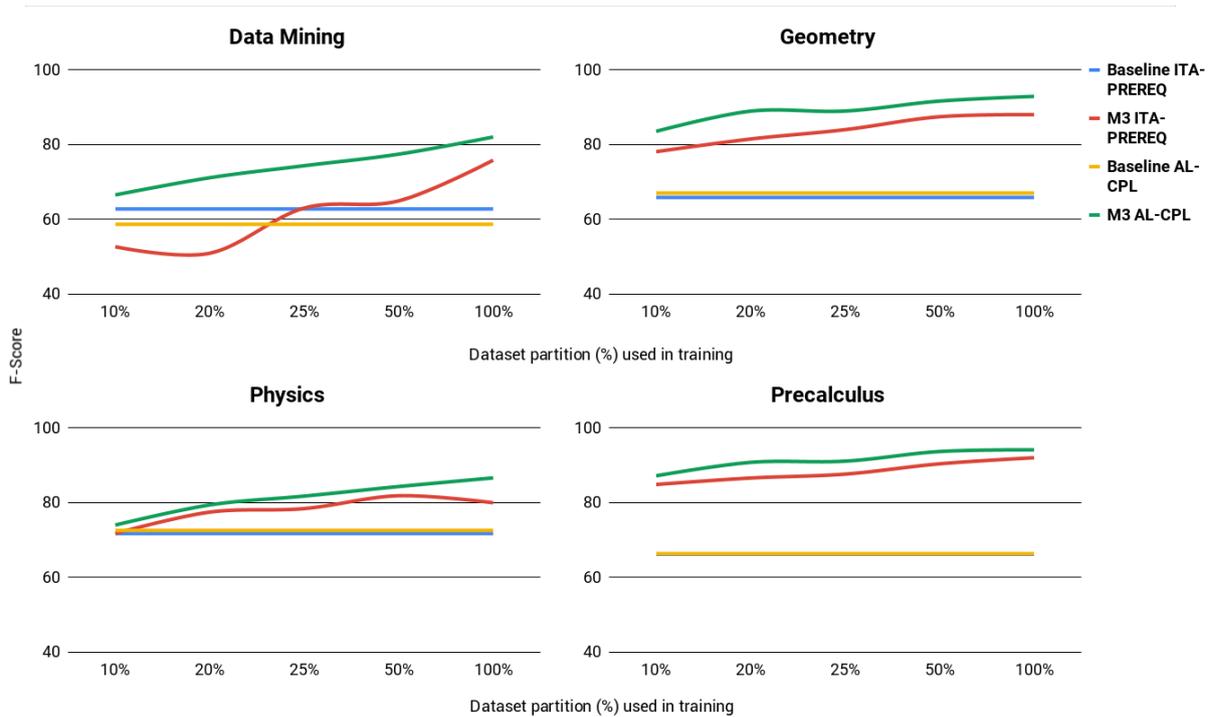


Figure 2: Incremental training strategy results for ITA-PREREQ and AL-CPL compared to the respective baselines.

(70.01% and 44.49% respectively).

However, considering that cross-domain experiments prove that further work needs to be done with this respect, in the next Section (4.2.1) we present another in-domain strategy based on incremental training with the aim of studying the impact of the training set size on the classifier performances.

4.2.1 Incremental Training Strategy

For the purposes of this paper, we describe incremental training strategy as the process of adding incrementally new concept pairs examples into the training set. Specifically, for each domain in the ITA-PREREQ and AL-CPL datasets we split the dataset in training and test set with 70% and 30% of the total examples, respectively.

We performed 5 experiments, feeding the M3 neural network model with different runs of 10%, 20%, 25%, 50% and 100% of the training set. All experiments, excluding the one with 100% of training samples, were performed using a k-fold cross validation strategy, with k equal to 10, 5, 4 and 2 according to the percentages of data samples previously defined.

Figure 2 reports results obtained for both ITA-PREREQ and AL-CPL datasets. As we can see, our model achieves good results even using limited portions of the ITA-PREREQ training data.

Specifically, for Geometry and Precalculus, even using only 10% of the training data the results we obtained are much higher than the baseline and they improve as the percentage of data samples in the training set increases. In respect to the Physics domain, we outperform the baseline by feeding our model with 20% of the training data. Using instead only 10% of the training data we obtained results comparable to those obtained by the baseline algorithm (71.78% and 71.79% respectively).

Data Mining is the only domain for which our classifier needs more training examples in order to obtain acceptable results. In fact, even if with 25% of the training set we can outperform the baseline outcomes, it is only with 100% of the examples that we are able to achieve satisfying results. This could be due to the fact that, as said previously, Data Mining is a more specialised topic with less clear prerequisite relations. Moreover, since Data Mining contains fewer concept pairs, it could be that a training set with only 10% or 20% of the concept pairs (38 and 77 respectively) is not sufficient for the network to identify regularities.

Results obtained for the AL-CPL dataset behave quite similarly, although we notice a faster increase in performances, especially for those domains that achieved lower results in the previous experiments (Data Mining and Physics).

5 Conclusion

In this paper we presented the results obtained on automatic prerequisite identification between LOs using a novel system based on neural network which exploits only linguistic features and does not rely on Wikipedia graph or LO metadata information. We performed our experiments on English and on a new Italian dataset, both in a in- and cross- domain scenario for four different domains.

The experiments demonstrated the effectiveness of our deep learning model and offer important insights into the exclusive use of linguistic feature on the task. The neural network achieved very good results for the in-domain setting, while we noticed a significant drop in performance for the cross-domain scenario. In the cross-domain setting, lexical features proved to be not well suited for the task, while global features obtained much better results, despite their simplicity. We thus think that further work needs to be done to investigate whether or not complex global features could improve the effectiveness of concept prerequisite learning models.

In the paper we also presented ITA-PREREQ, the first dataset annotated with prerequisite relation between concepts for Italian built starting from an English corpus (AL-CPL) with a completely automatic strategy.

The final goal would be to integrate this system as part of a educational design process, suggesting personalised learning paths, possibly in very distant domains from those used here, such as the humanities. We will address this lines of research in future work.

Acknowledgments

The work reported in the paper was partially supported by the 2year project (2018-2020) SchoolChain, *Soluzioni innovative per la creazione, la certificazione, il riuso e la condivisione di unità didattiche digitali all'interno del sistema Scuola*. Project funded by Regione Toscana (BANDO POR FESR 2014-2020).

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.
- Golshan B. & Papalexakis E. Agrawal, R. 2016. Toward data-driven design of educational courses: A feasibility study. *Journal of Educational Data Mining (JEDM)*, 8(1):1–21.
- Chiara Alzetta, Frosina Koceva, Samuele Passalacqua, Ilaria Torre, and Giovanni Adorni. 2018. Pret: Prerequisite-enriched terminology. a case study on educational texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Devendra Singh Chaplot, Yiming Yang, Jaime G Carbonell, and Kenneth R Koedinger. 2016. Data-driven automated induction of prerequisite structure graphs. In *EDM*, pages 318–323.
- Yetian Chen, José P González-Brenes, and Jin Tian. 2016. Joint discovery of skill prerequisite graphs and student models. In *EDM*, pages 46–53.
- Luis De-Marcos, José J Martínez, José A Gutiérrez, Roberto Barchino, and José M Gutiérrez. 2009. A new sequencing method in web-based education. In *2009 IEEE Congress on Evolutionary Computation*, pages 3219–3225. IEEE.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.
- Alexander R Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Wei Tai Ting, Robert Tung, Caitlin Westfield, and Dragomir R Radev. 2018. Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. *arXiv preprint arXiv:1805.04617*.
- Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.
- 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.
- Fabio Gasparetti, Carla Limongelli, and Filippo Sciarrone. 2015. Exploiting wikipedia for discovering prerequisite relationships among learning objects. In *2015 International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 1–6. IEEE.

- 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.
- Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling concept dependencies in a scientific corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 866–875.
- IEEE. 2002. Ieee standard for learning object metadata (draft). ieee standard 1484.12.1.
- Mikel Larranaga, Angel Conde, Inaki Calvo, Jon A Elorriaga, and Ana Arruarte. 2014. Automatic generation of the domain module from electronic textbooks: method and validation. *IEEE transactions on knowledge and data engineering*, 26(1):69–82.
- Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. 2019. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. *Proceedings of AAAI 2019*.
- 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. 2018a. Investigating active learning for concept prerequisite learning. *Proc. EAAI*.
- Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Lee Giles. 2017. Recovering concept prerequisite relations from university course dependencies. In *AAAI*, pages 4786–4791.
- Chen Liang, Jianbo Ye, Han Zhao, Bart Pursel, and C Lee Giles. 2018b. Active learning of strict partial orders: A case study on concept prerequisite relations. *arXiv preprint arXiv:1801.06481*.
- Nestor D Duque Méndez, Valentina Tabares Morales, and Rosa M Vicari. 2016. Learning object metadata mapping with learning styles as a strategy for improving usability of educational resource repositories. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 11(2):101–106.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Susan Nash. 2005. Learning objects, learning object repositories, and learning theory: Preliminary best practices for online courses. *Interdisciplinary Journal of E-Learning and Learning Objects*, 1(1):217–228.
- Liangming Pan, Xiaochen Wang, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Course concept extraction in moocs via embedding-based graph propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 875–884.
- Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in neural information processing systems*, pages 505–513.
- Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2018. Inferring concept prerequisite relations from online educational resources. *31st AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-19)*.
- Richard Scheines, Elizabeth Silver, and Ilya M Goldin. 2014. Discovering prerequisite relationships among knowledge components. In *EDM*, pages 355–356.
- 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.
- Kelvin Thompson and Francisca Yonekura. 2005. Practical guidelines for learning object granularity from one higher education setting. *Interdisciplinary Journal of E-Learning and Learning Objects*, 1(1):163–179.
- Argiris Tzikopoulos, Nikos Manouselis, and Riina Vuorikari. 2009. An overview of learning object repositories. In *Database Technologies: Concepts, Methodologies, Tools, and Applications*.
- Annalies Vuong, Tristan Nixon, and Brendon Towle. 2011. A method for finding prerequisites within a curriculum. In *EDM*, pages 211–216.
- 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.
- Shuting Wang, Alexander Ororbias, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. 2016. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the 25th acm international on conference on information and knowledge management*, pages 317–326. ACM.
- David Arnim Wiley. 2000. *Learning object design and sequencing theory*. Ph.D. thesis, Brigham Young University.

Yiming Yang, Hanxiao Liu, Jaime Carbonell, and Wanli Ma. 2015. Concept graph learning from educational data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 159–168. ACM.

Amal Zouaq, Roger Nkambou, and Claude Frasson. 2007. An integrated approach for automatic aggregation of learning knowledge objects. *Interdisciplinary Journal of E-Learning and Learning Objects*, 3(1):135–162.