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*

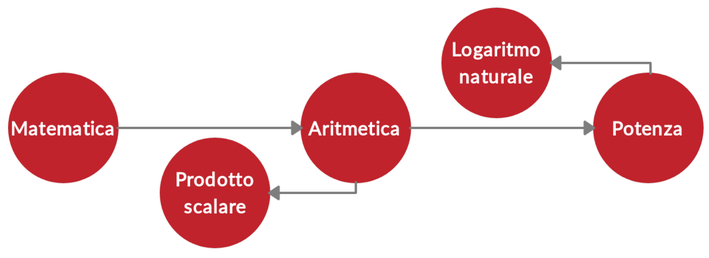
alessio.miaschi@phd.unipi.it, chiara.alzetta@edu.unige.it,
{francoalberto.cardillo,felice.dellorletta}@ilc.cnr.it

Research Issue

Learning Objects (LOs) are small and re-usable educational elements (e.g. lecture notes, multimedia content, presentations) with direct, succinct and homogeneous content. Uncovering educational relationships between LOs is a difficult and time consuming practice usually performed by domain experts when creating lectures and textbooks. The **prerequisite relation** is the most fundamental pedagogical relation: it defines what one needs to know before approaching a new content. In this work we present a deep learning method for prerequisite learning between concepts corresponding to LOs represented as Wikipedia pages.

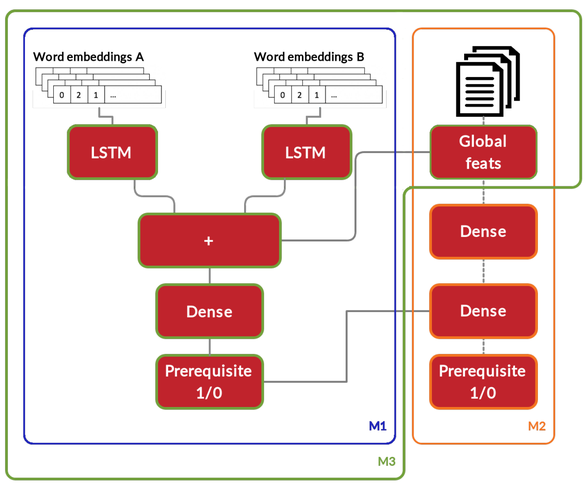
	AL-CPL			ITA-PREREQ		
Domain	**Concepts**	**Pairs**	**Prerequisites**	**Concepts**	**Pairs**	**Prerequisites**
Data Mining	120	826	292	75	429	154
Geometry	89	1,681	524	73	1,338	430
Physics	153	1,962	487	131	1,651	409
Precalculus	224	2,060	699	176	1,504	502
Total	586	6,529	2,002	455	4,922	1,495

Learning Objects Prerequisite Relations Example:



Prerequisite Learning Approach

MODEL



FEATURES

Lexical features: Word embeddings

Global features:
- In text: if B_t/A_t appears in *A*/*B*.
- Count: how many times B_t/A_t is mentioned in *A*/*B*.
- In first line: if B_t/A_t appears in *A*/*B*'s first line, i.e. *A*/*B* definition.
- In title: If B_t appears in A_t.
- Length: the number of words of *A*/*B*.
- Jaccard Sim.: the Jaccard similarity between *A* and *B*.
- Jaccard Sim. (Nouns): the Jaccard similarity between nouns appearing in *A* and *B*.
- RefD: the RefD metric between *A* and *B*.
- LDA Entropy: the Shannon entropy of the LDA vector of *A*/*B*.
- LDA Cross Entropy: the cross entropy between the LDA vector of *A*/*B* and *B*/*A*.

SETTINGS

- **Datasets:** three datasets: AL-CPL, its Italian version ITA-PREREQ, English Reduced, covering four domains, namely Data Mining, Geometry, Physics and Precalculus.

- **Classifiers:** three classification models:
  - M1 (only lexical features),
  - M2 (only global features),
  - M3 (combination lexical and global features).

  Each classifier is tested both in a in-domain and cross-domain scenario.

- **Baseline:** Zero Rule algorithm.

- **Evaluation:** in terms of F1
  - in-domain: 5-fold cross validation
  - cross-domain: leave-one-domain-out (training on three domains, testing on the fourth).

Results

IN-DOMAIN

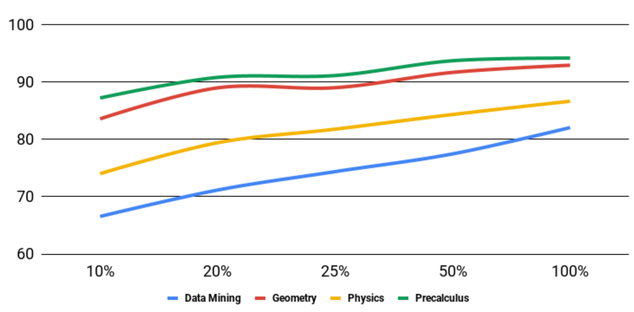
		Data Mining	Geometry	Physics	Precalculus	Avg.
ITA-PREREQ	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	72.45	86.89	79.28	90.53	82.28
	M2	64.25	85.27	76.26	89.02	78.7
	M3	**77.91**	**90.01**	**85.08**	**93.91**	**86.72**
English Reduced	Baseline	66.66	67.86	75.22	66.66	69.1
	M1	85.36	92.03	84.4	90.84	88.15
	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**

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

Incremental Training Experiments

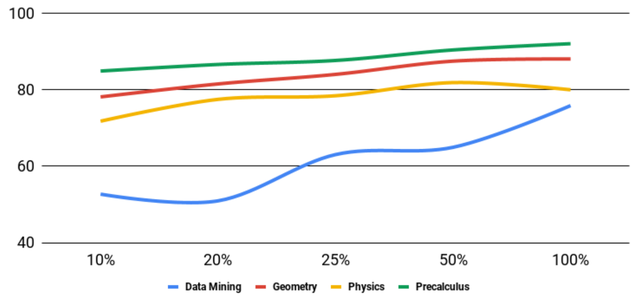
AL-CPL



Approach: incrementally adding new concept pairs examples into the training set of M3.

Experiments: we split the dataset in training (70%) and test set (30%). We performed 5 experiments, feeding the M3 neural network model with different runs of 10%, 20%, 25%, 50% and 100% of the training set. We used a k-fold cross as validation strategy, with k equal to 10, 5, 4 and 2 according to the percentages of data samples previously defined.

ITA-PREREQ



The 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019), Florence, Italy, August 2, 2019