Digging Into Prerequisite Annotation

Chiara Alzetta, Ilenia Galluccio, Frosina Koceva,
Samuele Passalacqua, Ilaria Torre

DIBRIS, University of Genoa, Italy
{chiara.alzetta,ilenia.galluccio,frosina.koceva}@edu.unige.it,
samuele.passalacqua@dibris.unige.it, ilaria.torre@unige.it

Abstract. Intelligent textbooks are often engineered with an explicit representation of their concepts and prerequisite relations (PR). PR identification is hence crucial for intelligent textbooks but still presents some challenges, also when performed by human experts. This may cause PR-annotated datasets to be inconsistent and compromise the accuracy of automatic creation of enhanced learning materials. This paper investigates possible reasons for PR disagreement and the nature of PR itself, with the aim of contributing to the development of shared strategies for PR annotation, analysis and modelling in textbooks.

Keywords: prerequisite relation · annotation · agreement.

1 Introduction

Fundamental functionalities of intelligent textbooks, such as content enrichment [23, 21, 9] and personalization [35, 31, 2], exploit the knowledge contained in the text to select relevant content and organise it according to a structure that reflects the prerequisite relations (PR) among concepts [17, 19]. PRs are utterly relevant in education since they establish which concepts are needed by a student to understand a further concept without loss of information or misconceptions [20, 15, 1, 14]. Despite their importance, PRs recognition, both in manual and automatic form, still presents many open challenges. PR manual annotation is in many cases a common way to achieve a PR-enriched intelligent textbook, but is also an essential step for developing automatic textbook content modelling. Systems for automatic PR learning (e.g. [22, 1]) often rely on manually annotated gold standards [29, 34, 15] for evaluating or training machine learning algorithms. Agreement rates [6] are often used to estimate annotations' reliability [8], but these measures are conditioned by many factors, including the design of the annotation task and the subjectivity of the phenomenon to be labelled [5, 28]. PR-annotations rarely achieve high agreement values unless the annotation task is extremely guided [18]. A broad literature discusses causes of disagreement in annotation [12, 11, 7], however this has not been fully explored for PRs. Our **goal**

is to investigate this matter and highlight possible causes of errors and issues, in order to provide a way to improve PR modelling for intelligent textbooks. This is useful especially when increasing the size of the dataset is not a practical way to reduce the impact of errors (as in case of high demanding and time-consuming annotation tasks like PR).

2 Experimental Study, Results and Discussion

Our analyses are conducted on PRET dataset described in [4], obtained by asking five domain experts to manually annotate PRs between the concepts mentioned in a computer science textbook chapter[10]. Experts had to read the text and in parallel annotate PRs among the encountered concepts. We stressed this point since our annotation goal is to model the textbook content rather than elicit the PR of a domain from the expert's mind. For the purpose of this paper, in order to investigate the variability of annotations, we performed an in-context annotation check, showing to each expert a subset of her/his annotations and asking to confirm or revise their annotation, after reading again the text. This process involved only "rare" annotations (concept pairs annotated only by him/her) since errors are known to be more probable among them [13]. In the end, the manual annotation resulted in 25 concept pairs annotated by all five experts, 46 annotated by four experts, 83 by three, 214 by two and 477 by only one annotator, for a total of 845 unique pairs.

2.1 Interpretation Variability and Annotation Errors. This section deals with the research question, i.e. which are the main causes of disagreement and errors. Our results cannot be generalised but are discussed in the light of the literature. During the annotation check mentioned above, we asked experts to specify the motivation for their revisions, among the following: *a) Not a Concept*: at least one concept of the pair is not a domain term; *b) Background Knowledge*: the PR derives form the expert's domain knowledge since it's not expressed in the text; *c) Too Far*: the concepts are too distant in the path; *d) Annotation Error*: mistake due to distraction; *e) Wrong Direction*: the concepts should be reversed; *f) Co-Requisites*: there is no dependency relation between the two concepts. The table in Fig.1 reports the distribution of these labels in the revised pairs. *Not a Concept* is the most recurrent problem for all annotators: common-usage terms like *channel* and *system* were added as domain concepts, but then revised. As it will be discussed below, this error does not seem mostly a trivial one, but an interpretation error commonly observed for educational concepts [33], whose boundaries verge on subjectivity [5]. Other error types are unevenly distributed: *Too Far* seems again related to interpretation and subjectivity issues, that may arise from the expert questioning his/her annotation style (e.g., annotating as PR also indirectly related concepts), *Background Knowledge* might be again accounted as annotation style error, however in this case it is not due to subjectivity, but to wrong interpretation of the task (i.e., the tendency to infer from other sources PRs that are not explicitly written in the text). The last three types seem all due to distraction. While distraction is inevitable (thus

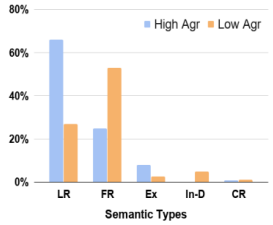
some kind of revision could be a good practice), subjectivity and annotation style errors are more tricky, but the possibility of revision could allow to obtain a more reliable annotation [26, 24]. Concept interpretation error will be analysed below more in depth since it is particularly critical for intelligent textbooks, even independently of PR relations, as discussed also in [33].

	REVISION		ERROR TYPES					
	#Rev PR (% over tot PR)	#Del PR (% over Rev PR)	Not a Concept	Background Knowledge	Too Far	Annotation Error	Wrong Direction	Co-PR
A1	175 (42.68%)	27 (15.43%)	37.04%	29.63%	3.70%	11.11%	7.41%	11.11%
A2	72 (25.35%)	16 (22.22%)	62.50%	0.00%	6.25%	12.50%	6.25%	12.50%
A3	190 (43.68%)	86 (45.26%)	72.94%	2.35%	17.65%	5.88%	0.00%	1.18%
A4	118 (42.30%)	44 (37.29%)	70.45%	9.09%	2.27%	13.64%	4.55%	0.00%
A5	109 (39.64%)	29 (26.61%)	55.17%	31.03%	10.34%	3.45%	0.00%	0.00%

Fig. 1. 'Revision' and 'Error Type' percentages.

Concept Interpretation. Despite the guidelines, the notion of concept was faced in different ways, using a fine vs coarse-grained interpretation. To understand the role played by concepts on agreement, we performed Pearson correlation between root and leaf nodes (those with zero in- and out-degree respectively) in annotations' graphs [16]. Our intuition is that graphs with high root nodes correlation suggest an agreement on what is intended as core knowledge, i.e. Primary Notions (PN) for the domain; similarly, correlation on leaves, reflecting the paths' Learning Outcomes (LO), indicates a common interpretation of concepts granularity. We noticed that revision helped to harmonise the annotations: while prior the revision we obtained on average a stronger correlation between root nodes (0.49) than leaves (0.29) ($p < 0.05$), on the revised graphs the correlations became similar (0.47 for roots and 0.46 for leaves with $p < 0.05$).

2.2 Semantic and Lexical Relationships in PRs. We investigate here if there are any semantic relations and linguistic patterns that can be identified as frequently occurring in PRs with high vs low agreement. As Fig.2 shows, lexical



LABEL	DESCRIPTION
Lexical Relation (LR)	Hypernyms (is-a), meronyms (part-of), taxonomic relations
Functional Relation (FR)	One of the two concepts uses the other, or is needed by it.
Example (Ex)	One concept is explained as an example of the other concept.
In-Depth (In-D)	The concept is used as in-depth description of another one.
Causal Relation (CR)	Cause-effect relation between two concepts.

Fig. 2. Semantic Relationship Type distribution and description for PRs annotated by 3 or more experts (*High Agr*) or 1 expert only (*Low Agr*).

relations are the most frequent semantic type among PRs with high agreement,

covering more than 60% of cases, while Functional Relations are more common among PRs with low agreement. This could be due to the fact that taxonomies, by definition, exploit a"dependency relation" to classify elements. On the other hand the Functional Relation is highly affected by the presentation of concepts in the text, which entails a text interpretation and a subsequent low agreement due to subjectivity. Such distributions highlight the need for annotation guidelines with clear examples of PRs not involved in lexical relations, since those types might be harder to identify along a text and could give raise to disagreement.

2.3 Holistic Nature of the Annotation Process. Previous work [29, 34, 36] mostly addressed PR annotation and evaluation as a pairwise combination of concepts; we also used this approach [3]. However, we believe that such approach over-simplifies the annotation and may result in misinterpretations of the relations contained in the text. Indeed, semantic annotations often result in (directed) graphs [27, 25, 30] in which each path is an interpretation of a relation that arises from reading the whole text (this explains their holistic nature) and should be evaluated accounting for those peculiarities. The commonly adopted pairwise evaluation of PRs misses the interdependence between concepts involved in a PR path and does not take into account the pedagogical characteristics of the annotated graph as a whole. Temporal relation processing may represent an interesting ground of comparison for PR, since precedence relation also shows a transitive and sequential nature. Researchers in both fields encounter similar limitations using traditional performance metrics used in information retrieval, e.g. precision, recall [32]. A common scenario in both fields is when three items A, B, C (concepts or events) are annotated by a rater such that $A < B$ and $B < C$, but another rater identifies the relation $A < C$: in such cases, traditional agreement metrics will fail to identify $A < C$ as a shared relation, even if it is an implicit consequence of the other two relations [32][2]. This suggests that agreement could be computed by considering transitive edges and path similarity in the two graphs. However, it is worth nothing that this approach involving transitive closure might result in considering as PR some relations that are too far in a path. This is an open issue, as well as the selection of proper metrics to compare PR graph similarities. In our current work we are testing metrics that work at different levels, among them Graph Edit Distance (GED), Vertex Edge Overlap (VEO), PageRank, inter-agreement in learning pahts based on the transitive characteristic of the PR (for details see [16]). We believe this research direction is a relevant contribution wrt the current approaches in order to compute not just agreement between annotators, but also between automatic methods for knowledge graph creation, highly useful in intelligent textbooks.

Conclusions In this paper we investigated different forms of disagreement on PR annotation that affect PR processing and management, and consequently textbook modelling. Our aim is thus to contribute to the modelling of PR structure in textbooks by developing a subjectivity-aware PR coding protocol. The analysis also opens the study of future research on using fuzzy PRs for structuring knowledge behind text instead of yes/no (binary) PRs.

[2] $<$ indicates both the temporal relation *before* and the prerequisite relation $\prec$

References

1. Adorni, G., Alzetta, C., Koceva, F., Passalacqua, S., Torre, I.: Towards the identification of propaedeutic relations in textbooks. In: International Conference on Artificial Intelligence in Education. pp. 1–13. Springer (2019)
2. Alpizar-Chacon, I., Sosnovsky, S.: Interlingua: Linking textbooks across different languages (2019)
3. Alzetta, C., Koceva, F., Passalacqua, S., Torre, I., Adorni, G.: Pret: Prerequisite-enriched terminology. a case study on educational texts. In: Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018 (2018)
4. Alzetta, C., Miaschi, A., Adorni, G., Dell'Orletta, F., Koceva, F., Passalacqua, S., Torre, I.: 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. vol. 2481. CEUR-WS (2019)
5. Amidei, J., Piwek, P., Willis, A.: Rethinking the agreement in human evaluation tasks (2018)
6. Artstein, R.: Inter-annotator agreement. In: Handbook of linguistic annotation, pp. 297–313. Springer (2017)
7. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
8. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? a meta-analytic investigation. Computational Linguistics **37**(4), 699–725 (2011)
9. Boulanger, D., Kumar, V.: An overview of recent developments in intelligent e-textbooks and reading analytics (2019)
10. Brookshear, G., Brylow, D.: Computer Science: An Overview, Global Edition, chap. 4 Networking and the Internet. Pearson Education Limited. (2015)
11. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational linguistics **22**(2), 249–254 (1996)
12. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of clinical epidemiology **43**(6), 543–549 (1990)
13. Fort, K., Nazarenko, A., Rosset, S.: Modeling the complexity of manual annotation tasks: a grid of analysis. In: International Conference on Computational Linguistics. pp. 895–910 (2012)
14. Gagne, R.M.: Learning hierarchies. Educational psychologist **6**(1), 1–9 (1968)
15. Gasparetti, F., De Medio, C., Limongelli, C., Sciarrone, F., Temperini, M.: Prerequisites between learning objects: Automatic extraction based on a machine learning approach. Telematics and Informatics **35**(3), 595–610 (2018)
16. Koceva, F., Alzetta, C., Galluccio, I., Passalacqua, S., Torre, I.: Prerequisite similarity metrics, http://teldh.dibris.unige.it/pr-similarity-metrics/
17. Labutov, I., Huang, Y., Brusilovsky, P., He, D.: Semi-supervised techniques for mining learning outcomes and prerequisites. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 907–915. ACM (2017)
18. Li, I., Fabbri, A.R., Tung, R.R., Radev, D.R.: What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. Proceedings of AAAI 2019 (2019)
19. Liang, C., Ye, J., Wang, S., Pursel, B., Giles, C.L.: Investigating active learning for concept prerequisite learning. Proc. EAAI (2018)

20. Manrique, R., Sosa, J., Marino, O., Nunes, B.P., Cardozo, N.: Investigating learning resources precedence relations via concept prerequisite learning. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). pp. 198–205. IEEE (2018)
21. Matsuda, N., Shimmei, M.: Pastel: Evidence-based learning engineering method to create intelligent online textbook at scale (2019)
22. Miaschi, A., Alzetta, C., Cardillo, F.A., Dell'Orletta, F.: Linguistically-driven strategy for concept prerequisites learning on italian. In: Proceedings of 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019) (2019)
23. Miller, B., Ranum, D.: Runestone interactive: tools for creating interactive course materials. In: Proceedings of the first ACM conference on Learning@ scale conference. pp. 213–214. ACM (2014)
24. Plank, B., Hovy, D., Søgaard, A.: Linguistically debatable or just plain wrong? In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers). pp. 507–511 (2014)
25. Poesio, M.: The mate/gnome proposals for anaphoric annotation, revisited. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004. pp. 154–162 (2004)
26. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc." (2012)
27. Ries, K.: Segmenting conversations by topic, initiative, and style. In: Workshop on Information Retrieval Techniques for Speech Applications. pp. 51–66. Springer (2001)
28. Saldaña, J.: The coding manual for qualitative researchers. Sage (2015)
29. Talukdar, P.P., Cohen, W.W.: Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. pp. 307–315. Association for Computational Linguistics (2012)
30. Tannier, X., Muller, P.: Evaluating temporal graphs built from texts via transitive reduction. Journal of Artificial Intelligence Research **40**, 375–413 (2011)
31. Thaker, K., Brusilovsky, P., He, D.: Student modeling with automatic knowledge component extraction for adaptive textbooks (2019)
32. UzZaman, N., Allen, J.: Temporal evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 351–356 (2011)
33. Wang, M., Chau, H., Thaker, K., Brusilovsky, P., He, D.: Concept annotation for intelligent textbooks. arXiv preprint arXiv:2005.11422 (2020)
34. Wang, S., Ororbia, A., Wu, Z., Williams, K., Liang, C., Pursel, B., Giles, C.L.: Using prerequisites to extract concept maps from textbooks. In: Proceedings of the 25th acm international on conference on information and knowledge management. pp. 317–326. ACM (2016)
35. Yin, C., Uosaki, N., Chu, H.C., Hwang, G.J., Hwang, J., Hatono, I., Tabata, Y.: Learning behavioral pattern analysis based on students' logs in reading digital books. In: Proceedings of the 25th international conference on computers in education. pp. 549–557 (2017)
36. Zhou, Y., Xiao, K.: Extracting prerequisite relations among concepts in wikipedia. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)