PREAP annotation manual for prerequisite relation annotation

Last update 2021-02-15

1 Preliminary Notes

The present document contains the instructions and recommendations for performing manual annotation of prerequisite relations in educational texts (e.g., textbooks), according to the PRErequisite Annotation Protocol (PREAP).

1.1 Intended Audience

This document is intended for annotators recruited for an Annotation Project. The manual contains Annotation Guidelines and Knowledge Elicitation Questions that annotators have to read and understand before starting the annotation task.

1.2 Terms and Definitions

- **Concept:** a concept is an abstract and general idea conceived in the mind. In education, concepts represent the building blocks of learning, namely what a student should understand in order to acquire new knowledge. Depending on learner's needs with respect to granularity, concepts can be very general (e.g., algebra, geometry, mathematics etc.) or very specific (e.g., radius, integer multiplication, fraction denominator). Either way, they are represented in texts as lexical entities constituted by a single or multi-word term.
- **Prerequisite Relations:** also called PRs, they are pedagogical relations that hold between pairs of educational concepts described in educational texts. These relations express a propaedeutic relationship, meaning that PRs express what should be understood first in order to avoid knowledge gaps when learning a new topic.
- **Corpus:** a corpus is a textual document, such as a textbook or any educational text. It can be enhanced with labels to become an annotated text, i.e. a text where certain information related to its content is made explicit through annotation.

- **Annotation:** in general terms, annotation consists of the process of adding comments, notes, explanations, or other types of external marks that can be attached to a (part of a) document. In this document, we refer to annotation as the manual process led by humans consisting of adding labels to a *textual corpus* in order to indicate the presence of prerequisite relations between two *concepts* mentioned in the corpus.
- **Annotation Protocol:** a systematic procedure defined by guidelines and specifications that specify how to obtain corpora enriched with explicit information regarding a certain phenomenon and that are designed to be reproduced on any unannotated texts at any time. PREAP is a protocol for annotating an educational text with PRs.
- **Annotation Project:** the set of tasks aimed at building an annotated text that includes explicit annotations about the phenomenon being studied (here, prerequisite relations in educational texts).
- **Project Manager:** the person or team leading the annotation project. The manager is in charge of taking decisions concerning the goals and settings of the annotation project.
- **Annotation Guidelines:** instructions and recommendations that indicate how to perform the annotation. They can be consulted at any time during annotation, but should be known in advance.
- **Annotator:** the person that performs the annotation on the corpus according to the guidelines and project principles.
- **Gold Standard Dataset:** the output of the annotation project. A Gold Standard Dataset (or *Gold-PR dataset* with reference to a Gold Standard annotated with PR relations) is a dataset annotated with PR relations following a systematic annotation procedure to produce high-quality annotations. It can be based on a single trusted annotation or obtained by combining multiple manual annotations into a single one.
  The Gold-PR could be exploited to i) obtain informative analysis of the annotated phenomenon; ii) train and test the performances of machine learning systems; iii) compare the manual annotations against those obtained using automatic systems for PR extraction to test their accuracies.

2 Annotation Manual

The annotation specifications comprise the annotation instructions and recommendations for performing PR annotation on texts, systematised within the *annotation manual*. The annotation manual is composed of two complementary resources: the *Annotation Guidelines* (AG), whose aim is to describe how the annotation process should be carried out in order to reduce inconsistencies in the annotations, and a list of *Knowledge Elicitation Questions* (KDE), aimed

at clarifying dubious cases through direct questions and examples and helping the annotators think over hard cases.

2.1 Annotation Guidelines

Annotation Guidelines (AG) for annotators concern different issues of the annotation process, grouped in four categories:

i) Concept identification (AG rec. 1-3);

ii) Text annotation (AG rec. 4-6);

iii) PR features and properties (AG rec. 7-9);

iv) Annotation revision (AG rec. 10-12).

I. Concept Identification

1. The goal of the annotation is identifying a prerequisite relation between two distinct terms of a textual corpus. The two terms represent domain concepts described in the text and can be referred to as target and prerequisite concepts.

2. A concept can be either a single or multi–word term extracted from the corpus.

3. Insert a prerequisite relation for a target concept if you think you need to know the information related to a different concept in order to understand what you are reading about the target concept. Each of the two concepts must be present either in the initial Terminology provided by the project manager or in the manual terminology built by you (i.e., the annotator) during the annotation process according to what option for concept annotation has been chosen from the project manager. In case one-shot annotation of concepts and PRs is permitted, if a concept is still missing in the terminology, add the corresponding term and then insert the relation.

II. Text Annotation

4. The relation must be inserted in the context (i.e., the sentence) where you find it. A concept could be mentioned more than once along the text, each time introducing novel information and recalling different concept(s). Make sure to add the prerequisite relation between two concepts exactly where the target concept description recalls the knowledge related to the concept you identified as prerequisite.

5. Build a concept pair only if a prerequisite relation does exist between the two: if you think that a relation between two concepts does not occur in the text, do not insert any relation.

6. *Trust the text*: you must annotate only concepts and relations that can be acquired from the text. Do not consider concepts and relations recalled from your background knowledge about the topic.

III. PR Features and Properties

7. A concept cannot be a prerequisite of itself: self prerequisites such as "*computer* is a prerequisite of *computer*" will not be allowed by the system.

8. Do not introduce loops in the annotation. Imagine that you have already annotated that: i) "fruit" is a prerequisite of "citrus", and ii) "citrus" is a prerequisite of "orange". By annotating that "orange" is a prerequisite of "fruit", you will create a loop.

9. Every time you insert a relation you must also define its weight. Allowed values comprise: *strong* (the prerequisite is absolutely necessary to understand the other term) and *weak* (the prerequisite is very useful but not strictly necessary).

IV. Annotation Revision

10. After completing your annotation, you should also perform an annotation check aimed at confirming your previously created PRs, or at revising them in case of mistake. By reading again the portion of text where you entered a relation, decide whether you want to confirm, delete or modify the pair.

11. Delete a prerequisite relation if you added it by mistake. Keep in mind, however, that you can delete one single instance of a pair at a time: if the same pair is annotated with the prerequisite relation in another part of the text, that will be preserved. If you think that ANY prerequisite relation between two concepts should be deleted, you must delete each of the relations having those two concepts.

12. Modify a prerequisite relation if you assigned it the wrong weight. You can modify the weight of the relation if you believe that text expressed a different relation strength than the one you originally assigned to the PR when creating it.

2.2 Knowledge Elicitation Questions

KEQs are designed to show examples involving commonly used terms to the annotators in order to build a shared understanding about the interpretation of PRs.

1. Which concepts (among those mentioned in the text) you need to master in order to understand the meaning of the target concept?

2. Which concepts are recalled to define the target concept?

3. Are other concepts mentioned in the same context (e.g., sentence or paragraph) of the target concept? If so, are they useful to understand the meaning of the target concept?

4. Does the target concept represent a special case of another concept mentioned in the text (e.g., *circumference*[target] is a special case of *ellipsis*[prerequisite]?

5. Does the target concept show a part-of relation with another concept mentioned in the text (e.g., the *elbow*[target] is a part of an *arm*[prerequisite])?

6. Does the target concept consists of sub-elements already mentioned in the text (e.g., *elbow*, *forearm* and *shoulder*[prerequisites] are parts of the *arm*[target])?

7. Is the target concept caused by another previously described concept (e.g., *rain*[prerequisite] causes *floods*[target])) or vice versa (e.g., *rain*[target] is caused by *low pressure*[prerequisite])? If so, which one? Try to follow the relation proposed by the text author to understand if a prerequisite relation exists.