PREAP specifications
Prerequisite Annotation Protocol for prerequisite relation annotation

Last update 2021-02-15

Preliminary Notes

The present document describes the steps and guidelines to carry out a project for the annotation of a corpus (e.g., textbook) with domain concepts and prerequisite relations (PRs) between concepts.

Intended Audience

This document is intended for the person in charge of leading an annotation project, i.e., the project manager of the annotation project. The document contains guidelines for all the steps of the annotation project, including the instructions for annotators recruited by the project manager to perform the manual annotation.

Content Overview

Section 1, **Terms and Definitions**, provides the preliminaries for understanding PREAP. Specifically, it provides the definition of the terms used along this document, thus we suggest to read this part carefully before reading the other sections and returning to it in case of doubts.

Section 2, **PREAP Prerequisite Annotation Protocol**, offers a detailed description of how to carry out an annotation project following PREAP principles, from selecting the text to be annotated to the creation of the final gold dataset. This section is aimed at providing the instructions and recommendations for managing an annotation project which relies on PREAP principles, thus its intended audience is the *project manager* (see below).

Section 3, **Annotation Manual**, comprises two complementary resources: the *Annotation Guidelines* and the *Knowledge Elicitation Questions*. Both are aimed to clarify how to identify and annotate prerequisite relations on texts. Thus, this section is intended specifically for *annotators*, but also for the *project manager* who is in charge of coordinating the whole project: the instructions in this section should be read and understood before starting the annotation task.

1 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 PREAP Annotation Protocol

2.1 Overview and Goals

The PRErequisite Annotation Protocol PREAP defines a systematic procedure aimed at building datasets by manually annotating educational texts with prerequisite relations. To this aim, the protocol is designed to support manual identification of the PRs while reading a textbook. The ultimate goal of this process is the creation of Gold-PR datasets. These datasets are highly valuable for training and testing automatic PR learning systems and to investigate the main aspects involved in the realisation of PR relations in textual data.

The annotation approach defined in PREAP is specifically designed to annotate PRs on educational texts aiming to identify manually recognised relations between concepts and, at the same time, being able to retrieve PR linguistic realisations in texts. The latter goal is particularly relevant for prerequisite relation identification if we consider that the content of a textbook is designed to guide students through a concept sequence designed to tackle relevant concepts and highlight their relations. The main goal of PREAP is to capture such concept sequence and structure through a ***textbook-oriented annotation*** process, i.e. an annotation strictly bound to the text.

The textbook-oriented annotation strategy has the advantage of making the annotations independent from any external database or knowledge structure and, at the same time, it reflects the teaching approach of the author of the textbook being annotated. As a consequence, the PR annotation approach defined by PREAP allows the creation of datasets that, potentially, could be used to investigate concepts organisation within the content of an educational text and if the PRs appear within recurrent linguistic patterns.

Here below we will describe in detail the main steps of an annotation project that follows the principles of PREAP annotation protocol.

2.2 Project Management

Prior to the text annotation phase, a *management phase*, supervised by a project manager who leads the whole annotation project, should define the project settings and goals.
The project manager shall be responsible for taking decisions about the following aspects.

- **Annotation Goal**: define the reason for starting the annotation project, i.e. define how the annotations will be used once they are created by annotators;
- **Corpus Selection**: choose the textbook to be annotated and prepare it for annotation (e.g., perform OCR if your text comes in PDF format, or remove figures if they are not relevant to the project);
- **Annotation setup**: decide where the annotation should be carried out, for example one might want to use an annotation tool like PRAT[1]. In that case, the manager should take care of any possible preliminary setup required by the chosen tool;
- **Annotators selection and training**: the first step is defining the annotator's profile: depending on project goal, the manager might want to recruit annotators having specific characteristics, such as a certain age or educational level. Once annotators are recruited, explain the task to them (either individually or as a group) and set up a pilot study to assess their understanding of the protocol principles.

[1] Our team developed PRAT as an interface for prerequisite annotation and has functions for quantitative and visual analysis.

2.3 PR Annotation Task

The main steps for performing PR annotation on textbooks can be summarised as follows:

1. Read the text and find the relevant domain concepts mentioned in the selected corpus;

2. Read the text and, if you encounter a concept that needs some prior knowledge to be understood, indicate its prerequisite concept(s) from the list of concepts found at step 1.

3. Revise the pairs you created reading again the portion of text where they were annotated.

In order to properly putting into practice such principles, an *Annotation Manual* is available (see Sec. 3), which systematically tackles different aspects of the text annotation phase. The core two, i.e. concept and PR identification, are discussed here below. Revision is explained in the next section.

2.3.1 Concepts Identification.

The first step of the annotation protocol consists of identifying the domain concepts mentioned in the text. This step can be tackled as an autonomous step of the annotation process as the project manager is in charge of taking preliminary decisions concerning how concepts should be extracted and used. PREAP admits the following approaches that address different needs of the annotation project:

a) *Manual identification:* this approach lets annotators manually identify which terms in the text correspond to domain concepts. In general it is recommended a two-steps annotation process where the first step returns the list of domain concepts and the second step returns the PR relations between them. One-shot manual annotation of concepts and PRs is prone to return a high number of overlapping concepts and a consequent huge amount of PRs with low agreement;

b) *Automatic identification:* this approach exploits an automatic term extraction system to acquire from the text a list of concepts that the annotators have to use as–it–is;

c) *Semi-automatic identification:* this approach uses an automatic term extraction system to acquire from the text the candidate concepts, which then require manually revision by relying on domain experts or on the manager's own domain knowledge to define the ultimate list of concepts. The annotators will use that list as-it-is.

Consider that, with option a), letting annotators freely add concepts without exchanging views with the other annotators, produces richer but less homogeneous annotations. If the manager decides to adopt option *b)* or *c)*, (s)he should take care of extracting the concepts from the text. When possible, discussion among experts to agree on the concepts list turns to be useful.

2.3.2 PRs Textbook-Oriented Annotation

PR annotation task consists in reading the text and pairing the concepts in the text if the annotator considers them as having a prerequisite relation. Annotating PRs meanwhile reading the text means that the annotators follow the flow of the text to identify concepts that are used by the author to explain a new concept. The details on how to distinguish a PR from other relations are reported in the Manual. Annotators do not have to label the absence of a PR, since it would make the annotation process hard to carry on (one would have to label at least $n(n-1)$ PRs, with n being the number of concepts). Annotating PRs and concepts together in the same process generally results in a dataset with sets of concepts and PRs that usually are very large and with some overlaps. Unless the goal is studying the linguistic realization of PRs, one-shot annotation of concepts and PRs is not recommended.

2.4 Annotation Revision

Manual annotation is known to contain errors due to misinterpretation of the text or the guidelines, but also to natural distraction. To address this problem and improve the internal coherence of the annotations, PREAP recommends "in-context revision", in order to check if the pairs created by an annotator should be kept in the annotation, modified or deleted.

The peculiarity of this revision phase is that, in line with the PR annotation approach, concept pair revision is performed by reading again the portion of textbook where the relation was entered (i.e., "in-context"). In other words, in order to revise her/his pairs, each annotator is required to read again the portion of text where she/he found the PR relation to check if a pair was inserted intentionally or by mistake. In case reducing the workload of this step is a priority, we recommend to revise at least those PRs identified by a low number of annotators.

2.5 Post-Annotation Procedures for Gold-PR Dataset Creation

Once annotation and revision are completed, the project manager should take care of the subsequent steps concerning the annotations evaluation and combination.

2.5.1 Agreement Evaluation

Agreement between annotations produced by two or more annotators is a measure of similarity and annotation reliability. Cohen's and Fleiss' kappa (k) are two of the most prominent metrics used for agreement evaluation and we recommend to use them on PR annotated texts as follow: the first is used between the annotations of pairs of annotators, the latter is computed between all the annotations produced for the same text. Since PR relations, as defined in PREAP, are characterized by transitivity, it is recommended to apply the k metric in a way that accounts for it. Specifically, it assumes that two annotators agree on the PR $A \prec C$ in both the following cases: *i)* both annotators manually created the pair $A \prec C$ and *ii)* one annotator created the pair $A \prec C$ and the other created the pairs $A \prec B$ and $B \prec C$.
The metric should thus be computed as follows. Given the list T of concepts used during annotation, consider as total items of the annotation task the list P of each pairs-wise combination p of concepts in T, regardless the relation direction (i.e., $A \prec B$ and $B \prec A$ are both included in P). For each annotator, consider as positive PR each p that is either manually created by the annotator or that can be derived for the transitive property. Consider p as a negative PR otherwise. Then, compute k for each pair of annotators using the following standard k equation, where P_o stands for the observed agreement (i.e., probability for an item to receive the same annotation by both raters), while P_e denotes the agreement expected by chance (i.e., the probability of each individual category).

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{1}$$

Using the same method to compute agreement supports comparability of agreement scores.

2.5.2 Gold-PR Dataset Creation and Revision

Gold standard datasets are intended to provide a generally accepted annotation of a phenomenon that can be looked at as accurate and reliable reference. Therefore, gold datasets are usually built by a pool of experts by performing a shared annotation or, more frequently, by combining their single annotations. Different criteria can be adopted to combine the annotations, ranging from intersecting to joining them. The former approach means taking only those annotations inserted by all the annotators, while the latter consists of including the annotations added by each annotator. The first approach maximises the precision with respect to the ground truth, while the latter maximises the recall. The former can be useful when the annotators are not expert, while the latter is suggested when the annotators are experts and the boundaries of the phenomenon have some fuzziness that are to be taken into account. In between of these two approaches, others can be defined that weight the number of annotators or specific features.

In order to release the dataset, a final check is recommended to avoid loops and inconsistencies. Tools for graph analysis can be used to support this phase, even though the correction of errors mostly requires manual revision.

2.5.3 Gold-PR Dataset Release

To support dataset usage and comparison with other datasets it is recommended that all the relevant information concerning the annotation process and the main statistics about the final Gold-PR dataset are carefully documented. In particular, the project manager should take care of publishing the list of concepts, the list of PR pairs of concepts, and the PR pairs in-context. Moreover, a description of the dataset using standard vocabularies is recommended. *Metadata* should include at least:
- dataset main information: name, description, URL, keywords, language, encoding format,
- corpus (textbook) information: bibliographic references, curators,
- annotation method: PREAP protocol and the features adopted concerning concept annotation method, combination method,
- dataset statistics: number of domain concepts involved in PR relations, number of PR relations, number of annotators, Kappa agreement score,
- information for accessing and using the dataset: license type, free/paid access.

3 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.

3.1 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.

3.2 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.