Extracting Dependency Relations from Digital Learning Content

Giovanni Adorni[1], Felice Dell'Orletta[2], Frosina Koceva[1(✉)], Ilaria Torre[1], and Giulia Venturi[2]

[1] Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy
{giovanni.adorni,ilaria.torre}@unige.it, frosina.koceva@edu.unige.it
[2] Istituto di Linguistica Computazionale Antonio Zampolli (ILCCNR), Pisa, Italy
{felice.dellorletta,giulia.venturi}@ilc.cnr.it

Abstract. Digital Libraries present tremendous potential for developing e-learning applications, such as text comprehension and question-answering tools. A way to build this kind of tools is structuring the digital content into relevant concepts and dependency relations among them. While the literature offers several approaches for the former, the identification of dependencies, and specifically of prerequisite relations, is still an open issue. We present an approach to manage this task.

Keywords: Prerequisite relationship · Concept extraction
Graph mining

1 Introduction

The 21th century is marked by the exponential growth of data and of digital contents. Digital libraries evolved from static storage and retrieval platforms to dynamic services to explore, exchange and share information and knowledge.

In this paper, our focus is on the potential role of digital libraries for education. The idea is that digital resources can not only be explored and shared but they can be coupled with services that support learning processes. This usually requires that content is extracted, structured and enriched with annotations. Since the objective is supporting learning, the extraction of relevant concepts has to be complemented with the identification of prerequisite relations among these concepts. This enables the building of services that, for example, enable to find pieces of knowledge in the text and to extract also the related propaedeutic concepts and resources that allow such information to be properly understood (prerequisite relations).

Manual annotation is of course the most effective approach, but it is time consuming and requires experts knowledge. Therefore, a challenge is the automatic learning of the knowledge structure of the content.

While several methods exist (e.g., [1,3]) to face the issue of concept extraction, the identification of prerequisite relations among concepts is still an open research problem. In this paper we present methods and approaches for facing this issue.

G. Serra and C. Tasso (Eds.): IRCDL 2018, CCIS 806, pp. 114–119, 2018.
https://doi.org/10.1007/978-3-319-73165-0_11

2 Research Issue and Background

The two main tasks for automatic concept map building are the concept extraction and the relations identification between concepts [7]. Even though there is a long-standing interest since at least 1971 Gagné's work on learning hierarchies [6], identifying prerequisite relations among concepts is an open issue.

The prerequisite relation between two concepts A and B is a dependency relation which represents what a learner must know/study (concept A), before approaching concept B. Thus, A is a propaedeutic concept, i.e. a requirement, for B and the learner should first understand A in order to understand B.

The prerequisite relation can represent a hyponymy or meronymy relation in the case where the hyponym/meronym concept is going to be further in-depth studied and therefor is itself a prerequisite to another concepts. The prerequisite relation usually requires experts to be evaluated since its semantics can be properly evaluated only by considering the whole graph and the learning goal.

Notation. In the following we provide the conventions and definitions that will be used along the paper. We define a document D as a textual resource. The output of the concept extraction is the terminology $T \in D$ with $t \in T$, where t is a domain-specific term, composed of one or more words (single nominal terms or complex nominal structures with modifiers). For each term, the process returns also its relevance $r = [0, 1]$ (see Sect. 3 for definition).

When D is structured into parts, sections (S), the output of the concept extraction can be $T \in D$ and $T \in S$ according to the needs. Subsections are managed as Sections. Thus we have concept-document and concept-section relationships. We denote these relationships as relevance functions $F(\cdot, \cdot)$ which take the concept and D/S as arguments and have the relevance r as output.

The final output of concepts and prerequisite relations extraction is a concept graph G. Similarly to [10], we represent G as a set of triples in the form $G = \{(t_1, t_2, p) | t_1, t_2 \in T, 0 \leq p \leq 1\}$, where p is the prerequisite relationship and can take a value from 0 to 1, indicating the strength of the prerequisite relation between t_1 and t_2 (where t_1 is prerequisite of t_2).

Term appearance in section is defined as a pair (t_i, s_j), $t_i \in T$ and $s_j \in S$.

3 Concept Extraction

Our approach to the identification of prerequisite relations was tested on the handbook entitled *Computer Science: An Overview: Global Edition*, G. Brookshear and D. Brylow, Pearson 2015. In order to identify relevant concepts within the considered book, we exploited Text–To–Knowledge (T2K^2) [3], a software platform developed at the Institute of Computational Linguistics "A. Zampolli" of the CNR in Pisa. T2K^2 relies on a battery of tools for Natural Language Processing, statistical text analysis and machine learning which are dynamically integrated to provide an accurate representation of the linguistic information and of the domain-specific content of multilingual text corpora. T2K^2 encompasses two main sets of modules, respectively devoted to carry

out the linguistic pre-processing of the acquisition corpus and to extract and organize the domain knowledge contained in the linguistically annotated texts. Each section of the considered handbook was automatically enriched (i.e. annotated) with linguistic information at increasingly complex levels of analysis, represented by sentence splitting, tokenization, Part-Of-Speech tagging and lemmatization. According to the methodology described in [2], the automatically POS–tagged and lemmatized input text is searched for candidate domain–specific terms denoting domain entities expressed by either single nominal terms (e.g. *internet, network, software*) or complex nominal structures with modifiers (typically, adjectival and prepositional modifiers), where the latter are retrieved on the basis of a set of POS patterns (e.g. adjective + noun, noun + preposition + noun) encoding morpho–syntactic templates for multi–word terms (e.g. *Internet Protocol, eXtensible Markup Language, client/server model*). The domain relevance of both single and multi–word terms t included in the extracted list T is weighted on the basis of the C-NC Value [5] aimed at assessing how much a term is likely to be conceptually independent from the context in which it appears. Accordingly, a higher C-NC rank is assigned to those multi-word terms that are more relevant for the domain of the document collection in input. The extracted domain-specific entities are organized according to co-occurrence relations, i.e., relations between entities co-occurring within the same context. The relevance of relations is weighted using the log-likelihood metric for binomial distributions as defined by [4]. According to this metric, for example, the term *Internet* is strongly related with *Internet Protocol addresses, Simple Mail Transfer protocol, message*, etc. The extracted relations between terms can be visualized in a 'knowledge graph' which can be exploited in a number of graph analyses. M1 in the next section is based on the knowledge graph.

4 Prerequisite Relationship Identification

In this paper we propose two methods for identifying candidate prerequisite relationships $(t1, t2, p)$, with $p \in [0, 1]$. The underlying principles are:
- Co-occurrence of two concepts is a necessary but not sufficient condition to identify the prerequisite relation. The principle can be extended from the sentence level to a section level.
- Temporal occurrence of terms and/or sections are taken into account to identify the direction of prerequisite relation, with different granularities.

Since the methods exploit these principles in different ways, they are designed to be finally combined in order to exploit the benefits of both the approaches.

Method 1 (M1) is based on temporal order and co-occurrence of terms. Steps:

- Building a list L of terms $t \in T$ ordered according to their temporal appearance in D where the term t has the first significant density (which can be compute with different methods, e.g. Burst Analysis).
- Transforming the undirected knowledge graph from Sect. 3 generated with log-likelihood metric into a directed graph G_1, where direction is derived from the ordered list of terms L.

Result: Candidate triples for prerequisite relations are the adjacent terms in G_1 (Fig. 1). The G_1 graph is represented as a $n\,x\,n$ matrix M_1, with $n = |T|$. Each element t_{ij} represents the weight p of the prerequisite relationship between terms t_i and t_j, with $p = [0, 1]$.

The strength of relationship p can be defined using different approaches as: NLP analysis, Lexical pattern and other heuristics.

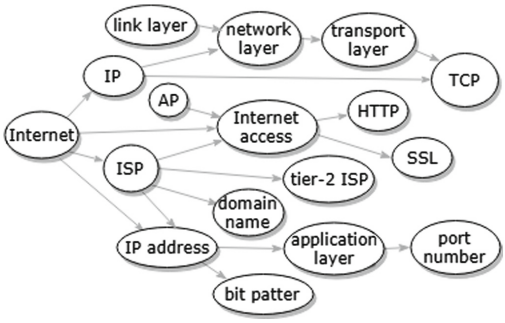

Fig. 1. Method 1 - Examples of candidate prerequisite relations

Method 2 (M2) is based on text structure D/S (Table of Content): The goal of this approach is to identify, for each term, the cluster of terms that are likely or unlikely to be in prerequisite relationship with the term. $TOC(s_i, s_j)$ represents the order $\prec$ of section i and section j, where $s_i, s_j \in S$. The application of the method is represented in the examples in Fig. 2. Steps:

- For each term $t \in T$, identifying the section s_i where the relevance function $F(t, q)$ has max value (i.e., identifying the section where the term has the higher relevance in the document); the assumption is that a concept is explained where it has maximum relevance.
- For each (t_v, s_i), where $v \neq u$, identifying the section s_j where the relevance function $F(t_v, s_j)$ has max value
  - (i) If $s_j \prec s_i \wedge \nexists\,(t_u, s_j)$, its unlikely that t_u is a prerequisite of t_v based on the principle that in s_j there should be at least one occurrence of the prerequisite (t_u), see Fig. 2 (i).
  - (ii) If $s_i \prec s_j \wedge \nexists\,(t_u, s_j)$ is likely that t_v is a prerequisite of t_u, since t_v is explained before t_u and it also co-occurs in s_i, see Fig. 2 (ii).
  - (iii) If $s_j \prec s_i \wedge \exists\ (t_u, s_j)$ there is some probability that t_v is a prerequisite of t_u, since they could be highly related concepts but not as prerequisite relationship. Similarly, if $s_i \prec s_j \wedge \exists\,(t_u, s_j)$ there is some probability that t_u is a prerequisite of t_v, for the same reason as in the previous point, see Fig. 2 (iii).
  - (iv) If $s_i = s_j$, thus t_v and t_u co-occur with maximum relevance in the same section, see Fig. 2 (iv), this means that the concepts are highly related but we cannot identify the prerequisite relationship unless further analysis is performed, such as: NLP, Lexical pattern and other heuristics.

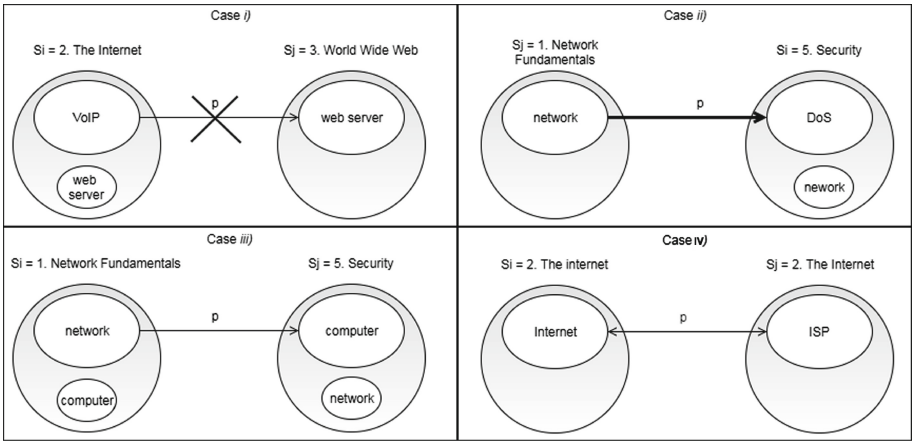

Fig. 2. Method 2 - Examples of candidate prerequisite extraction

Result: The candidate prerequisite relations are represented as a $n\ x\ n$ matrix M_2, with $n = |T|$. Each element t_{ij} represents the weight p of the prerequisite relationship between terms t_i and t_j, with $p = [0, 1]$. The implementation of the algorithm can apply values of p according to the rules above which can be tuned in order to fit the specific domain.

5 Discussion and Conclusion

In this section we discuss the proposed approach by comparing our methods with related approaches for concept and prerequisite extraction. An approach that exploits textbook internal information (TOC) to identify prerequisite relations is adopted in [10], even though they also exploit external knowledge (from Wikipedia) to extract the relevant concepts. Another approach that exploits Wikipedia is described in [8]. The authors define a metric (i.e., refD) that models the relation by measuring how differently two concepts refer to each other. In [9] the authors mine prerequisite relations among MOOC course concepts by defining three main features: semantic (incorporates wikipedia knowledge), contextual (similar to refD [8]) and structural distributional patterns.

Unlike the above cases, our approach exploits only features from the text (co-occurrence, term density, temporal and TOC ordering) for concept and prerequisite extraction, without using external knowledge. With respect to [10], while the authors exploit TOC title match and order coherence, we identify candidate prerequisite relation by the joint usage of not only TOC order (M2) but also the temporal concept density order (M1), thus providing a more granular method. Moreover, while in [10] the information overlap is calculated by using Wikipedia title match and similarity functions, we use concept-section order analysis (M2) to identify three specific cases of concept redundancy of which (ii) identifies prerequisite candidate conceptually similar to refD in [8] where

the sections in our case can be seen as the wikipedia articles in refD. Whereas most of the aforementioned methods for prerequisite extraction result in a concept hierarchy building, i.e. tree structure, the M2 (*iii*) give the bases towards a graph building by adding parallel prerequisite relations.

Enhancement of M1 can be made by introducing metrics based on concept bursting intervals (e.g. [11]) for building the list L. In addition, by analyzing more than one book (with the same subject), both methods can be improved by reducing biases due to the author's subjective choices in structuring the book. We are working on testing the methods and the mentioned enhancements.

Acknowledgements. The authors thank prof. Carlo Tasso for making available Distiller system for concept extraction during the initial experiments of the described methodology.

References

1. Basaldella, M., Chiaradia, G., Tasso, C.: Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction. In: COLING, pp. 804–814 (2016)
2. Bonin, F., Dell'Orletta, F., Venturi, G., Montemagni, S.: A contrastive approach to multi-word term extraction from domain corpora. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (2010)
3. Dell'Orletta, F., Venturi, G., Cimino, A., Montemagni, S.: T2k^2: a system for automatically extracting and organizing knowledge from texts. In: Proceedings of 9th International Conference on Language Resources and Evaluation, pp. 2062–2070 (2014)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. **19**(1), 61–74 (1993)
5. Frantzi, K., Ananiadou, S.: The C-value/NC value domain independent method for multi-word term extraction. J. NLP **6**(3), 145–179 (1999)
6. Gagné, R.M.: Learning hierarchies. In: Merrill, M.D. (ed.) Instructional Design: Readings, pp. 118–131. Prentice-Hall, Englewood Cliffs (1968, 1971)
7. Kowata, J.H., Cury, D., Boeres, M.: A review of semi-automatic approaches to build concept maps. In: Proceedings of the 4th Conference on Concept Mapping, pp. 40–48 (2010)
8. Liang, C., Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: EMNLP, pp. 1668–1674 (2015)
9. Pan, L., Li, C., Li, J., Tang, J.: Prerequisite relation learning for concepts in MOOCs. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, Long Papers, vol. 1, pp. 1447–1456 (2017)
10. 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 Conference on Information and Knowledge Management, pp. 317–326 (2016)
11. Yoon, W.C., Lee, S., Lee, S.: Burst analysis of text document for automatic concept map creation. In: Ali, M., Pan, J.-S., Chen, S.-M., Horng, M.-F. (eds.) IEA/AIE 2014. LNCS (LNAI), vol. 8482, pp. 407–416. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07467-2_43