Discovering Hierarchical Relationships in Educational Content

Marián Šimko and Mária Bieliková

Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Ilkovičova 3, 842 16 Bratislava
{simko,bielik}@fiit.stuba.sk

Abstract. Adaptive educational hypermedia necessitate semantic description of a domain, which is used by an adaptive engine to perform adaptation to a learner. The bottleneck of adaptive hypermedia is manual authoring of such semantic description performed by a domain expert mainly due to the amount of descriptions to be created. In this paper we present a method for automated discovery of is-a relationship, one of the most important relationships of conceptual structures. The method leverages specifics of educational content. The evaluation shows reasonable accuracy of discovered relationships reflecting in reduced domain expert's efforts in domain model creation

Keywords: adaptive hypermedia, domain model, semantics discovery, automatic relationship discovery, natural language processing

1 Introduction and Related Work

In order to enable adaptation during learning within an educational web-based system, a subject domain has to be properly semantically described. An adaptive engine utilizes *domain model* a representation of a domain conceptually describing resources to be a subject for adaptation.

Domain representation among different educational systems varies from conceptual maps to complex domain ontologies [8]. In most cases, core domain model consists of domain knowledge elements (concepts represented by relevant domain terms) and relationships between them. There are typically dozens of concepts and hundreds of relationships of various kinds in a domain model making manual creation and maintenance of such structure a demanding and very difficult task. It is important to seek for methods that facilitate domain model creation and reduce a teacher's (or an instructional designer's) efforts.

Automated domain model creation is possible by processing underlying textual content and/or a structure of a course. Much research has been devoted to text mining, however, to our best knowledge, only a small number of approaches focus on *educational* text mining, while considering its *specifics* such as domain specificity of vocabulary (it is natural to introduce new terms) or explanatory nature of language used (learning objects are richer in explanatory phrases).

There exist few approaches to automatic relationship acquisition in adaptive learning. The authors of adaptive system My Online Teacher developed a method for computing similarity between concepts by calculating correspondence weights computation between concepts attributes [5]. The idea is based on co-occurrence comparison of keywords and overall attributes' contents of concepts. Sosnovsky et al. aim at automated prerequisite and outcome relationships identification [10]. Based on predefined concept pattern detection, they extract concepts from learning objects on C programming language. An interesting example of automated metadata acquisition was performed in the case of adaptive vocabulary acquisition system ELDIT [2], where methods and techniques of natural language processing were employment in order to create relationships between vocabulary entries and their examples. In our previous work we devised a method for relevant domain terms relatedness computation based on statistical and graph processing of the domain model [12]. Zouaq and Nkambou present a two-step method for domain ontology learning from educational text, including concept rela-tionships [16]. The method is based on pattern-based semantic analysis and linguistic processing of educational content. A semi-automatic approach to domain model building is presented by Šaloun et al. [11].

Despite small number of approaches in the domain of adaptive learning, there is a lot of work in taxonomical relationship extraction in the field of ontology learning. The approaches can be according to Cimiano [3] divided into the three groups: lexico-syntactic patterns matching (e.g., [7]), leveraging distributional hypothesis (e.g., [1]), and co-occurrence analysis (e.g., [6]). These approaches form a solid basis for further adoption to educational domain to utilize the potential of educational text specifics.

In this paper we present a method for automated relationship discovery in educational content. We extend our previous research in the area of automated domain model acquisition [13] and we focus on hierarchy (is-a) relationship between relevant domain terms. It constitutes one of the most important types of relationships as it forms a skeleton of a domain representation. We present three techniques, each covering different linguistic aspect of educational content.

2 Method for is-a Relationship Discovery

Our method for is-a relationship discovery combines statistics- and linguistics-based approaches to data mining. It builds on preceding learning objects preprocessing and relevant domain terms identification steps, which are based on analysis of learning content [13].

The method incorporates three techniques, each consisting of different steps:

- Explanation phrase processing,
  - explanation phrase lookup,
  - relevant domain term overlap computation,
  - distance of overlapping tokens computation;
- Determination phrase processing,
  - determination phrase lookup,

    - relevant domain term overlap computation;
- Relevant domain term lexical analysis,
    - relevant domain term lexical overlap computation.

The final step of each technique is is-a relationship generation, where relationships are generated and a confidence for each relationship is derived. While focusing on a different language aspect, each technique yields its own set of is-a relationships. These sets are filtered and combined with respect to a relevance of particular technique in order to produce a single set of is-a relationships[1].

2.1 Explanation Phrase Processing

This technique is based on lexico-syntactic analysis of underlying text content. Our aim is to find explanation phrases in sentences and extract *explanation candidates* for is-a relation.

The main idea is to search for patterns, which indicate is-a relationship between relevant domain terms. Patterns are inspired by a seminal work of Hearst, where she proposed a set of general patterns for hyponymy relationship acquisition from unstructured text [7]. We adopt such patterns for an educational domain by incorporating lexical and syntactical constructions that are used with the intent to teach, explain or clarify (hence they involve a potential is-a relationship). Learning objects typically are represented by resources, which aim to introduce and expound certain phenomena related to subject domain. For instance, the pattern:

$$\text{Understand } \{\textit{something}\} \text{ to be/mean } \{\textit{something}\}$$

in a learning object may indicate is-a relationship. It is more likely to indicate is-a relationship in a learning object than in an ordinary text. In lexico-syntactic patterns definition we particularly focus on verb forms that indicate an effort of cognitive organization of objects in the sentence, e.g., to be, to understand, to constitute, to name, to represent, to divide, to belong to, to fall into, etc.

When matching patterns, not only word forms but also their morphological tags are matched reflecting into increased accuracy of match. For example, applying a rule consisting of explanation verb form "is termed as" bound with all nouns and adjectives in nominative case will match the sentence "*Biologically our species is termed as Homo sapiens*" indicating is-a relationship: *is-a*(Homo sapiens, species).

Explanation phrase processing is divided into the following steps:

1. *Explanation phrase lookup.* In this step we use predefined rules to match patterns adopted for educational text with learning objects content and extract so called explanation phrase candidates, which contain tokens satisfying morphological criteria defined by the rules.

[1] In fact, the final step of our method covers relationship combination, duplicates removal, and loop resolution. Since we aim to explore the three techniques, detailed description of these steps is beyond the scope of the paper.

2. *Relevant domain term overlap computation.* The goal of this step is to check if extracted explanation phrases contain relevant domain terms. We compare explanation phrase candidate tokens with relevant domain terms and compute lexical overlap between their word forms. Basically, more overlapping words, the higher overlap.
3. *Distance of overlapping tokens computation.* We recognize distance of a token as a measure representing how tight the token is bound to an explanation verb. The closer a token is, the more likely it is paradigmatically related to explanation head, which indicates is-a relationship.
4. *Is-a relationship generation.* We generate is-a relationship and compute confidence of relationship correctness based on (i) lexical overlap with available relevant domain terms, and (ii) distance obtained in previous steps.

2.2 Determination Phrase Processing

A characteristic of an explanatory text is that it expounds new topic and extends a vocabulary by introducing new terms. The main idea of this technique is based on an observation that newly defined relevant domain terms are relatively often accompanied by a term that clarifies or classifies relevant domain terms' meaning. Accompanying terms are nouns or noun phrases, as they are main holders of the meaning (in contrast with other lexical categories such as adjectives or adverbs, which qualify nouns). For example, from the sentence "*Here is an example which uses the predicate listp to check for a list.*" we deduce that *is-a*(listp, predicate). In this work we refer to such relation between two terms to as determination. A phrase composed of two (or more) such terms, i.e., nouns or noun phrases that collocate, we refer to as *determination phrase*. Determination phrases are typically bound with rather technical terms, which have roots in a language (natural or artificial) different from the language of a subject domain.

Is-a relationships indicated by determination phrases more likely refer to the concepts from lower parts of taxonomy as they connect a more specific term with a term of any level of specificity. We assume that there is reduced intent of a teacher to explicitly determine a more general term by accompanying it with another more general term.

Determination phrase processing consists of the following steps:

1. *Determination phrase lookup.* In the first step we look up all determination phrases in learning objects by matching the syntactical pattern:

$$NP_0\ NP_1\ [,\ NP_2\ \ldots [\text{and/or}]\ NP_n]$$

representing subsequently collocated noun phrases. NP_0 is supersumed and $NP_{1..n}$ are subsumed noun phrases. Further restrictions related to noun phrases' grammatical categories (e.g., number, case) are language-specific and may reflect different morphological and word-formation rules.
2. *Relevant domain terms overlap computation.* Similarly to the preceding part of the method, we check after matching the pattern if extracted determination phrases contain relevant domain terms. For each determination phrase

token we compute lexical overlap with relevant domain terms. Distance discrimination factor is not relevant here as the distance is constant for all overlapping relevant domain terms.

3. *Is-a relationship generation.* In this step we generate is-a relationship candidates by traversing all determination phrases in each document and computing is-a relationship confidence. We consider in computation the overlap with relevant domain terms and also the count of determination phrase occurrences. It is due to the fact that determination phrases occur more frequently across the whole text corpora. The number of occurrences of such phrases is directly proportional to the number of occurrences of technical (relevant domain) terms (since determination phrases more likely cover sequences of noun phrases containing technical terms).

2.3 Relevant Domain Term Lexical Analysis

This technique's basis follows from an observation that basic word forms of relevant domain terms, which form is-a relationship, often overlap lexically (e.g., data type vs. atomic data type). Based on an assumption that a longer form is a specification of a shorter form, we examine the extent to which two relevant domain terms overlap lexically.

Relevant domain term lexical analysis consists of the following steps:

1. *Relevant domain terms lexical overlap computation.* In this step we match lexical forms of relevant domain term representations. We compute lexical overlap between each pair of relevant domain terms in a course as follows:

$$overlap(rdt_i, rdt_j) = \frac{|L(rdt_i) \cap L(rdt_j)|}{|L(rdt_i) \cup L(rdt_j)|} \tag{1}$$

   where rdt_i is a relevant domain term and $L(rdt_i)$ is a set of token lemmas of rdt_i. For the example above holds: $overlap(datatype, atomicdatatype) = 0.6\bar{6}$. Lexical overlap is directly proportional to the number of tokens common for both relevant domain terms. Overlap is equal to 0 for lexically different terms and for equal ones it is 1.
2. *Is-a relationship generation.* We compute confidence of relationships based on relevant domain terms lexical overlap. We believe even a small overlap can indicate relatively reliable presence of is-a relationship between two relevant domain terms.

The proposed method for is-a relationship generation is based on comprehensive linguistic analysis of learning objects. It particularly considers specifics of educational content and the explanatory nature of learning material. It is based on an assumption that learning objects contain explanatory phrases (indicated by selected explanatory verbs) and determination phrases (a specific form of collocation of nouns or noun phrases). In addition, it utilizes lexical analysis of multiword relevant domain terms.

Although being language independent, the method's accuracy is strongly connected with language-specific patterns and lexico-syntactical matching rules that are used by a particular technique.

3 Evaluation

We evaluated our method in the domain of learning programming in a course of Functional programming lectured at the Slovak University of Technology in Bratislava.

3.1 Data and method application

The official learning material for functional programming consists of 79 explanatory learning objects on the functional programming paradigm and programming techniques in the Lisp language. The material is hierarchically organized into chapters and sections according to a printed textbook for the course. All learning materials are in Slovak. The course is available online in our adaptive educational system ALEF [14].

As the Functional programming course has already been involved in adaptive learning as a part of our previous research [12, 14], learning objects have assigned relevant domain terms defined by domain experts.

Before evaluating partial techniques for domain model acquisition, we preprocessed all 79 learning objects following the natural language processing pipeline consisting of tokenization, POS tagging, lemmatization and sentence-based segmentation. Then we applied our method for is-a relationship discovery. By applying all three techniques for is-a relationship extraction (explanation phrase processing, determination phrase processing, relevant domain term lexical analysis), we extracted 84, 92 and 55 is-a relationships, respectively. By making union from all is-a relationships we obtained a final set of 206 unique relationships, which were a subject of evaluation.

3.2 Results and discussion

We performed two-step evaluation of our method: a posteriori evaluation and comparison against the gold standard. In a posteriori evaluation we involved four domain experts to assess correctness of acquired relationships in order to set relevancies of the techniques accordingly. We achieved the best results for determination phrase processing, followed by relevant domain term analysis. Relationships acquired by explanation phrase processing ranked third, mainly due to the highest complexity of natural language processing. We used the information about technique correctness to update overall relationship confidences and to create a combined sorted list of is-a relationships. Providing more details on a posteriori evaluation is beyond the scope of this paper.

We evaluated the updated set of is-a relationships against the gold standard represented by a functional programming domain model created manually by a group of several domain experts independently of our method. The manually created domain model is employed in adaptive learning portal ALEF [14] as a part of educational activities at the Slovak University of Technology in Bratislava.

The gold standard consists of 162 relevant domain terms and 256 relationships in between. 128 of them are is-a relationships, 135 relevant domain terms

are involved in is-a relationship (at least one incident edge is is-a relationship). It is important to note that the relevant domain terms from the gold standard correspond to relevant domain terms utilized by our method.

In order to evaluate the validity of generated is-a relationships, we borrowed from the work of Mädche and Staab [9]. In their work they proposed two layers of taxonomy comparison: lexical and conceptual. At lexical comparison level the terminological overlap between two taxonomies is computed. At the conceptual comparison level semantic structures of taxonomies are compared.

As the relevant domain term sets are intentionally identical, lexical comparison level is not relevant here. We assess the structure of the generated is-a relationships only. We follow their approach and adopt the definition of semantic cotopy of concept in taxonomy and slightly change it for our purpose:

$$SC(rdt, DM) = rdt_j \in DM : isa_{DM}(rdt, rdt_j) \vee isa_{DM}(rdt_j, rdt) \quad (2)$$

where SC is semantic cotopy of relevant domain term rdt in the domain model DM. isa_{DM} is is-a relationship in domain model DM. Semantic cotopy of a relevant domain term rdt represents a set of all subsumed and supersumed relevant domain terms of rdt in a given domain model DM.

We utilize the notion of semantic cotopy and we define taxonomic precision and taxonomic recall measures as follows:

$$P_T(DM_{retr}, DM_{rel}) = \frac{\sum_{rdt \in DM_{uni}} |SC(rdt, DM_{retr}) \cap SC(rdt, DM_{rel})|}{\sum_{rdt \in DM_{uni}} |SC(rdt, DM_{retr})|} \quad (3)$$

$$R_T(DM_{retr}, DM_{rel}) = \frac{\sum_{rdt \in DM_{uni}} |SC(rdt, DM_{retr}) \cap SC(rdt, DM_{rel})|}{\sum_{rdt \in DM_{uni}} |SC(rdt, DM_{rel})|} \quad (4)$$

where P_T and R_T are taxonomic precision and recall of domain model DM_{retr} with respect of domain model DM_{rel}, respectively. DM_{retr} represents domain model containing is-a relationships generated by our method ("retrieved") and DM_{rel} represents the gold standard ("relevant"). $DM_{uni} = DM_{rel} \cup DM_{retr}$.

These are one of the strictest measures for quantitative comparison of two taxonomical structures as they fully consider transitivity of is-a relationship. If some erroneous is-a relationship occurs in the center of taxonomy, it affects not only incident relevant domain terms, but also all subsumed and supersumed relevant domain terms in a hierarchy.

Beside comparison of taxonomical structures, we further want to assess the method we proposed. Data that are used, i.e., learning object corpus that was processed, are an important factor of the method success rate. The method we proposed relies on the already defined set of relevant domain terms and it

assumes that they occur within the text. However, we found out that some relevant domain terms as defined by the gold standard creators did not occurred in the learning objects in the form we were able to process (e.g., they occur in non-processable content such as pictures, or contain special symbols not properly recognized during preprocessing step). As a result, such relevant domain terms cannot be involved in any relationship. Thus, we computed taxonomic precision and recall considering both (i) the gold standard, and (ii) the gold standard without relevant domain terms that could not be found in the learning objects. The results depicting precision and recall measures together with F-measure (denoted as P_T, R_T, F_T for (i) and P'_T, R'_T, F'_T for (ii)) are presented in Fig. 1.

We consider the obtained results of the evaluation very reasonable. Precision of the generated is-a relationship is very promising, recall is a bit lower than expected. Deeper insight into the results revealed that recall is to a certain extent affected also by underlying text corpora that is a source for processing.

We identified several reasons that could have a negative impact on the results of evaluation:

– language - Slovak language is inflective language containing considerable number of exceptions in morphology, word formation and also in phrasing. Our method could benefit from more precise preprocessing including constituent identification and anaphora resolution. The language issue reflects into pattern detection and reduces number of correctly matched patterns. However, since the proposed method is language-independent (albeit different patterns for distinct languages need to be defined), we anticipate much better results for less complicated languages such as English. As the explanation phrases processing yielded less satisfactory results, we may also focus on patterns refinement by creating stricter matching criteria.

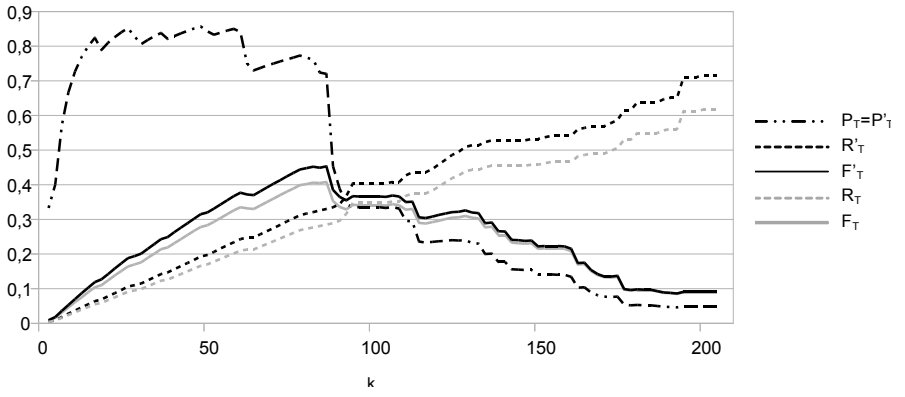

Fig. 1. Precision, recall and F-measure for top k generated is-a relationships when compared against the gold standard. We considered two gold standard variants: containing all relevant domain terms from the domain model (P_T,R_T,F_T) and only relevant domain terms that were recognized in the text (P'_T,R'_T,F'_T).

– size of text corpora - size of functional programming learning objects corpora is much smaller than corpora, where similar tasks from related fields such as ontology learning were applied. The less content, the lower the chance to extract is-a relationships. This drawback can be potentially reduced by processing additional, student generated content created during learning. When considering collaborative learning web-based environments, external learning resources assigned by students may be processed. As a result, new relationships may be extracted that can enrich or enhance existing set.
– the gold standard accuracy - unlike other approaches, we could not compare the results against extensively used gold standards such as WordNet. In order to be objective, we need to admit that accuracy of the gold standard we used is not perfect. Although the precision of the manually created domain model (in comparison with an non-existent perfect domain model) most probably cannot be doubted, the recall may be disputable. As the gold standard is a result of small group of domain experts, it may consider some valid relationships to be incorrect (or disputable - other domain experts opinions could be not uniform), it may reflect into decreased precision, recall and F-measure of automatically generated is-a relationships.

4 Conclusions

Automated metadata discovery is very important for adaptive web-based educational systems since necessary semantic descriptions of underlying resources are very hard to create and maintain manually. By devising our method for hierarchical relationship discovery in educational texts we extend the state-of-the-art in the educational text processing and educational metadata acquisition. We built on the preceding research in lexico-syntactical analysis of text while adopting to and leveraging specifics of educational content.

Evaluation of the method in the functional programming course showed that the method yields very promising results that can be used as a solid basis for supporting content and metadata authors by offering them sets of relationships to select from while designing an adaptive course.

The issue of automated metadata discovery is especially relevant in dynamically changing social learning environments [14, 15] with user-generated content (tags, comments, annotations) being created on daily basis. The ability to provide necessary descriptions in such environments is reduced even more and with no automated support for metadata creation they cannot fully benefit from advanced functionality such as recommendation or personalized search over user-generated content.

In our future work we will focus on discovery of other types of relationships. We aim to provide an integrated framework for educational content mining by following up the results of our previous works [12, 13]. While considering adaptive web-based learning 2.0, we will also research how user-generated content can supplement content provided by teachers and to what extent user-generated content (and which forms) could facilitate and improve metadata discovery from text content that it is assigned to.

Acknowledgments. This work was supported by grants No. VG1/0675/11, VG1/0971/11 and it is a partial result of the Research and Development Operational Program for the projects SMART, ITMS 26240120005 and SMART II, ITMS 26240120029, co-funded by ERDF.

References

1. Bisson, G., Nedellec, C., Canamero, L.: Designing clustering methods for ontology building The Mo'K workbench. In Proc. of the First Workshop on Ontology Learning OL'2000, CEUR-WS, Vol. 31, pp. 13-19 (2000)
2. Brusilovsky, P., Knapp, J., Gamper, J.: Supporting teachers as content authors in intelligent educational systems. In Int. Journal of Knowledge and Learning, 2006, vol. 2, no. 3/4, pp. 191–215 (2006)
3. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag, ISBN: 978-0-387-30632-2. 347p (2006)
4. Cimiano, P., Hotho, A., Staab, S. Learning concept hierarchies from text corpora using formal concept analysis. In J. of Artificial Intelligence Research. Vol. 24, pp. 305–339 (2005)
5. Cristea, A.I., de Mooij, A.: Designer Adaptation in Adaptive Hypermedia Authoring. In Proc. of the Int. Conf. on Information Technology: Computers and Communications ITCC'03. IEEE, pp. 444–448 (2003)
6. Fotzo, H. N., Gallinari, P.: Learning generalization/specialization relations between concepts: application for automatically building thematic document hierarchies. In Proc. of the 7th Conf. on Computer-Assisted Inf. Retr., CID, pp. 143–155 (2004)
7. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th Int. Conf. on Computational Linguistics. ACL, pp. 539–545 (1992)
8. Henze, N., Nejdl, W.: A Logical Characterization of Adaptive Educational Hypermedia. New Review of Hypermedia and Multimedia, 10(1), pp. 77–113 (2004)
9. Mädche, A., Staab, S.: Measuring similarity between ontologies. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Springer, pp. 15–21 (2002)
10. Sosnovsky, S., Brusilovsky, P., Yudelson, M.: Supporting Adaptive Hypermedia Authors with Automated Content Indexing. In Proc. of 2nd Int. Workshop on Authoring of Adaptive and Adaptable Educ. Hypermedia, pp. 380–389 (2004)
11. Šaloun, P., Velart, Z., Klimanek, P.: Semiautomatic domain model building from text-data. In Proc. of Sixth Int. Workshop on Semantic Media Adaptation and Personalization SMAP 2011. IEEE Computer Society, pp. 15–20 (2011)
12. Šimko, M., Bieliková, M.: Automatic Concept Relationships Discovery for an Adaptive E-course. In Barnes, T., Desmarais, M., Romero, C., Ventura, S. (Eds.). Proc. of 2nd Int. Conf. on Educational Data Mining, EDM 2009, pp. 171–179 (2009)
13. Šimko, M., Bieliková, M.: Automated Educational Course Metadata Generation Based on Semantics Discovery. In LNCS 5794, Proc. of European Conf. on Technology Enhanced Learning, EC TEL 2009. Springer, pp. 99–105 (2009)
14. Šimko, M., Barla, M., Bieliková, M.: A Framework for Adaptive Web-based Learning 2.0. In Reynolds, N., Turcsányi-Szabó, M. (Eds.). KCKS 2010, IFIP Advances in Information and Com. Technology, Vol. 324. Springer, pp. 367–378 (2010)
15. Tvarožek, J.: Bootstrapping a Socially Intelligent Tutoring. Information Sciences and Technologies Bulletin of the ACM Slovakia. 3(1), pp. 33–41 (2011)
16. Zouaq, A., Nkambou, R.: Building Domain Ontologies from Text for Educational Purposes. In IEEE Trans. Learn. Technol., Vol. 1, Issue 1, pp. 49–62 (2008)