Automated Assistance for Data Modelers combining Natural Language Processing and Data Modeling Heuristics: A Prototype Demonstration*

Benjamin Ternes[1], Kristina Rosenthal[1], and Stefan Strecker[1]

Enterprise Modelling Research Group, University of Hagen, Hagen, Germany,
{benjamin.ternes,kristina.rosenthal,stefan.strecker}@fernuni-hagen.de

Abstract. Identifiers of model elements convey semantics of conceptual models essential to interpretation by human viewers. Prior research shows that devising meaningful identifiers for model elements challenges data modelers from early learning stages to advanced levels of modeling expertise, constituting one of the most common difficulties data modelers face. We demonstrate the Automated Assistant, an integrated modeling tool support component combining natural language processing techniques and data modeling heuristics to provide data modelers with modeling-time feedback on identifying and signifying entity types, relationship types, and attributes with meaningful identifiers. Different from other approaches to automating assistance for data modelers, the Automated Assistant implementation does not rely on fixed reference solutions for modeling tasks as it processes (m)any natural language descriptions of modeling tasks. We report on the current state of prototype development, discuss the Automated Assistant implementation and outline future work.

Keywords: Conceptual Modeling · Data Modeling · Modeling Tool · Natural Language Processing · Process-oriented Feedback

1 Introduction

Model element identifiers (labels), e. g., for entity types, relationship types and attributes in an Entity-Relationship (ER) diagram, carry and convey semantics important to sensible interpretation of conceptual data models by human viewers, semantics transcending the formal semantics of model elements [12]. Devising meaningful, appropriate, and expedient identifiers is a prerequisite for comprehensible and usable conceptual models [7]. Empirical research shows that data modelers oftentimes face difficulties devising meaningful identifiers for model elements from early learning stages to advanced levels of modeling expertise, described as one of the most common difficulties beginning learners face [1,11]. To

support learners of data modeling in devising meaningful identifiers for model elements, we design and implement the Automated Assistant, a research prototype that provides modeling-time feedback to data modelers on their choice of model element identifiers based on the natural language description of the respective modeling universe of discourse (cf. [14]).

The primary use case of the Automated Assistant is a learning context in which learners of data modeling are asked to create an ER diagram based on a textual description of a modeling task—a common learning scenario in many settings. Research, design and prototype implementation of the Automated Assistant combine and build on research on data modeling heuristics for identifier formulation from natural language descriptions [2,15] and on research on Natural Language Processing (NLP) [3,4]. Different from prior work (e. g.[9,15,5,8]), the Automated Assistant does not require the time-consuming ex ante construction of (multiple) reference solutions to a specific modeling task description but works on (m)any natural language descriptions of a modeling task (as long as they comply with English grammar).

The Automated Assistant prototype described in the present work substantially extends an earlier prototype implementation (cf. [14]). Specifically, we substantially extend the number of implemented data modeling heuristics for better detection of compound nouns signifying entity types as well as for detecting associated attributes, and we enhance the earlier prototype implementation by an identification component for data types of attributes. Moreover, the prototype implementation now presents generated candidates for element identifiers color highlighted to the modeler.

2 The Automated Assistant Prototype Overview

The Automated Assistant prototype implementation builds on TOOL, a modeling tool web application [13]. TOOL provides a graphical data modeling editor which implements a variant of the Entity-Relationship Model [10] to create ER diagrams. The Automated Assistant hooks into the graphical editor to assist data modelers at modeling-time with suggestions on signifying element identifiers. The Automated Assistant implementation utilizes the Stanford CoreNLP toolkit [6] which provides a pipeline framework for deriving linguistic annotations from natural language text input, in combination with heuristics rules for data modeling that we adapt and refine to the in-browser ER modeling variant. Moreover, the Automated Assistant implements widely accepted naming conventions for ER modeling (e. g., [2]).

In a nutshell, the prototype implementation of the Automated Assistant essentially applies a two-step procedure: First, starting from a (1) (natural language) modeling task description, a list of potential candidates for meaningful identifiers for entity types, associated attributes, and relationship types is generated. Then, based on the results of step (1), feedback (2) is provided to modelers at modeling-time in the graphical modeling editor on their choice of model element identifiers and data types—potential candidates for meaningful entity

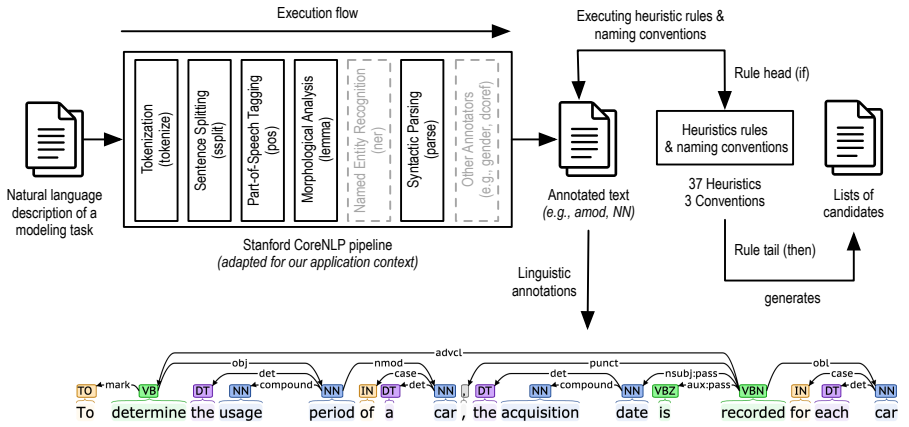

Fig. 1. Overview of the adapted annotation pipeline to generate a list of candidates for entity types, associated attributes, and relationship types (exemplary linguistic annotations generated using the the Stanford CoreNLP toolkit at https://corenlp.run).

type identifiers as well as corresponding attributes, and data types are color highlighted in the textual description displayed to the modelers. To generate a list of potential candidates based on natural language description (cf. step 1), the pipeline of the CoreNLP toolkit is adapted to the primary use case of the Automated Assistant (see Fig. 2), and enhanced in several design iterations by conscientiously revising the applied heuristic rules for classifying certain patterns of statements of the English language for identifying identifiers of model elements (e. g., [2]). The identification of potential candidates mainly relies on part-of-speech tagging and statistical dependency parsing (e. g., words which modify nominal phrases—adjectival modifiers) and to establish (grammatical) relations between words (e. g., by typed dependencies). The Automated Assistant utilizes multiple NLP techniques (built into the CoreNLP implementation) for the analysis of modeling task descriptions in a specific order: tokenization, sentence splitting, part-of-speech tagging, morphological analysis, and syntactic parsing (see Fig. 2). The subsequent identification of candidates for identifiers for entity types, associated attributes, and relationship types builds on linguistic annotations returned from CoreNLP as a semantic graph consisting of lists of tuples. Starting from these linguistic annotations, the revised prototype implementation applies 37 heuristic rules (13 on relationship types, 13 on entity types, and 11 on associated attributes) and 3 naming conventions for formulating model element identifiers to match them with the syntactical functions of the annotation pipeline. The heuristic rules and naming conventions are reconstructed, adapted and refined from a comprehensive review of prior work (the adapted annotation pipeline is described in further detail in [14]).

From a modeler's perspective, the Automated Assistant suggests feedback (cf. step 2), i. e., whenever a modeler devises identifiers of a model elements on the modeling canvas of the graphical modeling editor (see Fig. 2). The provided feedback aims at supporting modelers in developing an understanding of how to identify and signify model elements meaningfully, and, in particular, to encourage modelers to rethink their choices for model element identifiers based on sensible auto-generated suggestions of the Automated Assistant. Hence, the Automated Assistant provides feedback in three categories: (a) 'Great' for positive feedback, e. g., if a label matches an entity type of the generated candidate list (in green); (b) 'Reconsider' for a label that is not mentioned in the modeling task description or that is not identified as a possible candidate for an entity type, attribute, or relationship type (problem-oriented feedback; in orange); (c) 'Convention' if the input does not follow the implemented naming conventions, e. g., if an entity type label begins with a lowercase letter, or a spelling error is examined (neutral feedback; in yellow). As a major usability enhancement to an earlier version of the Automatic Assistant (cf. [14]), we have further extended the integrated modeling tool support to color highlight possible entity types and associated attributes in the textual description, e. g., when there is uncertainty or when the modeler wants to compare them to the chosen model element identifiers on the modeling canvas (see Fig. 2, right side). Compared to earlier prototype implementations, the Automated Assistant now provides more accurate feedback on data types of attributes based on named entity recognition. A short video demonstrator of the Automated Assistant is available at: https://video.fernuni-hagen.de/Play/896

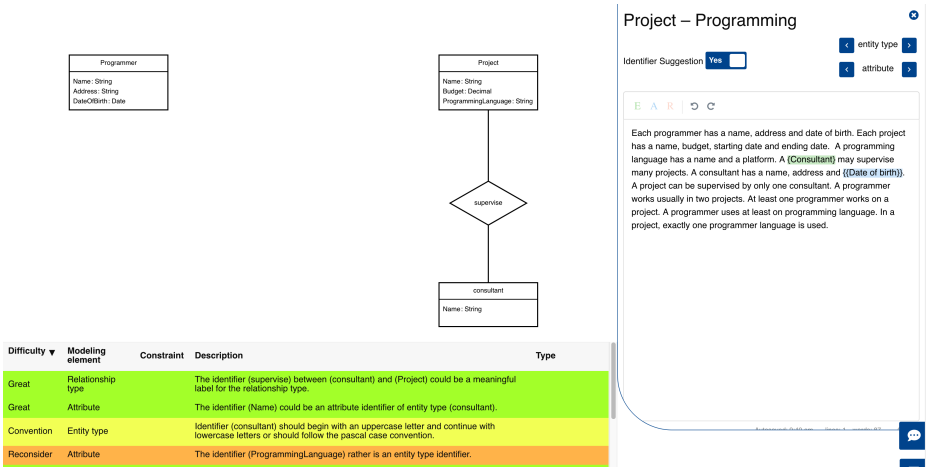

Fig. 2. Feedback on the modeler's choice of model element identifiers ('consultant', 'ProgrammingLanguage', etc.) provided by the Automated Assistant (highlighted in color below ER diagram and on the right side in the task description) based on the processing of the natural language description of the task shown on the right by the Automated Assistant.

3 Discussion and Outlook

The Automated Assistant provides modeling-time feedback to modelers on their choice of model element identifiers based on natural language processing of modeling task descriptions to assist in overcoming one of the most common (learning) difficulties data modelers face, i. e., devising meaningful and expedient identifiers for model elements [1,10]. Evaluation of the feedback generated by the Automated Assistant from three modeling task descriptions of increasing length and complexity suggests that the generated feedback is similar to human (e. g., instructor's) advice, and helpful to learner of data modeling.

Compared to earlier prototype implementations (see [14]), precision and recall for five demonstration cases, i. e., modeling tasks used in teaching conceptual data modeling, improved in recognizing entity type and relationship type identifiers from the text descriptions by an average of 0.045 with the largest increase in attribute identification (entity types: precision 0.88 (from 0.84), recall 0.97 (from 0.94); attributes: precision 0.86 (from 0.79) 0.86, recall 0.79 (from 0.71); relationship types: precision 0.84 (from 0.84), recall 0.74 (from 0.69)). The newly added identification of data types achieves a precision of 1.00 and a recall of 0.66. In terms of precision and recall, the current prototype implementation performs on par or better than related approaches (e. g., [5,9]), especially with respect to finding candidates for entity type identifiers, with prior research demonstrating values for recall from 0.92 [9] to 0.95 [8]. Closer inspection of precision and recall results show, as expected, the heuristic nature of the rules applied to recognize relationship types, entity types, attributes, and corresponding data types—and of the NLP techniques we use. Still, compound nouns with three (and more) words such as 'social security number' are, for example, not always properly recognized—a limitation we have already addressed by revising the implemented data modeling heuristics for entity types and attributes but which needs further work. For example, further prototype evaluation exemplifies that the identification of inherent relationships of compound nouns should be improved posing a particular challenge as there are many ways to combine them (see [5]). Moreover, we are currently implementing heuristic rules for identifying generalization hierarchies in textual descriptions to provide data modelers with automated assistance in modeling generalization relationships. Future work on the Automated Assistant will also address the identification and recognition of synonyms in the textual description by utilizing word embedding approaches such as WordNet.

We are currently planning for further systematic evaluation by empirical studies investigating how well the feedback provided by the Automated Assistant supports learners of data modeling in devising meaningful and appropriate identifiers for model elements. Building on prior mixed-methods research into data modeling processes [10,11,13], modelers will be observed from complementary perspectives, e. g., based on the think-aloud method and surveying modelers on their perception of the feedback as provided by the Automated Assistant. Subsequently, we are planning to apply the Automated Assistant in an introductory course on data modeling with 200+ students per semester to collect feedback on the support provided by the Automated Assistant.

References

1. Batra, D., Hoffer, J.A., Bostrom, R.P.: Comparing representations with relational and EER models. Communications of the ACM **33**(2), 126–139 (1990)
2. Chen, P.P.: English Sentence Structure and Entity-Relationship Diagrams. Information Sciences **29**, 127–149 (1983)
3. Chowdhury, G.: Natural language processing. Annual review of Information science and technology **37**, 51–89 (2003)
4. Liddy, E.D.: Enhanced Text Retrieval Using Natural Language Processing. Bulletin of the American Society for Information Science and Technology **24**(4), 14–16 (2005)
5. Lucassen, G., Robeer, M., Dalpiaz, F., van der Werf, J.M.E., Brinkkemper, S.: Extracting conceptual models from user stories with Visual Narrator. Requirements Engineering **22**(3), 339–358 (2017)
6. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. Baltimore, Maryland (2014)
7. Mendling, J., Reijers, H.A., Recker, J.: Activity labeling in process modeling: Empirical insights and recommendations. Information Systems **35**(4), 467–482 (2010)
8. Omar, N., Hanna, P., Mc Kevitt, P.: Heuristics-based entity-relationship modelling through natural language processing. In: 15th Artificial Intelligence and Cognitive Science Conference. pp. 302–313. Dublin, Ireland (2004)
9. Omer, M., Wilson, D.: Implementing a Database from a Requirement Specification. International Journal of Computer and Information Engineering **9**(1), 33–41 (2015)
10. Rosenthal, K., Strecker, S.: Toward a taxonomy of modeling difficulties: A multimodal study on individual modeling processes. In: 40th International Conference on Information Systems, ICIS 2019, Munich, Germany, December 15–18, 2019 (2019)
11. Rosenthal, K., Strecker, S., Pastor, O.: Modeling difficulties in data modeling: Similarities and differences between experienced and non-experienced modelers. In: 39th International Conference on Conceptual Modeling, ER 2020, Vienna, Austria, November 3–6, 2020. LNCS, vol 12400. pp. 501–511 (2020)
12. Speaks, J.: Theories of Meaning. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2019 edn. (2019)
13. Strecker, S., Rosenthal, K., Ternes, B.: Studying Conceptual Modeling Processes: A Modeling Tool, Research Observatory, and Multimodal Observation Setup. In: Gimpel, H., Krämer, J., Neumann, D., Pfeiffer, J., Seifert, S., Teubner, T., Veit, D., Weidlich, A. (eds.) Market Engineering – Insights from Two Decades of Research, pp. 99–111. Springer, Cham (2021)
14. Ternes, B., Rosenthal, K., Strecker, S.: Automated Assistance for Data Modelers: A Heuristics-based Natural Language Processing Approach. In: 29th European Conference on Information Systems, ECIS 2021, Marrakech, Morocco, June 15–17, 2021 (2021)
15. Tjoa, A.M., Berger, L.: Transformation of requirement specifications expressed in natural language into an EER model. In: Elmasri, R., Kouramajian, V., Thalheim, B. (eds.) Proceedings of the 12th International Conference on the Entity-Relationship Approach, ER 1993, Arlington, Texas, USA, December 15–17, 1993. LNCS, vol 823. pp. 206–217 (1994)