Natural Language Generation from SNOMED Specifications

Mattias Kanhov*, Xuefeng Feng, and Hercules Dalianis

Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden
Address for Correspondence: *kanhov@dsv.su.se

Abstract. SNOMED (Systematized Nomenclature of Medicine) is a comprehensive clinical terminology that contains almost 400,000 concepts, since SNOMED is a formal language; it is hard to understand for users who are not acquainted with the formal specifications. Natural language generation (NLG) is a technique utilizing computers to create natural language descriptions from formal languages. In order to generate descriptions of SNOMED concepts, two NLG tools were implemented for the English and Swedish version of SNOMED respectively. The one for English used a natural language generator called ASTROGEN to produce description texts. This tool also applied several aggregation rules to make the texts shorter and easier to understand. The other tool used C#.Net as the programming language and applied a template-base generation technique to create concepts explanation in Swedish. As a base line same SNOMED concepts were presented in a tree structure browser.

To evaluate the English NLG system, 19 SNOMED concepts were randomly chosen for the generation of text. Ten volunteers participated in this evaluation. Five of them estimated the accuracy of the texts and others assessed the fluency aspect. The sample texts got a mean score 4.37 for accuracy and 4.47 for fluency (max 5 score).

To evaluate the Swedish NLG system, five concepts were randomly chosen for the generation of texts. In parallel two physicians with knowledge in SNOMED created manually natural language descriptions of the same concepts. Both manual and system generated natural language descriptions were evaluated and compared by in total four physicians. All respondents scored the manual natural language descriptions the highest in average 83 of 100 scores while the system generated natural language texts obtained around 68 of 100 scores. All three respondents unanimously except one respondent (scoring 7 of 10) preferred the system-generated text.

This paper presents a possible way using Natural Language Generation to explain the meaning of SNOMED concepts for people who are not familiar with SNOMED formal language. The evaluation results indicate that the NLG techniques can be used to implement this task.

Keywords: English; Evaluation; Formal Specification; Natural Language Generation; SNOMED; Swedish

Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth 2012 Workshop, edited by Hanna Suominen

1 Introduction

For hundreds of years, physicians and health personnel have used Latin and Greek terms exclusively when describing symptoms, diseases and body parts; but today more and more words from local languages are used. Therefore SNOMED CT,[1] containing almost 400,000 concepts, has been defined as a lingua franca in medicine. However, SNOMED is a formal language that is difficult to understand for persons not trained in formal specifications. Natural language generation (NLG) is the technique whereby we let a computer generate a piece of natural language text describing some artefacts or events. NLG uses the same techniques as a human would use to produce text.[2] Early work in generating natural language from formal specification was carried out by Black[3] and Rolland & Proix.[4] Liang et al.[5] used an ontology verbaliser to generate natural language expression for SNOMED concepts in both English and Chinese. How can we make the medical concepts expressed in SNOMED useful for medical domain experts such as nurses and physicians who are not familiar with the SNOMED formalism? Can one use automatic NLG to describe SNOMED?

2 Materials and Methods

SNOMED was first created for English but is today also available in languages such as Spanish, French, Dutch, Danish and Swedish. Dalianis[6] described a method to validate formal specifications using NLG and specifically using aggregation to compact the text. CliniClue Xplore[7] can be used for browsing the SNOMED terminology in a tree structure. We have constructed two NLG tools for English SNOMED and for Swedish SNOMED, respectively. The English one uses a natural language generator called ASTROGEN[6] to generate the description for disease concepts in SNOMED. Several aggregation rules can be applied in this tool, including syntactic aggregation, bounded lexical aggregation and unbounded lexical aggregation. By applying the aggregation rules, the descriptive texts are much shorter and easier to read. The Swedish one is an application written in C# .NET which uses a template-based generation process to produce a short description of diseases and disorder. The system utilizes syntactic aggregation in order to make the text more compact while still conveying the medical content to the user. Both systems use an interface where the user enters a SNOMED concept and then asks the system to generate a natural language description of it (see Figures 1 and 2).

3 Results

The evaluation of the Swedish NLG system was performed by randomly choosing five diseases and letting the system generate sample descriptions for these diseases. We used four physicians as respondents. To have reference descriptions to compare with the generated texts from the NLG system, two of the physicians (who had knowledge in SNOMED) were asked to use the CliniClue Xplore printout for the same five concepts (the same input the NLG system uses) and write manual descriptions using these data. They then evaluated each other's texts together with the system

generated texts and were asked about the content, language and usefulness for the disease or disorder in the texts. They also evaluated the content of SNOMED by viewing CliniClue Xplore printouts in order to find out whether there was enough information to explain a disease or disorder fully and to see what type of presentation the respondents preferred. All respondents scored the manual descriptions the highest in average 83 of 100 scores while the system-generated texts obtained around 68 of 100 scores. All three respondents unanimously except one respondent (scoring 7 of 10) preferred the system-generated texts to the CliniClue Xplore printouts for explaining disease concepts.

The common cold has the causative agent which is/are the virus.
The common cold is/has the courses.
The common cold has the episodicities.
The common cold has finding site of the upper respiratory tract structure.
The common cold is a kind of the viral upper respiratory tract infection.
The common cold has the pathological process which is/are the infectious process.
The common cold has the severity level as the severities.
The common cold is also called the acute coryza, the acute infective rhinitis, the acute nasal catarrh, the acute nasopharyngitis, the acute nasopharyngitis, nos, the acute rhinitis, the cold, the head cold, the infective nasopharyngitis, the infective nasopharyngitis, nos and the infective rhinitis.

Fig. 1. English NLG using aggregation from the SNOMED concept of the common cold

Förkylning är en virusinfektion i övre luftvägarna. Orsaken till sjukdomen är virus. Sjukdomen finns i övre andningsvägar.

Fig. 2. Swedish NLG using aggregation from the SNOMED concept of the common cold

4 Discussion

The respondents would like not to have the English explanations marked in red in Figure 1, because they thought that those explanations were not helpful for understanding the SNOMED concepts. The Swedish NLG did not generate these explanations. Moreover, the Swedish SNOMED does not contain any synonyms, therefore the Swedish NLG is also shorter than the English NLG. Liang et al.[8] applied Rhetorical Structure Theory (RST) to structure the relationships of SNOMED, and created paragraphs about the contents of SNOMED in natural language. Sundvall et al.[9] has constructed a tool called TermViz as an add-on to CliniClue to visualize SNOMED CT but it is customized for a certain domain. However, no one has evaluated their results.

5 Conclusion

This research supplies a possible method (natural language generation) to express the concepts of SNOMED in English and Swedish for health-care professionals who do not know the SNOMED formalism. According to the evaluation results, the explana-

tions of concepts in SNOMED generated by using NLG techniques were readable and useful for understanding the concepts.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments and suggestions.

References

1. IHTSDO, 2010. SNOMED CT User Guide 2010, <http://www.ihtsdo.org.>[Accessed 15 May 2012].
2. Reiter, E. and Dale, R., 2000. Building natural-language generation systems, Cambridge University Press.
3. Black, W. J., 1987. Acquisition of conceptual data models from natural language descriptions. Proceedings of the 3rd Conference on European Chapter of the Association for Computational Linguistics, (ACL), pp. 241-248.
4. Rolland, C. and Proix, C., 1992. A natural language approach for requirements engineering. Advanced Information Systems Engineering, Springer, pp. 257-277.
5. Liang, S.F., Stevens, R. and Rector, A., 2011. OntoVerbal-M: a multilingual verbaliser for SNOMED CT. Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011) in conjunction with the International Semantic Web Conference (ISWC 2011), pp. 13-24.
6. Dalianis, H. 1999. Aggregation in Natural Language Generation, Journal of Computational Intelligence, Volume 15, Number 4, pp. 384-414, November 1999.
7. CliniClue, 2011. The Clinical Information Consultancy. CliniClue Xplore. Available at: `http://www.cliniclue.com/software` [Accessed 9 April 2012].
8. Liang, S.F., Scott, D., Stevens, R. and Rector, A., 2011. Unlocking medical ontologies for non-ontology experts. Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pp. 174-181.
9. Sundvall, E., Nyström, M., Petersson, H. and Åhlfeldt, H., 2006. Interactive visualization and navigation of complex terminology systems, exemplified by SNOMED CT. Proceedings of Medical Informatics Europe (MIE), 27-30 Aug, pp. 851-856.