Extraction of Specific Nursing Terms Using Corpora Comparison

Guoqian Jiang, MD, Ph.D., Hitomi Sato, RN, Akira Endoh, Ph.D., Katsuhiko Ogasawara, Ph.D., Tsunetaro Sakurai, MD, Ph.D.
Department of Medical Informatics, Hokkaido University Graduate School of Medicine, Sapporo, Japan

Abstract

The purpose of the study is to develop and evaluate a bottom-up approach to extract the specific nursing terms from textual nursing records using corpora comparison. A nursing records corpus was developed as the target corpus, and a newspaper corpus and a medical literature abstracts corpus were developed as the reference corpora. Two filters were established to extract the technical terms and the relative frequency ratio was used for corpora comparison. The issues related to the improvement of both the algorithms and the evaluation methods were discussed.

Introduction

With the increasing computerization of nursing records in Japan, the standardization issues of nursing terminology are becoming urgent. A national standardized classification of nursing terms (named Nursing Master) has been released by the MEDIS-DC since November, 2003. In current stage, the Nursing Master contains only nursing action terms with a simple 4-layer classification. In addition, a Japanese version (beta 2) of ICNP was made and released by Japan Nursing Association since 2002. The ICNP is a classification of nursing phenomena, actions and outcomes, and provides a terminology for nursing practice that serves as a unifying framework into which existing nursing vocabularies and classifications can be cross-mapped to enable comparison of nursing data. The common feature of these classifications could be considered as that they are all developed using a top-down approach [1]. Ideally, the development of terminology system requires a close coordination between bottom-up and top-down approaches as bottom-up approaches help the expansion of the content within nursing terminology system [2]. The purpose of the study is to develop and evaluate a bottom-up approach to extract the specific nursing terms from textual nursing records using corpora comparison.

Methods

A nursing corpus containing 15 months of nursing records collected from nursing information system in Hokkaido University hospital was developed as the target corpus. A newspaper corpus was used as the reference corpus, and a medical corpus containing 52,144 textual abstracts retrieved from a Japanese literature database was developed as another reference corpus. The plain texts of each corpus were all tagged with part of speech using a Japanese morphological analysis system named ChaSen. Justeson and Katz' part-of-speech filter was used to extract the technical terms and a first-word filter was developed to exclude the non-terms. And the relative frequency ratio with the heuristic cutoff values was used for corpora comparison. A validation was taken by matching the extracted terms with the medical modifiers. The precision was measured in two ways, i.e. an automatic way through matching the extracted terms with the terms in ICNP, and an expert-based way through using the knowledge of 3 expert nurses. The recall was not measured in this study.

Results & Discussions

The nursing corpus was composed of 5,193,662 tokens with 19,801 token types, whereas for the newspaper corpus was composed of 13,672,008 tokens with 118,240 token types and for the medical corpus 10,727,352 tokens with 45,612 token types. Using the Justeson and Katz' part-of-speech filter, 1,341,547 terms were extracted from both the nursing and newspaper corpora (newspaper group) and 1,050,289 from both the nursing and medical corpora (medical group). Using the relative frequency ratio with the heuristic cutoff values and the first-word filter, 9,326 terms were extracted from the newspaper group and 7,590 from the medical group. For the validation process, we compared the matched number with 1807 medical modifiers between the two groups. The matched number (329) in the newspaper group was significantly higher than that (173) in the medical group.

5,800 terms were common to both groups. We matched the common terms with all terms in ICNP and only about 5% (280) of the common terms were matched. The expert-based precision (n=100) was calculated as 78.2%.

Conclusion

The corpora comparison is useful to extract the specific terms in nursing domain that possibly expanding the contents of current nursing terminology system (e.g. ICNP). The further studies should focus on the refinement of the terminology extraction tool and the development of the effective methods for evaluation.

References

1. Jiang G, Sato H, Endoh A, Ogasawara K, Sakurai T. Developing a support tool for describing the nursing practice in Japan with ICNP using Protégé-2000. Proceedings of Joint Conference of Medical Informatics in Japan. 2004; 24(suppl.): 1204-1205.
2. Harris MR, Savova GK, Johnson TM, Chute CG. A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. J Biomed Inform. 2003;36: 250-259.