MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Challenges for Integrating Korean Medical Vocabulary into the UMLS

[a]Jinwook Choi M.D. PhD., [a]Seung-Bin Han, [a]Seunghee Kim, [b]William T. Hole, M.D., [b]Suresh Srinivasan

[a]Department of Biomedical Engineering, College of Medicine, Seoul National University, [b]National Library of Medicine, Bethesda, MD, USA

Abstract

We present the problems we faced and their solutions when we tried to add Korean terms to the Metathesaurus of the Unified Medical Language System (UMLS MT). Challenges to integrating Korean into the UMLS MT are analyzed; such as character set incompatibility, incomplete translation; matching UMLS terms and concepts. Suggestions are given to improve the integration of Korean terminology into the UMLS. Our work is expected to be a useful precedent to merge medical vocabularies in other languages into the UMLS MT.

Keywords:

Unified Medical Language System, Korean language

Introduction

Multilinguality is a major issue for the international use of medical terminologies since it is a way to share and reuse knowledge. The goal of National Library of Medicine's long-term UMLS® project is to help health professionals and researchers use biomedical information from different sources, to facilitate information sharing and to enable accurate transmission of medical data or knowledge [1].

Challenges

We present an analysis of some of the problems encountered in the UMLS when integrating languages other than English especially the Asian languages. The following problems currently limit integration and use of the UMLS in Korean.

Character set incompatibility

The character set used to represent characters in UMLS terms is 7-bit ASCII, while the Korean character code systems are based on a 2-byte code systems (DBCS). It is used all 16-bits to represent Korean characters in Hangul. We use the EUC-KR as the character set. Ways of converting code sets to 7-bit are complicated and it might be very difficult and inefficient to map each character in the KSC 5601 to a 7-bit ASCII character. Unicode is a 16-bit character set, which represents virtually all existing character sets. Unicode 3.0[2] has 11,172 Korean characters that are used in most Korean OS systems. A truly multilingual UMLS would be possible by introducing Unicode. However, we must note that not all computer systems or applications currently support Unicode characters.

Quantitative and qualitative issues related to translation

While UMLS concepts come from more than 100 vocabularies, the UMLS MT has not any Korean terms. MeSH, the Medical Subject Headings, is a comprehensive medical thesaurus which indexes documents in the MEDLINE database [3]. Korean librarians have attempted to translate the MeSH since 1995. However, the translations were incomplete because only some Korean terms were included, and it was just translated into a Korean dictionary. A future translation should be done in collaboration with the NLM [4].

Approach determination for integration

In order to determine what approach is appropriate to integrate Korean terms in the UMLS MT between the concept and term levels, we have evaluated coverage of the UMLS as compared with Korean medical concepts, and to identify differences. Then, we identified the best way of integrating the Korean medical vocabulary into the UMLS MT at the term level.

Adding Korean terms in the UMLS Metathesaurus

The easiest way to rapidly add Korean terms to the UMLS is to merge the Korean translation of a vocabulary which already exists in the MT in English. Unlike merging a new English vocabulary, integrating a translation of a vocabulary already in the MT is usually less challenging than integrating a new terminology because it is integration at the term level and not at the concept level. As first step in integrating Korean medical vocabulary into the UMLS MT, the Korean version of ICD-10 (KCD) was chosen since the KCD has the very same structure as the English one. In this manner, unique concept identifiers of UMLS were given to more than 22,000 Korean terms of KCD.

Conclusions

In order to offer reasonable coverage of the medical domain, more vocabularies such as MeSH and ICD-9-CM should be translated into Korean. Once completed, adding new Korean terms to already translated concepts would better reflect the diversity of the biomedical language.

References

[1] UMLS Knowledge Sources. (14th ed.) Bethesda (MD): National Library of Medicine, 2003AA

[2] The Unicode Consortium. The Unicode Standard, Version 3.0 www.unicode.org

[3] Nelson S, Olson N, Fuller L, Tuttle M, Cole W, Sherertz D. Identifying concepts in medical knowledge. Medinfo 1995:33-6.

[4] Nelson, Stuart J.; Schopen, Michael; Schulman, Jacque-Lynne; Arluk, Natalie, An Interlingual Database of MeSH Translations. 8th International Conference on Medical Librarianship; 2000 Jul 4; London, UK.

Address for correspondence E-mail: jinchoi@snu.ac.kr