Integrating Korean Medical Vocabulary into the UMLS: The Challenge and Process

[a]Seung-Bin Han, [a]Miyoung Kwak, [a]Boyoung Kim, [a]Jinwook Choi, [b]Seungsik Kang, [c]William T. Hole, [a]Suresh Srinivasan

[a]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, Korea; [b]Division of Computer Science & Engineering, Kookmin University, Seoul, Korea;
[c]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 Metathesarus of the Unified Medical Language System (UMLS MT). Challenges to integrating Korean into the UMLS MT are analyzed, such as matching UMLS terms and concepts, incomplete translation, character set incompatibility, and absence of lexical resources to create Korean indexes. 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:
UMLS Metathesaurus (MT),Korean medical vocabulary, ICD-10, KCD, Unicode

Introduction

Multilinguality is a major issue for the international use of medical terminologies since it is a way to share and reuse knowledge. However, the use of medical knowledge does not always require language-dependent support. Multilinguality is a common feature of all major medical terminologies (UMLS, SNOMED International, GALEN, ICD, ICPC). Although all of these terminologies have a language independent underlying model, only a few of them offer real multilingual support [1]. The most translated terminologies are the ICDs; ICD-10 [2] has been translated into more than twenty languages. GALEN, the youngest of the medical terminologies, is multilingual by design and currently provides seven European languages (French, Italian, Dutch, German, Finnish, Swedish and English)[3]. SNOMED International and portions of the UMLS are available in an increasing number of languages. We believe no terminology yet offers broad coverage, a strong underlying model and multilingual support across character sets [4]. Controlled terminology is key to the utility of Clinical Patient Record (CPR) systems [5-7]. Despite dramatic developments in medical knowledge and IT technologies in Korea, no Korean language is contained in medical terminology systems, except for ICD. We have therefore made a special effort to include the Korean language (Hangul) in the UMLS Metathesaurus (MT). In this paper, we present the process of integration, problems faced, and solutions for inclusion of Korean into the UMLS.

Background

The goal of NLM's long-term Unified Medical Language System® (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 [8].
There are three major components in the UMLS Knowledge Sources:

1. The UMLS Metathesaurus® (MT) contains information about biomedical concepts and terms from the many controlled vocabularies and classifications used for patient records, administrative health data, bibliographic and full-text databases, and expert systems. It is organized by concept or meaning. The 2003AA edition of the MT includes 875,255 concepts and 2.14 million concept names in its source vocabularies.

Two important tables in the MT are used for integration. The MRCON table is composed of several fields. The Concept Unique Identifier or CUI is one key identifier. CUIs connect all names for a given concept, and are used to link concepts in relationships, and identify all attributes of a concept. Lexical Unique Identifiers (LUIs), identify and link all concept names that are lexical variants of each other, and String Unique Identifiers (SUIs), identify and link all identical strings. The string field, STR is used for storing the concepts names. The MRSO table has a source abbreviation field (SAB) and a CODE field for each entry in the source vocabulary and shares the three unique identifiers in MRCON.

2. The Semantic Network, through its semantic types, provides a consistent categorization of all concepts represented in the MT. The links between the semantic types provide the structure for the Network and represent important relationships in the biomedical domain.
3. The SPECIALIST Lexicon (SL) is an English language lexicon with many biomedical terms. Information for each entry includes base form, spelling variants, syntactic category, inflectional variation of nouns and the conjugation of verbs. This information is used by the lexical tools. Lexical Unique Identifiers (LUIs) are defined in English by the 'luinorm' flow through a SPECIALIST lexical program called lvg. The tools also identify words and strings used in the Metathesaurus

indexes.

It would greatly assist research of Korean medical informatics if the UMLS contains a Korean word index table in Korean characters, in addition to the existing 15 indexes. This addition will require a Korean lexicon and programs.

Analysis of 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:

What approach is appropriate for adding Korean terms to the UMLS MT: concept level & term level

Since the UMLS MT is composed of concepts and interconcept relationships, in principle, it is a language-independent representation of medical knowledge. However, words are used to name concepts, and, at the term level, the MT is language-dependent. Therefore, we need to evaluate coverage of the UMLS as compared with Korean medical concepts, and to identify differences. Also, we must identify the best way of integrating the Korean medical vocabulary into the UMLS MT at the term level.

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 [9]. MeSH is a valuable component of the UMLS MT, and many MeSH translations appear in the MT. . 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 [10]. A future translation should be done in collaboration with the NLM [11].

Character set incompatibility

The proper representation of Korean characters in MT files is a major issue. The character set used to represent characters in UMLS terms is 7-bit ASCII including transliterated 8-bit characters from a number of Western European languages.
Clearly a Unicode representation of these 8-bit characters is preferable, and it is essential for Korean characters which are very different from western alphabetic characters. The Korean character code systems are based on a 2-byte code systems (DBCS). They use all 16-bits to represent Korean characters in Hangul. Previously, we have used the EUC-KR as the character set. This is compatible with the KSC5601-1987 and the ASCII code system. 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. Finally, retrieval and searching across character sets is a difficult problem which will be solved with Unicode [12] and character equivalence tables. Character equivalence tables have yet to be built. Further, Unicode must be implemented for all users of differing operating systems. Such support is still evolving.

Lexical tools to create the Korean indexes

Lexical matching techniques require lexical items to be identified and transformed into their base form. Then, derived forms can be computed. LVG computes inflectional and derivational variants by applying a set of rules and facts to the base forms [13]. These lexical resources do not currently exist for Korean in the UMLS. Natural language processing also requires other lexical tools like a stemmer and a lexical analyzer, which are language dependent. Korean language and its rules and facts for variant generation are complicated and very different from other alphabet based languages, so we must develop lexical tools to create the Korean indexes and to assign LUIs to terms in our vocabularies.

Possible Solutions and the integration Process

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 estimated the concept coverage and capability of the UMLS as proposed for its application into the CPR system in Korea. As a preliminary study, we examined how many concepts in the current UMLS can cover the existing clinical terms presently used in the Korean medical records. To do so, we tested the mapping between ‘chief complaints’ extracted from the discharge records of Seoul National University Hospital (SNUH) and the UMLS ‘Sign or Symptom’ concept terms. Thirty-five percent of chief complaints of the SNUH were conceptually matched with the UMLS ‘Sign or Symptom’ concepts. Most of fifth-eight percent were matched with ‘Disease or Syndrome’ concepts of UMLS rather than matched with ‘Sign or Symptom’ concepts of UMLS MT. Since these terms are representing such that diagnosis, operation names, clinical laboratory test names, not Sign or Symptom related terms. The rest terms of seven percent were not found in the UMLS MT or those terms that used to special circumstances of a tertiary hospital in Korea.

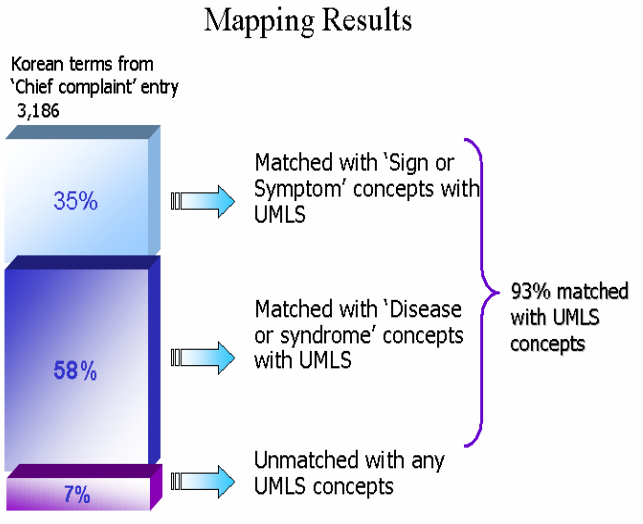

Figure 1 - Concept coverage of the UMLS compared with Korean medical concepts

93% of the Korean medical concepts were covered by UMLS. 7% of unmatched concepts were caused by special circumstances in the Korean tertiary system. (Figure 1) We presented an analysis of some of the differences between the two terminology sets and the issues requiring consideration for integration in a study presented at 2002 AMIA annual conference. [14, 15] Finally, we decided to process the Korean vocabularies into UMLS MT at the term level rather than at the concept level rather than to add or create new concepts in MT.

Adding Korean terms in the UMLS MT

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. One simple solution would use ICD codes for the mapping. ICD codes (as well as the codes in source vocabularies in general) are recorded in the MT in the MRSO.CODE field. For example, the ICD code for Addison's disease is E27.1. Searching for 'E27.1' in MRSO retrieves several strings, all attached to the same concept:

C0001403|L0001403|S1921523|MTHICD9|ET|255.4|0|
C0001403|L0494940|S0718028|ICD10AM|PT|E27.1|3|
C0001403|L0494940|S0718028|**ICD10**|PT|**E27.1**|3|
C0001403|L1288823|S1530769|DMDICD|PT|E27.1|1|

The key combining the first three fields (CUI, LUI and SUI) can be used to join with the MRCON table, for example:

C0001403|ENG|S|L0494940|PF|S0718028|Primary adrenocortical insufficiency|3|
C0001403|GER|S|L1288823|PF|S1530769|Primaere Nebennierenrindeninsuffizienz|1|

If the Korean string for 'Primary adrenocortical insufficiency' is

원발성 부신피질 기능부전, and its code is still

E27.1, new term and string identifiers need to be created for this string, but no concept identifier, since the concept already exists. The result would be like follows:

MRSO → C0001403|**L9999999**|**S9999999**|**KCD3**|PT|
E27.1|3|
MRCON →C0001403|KOR|S|L9999999|PF|S9999999|

원발성 부신피질 기능부전 |3|

We have received the ranges for the SUIs and LUIs for Korean strings from the NLM. There is room for 100,000 of each: SUI range: S + 7 digits ranging from 9900000 - 9999999, i.e. S9900000 - S9999999, and similar LUI ranges. We are free to assign our own unique identifiers SUI, LUI for Korean language, and link them to the universal Concept Unique Identifiers, CUIs. At present, we believe that most of KCD3 terms are already in normal form, so we just assigned LUIs by copying the SUIs.

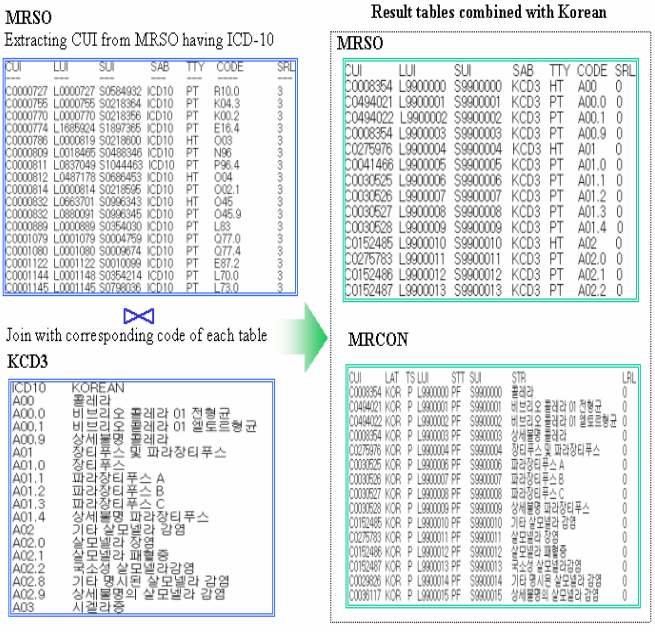

Figure 2- Table joining with ICD10 code and the result tables combined with Korean terms

To map KCD3 terms into ICD10 in MT, first, we extracted all CUI and source code CODE of ICD10 in MRSO, the datarows numbers 13,505. And the Korean medical vocabularies of KCD3 have 14,426 rows. Second, CUIs were given to each Korean strings of KCD3, by joining

ICD10 and KCD3 with same CODE. The resulting counts of this joining are 13,154. Third, we just gave the SUIs and LUIs numbers for each Korean strings according to alphabetical order of ICD10. Figure 2 shows a simplified implementation. In this manner, Korean strings would become synonyms of the corresponding concepts in the MT, and be stored, as all other strings in the MRCON and MRSO table.

Adapting Unicode system

Unicode is a 16-bit character set, which represents virtually all existing character sets. Looking toward the future, as the system becomes more international, the UMLS will need to evolve beyond the 7-bit ASCII character set. NLM anticipates being able to distribute the Metathesaurus in UTF-8 and perhaps later in UTF-16. Unicode UTF-8 represents the 16-bit characters in a sequence of bytes in a way that can be safely handled in all traditional file systems.
We are currently using a KSC5601-1987 character set and a Unicode 2.0 character set in multilingual software. The 11,172 characters are used in most Korean Windows OS systems. There are mapping tables from KSC5601-1992 and others to Unicode 3.0 at www.unicode.org, Unicode 3.0 has the same range as Unicode 2.0, without any changes. To validate the Unicode 3.0, we tested compatibility of character code in different systems. We sent sample files in Metathesaurus format saved as Unicode and UTF-8 to the NLM where it was determined that the files and characters were correctly represented. We must note that not all computer systems or applications currently support Unicode characters.

Adapting Unicode system

Unicode is a 16-bit character set, which represents virtually all existing character sets. Looking toward the future, as the system becomes more international, the UMLS will need to evolve beyond the 7-bit ASCII character set. NLM anticipates being able to distribute the Metathesaurus in UTF-8 and perhaps later in UTF-16. Unicode UTF-8 represents the 16-bit characters in a sequence of bytes in a way that can be safely handled in all traditional file systems.
We are currently using a KSC5601-1987 character set and a Unicode 2.0 character set in multilingual software. The 11,172 characters are used in most Korean Windows OS systems. There are mapping tables from KSC5601-1992 and others to Unicode 3.0 at www.unicode.org, Unicode 3.0 has the same range as Unicode 2.0, without any changes. To validate the Unicode 3.0, we tested compatibility of character code in different systems. We sent sample files in Metathesaurus format saved as Unicode and UTF-8 to the NLM where it was determined that the files and characters were correctly represented. We must note that not all computer systems or applications currently support Unicode characters.

Discussion

This paper described four major challenges and the integration process when we tried to add Korean terms into the UMLS MT. The process of integration and its possible solutions were based on the suggestions of UMLS staff of NLM through the exchange of much e-mail since 2001 and the meeting at the 2002 AMIA conference. Integration of Korean into UMLS MT is under way. Most of problems and solutions proposed in this paper could be applied to other non-English languages as well--especially Asian languages.
A truly multilingual UMLS would be possible by introducing Unicode. However, it may take time for every computer system to support Unicode.
We mapped KCD3 into the ICD10 in the MT at the term level. However, this may not be the final solution because KCD3 was recently substantially revised to KCD4. We may need to maintain both versions of KCDs in the MT in the future since most clinicians are more familiar with KCD3. We must work with the committee of Korean medical vocabularies to determine which version is preferred or which term is preferred.
In out initial effort, we avoided dealing with LUIs by copying the SUIs. The assigning of an LUI to each Korean medical term in the next version would be depends on the lexical sources for Korean. We have noted a useful lexical resource for analyzing Korean morphology called HAM [16]. The HAM, a morphological analyzer, was used to extract the indexing terms and to find out the lexical category and the base form of each lexical item.
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.
This is the first trial to integrate Korean language even among Asian countries that used different character sets. We hope our work is a useful precedent to merge other Korean medical vocabularies into the UMLS Metathesaurus and for others to integrate their own terminologies into international systems.

Acknowledgement

We gratefully acknowledge Dr. William T. Hole, Suresh Srinivasan and Olivier Bodenreider (UMLS staff, NLM) for supporting us sincerely. Our work is sponsored by the Ministry of Health and Welfare in Korea.

References

[1] Bodenreider O, McCray A. From French vocabulary to the UMLS: A preliminary study. Medinfo 98; 670-674
[2] International Classification of Diseases version 10, World Health Organization (WHO)
[3] www.opengalen.org/

[4] The Systematized Nomenclature of Medicine (SNOMED), College of American Pathologists (CAP), http://www.snomed.org/
[5] Chute C. et al., The content coverage of clinical classifications. JAMIA 1996; 3: 224-233
[6] Elkin P, Steven HB, Chute CG. Guideline for health informatics: Controlled health vocabularies-vocabulary structure and high-level indicators. Proc Medinfo 2001; 191-195
[7] Cimino JJ. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994; 1; 35-50
[8] UMLS Knowledge Sources. (14th ed.) Bethesda (MD): National Library of Medicine, 2003AA
[9] Nelson S, Olson N, Fuller L, Tuttle M, Cole W, Sherertz D. Identifying concepts in medical knowledge. Medinfo 1995:33-6.
[10] Korean MeSH, Medical research information center. www.medric.or.kr/
[11] 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.
[12] The Unicode Consortium. The Unicode Standard, Version 3.0 www.unicode.org
[13] McCray A, Srinivasan S, Browne A. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9
[14] SB Han, J Choi. A Comparative Study on the 'Sign or Symptom' Concepts of the UMLS and Clinical Terms in Korean Medical Records, JKOSMI 2001; 7; 4; 1-10
[15] SB Han, J Choi, et al. Analysis of discordance on the 'Sign or Symptom' concepts between the UMLS and clinical terms in Korean medical records. Proc AMIA Annu Fall Symp 2002: 1039
[16] S Kang, Analysis of Korean morphology and information retrieval. Hongrung Science 2002.

Address for Correspondence

Jinwook Choi, M.D., PhD. is an assistant professor of the Department of Biomedical engineering at Seoul National University School of Medicine. He is currently involved in international standardization of medical information for ISO/TC215. His major research interests are hospital information systems, medical terminologies, computerized patient record, HL7 and many other health related issues. Interested readers may contact the author, via either jinchoi@snu.ac.kr or Department of Medical Informatics at Seoul National University School of Medicine, 28Yongon-Dong Chongno-Gu , Seoul Korea, 110-799