Abstract: The NLM/AHCPR Large-Scale Vocabulary Test data set was used to investigate the nature of ad hoc relationships formed when users mapped broader to narrower terms as the closest available conceptual match. Among the 1,162 term pairs examined, the most common semantic differences were based on anatomical or functional distinctions between the terms. As expected, modification accounted for the majority of syntagmatic specifications characterized, but hierarchical specification accounted for over a fifth of the mappings.

1. Introduction

There is increasing interest in alignment and integration of knowledge structures, both within and among systems and subject domains. Although automated approaches can work well where source and target terms form close lexical matches, mapping concepts from one terminology to another remains largely a labor-intensive, cognitively demanding task, with results often far from perfect or even satisfying.

When no exact conceptual match exists for a source term in a target vocabulary, one must consider what constitutes the next best option, that is, which term is nearest or closest in meaning. Intuitively, it seems most natural to assess first the possibility of mapping the source term to a broader term. One might decide what classes the concept is most likely to be a member of and search for those classes. Alternatively, if the term is complex, one might consider which aspect of the term is most specific and strip it away; this would probably be some modifying word or phrase, such as an adjective or prepositional phrase. However, if no satisfying broader target term exists, one must consider whether to map to a term that is conceptually narrower or is related in some other way. While “mapping up” seems cognitively straightforward, “mapping down” or “across” is likely far less so. This paper investigates the basis for decisions to “map down” by examining the nature of relationships between broader source terms and the narrower target terms to which they are mapped as closest in meaning.

In 1996, the U.S. National Library of Medicine and the Agency for Health Care Policy and Research performed a distributed national experiment known as the Large Scale Vocabulary Test (LSVT) (Humphreys, McCray, & Cheh, 1997). It aimed to determine the degree to which a combination of existing machine-readable health terminologies covered the concepts and terms needed for a comprehensive controlled vocabulary for health information systems. During a 6-month period, some 63 participants compared more than 40,000 terms from a variety of health-related terminologies with the UMLS Metathesaurus through a specialized Internet-based interface. Testers (and reviewers) were asked to map their source terms to the UMLS target term most closely related in meaning. Of the 41,127 terms submitted, exact matches were found for 58%; 28% were mapped to broader terms, 3% to narrower terms, and 10% to terms related in some other way, while no match was found for 1%.

Most of the terms in the LSVT data set are noun phrases consisting of a head noun plus optional pre-modifiers (typically adjectives or other nouns) and optional post-modifiers (typically prepositional phrases or various types of clauses). A recent study (McCray & Browne, 1998) investigated the 11,387 LSVT source terms that were narrower in meaning than the UMLS concepts to which they were mapped using increasingly sophisticated lexical processing techniques to recognize the modifiers, which affect a term’s intrinsic clinical meaning.
Processing the terms involved various combinations of string matching, ignoring case, uninveting inverted terms, generating uninflected forms, eliminating a small number of stop words, and incorporating similarly processed UMLS synonyms for each term. Most of the members of the NT-BT mapping pairs differed only by modification. Pre-modifiers were most common and tended to be simple, while post-modifiers were less frequent and more complex: 63% differed only by some form of modification.

The present research analyzed the 1,162 term pairs from the LSVT in which source terms were broader in meaning than the UMLS concepts to which they were mapped, in an effort to characterize the ad hoc semantic and structural relationships formed by users. A secondary objective was to compare these relationships formed when users mapped down from broader to narrower terms to those observed in the earlier study of narrower to broader term mappings.

2. Methodology

All 1,162 term pairs that had been designated by the users as “broader than” mappings were extracted from the full LSVT data set for analysis. After manual replication of the lexical processing techniques employed by McCray & Browne (1998), each BT-NT pair was examined to determine how the broader source term differed from the narrower target term to which it had been mapped. These differentia were analyzed and characterized in two ways. First, they were assigned to semantic categories based strictly on their meaning in the context of the common aspects of the two terms, and then they were assigned to structural categories reflecting their syntactical role in the noun phrases.

A small number of BT-NT mappings were eliminated from the analysis because they were duplicates (n=36); had been erroneously classified, i.e., appeared to be NT-BT rather than BT-NT mappings (n=18); or contained so-called clinical qualifiers (n=25), noun phrases such as “status post” or “history of” indicating some operational or administrative function (Chute & Elkin, 1997). The remaining 1,103 BT-NT pairs formed the basis for subsequent analysis.

3. Results

A total of 805 BT-NT pairs were found to differ on the basis of a single semantic category and a single structural category. (An additional 119 term pairs had more complex semantic differences, and were assigned to two or more categories; these also tended to exhibit more complex structural behavior and were not further analyzed in this study.) The results are considered in two main ways, first, according to structural or syntactic distinctions between the broader and narrower terms, and then, according to the semantic differences. A third type of distinction was based on hierarchical specification. Other categories that were observed in the sample as a whole, but which were not analyzed in further detail were those in which the broader and narrower terms shared some lexical material (8.5%), and those which seemed to be associatively related (5.2%).

3.1 Syntagmatic Specification

In syntagmatic specification, the semantic meanings of the terms have been modified by virtue of changes in the syntax or structure of the noun phrases. The structural categories with relative frequencies, along with examples of the BT-NT mappings in each case are as follows.

- Pre-modifiers (52.4%)
  
  RHINITIS → ALLERGIC RHINITIS
  CISTERN → CISTERNA MAGNA
These examples demonstrate pre-modification of the head noun by an adjective. The location of the adjective before or after the head noun in the noun phrase does not change the nature of the pre-modification.

- Complex compounds (16.5%)
  
  HEADACHE, FRONTAL OR PERIORBITAL → FRONTAL HEADACHE  
  USE OF DRUGS, ALCOHOL, OR TOBACCO → DRUG USAGE  
  HEPATIC DISEASE, CHRONIC → CHRONIC LIVER DISEASE AND CIRRHOSIS  
  GASTRIC RESECTION → RESECTION OF STOMACH AND GASTROJEJUNAL ANASTOMOSIS

This class contains source terms that are broader than their target terms because of compounding devices in either the source or the target terms. When in the broader term, this is typically indicated by “or,” with the term mapped to a narrower term that is a component of the broader term, such as in the first two examples. When in the narrower term, this is usually indicated by “and,” with the broader term lacking the additional element, such as in the latter two examples.

- Post-modifiers (9.1%)
  
  FRACTURE, COMPRESSION → COMPRESSION FRACTURE OF SPINE  
  PARTICULATE MATTER → PARTICULATE MATTER IN SYNOVIAL FLUID

Post-modifiers were most commonly prepositional phrases, such as in these examples.

- Neoclassical compounds (0.6%)
  
  LITHOTOMY → CHOLELITHOTOMY

Neoclassical compounds are morphological structures in which individual morphemes of typically Greek or Latin origin are joined together to form a single word. In this example, the addition of the anatomical prefix CHOLE- further specifies the procedural term LITHOTOMY.

3.2 Semantic categories

The semantic categories with relative frequencies, along with examples of the BT-NT mappings in each case are as follows.

- Anatomical structures, regions, spaces, and substances (28.7%)
  
  SINUS DRAINAGE → NASAL SINUS DRAINAGE  
  SUN-DAMAGED SKIN → SUN-DAMAGED SKIN ON FACE

In the first example, the head of the noun phrase is DRAINAGE, which is already modified by the anatomical structure SINUS; it is further specified anatomically by the pre-modifier NASAL. In the second, the head noun SKIN, already modified by SUN-DAMAGED, is also further specified to an anatomical region, this time by the post-modifier ON FACE.

- Functional, dysfunctional, and etiological concepts (27.7%)
  
  CERVICAL MYALGIA → DRUG-INDUCED CERVICAL MYALGIA  
  CARDITIS → VIRAL CARDITIS

In the first example, the head noun is MYALGIA, which is already modified anatomically by CERVICAL; it is further specified by the etiological pre-modifier DRUG-INDUCED. Likewise
in the second example, the head noun CARDITIS is specified etiologically by the pre-modifier VIRAL.

- Qualitative concepts (12.6%)

  FITNESS → PHYSICAL FITNESS
  TRAIT → PERSONALITY TRAIT

  In both of these examples, the head nouns are further specified by broad qualitative pre-modifiers: FITNESS by PHYSICAL and TRAIT by PERSONALITY.

- Procedures, measurements, and devices (8.8%)

  ILEAL BIOPSY → ENDOSCOPIC BIOPSY OF ILEUM
  COAGULATORS → COAGULATORS, LASER

  In the first example the head noun of the source term, BIOPSY, is modified anatomically by ILEAL. The target term has a different syntactic structure, but is made narrower semantically by the pre-modifier ENDOSCOPIC, which specifies the type of biopsy procedure. The second example specifies what kind of COAGULATORS with the pre-modifier LASER.

- Temporal concepts (8.1%)

  FATIGUE SYNDROME → FATIGUE SYNDROME, CHRONIC
  CLAUDICATION → INTERMITTENT CLAUDICATION

  Both of these noun phrases are examples in which the head noun, a dysfunction term, is further specified by a temporal pre-modifier: FATIGUE SYNDROME by CHRONIC, and CLAUDICATION by INTERMITTENT.

- Spatial and locative concepts (7.2%)

  LUNG → RIGHT LUNG
  PANCREATIC MARGIN → ANTERIOR MARGIN OF PANCREAS

  In the first case, the anatomical structure LUNG is specified spatially by the pre-modifier RIGHT. In the second, the head noun of the source term is MARGIN, which is modified anatomically by PANCREATIC. The target term has a different syntactic structure, but is made narrower semantically by the spatial pre-modifier ANTERIOR.

- Quantitative concepts (4.1%)

  PULMONARY NODULE → MULTIPLE PULMONARY NODULES
  FOVEAL REFLEX → ABSENT FOVEAL REFLEX

  In both of these examples, the head nouns are pre-modified quantitatively: PULMONARY NODULES by MULTIPLE, and FOVEAL REFLEX by ABSENT.

- Chemicals and biological agents (2.9%)

  CARBIDE → TUNGSTEN CARBIDE
  BUBBLE → BUBBLE, AIR

  Both of the head nouns in these examples are pre-modified by specific chemically or biologically active agents: CARBIDE by TUNGSTEN and BUBBLE by AIR.
3.3 Hierarchical specification

A surprisingly substantial proportion of the mappings required that the test participants use reasoning based on hierarchical inferencing.

- Hierarchical inferencing (21.4%)

KERATITIS, BACTERIAL → GONOCOCCAL KERATITIS
PRODUCTS FOR SAFE SEX → CONDOM

In the first example, it is necessary to know that gonococcus is a type of bacterium. The second example is a little more complex. To know that a CONDOM is a member of the class of objects that are PRODUCTS FOR SAFE SEX, it is necessary to know that a condom is used in the prevention of sexually transmitted diseases, which prevention is an example of safe sexual practices.

4. Discussion

The cognitive task of mapping down was sufficiently distinct from mapping up to result in different types of choices, and thus, different patterns of inter-term relationships. This may be seen primarily in the preponderance of anatomical and functional semantic differences, and in the surprising number of mappings based on hierarchical rather than syntagmatic specification.

Direct comparisons between the results of this study with that of McCray & Brown (1998) must be made with caution due to differences in focus and methodology. Most important, the current study emphasizes semantics rather than structure, which is reflected primarily in the manual versus automated approach; even the most sophisticated lexical processing techniques currently available are not yet capable of fine semantic analysis. However, it is worth noting that, as in the earlier study of mappings of narrower to broader terms, most pairs in this study differed primarily by modification. Indeed, an almost identical proportion of the differences between broader and narrower terms was due to some sort of linguistic modification (i.e., pre-modification, post-modification, or both) in this study (61.5%), as was reported by the earlier study (63%).

The identification of hierarchical specification seems to be an area in which the two studies differed. In contrast to the 21.4% found here, their 5.9% “residue” included both synonymous terms that were not in the UMLS and as well the cases in which the hierarchical relationship was not overtly signaled by linguistic modification. However, it is also possible that the class of term pairs characterized by shared lexical material (31%) in the earlier study also contained some of these hierarchical specification relationships.

The results of this study should help illuminate the cognitive processes underlying tasks of mapping between and among vocabularies, and thus inform methods for designing, aligning, and integrating knowledge structures. It appears that hierarchically structured knowledge bases may be used for inferential reasoning to profitably supplement sophisticated lexical processing techniques.

5. References

