Assessing the consistency of a biomedical terminology through lexical knowledge

Olivier Bodenreider[a], Anita Burgun[b], Thomas C. Rindflesch[a]
[a] U.S. National Library of Medicine, Bethesda, Maryland, USA
[b] Laboratoire d'Informatique Médicale, University of Rennes, France

Abstract:

***Objective**: In this paper, we investigate the use of lexical knowledge for determining consistency in biomedical terminologies. We focus on adjectival modification as a way of assessing the systematic use of linguistic phenomena to represent similar lexical or semantic features in the constituent terms of a vocabulary. **Methods**: Terms consisting of one or more adjectival modifiers followed by a head noun are selected from disease and procedure terms in SNOMED. After one modifier is extracted from the term, the remaining head noun – along with the other modifiers, if any – forms the context of this term. Modifiers sharing the same context are clustered together and ranked by frequency. For a pair (m1, m2) of frequently co-occurring modifiers, two terms m1c and m2c are created by systematically associating each modifier with the context in which at least one of the modifiers appears, called c. The existence of m1c and m2c is checked in both the vocabulary studied and the entire UMLS Metathesaurus, as well as the existence of the term corresponding to the context alone. Finally, relationships between m1c and m2c and between each of these terms and their context c are studied. **Results**: Four pairs of modifiers were studied: (acute, chronic), (unilateral, bilateral), (primary, secondary), and (acquired, congenital). The numbers of contexts studied for each pair ranged from 73 to 974. The percentage of contexts associated with both modifiers ranged from 5% to 50% in SNOMED and from 10% to 60% in UMLS. The presence of the context term varied from 31% to 64% in SNOMED and from 43% to 79% in UMLS. Finally, 172 occurrences (9%) of synonymy between a modified term and the context term were found in SNOMED. 145 such occurrences (8%) were found in the entire Metathesaurus. **Discussion**: The application of this method to discovering inconsistencies in a vocabulary is discussed, as well as differences among the different pairs of qualifiers studied. Examples or inconsistencies are presented and their consequences in terms of knowledge representation are discussed.*

INTRODUCTION

Large biomedical terminologies are usually the result of a team effort sustained over a long period of time. Although computerized editing environments may greatly facilitate the development of such terminologies, it is not uncommon to find inconsistencies in the terms or in the relationships among terms [1, 2]. The use of description logics (DL), for example, to analyze a vocabulary may help detect and fix semantic inconsistencies by automatically classifying the concepts (e.g., by comparing the expected classification to that proposed by the system). However, a significant amount of manual work is usually required for entering the terms into a DL-based system. Moreover, lexical phenomena that do not influence the semantics of a term may still fail to be caught by such systems.

In a previous study, we applied lexical knowledge to suggest hyponymic relationships among medical terms [3]. More precisely, we used the property that adjectival modifiers usually introduce a hyponymic relationship to suggest possible hyponymic relationships between modified and unmodified terms (e.g., *secondary cardiomyopathy* and *cardiomyopathy*). We found that less than half of the hyponymic relationships suggested by this method were actually recorded as hierarchical relationships in the Unified Medical Language System® (UMLS®). This method was used to suggest some 20,000 possibly missing relationships to be reviewed by UMLS editors. We also argued that patterns based in particular on additional knowledge about the modifiers might help assess certain hyponymic relationships automatically. For example, if *chronic ischemic enteritis* is a hyponym of *ischemic enteritis*, knowing that *acute* is an antonym of *chronic* allows the inference that *acute*

Proceedings of the Workshop on Natural Language Processing
in Biomedical Applications (NLPBA'2002) 2002:77-83.

ischemic enteritis is also a hyponym of *ischemic enteritis*.

Following-up on this study, we decided to apply lexical knowledge to the analysis of biomedical terminologies, with the aim of assessing the consistency of a terminology. In other words, our hypothesis is that lexical knowledge may help discover inconsistencies in a vocabulary, either lexical (inconsistent use of linguistic phenomena in terms) or structural (inconsistent organization of the terms). The goal of this study is not to automatically assess consistency. Rather, we propose an unsupervised method to detect potential inconsistencies, which can support and focus the effort of human editors of a medical vocabulary.

The method we suggest is based on word affinities derived from a corpus, and reuses techniques described by Grefenstette [4]. In our study, the corpus is a biomedical terminology, and word affinity is restricted to the adjectival modification of nouns. Although Nazarenko et al. also applied these techniques to a biomedical terminology [5], their goal was to identify semantic classes, not to assess consistency.

MATERIAL AND METHODS

The method may be summarized as follows. Starting with a list of terms, a syntactic analysis of the terms supports the identification of adjectival modifiers. The analysis is restricted to simple terms constituted of one or more modifiers followed by a head noun. After a modifier is extracted from the term, the remaining head noun – along with the other modifiers, if any – forms the context of this term. Modifiers sharing the same context are clustered together and ranked by frequency. Pairs of frequently co-occurring modifiers (i.e., occurring in the same context) are established. For a given pair of modifiers (m_1, m_2), terms are created by associating each modifier with the context c in which either one was detected (m_1c, m_2c). The existence of m_1c and m_2c is checked in both the vocabulary studied and the entire UMLS Metathesaurus, as well as the existence of the term corresponding to the context alone. Relationships between m_1c and m_2c and between each of them and their context c are studied.

Material

The UMLS Metathesaurus[1] (12th edition, 2001) contains over 1.5 million terms drawn from more than fifty medical vocabularies, and organized in some 800,000 concepts. A concept is defined as the set of synonymous terms corresponding to a single meaning. Conversely, terms are names for concepts [6]. In order to address the large size of the Metathesaurus, we limited our study to terms from SNOMED International[2] (version 3.5, 1998), one of the source vocabularies in the UMLS. We further selected from SNOMED terms from two major components of clinical medicine: diseases and procedures. We also removed from this set section headers, which often contain metadata. The notation "NOS", meaning "not otherwise specified", was removed from the terms. Finally, we excluded all terms containing a comma (10% of our original set). Commas usually signal a permuted form (e.g., *glucose measurement, urine*) or, more generally, a complex term (e.g., *patient transfer, in-hospital, unit-to-unit*) whose structure is usually not suitable for natural language processing tools. Our final list contains 65,124 terms (39,997 disease terms and 25,1274 procedure terms), corresponding to 41,842 concepts in SNOMED and 43,627 concepts in the Metathesaurus.

Identifying adjectival modifiers

The study of adjectival modification in the SNOMED terms under consideration was based on an underspecified syntactic analysis [7] that draws on a stochastic tagger [8] as well as the SPECIALIST Lexicon, a large syntactic lexicon of both general and medical English that is distributed with the UMLS. Although not perfect, this combination of resources effectively addresses the phenomenon of part-of-speech ambiguity in English, and, for example, correctly identifies *open* as an adjective (rather than a verb) in the term *open wound*. The resulting syntactic structure identifies the head and modifiers for the noun phrase analyzed. Each modifier is also labeled as being either adjectival, adverbial, or nominal. Although all types of modification in the simple English noun phrase were labeled, only adjectives were selected for further analysis in this study. For example, the term *male erectile disorder* was analyzed as:

[[modifier(*male*,adj)],
[modifier(*erectile*,adj)],
[head (*disorder*,noun)]].

This syntactic analysis was used to restrict the original set to terms consisting of at least one adjectival modifier followed by possibly other modifiers and a head noun. This specification excludes both simple terms (e.g., one isolated noun) and complex terms, not suitable for our analysis. 14,958 terms were considered for further analysis.

Establishing a list of adjectival modifiers and their contexts

For each adjectival modifier found in a term, we created a context made from the remainder of the

[1] umlsks.nlm.nih.gov

[2] www.snomed.org

term once the modifier was removed. Context words were lower-cased and sorted by alphabetical order to help identify similar contexts. For example, from the term *primary lacrymal atrophy*, we identified the modifier *primary* associated with the contex *atrophy lacrymal*, and the modifier *lacrymal* associated with the context *atrophy primary*. 20,176 (modifier, context) structures were created, corresponding to 3721 unique adjectival modifiers and 11,991 unique contexts.

Computing the co-occurrence of modifiers

The (modifier, context) structures were analyzed in order to identify pairs of modifiers frequently associated with the same context. From the two (m_1, c) and (m_2, c) structures created from the terms m_1c and m_2c where m_1 and m_2 represent two distinct modifiers and c represents their common context, the pair of co-occurring modifiers (m_1, m_2) is recorded. The frequency of co-occurrence for (m_1, m_2) is equal to the number of times m_1 and m_2 share a common context. For example, the context *atrophy lacrymal* is associated with the modifiers *primary* and *secondary*. Therefore, the pair of co-occurring modifiers (primary, secondary) is recorded for this context. The same pair of modifiers is associated with many other contexts, such as *amyloidosis* (in *primary amyloidosis* and *secondary amyloidosis*). The total frequency of co-occurrence for the modifiers (*primary*, *secondary*) is 45. In other words, these two modifiers share 45 distinct contexts. 40,883 pairs of co-occurring modifiers (m_1, m_2) were recorded, with frequency of co-occurrence ranging from 1 to 208. Only 495 pairs have a frequency of 5 or more.

Transforming terms

The existence of a pair of co-occurring modifiers (m_1, m_2) means that the two modifiers share at least one common concept c. However, m_2 may not be systematically associated with all the contexts associated with m_1. For a given pair of co-occurring modifiers (m_1, m_2), we created possible terms by associating each modifier with all the contexts in which the other modifier from the pair was detected. For example, using the (*primary*, *secondary*) pair, contexts associated with *primary* include *ovarian failure* and *amyloidosis* and contexts associated with *secondary* include *hyperprolactinemia* and also *amyloidosis*. The following six terms are created:

- *primary ovarian failure,*
- *secondary ovarian failure,*
- *primary amyloidosis,*
- *secondary amyloidosis,*
- *primary hyperprolactinemia*, and
- *secondary hyperprolactinemia*.

Looking-up transformed terms in SNOMED and the UMLS

The terms created in this way were mapped to the UMLS (and therefore also to SNOMED) by first attempting an exact match between the input term and Metathesaurus concepts. If an exact match failed, normalization was then attempted. This process makes the input and target terms potentially compatible by eliminating such inessential differences as inflection, case and hyphen variation, as well as word order variation. Moreover, the mapping is considered successful only if the concept mapped to is semantically compatible with the original term. Knowing that original terms are diseases (or procedures), mapping to concepts whose semantic type does not correspond to a disease (or a procedure) results in a failure.

Analyzing the relationships among terms associated with a pair of modifiers

Two terms m_1c and m_2c sharing the same context c and differing only by one adjectival modifier (m_1 or m_2) are expected to be represented as siblings and to be in direct hierarchical relationship with the context c. Such a representation is expected to be found in both the original vocabulary studied and the UMLS. The hierarchical features of SNOMED codes were used to calculate the relationship between two SMOMED terms. For example, the terms *bilateral vasotomy* (P1-7A124) and *unilateral vasotomy* (P1-7A122) were considered siblings because their codes share all digits but the last one. They can also be seen as descendants of *vasotomy* (P1-7A120), whose code ending with 0 denotes a higher level in the SNOMED hierarchy. In the UMLS Metathesaurus, two concepts were considered in direct hierarchical relationship if related by means of parent/child (PAR/CHD) and broader/narrower (RB/RN) relationships, and siblings if they shared at least one common first-generation ancestor.

RESULTS

From the most frequent pairs of co-occurring modifiers, we selected four pairs for further analysis: (*acute*, *chronic*), (*unilateral*, *bilateral*), (*primary*, *secondary*), and (*acquired*, *congenital*). For each pair (m_1, m_2), the number of contexts associated with at least one of the modifiers, the presence in the terminology of the modified terms (m_1c, m_2c) and of the context (c), and the nature of the relationship between the two modified terms (m_1c / m_2c) and between the modified terms and the context (m_1c / c, m_2c / c) are summarized in Table 1.

		m_1: *acquired* m_2: *congenital* (N = 974)				m_1: *acute* m_2: *chronic* (N = 608)				m_1: *primary* m_2: *secondary* (N = 187)				m_1: *unilateral* m_2: *bilateral* (N = 73)			
		SNOMED		UMLS		SNOMED		UMLS		SNOMED		UMLS		SNOMED		UMLS	
present	both	52	5%	97	10%	208	34%	244	40%	45	24%	69	37%	37	51%	44	60%
	m_1c only	100	10%	76	8%	203	33%	190	31%	78	42%	67	36%	22	30%	18	25%
	m_2c only	822	84%	801	82%	197	32%	174	29%	64	34%	51	27%	14	19%	11	15%
	context	306	31%	418	43%	324	53%	399	66%	119	64%	147	79%	41	56%	50	68%
m_1c and m_2c siblings		10	1%	51	5%	142	23%	225	37%	29	16%	64	34%	41	56%	41	56%
relationship of m_1c or m_2c to c	child	44	4%	181	17%	300	37%	294	35%	90	39%	101	39%	42	38%	41	35%
	siblings	18	2%	93	9%	78	10%	239	28%	16	7%	65	25%	22	20%	32	27%
	synonyms	99	10%	82	8%	38	5%	26	3%	29	13%	24	9%	2	2%	1	1%
	none	865	84%	715	67%	400	49%	293	34%	97	42%	66	26%	44	40%	43	37%

Table 1 – Characteristics of four pairs of co-occurring modifiers (m_1, m_2)

The pair (*acquired*, *congenital*) will be used to illustrate the results. 974 contexts are associated with either modifier of the pair. Both modified terms are present in SNOMED in 52 cases, and in the UMLS in 97 cases (e.g., *acquired spondylolisthesis*, *congenital spondylolisthesis*). Terms modified by *congenital* only (e.g., *congenital bronchiectasis*) are more frequent (822 in SNOMED) than those modified by *acquired* only (e.g., *acquired epidermolysis bullosa*, 100 in SNOMED). Their contexts (e.g., *epidermolysis bullosa*) are present in SNOMED in 306 cases and in the UMLS in 418 cases. The terms modified by *acquired* and *congenital* are not frequently represented as siblings in SNOMED (10 cases). For example, *acquired keratoderma* (D0-22310) and *congenital keratoderma* (D4-40130) are represented in two separate branches of the disease hierarchy in SNOMED. Moreover, the relationships between modified terms and their context also contribute to the characterization of a pair of modifiers. Most terms modified by *acquired* and *congenital* do not have any paradigmatic relationship represented with their context. For example, although *keratoderma* exists as a concept in the Metathesaurus, there is no relationship between *acquired keratoderma* or *congenital keratoderma* and *keratoderma*. In 44 cases, the relationship is hierarchical (e.g., between *congenital porphyria* and *porphyria*). In 18 cases, the modified term and its context are siblings (e.g., *congenital Addison's disease* and *Addison's disease*). Finally, in 99 cases, they are considered synonyms in SNOMED (e.g., *acquired polycythemia* and *polycythemia*).

DISCUSSION

Ontological perspective

Classically, in the Ogden-Richards triangle, there is a distinction between the symbol (here, the term), the concept named by the term, and the referent ("thing in the world") referred to by the concept and for which the term stands [9]. In this study, the terms m_1c and m_2c modified by a pair of modifiers (m_1, m_2) and their context (c) are terms used to name concepts. In the simplest case, the three distinct terms m_1c, m_2c and c stand for three referents, referred to by three concepts. Indeed, we found many occurrences of this representation. In this case, the context represents generic knowledge, while the modified terms bear some kind of specification. The context c is in hierarchical relationship with both m_1c and m_2c, and therefore, m_1c and m_2c are siblings. In many cases, however, more than one symbol is available to name a concept (synonymy). Sometimes, the same symbol is used to name several concepts (polysemy). While synonymy and polysemy are well-known linguistic phenomena, other associations among terms, concepts and referents may be found as well. Namely, the following situations may occur and will be discussed: missing referent, missing concept, and missing symbol.

Missing referent: In this experiment, we artificially created terms by associating modifiers with contexts, knowing that some of these associations may not actually stand for an existing referent. For example, a *congenital cleft hand* results from a developmental anomaly and no other circumstance later on in life can cause the same condition. In other words, there is no such referent as an *acquired cleft hand*. Therefore, the term *acquired cleft hand*

and the concept it could name are purposely and correctly missing from medical terminologies. Moreover, because the only possible circumstance for a cleft hand to occur is *congenital*, there is no need for a generic term *cleft hand*. The existence of only one specialized concept suppresses the need for a generic concept. *Acute copper deficiency* provides another example of a referent that does not exist. In this case, however, the generic term *copper deficiency* does exist, but symbolizes the same meaning as *chronic copper deficiency*.

Domain knowledge is needed to distinguish between a referent that does not exist and failure to represent an existing referent. The use of the UMLS partially supplies this knowledge. Modified terms and contexts are consistently more likely to be found in the UMLS than in SNOMED, which is not surprising since, by design, the UMLS is both broader in scope and more granular. Terms present in the UMLS but not in SNOMED may indicate that the referent does actually exist while SNOMED lacks a term to name it. For example, the generic term *hearing disorder* is present in the UMLS, but not in SNOMED, although *congenital hearing disorder* is present in SNOMED.

Missing concept: The absence of a concept in the UMLS may also result from an incomplete representation of the world, and, once again, domain knowledge is needed to find out. For example, although *congenital pneumonia* and *pneumonia* are represented in both systems, there is no *acquired pneumonia* in SNOMED or in the UMLS. In this case, *acquired pneumonia* is the most common form of the disease, so common that the generic term *pneumonia* is used to represent the prototypical term. Many cases of this phenomenon can be found.

Incomplete knowledge representation may result in inaccurate reasoning. For example, the prototypical form of meningocele, *congenital meningocele*, is clustered together (in the same concept) with the generic term *meningocele*. As a consequence, *acquired meningocele* is correctly represented as a child of the generic term, but wrongly represented as a child of *congenital meningocele*, allowing properties of congenital meningoceles to be falsely inherited by acquired meningoceles. The term *congenital meningocele* symbolizes a concept distinct from that named by the generic term and should therefore be represented as a distinct concept.

Missing symbol: In some cases, the absence of a term in SNOMED simply results from a lack of synonymy being represented. For example, the term *primary polycythemia* does not exist in SNOMED, but the concept it symbolizes does, simply named by a different term (*polycythemia vera*). The synonymy between the two terms is recognized in the UMLS.

Lexical inconsistency

Besides discrepancies in knowledge representation that could be detected by description logics-based analyses as well, this study also revealed lexical inconsistencies. For example, the two SNOMED terms *primary open angle glaucoma* and *secondary open-angle glaucoma* are hyphenated differently. Also, some but not all terms modified by *bilateral* exhibit a plural mark while the term modified by *unilateral* is often, but not always, singular. The systematic creation of synonyms based on of spelling variants (e.g., *anemia* / *anaemia*) could have been tested as well.

CONCLUSION

In this study, we used lexical knowledge to assess the consistency of a biomedical terminology, not only from the perspective of knowledge representation, but also for checking the consistent use of linguistic phenomena in terms. This method alone is certainly not sufficient for ensuring consistency, and we reaffirmed the need for domain knowledge. However, we believe that it can be useful to limit and focus the effort of the human editors of biomedical terminological systems. In the future, we would like to generalize this approach based on adjectival modification (e.g., *hepatic carcinoma*) to other kinds of modifiers, especially nominal (e.g., *liver carcinoma*) and prepositional (e.g., *carcinoma of the liver*).

References

1. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994;1(1):35-50.
2. Schulz S, Hahn U. Medical knowledge reengineering-converting major portions of the UMLS into a terminological knowledge base. Int J Med Inf 2001;64(2-3):207-21.
3. Bodenreider O, Burgun A, Rindflesch TC. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. Proceedings of TIA'2001 "Terminology and Artificial Intelligence" 2001:11-21.
4. Grefenstette G. Corpus-derived first, second and third-order word affinities. In: EURALEX; Amsterdam; 1994.
5. Nazarenko A, Zweigenbaum P, Bouaud J, Habert B. Corpus-based identification and refinement of semantic classes. Proc AMIA Annu Fall Symp 1997:585-9.

6. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1-2):193-201.
7. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. In: Proceedings of the 6th Applied Natural Language Processing Conference. San Francisco: Morgan Kaufmann Publishers; 2000. p. 188-95.
8. Cutting DR, Kupiec J, Pedersen JO, Sibun P. A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing. 1992:133-140.
9. Ogden CK, Richards IA. The meaning of meaning : a study of the influence of language upon thought and of the science of symbolism. [8th]. ed. New York: Harcourt Brace; 1946.

Address for correspondence:

Olivier Bodenreider
National Library of Medicine, MS 43
8600 Rockville Pike
Bethesda, MD 20894 – USA

e-mail: olivier@nlm.nih.gov