Characterizing the definitions of anatomical concepts in WordNet and specialized sources

Olivier BODENREIDER
National Library of Medicine
8600 Rockville Pike
Bethesda, MD, 20894
olivier@nlm.nih.gov

Anita BURGUN
Laboratoire d'Informatique Médicale
Avenue du Pr Léon Bernard
35043 Rennes Cedex, France
Anita.Burgun@univ-rennes1.fr

Abstract

Objectives: The objective of this study is to characterize the definitions of anatomical concepts in a general terminological system (WordNet) and a domain-specific one (a medical dictionary). **Methods**: Definitions were first classified into five groups with respect to the nature of the definition. The principal noun phrase (or head) of the definiens was then compared to the definiendum through a reference hierarchy of anatomical concepts. **Results**: This study confirms the predominance of genus-differentia definitions for anatomical terms. Hierarchical relationships are, as expected, the principal type of relationships found between the definiendum and the head of the definiens. **Discussion**: Differences in the characteristics of the definitions between WordNet and medical dictionaries are presented and discussed.

Introduction

We are interested in characterizing the definitions of medical terms in various sources in order to get a better understanding of their structure. Our ultimate goal, though, is to obtain a representation of the definitions in a formalism such as conceptual graphs in order to compare definitions from various sources. This study is part of a larger project aimed at comparing definitions of medical terms in specialized sources such as medical dictionaries with those in general resources such as WordNet®. In other words, our goal is to compare definitions of medical terms for health professionals and for users of consumer health applications.

Although not completely unrelated to them, the task of characterizing definitions is quite different from other tasks in which definitions are involved, especially acquiring definitions from a corpus (see Klavans and Muresan (2000) for an application to the medical domain) or acquiring an ontology from definitions as proposed by Shaikevich (1985).

1   Background

1.1   Kinds of lexical definitions

As in most dictionaries, the definitions in both medical dictionaries and WordNet are made of two parts: the term to be defined (or *definiendum*) followed by or linked to the expression used to define it (or *definiens*). Besides relying on synonymy or antonymy, i.e. linking a term to its synonym or opposite, several methods can be used to create dictionary definitions, also called lexical definitions. In a *Genus-Differentia* definition, the definiendum is described first by a broader category, the genus, then distinguished from other items in that category by differentia. Although a similar method has been long used to classify living organisms, its application extends beyond the domain of biology. Other kinds of definitions include those describing the cause or the function of the definiendum [Swartz (1997)].

1.2 Definitions in WordNet

WordNet, an electronic lexical database, has been developed and maintained at Princeton University since 1985 [Fellbaum (1998)]. Sets of synonymous terms, or synsets, constitute its basic organization. The current version (1.7) integrates about 100,000 synsets. Several types of relations between synsets are recorded in WordNet, including hyponymy and meronymy. In addition, each synset has a definition (or gloss) that defines the synset.

WordNet has been compared to specialized knowledge sources, including sources in the biomedical domain see for example Burgun and Bodenreider (2001). Comparison often relies on the semantic structure of WordNet, i.e., the relationships represented in WordNet (synonymy among terms and hierarchical relationships among synsets). However, as noted by Harabagiu and Moldovan (1998), the definitions also represent an interesting source of knowledge. They propose to transform the definitions into directed acyclic graphs whose nodes are WordNet synsets and whose links are lexical relations. While the somewhat stereotyped structure of most WordNet glosses is expected to facilitate the analysis, they acknowledge that ambiguity, both lexical and semantic, is likely to represent a major difficulty.

1.3 Definitions of anatomical terms

We selected anatomy as the domain of our study because it is central to the larger biomedical domain and to some extent part of the general domain. Not surprisingly, anatomy is well-represented in WordNet, where the synset "body part" has 1785 hyponyms, direct and indirect. Specialized resources such as the University of Washington Digital Anatomist symbolic knowledge base (UWDA) created by Rosse et al. (1998) constitute authoritative resources useful for establishing a list of anatomical terms. In addition, UWDA will also be useful for building the lattice of anatomical concepts needed for analyzing conceptual graphs later in this project. UWDA, however, could not be used as a source of definitions in this study because it provides definitions only for some high-level concepts.

Anatomy inherently results from observation, sometimes long before the entity observed can be named and classified. As a consequence, in anatomical definitions generally, what is attached to a lexical entry is still sometimes a description, useful for locating the physical entity while observing or dissecting, rather than a definition, for locating the concept in semantic space. For example, a description of "Adrenal gland" may refer to its shape, color and location relative to the kidney. Depending on the source, descriptions are either free-text or structured. For example, a template for the description of nerves includes information about their origin, their distribution and their branches.

A major kind of lexical definition, which includes definitions of anatomical concepts, is the traditional genus-differentia definition, in which the genus is often a broad category such as "organ" or "muscle" and the differentia can be, among others, a location (e.g., "situated near the kidney") or a function (e.g., "that carries blood from the heart to the body"). In addition to genus-differentia definitions, in which the definiendum is *by definition* in a taxonomic relationship with the definiens, various kinds of definitions may be found for anatomical terms. These include definitions by meronymy, in which the definiendum is in a 'part of' relationship with the definiens and definitions emphasizing a property or a function, expressed by a general term, instead of a genus. Examples of the various kinds of descriptions and definitions found for anatomical terms are given in Table 1.

	Subcategory	Example
Definition	Genus-Differentia	**Tarsal bone**: the seven bones of the ankle
	Meronymy	**Small intestine**: the proximal <u>portion of</u> the intestine
	Property	**Diaphragm**: a muscular <u>partition</u> separating the abdominal and thoracic cavities
Description	Free-text	**Adrenal gland**: a flattened body situated in the retroperitoneal tissues at the cranial pole of each kidney
	Structured	**Soleus muscle**: • <u>origin</u>, fibula, popliteal fascia, tibia; • <u>insertion</u>, calcaneus by tendo calcaneus; • <u>innervation</u>, tibial; • <u>action</u>, plantar flexes ankle joint

Table 1 – Categories of descriptions and definitions found for anatomical terms.

2 Material

2.1 Source of anatomical terms

Starting with approximately 4000 concepts in UWDA, we used the term listed in UWDA as the "preferred term" for each concept. This is the term used in most anatomy textbooks, as opposed to, say, clinical variants. We then filtered out terms corresponding to highly specialized concepts, not likely to be found in a general resource. We used filters based on the presence in the term of adjectival or prepositional modifiers indicative of the specialization of the term (e.g., left / right, anterior / posterior, mention of a particular vertebra, finger or toe). For example, "median nerve" belongs to our list while "right median nerve" was filtered out. Names for specific joints (e.g., "Calcaneocuboid joint") and ligaments (e.g., "Patellar ligament") were also filtered out, leaving mostly muscles (e.g., "Biceps brachii") and nerves (e.g., "Sensory nerve") in addition to organs such as heart and lung and organ categories such as gland and muscle. Applying these filters, we selected 420 terms (about 10%) suitable for further analysis.

2.2 Source of definitions

We used WordNet (1.7) as the general resource (using WordNet glosses as definitions) and Dorland's medical dictionary (27th edition) as the specialized resource.

Out of the 420 anatomical terms selected, 134 were defined in WordNet and 213 in Dorland's. The definitions of the 117 anatomical terms found in both sources were finally selected as the material for this study.

3 Methods

3.1 Resolving ambiguity

Ambiguity was found in both WordNet and Dorland's when trying to map anatomical terms to these resources.

Anatomical terms were mapped to WordNet using the standard *wn* function. When the mapping resulted in multiple senses, having anatomy as a target helped selecting the correct sense, i.e., the synset with "body part" in its hypernyms. In the rare cases of mapping to multiple hyponyms of "body part", the synset at the deepest level of the hierarchy was selected. The definitions of the few anatomical terms mapped to WordNet but outside the hierarchy of body parts (e.g., "intervertebral disc") were not used in this study.

Like many dictionaries, Dorland's lists definitions for the multiple senses or usages of a lexical entry as numbered definitional items. When an entry had multiple definitions, the correct one was selected manually.

3.2   Preparing the definitions

The definitions of anatomical terms in WordNet are often limited to one sentence and were processed entirely. By contrast, Dorland's definitions are often encyclopedia definitions. For this reason, only the first sentence of Dorland's definitions was considered in this study.

3.3   Classifying the definitions

The definitions were analyzed manually by the two authors, using the following strategy to classify them with respect to the kind of their definiens. The first issue was to distinguish between definition and description as they were defined in section 1.3. Then, definitions were classified in the following three subcategories: genus-differentia definition, definition by meronymy and definition based on a property. Descriptions were classified in two subcategories: free-text descriptions and structured descriptions. Definitions whose definiens did not fit any of these kinds were marked for separate analysis.

When the two authors disagreed about the classification of a definition, it was analyzed again until a consensus was reached.

3.4   Analyzing the relationship of the definiendum to the definiens

We used the MetaMap program developed by Aronson (2001) to map the definiens to the Unified Medical Language System® (UMLS®) [Lindberg et al. (1993; UMLS (2001)]. As a result, we extracted all biomedical concepts from the definiens, allowing us to access properties such as their semantic category and their relationships to other concepts. In addition, we took advantage of the shallow syntactic analysis provided by MetaMap in order to identify the first noun phrase (or head) of the definiens. When the noun in the first noun phrase was "pair" (e.g., "the twelve pairs of nerves connected with the brain"), the next noun phrase was used as the head. A similar correction was used to systematically prevent some adjectives from being interpreted as nouns (e.g., "longest" in "the longest and thickest bone of the human skeleton").

Since the definiendum comes from UWDA, which is one of the constituent vocabularies in the UMLS and the concepts extracted from the definiens by MetaMap are also UMLS concepts, the various kinds of relationships recorded in the UMLS can be exploited to compute whether medical concepts from the definiens (especially the head) are related to the definiendum. The following relationships were sought between the concepts corresponding to the definiendum and the head of the definiens: ancestor, descendant, sibling, other (usually associative) relationship. In addition, the relationship between the definiendum and the head was considered to be synonymy when the two terms mapped to the same UMLS concept.

3.5   Comparing the two approaches

In order to study whether there is a relationship between the two methods of characterization (class of definition and relationship of the definiendum to the head of the definiens), we built a table of contingency to summarize the cross-classification of the definitions into these two characteristics.

4   Results

4.1   Classification of the definitions

The distribution of the definitions into the various classes introduced in Table 1 is summarized in Table 2. While a large majority of the 234 definitions examined correspond to true definitions, some 12% of them are actually anatomical descriptions, structured or not. Not surprisingly, two thirds of the definitions follow the Aristotelian pattern of genus and differentia.

In eight cases, the definition did not meet any of the classification criteria. Five of these cases involved the definition of an adjective by Dorland's rather than that of the corresponding noun (e.g., "pisiform: resembling a pea in shape and size" instead of the wrist bone called "pisiform"). Other outliers included one reference to a table, one reference to a synonym, and the definition of a subentry that is not valid outside the context of the entry ("small bone: one whose main dimensions are approximately equal").

	Subcategory	N	%
Definition	Genus-Differentia	155	66
	Meronymy	14	6
	Property	30	13
Description	Free-text	13	6
	Structured	14	6
Other		8	3
	Total	234	100

Table 2 – Categories of descriptions and definitions found for anatomical terms.

4.2 Relationship of the definiendum to the definiens

The distribution of the relationship of the definiendum to the head of the definiens as defined in section 3.2 is summarized in Table 3. In two cases corresponding to the definition of an adjective instead of that of the wrist bone qualified by this adjective (e.g., "pisiform"), no concept could be identified by MetaMap from the definition. The total number of relationships studied between the definiendum and the head is thus 232 (out of the 234 definitions).

Examples of synonymy between the definiendum and the definiens include "Axis: the second cervical vertebra" and "maxilla: the upper jawbone in vertebrates". Although these definitions meet the criterion for a genus-differentia definition, the relation of the definiendum to the definiens is actually synonymy rather than hyponymy, the two terms being clustered into the same UMLS concept.

Hierarchical relationships are the principal type of relationships found between the definiendum and the head of the definiens.

Although descendant relationships usually denote an error in the mapping of the definiens to UMLS concepts, some definitions use holonymy (the inverse of meronymy) to relate the definiendum to the definiens, for example, in "nerve: any bundle of nerve fibers running to various organs and tissues of the body".

Finally, sibling relationships between the definiendum and the definiens either correspond to a kind of definition other than genus-differentia or denote some potential knowledge representation issue in the UMLS (e.g., although the patella is indeed "a triangular sesamoid bone", no medical vocabulary in the UMLS records any hierarchical relationship between the concepts "patella" and "triangular bone").

Relationship	N	%
Synonymy	8	3
Ancestor, first-generation	82	35
Ancestor, other	48	21
Descendant	3	1
Sibling	19	8
Other (usually associative)	0	0
None	72	31
Total	232	100

Table 3 – Relationship between the definiendum and the head of the definiens.

4.3 Comparison between the two approaches

Table 4 summarizes the cross-classification of the definitions into the two characteristics studied: class of definition and type of relationship of the definiendum to the head of the definiens.

Since *by definition* of genus-differentia definitions the genus is a broader category compared to the definiendum, the ancestor relationship is logically predominant in the genus-differentia definitions. However, the number of definiens mapping to an ancestor of the definiendum is slightly less than the number of genus-differentia definitions. In addition, not all hierarchical relationships in UMLS are taxonomic, therefore it is not surprising to find that some relationships listed as ancestor actually correspond to meronymic definitions.

For most definitions based on a property, there is usually no relationship found in the UMLS between the definiendum and the property. The concept used to represent the property is often general (e.g., "sac", "tube") while some concept more specific to the domain of anatomy could have been used instead (e.g., "saccular viscus", "tubular viscus").

Almost the same thing could be said about the descriptions, especially free-text descriptions where the head of the definiens is a general term (e.g., "structure", "unit", "mass"), not related to the definiendum.

	Definition			Description			
	Genus-differentia	Meronymy	Property	Free-text	Structured	Other	Total
Synonymy	5		1		1	1	8
Anc., 1st gn	77	4	1				82
Anc., other	46		1	1			48
Descendant	1	2					3
Sibling	11	1	5	1	1		19
None	15	7	22	11	12	5	72
Total	155	14	30	13	14	6	232

Table 4 – Cross-classification of the definitions into the two characteristics studied.

5 Discussion

5.1 General versus specialized resources

Not only different characteristics of the definitions can be compared to each other, but it is also possible to take advantage of the characteristics to profile a source of definitions or to compare several sources. For example, Table 5 shows the classification of the definitions reported in section 4.1 but analyzed separately for each source.

Although the number of definitions is too small and the domain too limited to draw any definitive conclusion, it is remarkable that, for example, WordNet actually has some structured technical descriptions (e.g., "large intestine: beginning with the cecum and ending with the rectum; includes the cecum and the colon and the rectum; extracts moisture from food residues which are later excreted as feces").

This study also confirms the predominance of genus-differentia definitions in both general and specialized resources, although anatomical descriptions are more often found in Dorland's than in WordNet.

		WordNet	Dorland's
Definition	Genus-Dif.	75%	57%
	Meronymy	11%	15%
	Property	7%	5%
Description	Free-text	3%	8%
	Structured	3%	9%
Other		0%	7%
	Total	100%	100%

Table 5 – Categories of descriptions and definitions in two different sources.

5.2 Ontological perspective

In some cases, the definitions in both systems are different predicates that correspond to equivalent sets of objects. For example, gland may be defined as an "aggregation of cells, specialized to secrete or excrete materials not related to their ordinary metabolic needs". In other cases, however the definitions in both systems correspond to different sets of objects. For example, "salivary glands" in Dorland's include the three major glands (parotid, sublingual, and submandibular), as well as numerous small glands in the tongue, lips, cheeks, and palate. By contrast, in WordNet, salivary glands are "the three pairs of glands…", implicitly the major ones, thus virtually excluding the minor salivary glands. In this example, the term in Dorland's is generic while the term in WordNet actually corresponds to "major salivary glands".

Conclusion

Characteristics of the definitions of terms, especially from several sources, represent valuable information. Among other things, this study confirmed the predominance of genus-differentia definitions for anatomical terms in both WordNet and specialized resources. This knowledge is expected to help perform the deeper analysis needed for representing the definitions in a formalism suitable to their comparison.

Acknowledgments

References

Aronson, A. R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program*. Proceedings of AMIA Annual Symposium, 17-21.

Burgun, A., and Bodenreider, O. (2001). *Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System*. Proc NAACL Workshop, "WordNet and Other Lexical Resources: Applications, Extensions and Customizations", 77-82.

Fellbaum, C. (1998). *WordNet : an electronic lexical database* (Cambridge, Mass, MIT Press).

Harabagiu, S. M., and Moldovan, D. I. (1998). Knowledge processing on an extended WordNet. In "WordNet: An Electronic Lexical Database", C. Fellbaum, ed. (Cambridge, Massachusetts, MIT Press), pp. 379-405.

Klavans, J. L., and Muresan, S. (2000). *DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text*. Proceedings of AMIA Annual Symposium, 1049.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). *The Unified Medical Language System*. Methods Inf Med *32*, 281-291.

Rosse, C., Mejino, J. L., Modayur, B. R., Jakobovits, R., Hinshaw, K. P., and Brinkley, J. F. (1998). *Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base*. J Am Med Inform Assoc *5*, 17-40.

Shaikevich, A. Y. (1985). *Automatic Construction of a Thesaurus from Explanatory Dictionaries*. Automatic Documentation and Mathematical Linguistics *19*, 76-89.

Swartz, N. (1997). *Definitions, dictionaries and meanings*. http://www.sfu.ca/philosophy/swartz/definitions.htm *[Dec. 1, 2001]*.

UMLS (2001). *UMLS Knowledge Sources*, 12th edn (Bethesda (MD), National Library of Medicine).