Proceedings of AMIA Annual Symposium 2006:574-578.

Using WordNet to Improve the Mapping of Data Elements to UMLS for Data Sources Integration

Fleur Mougin[a], Anita Burgun[a], M.D.,Ph.D., Olivier Bodenreider[b], M.D.,Ph.D.
[a]EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I, France
[b]National Library of Medicine, Bethesda, Maryland

{fleur.mougin,anita.burgun}@univ-rennes1.fr, olivier@nlm.nih.gov

Each biomedical system has its own way of naming the pieces of information it contains, i.e., of defining its data elements (DEs). Integrating DEs facilitates the integration of biomedical resources. However, the mapping of DEs to the UMLS is ambiguous in many cases, when any correspondence is found at all. We propose to evaluate the potential contribution of a more general terminology: WordNet. Our method is based on synonyms, definitions, and structural properties of the terminologies. We applied it to a set of 474 DEs extracted from eleven biomedical sources. We show that WordNet can improve the direct mapping of DEs to UMLS when used to validate and disambiguate UMLS direct mappings. WordNet can also help identify indirect mappings of DEs to the UMLS.

INTRODUCTION

Because most biomedical systems have been developed independently of each other, they do not have a common structure, nor do they share a common data dictionary or data elements (DEs). A DE is a basic unit of information (called *attribute* in database parlance), having a unique meaning and distinct values, (called *instances* in databases)[1]. Examples of DEs in the biomedical domain include Gene Symbol and Pathology Name. The corresponding value sets would be the set of gene symbols (e.g., in a given model organism) and a list of diseases, respectively.

In practice, the major barriers to data integration are the heterogeneity of database schemas and the disparity of DEs across systems. The general framework of this paper is the integration of DEs in support of the integration of biomedical resources.

In a previous study [1], we used the Unified Medical Language System® (UMLS®) [2] for mapping DEs coming from separate Web resources to a biomedical terminology in order to integrate them. Toward this end, we attempted to find an exact match and a normalized match, by using existing lexical tools [3]. Finally, when no match was found, an approximate match was attempted using MetaMap, a program which maps text to concepts in the Metathesaurus [4].

The output of this mapping consists of the list of Metathesaurus® concepts for each DE, along with their semantic types, textual definition (when provided), synonymous terms, and ancestors.

The outcome of the mapping of DEs to the UMLS can be summarized as follows:

- *Unique match*. For example, the DE Additional cdna sequence is mapped to the concept *DNA, Complementary* by approximate match.
- *Multiple matches.* For instance, the DE Protein results in an exact match to three UMLS concepts: *Protein*, *Protein measurement*, and *Protein location.*
- *No match*. Some DEs are simply not mapped to any UMLS concepts, because they are not specific to the biomedical domain and need to be represented at a more general level. Examples of such DEs include features, keywords, and domains.

We propose to improve the mapping to the UMLS by using supplementary information. Our hypothesis is that general resources such as WordNet [5], an online lexical database of general English, could provide a complementary coverage of the domain described by the DEs under investigation. Previous studies have underlined common characteristics existing between the UMLS and WordNet [6] and proposed strategies for aligning them automatically and accurately.

By exploiting the properties of WordNet (WN), we expect to improve the mapping of DEs to the UMLS in the following ways. In case of unique matches, WN would help validate the UMLS mappings. This can be especially useful when MetaMap resolves acronyms (e.g. *cDNA*, as illustrated above), which is often error-prone. For multiple matches, WN would contribute external information useful for disambiguating UMLS mappings. Finally, WN would help identify indirect mappings to the UMLS when no direct UMLS mapping was found.

The objectives of this study are to validate and disambiguate the direct mappings of DEs to the UMLS using information from WN. Additionally, we propose to identify indirect mappings to the UMLS (through WN) for those DEs for which no direct match was found.

[1] http://www.atis.org/tg2k/_data_element.html

MATERIALS

Extracting data elements.
Our test set consists of data elements extracted from eleven Web-accessible biomedical sources, selected to be representative of the different kinds of resources found in the biomedical domain. Some of them contain information about genes: GeneCards, Entrez Gene, Geneloc, Genew (the HGNC database), and HGMD, others about proteins: Swiss-Prot, PDB, HPRD, InterPro or diseases: OMIM. Our application is not targeted to a particular model organism so we also included the resource MGI, which provides various kinds of information about mice (see the table in annex for links to these resources).

Creating a set of terms for querying sources. We first assembled a set of biomedical terms to be used as query terms in the data sources under investigation. These terms were extracted manually from a reference resource in the domain of medical genetics: the Genetics Home Reference. We then constituted our data set by selecting a random sample of 100 terms such as gene symbols (e.g. *HFE*, *BRCA1*) and pathologies (e.g. *hemochromatosis*, *breast cancer*).

Acquiring data elements from sources. The sources used in this study are Web-interfaces to biological databases, automatically generated by program. Therefore, it is expected that most pages of a given source share a common organization and presentation. We take advantage of this feature for identifying recurring terms throughout Web pages, which, we hypothesize, correspond to data elements. In practice, we developed a program for querying systematically the eleven sources through their query URL. For each source, a set of 100 HTML pages corresponding to entries from the set of biomedical terms is created. After eliminating the header and footer, the elements common to at least 75% of the HTML pages are extracted automatically. This selection results in eliminating specific information (e.g., a given gene name), while keeping general information (e.g., the term *Gene Name*). Examples of data elements extracted from the source Genew are Approved Symbol and Previous Names.

Integrating data elements through WordNet.
The data elements (DEs) extracted from various resources tend to be heterogeneous. In fact, each source often has its own way to name the DE it uses. For instance, the DE for pathological conditions is named Disorders in GeneCards, but Disease in HPRD. We previously proposed to exploit knowledge from UMLS for resolving the heterogeneity of DEs through linguistic approaches. We expand this work by exploiting a more general terminological resource, WordNet. WordNet is organized into sets of synonymous terms (verbs, nouns, adjectives, and adverbs), called synsets, each of which representing one lexical concept. The database contains about 150,000 lexical items organized in over 115,000 synsets. Synsets are organized into a hierarchy. Ancestors and descendants are called hypernyms and hyponyms, respectively, in WordNet parlance. Version 2.1 is used in this study.

METHODS

Our method can be summarized as follows. Starting from the mapping of DEs to UMLS obtained from a previous experiment, as described in the introduction, we first perform a similar mapping to WordNet (WN). We then exploit WN properties to validate unique matches to UMLS and disambiguate multiple matches. Finally, we attempt to find indirect mappings to UMLS through WN.

Mapping DEs to WordNet. In order to map DEs to WN, we use the *wn* program to associate terms with synsets. When a DE consists of more than one word, we map it to the longest spanning syntagm in WN. For instance, the DE Mus Musculus is mapped to the synset *mus_musculus#n#1* rather to the two synsets *mus#n#2* (type genus of the Muridae) and *musculus#n#1* (muscle). When multiple matches are found in WN, we use the context of the synsets for disambiguation purposes. In practice, we favor synsets whose definition or hypernyms contain predefined keywords related to the biomedical domain (e.g. word bases such as *biologic*, *medic*, *genetic*, *chromosom*). For example, as shown in figure 1, the synset selected for the word *transcription* is the second one because of the presence of the biomedical term *genetics* in its definition. Finally, we filter WN candidate synsets according to the syntactic category. For instance, in the DE detailed genetic map, the word *detailed* has three candidate synsets: one adjective and two verbs. Based on the syntactic analysis of the DE, only the adjective is selected here. The mapping to WN is fully automated and results for each DE in a list of synsets, along with their definition, synonyms, and hypernyms.

1. (n) transcription, written text (something written, especially copied from one medium to another, as a typewritten version of dictation)
2. **(n) transcription (genetics) the organic process whereby the DNA sequence in a gene is copied into mRNA; the process whereby a base sequence of messenger RNA is synthesized on a template of complementary DNA)**
3. (n) transcription (a sound or television recording)
4. (n) arrangement, arranging, transcription (the act of arranging and adapting a piece of music)
5. (n) recording, transcription (the act of making a record)

Figure 1: Candidate synsets for the word "transcription" (sense 2 in bold face corresponds to the medical meaning)

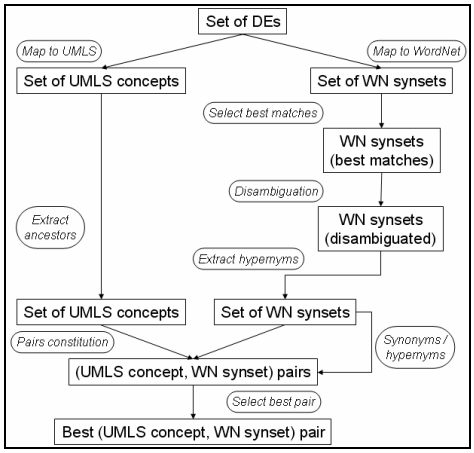

Figure 2: Direct and indirect mappings (through WordNet) of DEs to the UMLS

Validating unique mappings to UMLS. Assuming a mapping to WN is found, this mapping itself is either unique or multiple.

Unique mapping to WN. If the mapping to WN is unique, we exploit the properties of the candidate synset to validate the mapping to the UMLS. Toward this end, we compare the concept and synset according to the following criteria, in this order: 1) Similarity of their definitions, 2) Presence of common synonyms, and 3) Presence of common ancestors.

For criteria 2 and 3, we map the synonyms and hypernyms of the synset in WN to the UMLS through exact and normalized matches.

Multiple mappings to WN. When several mappings to WN are found, this indicates that the synset is ambiguous or only partially represented in WN. In both cases, the mapping to WN cannot be used for validating the mapping to UMLS automatically. For example, the DE Northern Blot, fully and correctly mapped to the UMLS concept "Northern Blot", is partially mapped to the two WN synsets "northern" and "blot".

Disambiguating multiple mappings to UMLS. In order to disambiguate the multiple mappings of a DE to the UMLS, we map it to WN, resulting in one or more synsets for this DE. We then associate pairwise the UMLS concepts and WN synsets, respectively, and select the best (concept,synset) pair using the similarity criteria described above for the validation of unique mappings.

Identifying indirect mappings to UMLS through WordNet. For those DEs for which no mapping to UMLS concepts was found (i.e., when the only mapping candidates are to WN synsets), we try to find an equivalent UMLS concept not from the DE itself, but from its mapping to WN. Starting from the WN synset(s) mapped to, we first attempt to map each of the synonyms in the synset(s) to the UMLS, using exact and normalized matches as before. If no synonym is mapped to UMLS, we start an equivalent mapping process for the direct hypernyms of the synset(s).

The resulting concepts constitute candidates for indirect mappings of the DEs to UMLS through WN.

The last three processes are completely automated but their results need to be checked for accuracy as no threshold for the similarity criteria has been determined yet. The whole process is illustrated in Fig. 2.

RESULTS

474 distinct DEs (548 tokens) were extracted from the eleven selected sources. Most of them were successfully mapped to WN. We provide the details of the mapping to WN with respect to the original mapping to UMLS and we analyze the contribution of WN to improving the mapping of DEs to UMLS. The first author, a bioinformatician, performed the evaluation by checking the results manually.

Mappings to UMLS vs. WordNet. Out of the 474 DEs, 387 (82.1%) were mapped to UMLS and 394 (83.1%) to WN. As illustrated in Table 1, in both UMLS and WN, when a mapping is found, it is unique in roughly half of the cases. The DEs mapped to the UMLS only include SNPs (Polymorphism, Single Nucleotide), rt-pcr (Reverse Transcriptase Polymerase Chain Reaction), and Micro-lesions. This finding is not surprising since these DEs are specific to the biomedical domain. Conversely, examples of DEs mapped to WN only include Homology, Lineage, Products, Pathways, Transcripts, and Motifs. Overall, 30 DEs (6.3%) were mapped to neither the UMLS, nor WN, including Paralogs, Ortholog, and Exuns.

Validating unique mappings to UMLS. WN provided supporting evidence for validating 82 unique mappings of DEs to UMLS (43.9%). More specifically, 68 were validated by exploiting definition similarity, 2 with synonyms, and 12 using ancestors. Following are some examples of mappings validated with respect to the type of evidence supporting the validation.

- The mapping of the DE mRNA sequence to the concept *RNA, Messenger* is validated by the synset *mrna#n#1* because of the similarity in their definitions (51.9%). Common elements in definitions include *nucleus*, and *RNA*.
- The mapping of the DE Duplication to the concept *Duplication* is validated by the synset *duplication#n#1* because they share a synonym: *Duplicate*.
- The mapping of the DE Length to the concept *Length* is validated by the synset *length#n#1* because they share the two common ancestors *Dimensions* and *Attribute*.

73 cases (39.0%) of unique mapping to UMLS could not be validated automatically by mapping to WN, because these DEs mapped to multiple WN synsets. For example, the DE Gene Function was mapped to only one UMLS concept *Gene Function*, but to four synsets in WN. Here, the mapping to WN still benefits the validation process by helping the experts focus on these cases. Finally, 32 unique mappings to UMLS (17.1%) could not be validated through their mapping to WN (16 unique and 16 multiple) because no common features could be found between the concept and synset(s) associated with the DE.

Disambiguating multiple mappings to UMLS. 95 multiple mappings of DEs to UMLS (47.5%) were successfully disambiguated with WN. Nearly all of them resulted from processing the definitions (94 compared to only one for the ancestors). One such example is the mapping of the DE Protein. Initially, it resulted in three concepts: *Protein*, *Protein measurement*, and *Protein location*. Through the mapping to the synset *protein#n#1*, we selected the concept *Protein* because of the similarity in their definitions.
74 multiple mappings (37.0%) could not be disambiguated because there was more than one WN candidate synset or no best (concept,synset) pair could be selected. The remaining 31 mappings of DEs to UMLS (15.5%) were not disambiguated because no mapping to WN was found.

Identifying indirect mappings to UMLS through WordNet. Overall, additional indirect mappings of to UMLS were identified through WN for 36 of the DEs with no direct mapping to the UMLS. Of these, 10 were unique and valid, and 26 ambiguous.
By exploiting synonymy in WN, 16 indirect mappings of DEs to UMLS were suggested. For instance, no direct mapping to the UMLS was identified for the DE topology, because no UMLS concept has *topology* as a synonym. However, this DE is mapped to the synset *topology#n#2*, of which one synonym is *regional anatomy*. Unlike *topology*, *regional anatomy* can be mapped to the UMLS. The DE topology can thus be mapped to the UMLS concept *Regional anatomy*, through a synonym from WN.
Using direct ancestors in WN, 21 indirect mappings to UMLS were found. For example, the DE Product was mapped to the synset *product#n#4*. This synset has no synonym but its direct hypernym *Chemical Substance* is a UMLS concept, which thus constitutes a potential UMLS mapping of the DE Product.

Table 1 summarizes the results of the mapping of DEs to UMLS and WN, with respect to mapping categories. The numbers in bold corresponds to those cases where WN contributed to improve the mapping to UMLS.

Table 1: Number of mappings to UMLS and WN for each category (bold numbers are cases where WN was useful)

		WordNet			
		Unique	Multiple	None	*Total*
UMLS	Unique	**82** + 16	**73** + 16	0	*187*
	Multiple	**95**	**74**	31	*200*
	None	**10**	**26** + 21	30	*87*
	Total	*203*	*210*	*61*	*474*

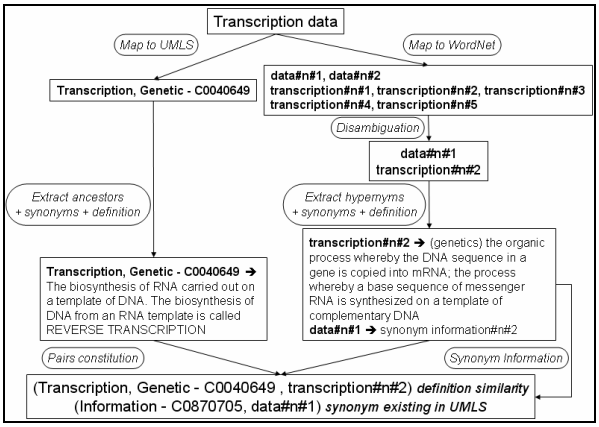

Figure 3: Example of the mapping process for the DE Transcription data

Example. In order to illustrate the contribution of WN, we describe the mapping to the UMLS of the DE Transcription data extracted from the source GeneCards (Figure 3). In the UMLS, a partial match is found to *Transcription*. In WN, two partial matches are found: to five synsets for *transcription* and to two synsets for *data*. The disambiguation process of *Transcription* is illustrated in Figure 1, resulting on the selection of the synset *transcription#n#2*. The synset *data#n#1* is chosen over data#n#2 because of the presence of its synonym *information* in the set of DEs (context). From the two independent mappings, it is now possible to:
- i) confirm that the mapping to the concept *Transcription, Genetic* is correct given the similarity observed in the definitions;
- ii) propose an indirect mapping of the word *data* to the concept *Information*, through the synset *data#n#1* which maps to the original DE and has a synonym in the UMLS.

DISCUSSION

Findings and limitations
Overall, for the 474 DEs under investigation, the mapping to WN contributed to validate 82 mappings to UMLS and to disambiguate 95. Additionally, WN facilitated the manual validation of 73 mappings and the disambiguation of 74. Finally, 36 indirect mappings of DEs to the UMLS were identified through WN, when no direct mapping to UMLS could be found. The use of WN for the purpose of disambigu-

ating mappings to UMLS is of particular interest in the context of automatic mapping strategies for data sources integration. As noted earlier, thresholds for similarity criteria have not been established yet and the mappings still require some degree of manual validation. The validity of the mappings was evaluated by one person only (FM). An independent evaluation would be required to confirm our results.

As shown in the results, the exploitation of synonyms in our method was of limited interest. In fact, only two mappings could be validated using synonyms. This can probably be explained by the small number of synonyms present in WN, especially compared to large terminological systems such as the UMLS.

Another finding is the relatively low similarity between some definitions. In fact, definitions in the UMLS tend to be rather long (cf. Fig. 3), resulting in a small percentage of common words with shorter definitions in WN. For example, the similarity observed between *Protein* and *protein#n#1* (34.8%) does not do justice to the fact that their definitions share five relevant elements (*organic*, *group*, *amino acids*, *living cells*, and *polymer*). This information, however, is sufficient to select the UMLS concept *Protein* over the two other candidate concepts.

Most indirect mappings proposed by WN are ambiguous (70.3%). For instance, the DE contributor, not mapped directly to the UMLS, is mapped to two synsets: *contributor#n#1*, whose direct hypernym *Donor* exists in the UMLS, and *contributor#n#2* whose direct hypernyms *Writer* and *Author* are also found in the UMLS. In this case, a manual review is necessary to select which one, if any, of the proposed indirect mappings is correct.

Future work

Indirect mappings of DEs to UMLS through their values. Some DEs remain unmapped to the UMLS even through synonyms and hypernyms in WN. We plan to define an alternative approach to mapping DEs to the UMLS, by mapping not the DEs themselves to WN, but their associated values. For example, the DE homology present in Entrez Gene is mapped to WN (synset *homology#n#1*) but not to the UMLS. However, its values include *Mouse*, *Rat*, and *Human* indicating that this DE gives information about organisms (among which some variant of a gene is shared). The DE homology could be associated to the DE Organism existing in Swiss-Prot, by analysing these DE values.

Exploiting structural properties. We use the structural properties of the UMLS and WN to validate and disambiguate the mappings of DEs to the UMLS as well as to identify new mappings. Mork also used structural properties to align representations of anatomy [7]. In our study, the exploitation of ancestors was useful to validate 12 original mappings to the UMLS. Moreover, it provided 21 new (indirect) mappings of DEs to the UMLS. However, our method is currently limited to exploiting ancestors. Additionally, we want to exploit descendants to search for additional entities that are common to UMLS and WN.

In summary, we found the mapping to WN to be useful not for improving UMLS mappings with fully automated solutions but for providing substantial assistance to the humans curating them (through the validation of unique mappings, the disambiguation of multiple mappings, and the identification of new mappings).

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Mougin F, Burgun A, Bodenreider O. Data integration through data elements: Mapping data elements to terminological resources. Proc Symp on Semantic Mining in BioMedicine. 2006: 52-59
2. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med; 1993;32(4):281-291
3. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;:235-239
4. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21
5. Fellbaum C. (editor) WordNet: An Electronic Lexical Database. MIT Press, 1998, Cambridge, MA
6. Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. Proc Workshop on WordNet and Other Lexical Resources. 2001:77-82
7. Mork P, Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. International Conference on Data Engineering. 2004: 787-790

Electronic resources used in this study

GeneCards	http://bioinformatics.weizmann.ac.il/cards/
Entrez Gene	http://www.ncbi.nlm.nih.gov/entrez/
Geneloc	http://genecards.weizmann.ac.il/geneloc/
Genew	http://www.gene.ucl.ac.uk/nomenclature/
HGMD	http://www.hgmd.org/
Swiss-Prot	http://www.expasy.org/sprot/
PDB	http://www.rcsb.org/pdb/
HPRD	http://www.hprd.org/
InterPro	http://www.ebi.ac.uk/interpro/
OMIM	http://www.ncbi.nlm.nih.gov/entrez/
MGI	http://www.informatics.jax.org/
GHR	http://ghr.nlm.nih.gov/
WordNet	http://wordnet.princeton.edu/