Toward Constructing the National Cancer Institute Thesaurus Derived WordNet (ncitWN)

Amanda Hicks
University of Florida
Florida, USA
aehicks@ufl.edu

Selja Seppälä
University College Cork
Cork, Ireland
selja.seppala@ucc.ie

Francis Bond
Nanyang Technological University
Singapore
bond@ieee.org

Abstract

We describe preliminary work in the creation of the first specialized vocabulary to be integrated into the Open Multilingual Wordnet (OMW). The NCIt Derived WordNet (ncitWN) is based on the National Cancer Institute Thesaurus (NCIt), a controlled biomedical terminology that includes formal class restrictions and English definitions developed by groups of clinicians and terminologists. The ncitWN is created by converting the NCIt to the WordNet Lexical Markup Framework and adding semantic types. We report the development of a prototype ncitWN and first steps towards integrating it into the OMW.

1 Introduction

The Global Wordnet Grid (GWG) is a platform created to join together wordnets by linking them to a central registry of concepts, using the Collaborative Interlingual Index (CILI) as a pivot. Data in the GWG is linked following an 'onion model', with 'a core of concepts shared by many wordnets', validated by the community, and axiomatized through ontologies. The core extends to a middle layer with fewer shared wordnets and out to a layer of concepts mapped to only a single wordnet. An external layer contains synsets defined in project wordnets that do not fulfill the CILI inclusion criteria. One of the advantages of the GWG is that the resource is no longer limited to networks of single-word units, but is now open to phrasenets (frequent adjective-noun, noun-prep-noun, and verb-object combinations, as well as proverbs, idioms, and compounds). This feature creates the possibility to link wordnets to domain-specific terminologies, which often include multi-word expressions. The Open Multilingual Wordnet (OMW) is the reference instantiation of the GWG (Bond et al., 2016) adding the constraint that all member wordnets must be open according to the open definition.[1]

To date no specialized terminologies have been included in the OMW. Consequently, there is no established procedure for mapping technical concepts to the CILI nor for determining whether a technical concept ought to be indexed in the CILI. We report a preliminary biomedical wordnet based on the National Cancer Institute Thesaurus (NCIt) called the NCIt Derived Wordnet (ncitWN) and preliminary mappings to the CILI. By mapping the NCIt to the CILI and thereby integrating it into the OMW, we are developing the first specialized vocabulary mapped to the CILI. The two outcomes will be: (i) the NCIt mapped to the CILI and integrated into OMW, but just as significantly (ii) groundwork for a method to reliably integrate open and freely available specialized terminologies with these lexical resources. This work is a first step toward realizing the goals outlined in Smith and Fellbaum (2004).

2 Resources

2.1 The Collaborative Interlingual Index

The CILI is implemented as a collaborative open-source software based on the best-practices of the Semantic Web – persistent IDs, Creative Commons Attribution 4.0 (CC BY) license allowing redistribution, and versioning (Bond et al., 2016). It integrates and extends the list of concepts in the OMW, including all concepts in Princeton WordNet of English (PWN) (Fellbaum, 1998). Each concept in the CILI is described with a unique definition in English. Currently, most of these definitions are derived from PWN 3.0. The CILI is compatible with the two schemas (Wordnet-LMF/lemon) (Vossen et al., 2016; Mc-

[1] The Open Definition, http://opendefinition.org (October 28, 2017).

Crae et al., 2014) used for encoding individual wordnets. The Semantic Web identifiers conform to the standards being adopted for encoding and integrating biomedical terminologies and ontologies (Ruttenberg et al., 2007; Schuurman and Leszczynski, 2008) and allow the CILI to be linked to ontologies and domain-specific terminologies. The CILI's open collaborative framework includes rules, tools, and safeguards to support high quality, agreed-upon mappings of wordnets to the CILI (Bond et al., 2016).

2.2 Princeton Wordnet

In order to get lemmas and domain information for English, we use the Princeton WordNet of English 3.0 (Fellbaum, 1998). Synsets are grouped into 45 **lexicographer files** which we use as coarse domains (for example, `noun.artifact` contains nouns denoting man-made objects). PWN also has explicit domains linked by the `domain-category` relation, which we intend to use in future work.

2.3 The National Cancer Institute Thesaurus and the UMLS Metathesaurus

The NCIt is a reference terminology developed by the National Cancer Institute that covers over 118,000 concepts and is available in the Web Ontology Language (OWL) (Sioutos et al., 2007). Although initially developed to support research and data management in the domain of cancer, it also includes concepts of general biomedical interest that are not specific to cancer, such as a robust typology of diseases, procedures, and adverse events. Each concept in the NCIt is associated with a unique identifier, a preferred term, and synonyms. Many terms also include an English definition, a description logic definition, and cross-references to other terminologies. The English definitions are developed by groups of clinicians and terminologists. The clinicians are often from the international English speaking community (USA, UK, Australia). The NCIt is released under a special license. We communicated with the creators and maintainers and have ensured that the ncitWN and its inclusion in the GWG is in compliance with their license.[2] Terms are also classified using the Unified Medical Language System (UMLS) Semantic Types: there are 127 semantic types linked in an is-a hierarchy.

[2]NCI THESAURUS Terms of Use, `https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/NCI_THESAURUS_license.txt` (October 28, 2017).

The NCIt is included in the Unified Medical Language System Metathesaurus, a biomedical thesaurus that links approximately 200 biomedical terminologies to an index of concepts (Schuyler et al., 1993). In this respect, the UMLS Metathesaurus can be viewed as a domain specific analogue of the Open Multilingual Wordnet (OMW). The UMLS Metathesaurus also includes translations of some of its source vocabularies into languages other than English. It is available in two data formats, the Rich Release Format and the Original Release Format. Semantic types such as "Drug" have been added to the UMLS Metathesaurus to impose more structure and to organize concepts (National Library of Medicine, 2009).

2.4 Wordnet-Lexical Markup Framework

Wordnet-Lexical Markup Framework is a wordnet-implementation of the Lexical Markup Framework (Francopoulo et al., 2006) (LMF), an ISO standard for NLP lexicons and Machine Readable Dictionaries based on the eXtensible Markup Language (XML) format. It encodes linguistic knowledge of the lexicalized concepts represented in the wordnets and supports integration of wordnets with OMW (Morgado da Costa and Bond, 2015; Vossen et al., 2013; Bond and Foster, 2013). Although no domain specific resources have been integrated into the CILI to date, this schema is well suited for the integration of an external resource such as the NCIt. Wordnet-LMF allows for a greater inventory of semantic relations than the NCIt currently contains, including entailment, part-whole relations, and derivations.

3 Methods

3.1 Convert NCIt to Wordnet-LMF

We have written a simple program (in Python 3) to reformat the NCIt as a wordnet (ncitWN). It filters out obsolete and retired concepts, creates the necessary metadata, and builds a wordnet. The conversion process is based on a few assumptions, to be tested further: (1) all concepts are lexicalized as nouns, and (2) the child-parent relationship in the thesaurus can be modeled as simple hypernymy.

The UMLS Semantic Types could be modeled as external links or as links within the wordnet (as PWN does). Currently we add them as metadata on each synset (using `dc:type`).

We validate the ncitWN data format with (1) the LMF Document Type Definition, which validates

the XML representation of the Wordnet-LMF documents (Vossen et al., 2016) and (2) the OMW's online tool (Morgado da Costa and Bond, 2015; Tan and Bond, 2011) that detects content violations such as duplicate or missing definitions.

3.2 Map ncitWN to the CILI

We have tested the feasibility of mapping the ncitWN to the CILI using two approaches.

The first, automatic, approach uses the prototype Wordnet-LMF formatted version of NCIt to automatically generate candidate mappings to the CILI using lemma overlap and compatibility of UMLS Semantic Types with WordNet coarse domains. The score is the sum of the Jaccard similarity calculated over lemma overlap with a boost of 0.1 each time there is a match between the wordnet coarse domains and the UMLS Semantic Type, based on a simple table of equivalences.

For example consider the following match:

- ***mask*** (NCIt) "A protective covering worn over the face, or an apparatus for administering anesthesia or oxygen through the nose or mouth" «Manufactured Object» (C86570)

- ***mask***$_{n:4}$ (PWN) "a protective covering worn over the face" «noun.artifact» (`i56041`)

Here the overlap in lemmas is 100% ({***mask***} vs. {***mask***}) and «Manufactured Object» matches «noun.artifact» so the score is 1.1. The equivalence table was made by first matching only lemmas and, assuming that all 100% matches were good, linking the UMLS Semantic Type and PWN coarse domain. All matches of semantic types with more than 500 examples were taken to be good. An inspection of the less frequent matches showed many to be good, this mapping should be revised in subsequent work.

We manually evaluated a sample of the automatically produced matches with a match score $> .75$. The annotation scheme is summarized in Table 1. '0' is not used for mapping, but was nevertheless used to annotate candidate matches. These annotations will be used to generate heuristics for refining match scores, thereby expediting the mapping process.

Note that an annotation of '0' does not indicate that there is no relation between the NCIt and PWN term, but only that there is no hierarchical relation. There might be non-hierarchal relations, e.g., linguistically derived from, that may be incorporated in future work.

Annotation	Meaning
eq	equivalence
spec	hyponym of
gen	hypernym of
0	no hierarchy relation

Table 1: Annotations for candidate matches from ncitWN synset to PWN synset

The second approach was a manual analysis of PWN 3.0's coverage of the NCIt. We randomly selected 94 concepts from the NCIt, stratified according to whether the concept was a root, middle, or leaf node (respectively, 19, 37, and 38 concepts). We then searched for candidate mappings through lemmas in PWN and evaluated the match based on the corresponding definitions in the CILI. The manual coverage analysis was based on NCIt preferred terms and excludes synonyms. Preferred terms that take the form of boolean expressions such as 'Diagnostic or Prognostic Factor' were decomposed into their component expressions, which were used for searching candidate mappings. Thus, for 'Diagnostic or Prognostic Factor', we restored the elliptical noun and obtained two multiword expressions (MWEs) for which we searched candidate mappings, i.e., 'Diagnostic Factor' and 'Prognostic Factor'.

We distinguish six matching scenarios summarized in Table 2 and illustrate them with examples below.

Annotation	Meaning
0	no match
1	exact match
2	full match
3	partial match of MWE
4	preferred term with partial match
5	not suitable to map to CILI

Table 2: Annotation scheme for candidate matches from NCIt terms to PWN synsets

The coverage analysis was carried out in several steps (see Figure 1). In step 1, we determined whether the NCIt preferred term had a match in the PWN lemmas. If it did not, we annotated it with '0'. If there was a match, in step 2, we compared the NCIt and CILI definitions. If both the lem-

mas and the definitions matched, we considered them an exact match ('1'); if the lemmas matched but the NCIt definition was either more specific or broader than the CILI definition, the NCIt preferred term has a partial map ('4'); if the NCIt term and definition were NCI-specific, the concept was not suitable to be mapped to the CILI ('5'). If none of these options applied and the NCIt term was an MWE, in step 4, we decomposed the MWE into its parts and searched each word individually. In case of a match, we determined whether the CILI definition for the matched PWN lemma corresponds to the compositional meaning of the word in the NCIt MWE. If the meaning and the definition matched, we assigned '1', otherwise '0'. In step 5, we assigned an annotation to the NCIt preferred MWE by considering all the individual annotations assigned to each word composing the MWE.

Examples of matching and non-matching cases:

0. NCIt ***Archaea*** (C61092) is not in PWN.

1. NCIt ***Area*** (C25244) and PWN ***area***$_{n:6}$ (`i63937`) have identical definitions.

2. NCIt ***Breast Cancer Prognostic Factor*** (C19601) has no exact match in PWN but its parts do. The individual annotations assigned to each matched part of the MWE ('breast cancer': 1; 'prognostic': 1; 'factor': 1) allow us to assign the global annotation '2' to the preferred term.

3. NCIt ***Ito Cell Tumor*** (C80350) has no exact match in PWN and only two out of the three words composing the MWE are in PWN with the same meaning ('cell': 1; 'tumor': 1; 'ito': 0). These individual annotations allow us to assign the annotation '3' to the preferred term.

4. NCIt ***Acclimatization*** (C68767), defined as "The physiological process through which an organism grows accustomed to a new environment", has a narrower definition than the CILI definition corresponding to PWN ***acclimatization***$_{n:1}$,"adaptation to a new climate (a new temperature or altitude or environment)" (`i107289`).

5. NCIt ***NCI Administrative Concept*** (C28389) and its definition are specific to the NCI, therefore not suitable for mapping to the CILI.

4 Results

Automatically generating candidate mappings based on lemma overlap and compatibility of UMLS Semantic Types with WordNet domain-category types resulted in 47,464 candidates (out of 118,000), of which 6,028 had a match score $> .75$: this means that either all lemmas overlap or else most lemmas overlap and the domains are compatible. An additional 10,454 matches had a score in the range $.75 > .5$.

To date we have checked 570 of the 6,028 candidates with a match score $> .75$. The results are summarized in Table 3.

Annotation	**Number**	**%**
eq - equivalence	369	64.7
spec - hyponym of	21	3.7
gen - hypernym of	33	5.8
0 - no hierarchy relation	147	25.8
evaluated candidates	570	100.0

Table 3: Candidate matches evaluation results

These mappings suggest further heuristics for automatically mapping concepts and refining the match score in future work, thereby expediting mapping and evaluation. Some sample heuristics are listed below.

- Add a score for the similarity of the definitions, e.g., if the Jaccard distance of the definitions is $> .90$, map with 'eq'.
- If the UMLS Semantic Type is 'Manufactured Object' and the PWN synset is a verb, annotate the pair with '0'.

In the manual analysis of PWN 3.0's coverage of the NCIt, we found that 20.2% of the NCIt concepts had an exact match in PWN (and therefore also the CILI), 11.7% had no match in PWN, and 47.9% had a matching head noun, suggesting a suitable child concept of a synset in PWN. Of the 19 top nodes in the NCIt hierarchy,[3] three were exact matches and 11 had head nouns that were an exact match in PWN, suggesting a parent/child link.

5 Future Work and Discussion

The coverage analysis and the initial evaluation of the match candidates have brought to light several concrete examples in which guidance is needed to

[3] We exclude the node 'Retired Concept' from the count.

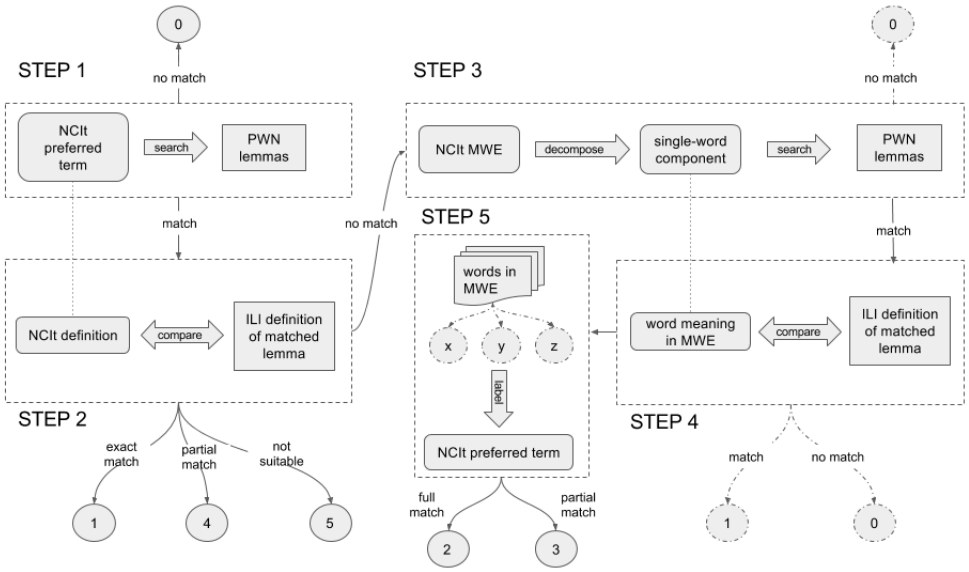

Figure 1: Steps of the manual coverage analysis

integrate specialized terminologies with the CILI. First, the NCIt contains dot objects and other cases of systematic polysemy that are sometimes distinguished in WordNet and would therefore have different relevant concepts in the CILI. For example, NCIt ***Cherry*** (C65311) does not have a proper definition but has two UMLS Semantic Types, fruit and plant, suggesting it can refer to a cherry tree or the fruit of a cherry tree. The candidate match in PWN is ***cherry***$_{n:2}$ (`i103308`) which is clearly defined as the tree, not the fruit. A matching strategy for such cases ought to be developed.

Second, we have encountered some cases where the core definition is the same, but exemplars or typical cases are different. In both examples below, an overlay is characterized as something to be applied over an object or surface.

- ***Overlay*** (NCIt) "A device designed to be applied over an object, typically for protection or identification" (C50093)

- ***overlay***$_{n:2}$ (PWN) "a layer of decorative material (such as gold leaf or wood veneer) applied over a surface" (`i56837`)

However, the NCIt characterizes an overlay as typically for protection or identification and PWN considers an overlay to be decorative. It is unclear whether these are similar enough to be considered equivalent, whether the NCIt concept should be considered a hypernym of the PWN synset (and therefore the corresponding CILI concept), or whether the typical functions, though not a necessary component of the definition, nuance the meaning sufficiently for no hierarchy relation to be added between the two.

Third, we find that some concepts are probably equivalent, but different definition writing criteria result in a narrower definition in PWN. Consequently, it is unclear whether the NCIt concept is a hypernym of the PWN sysnet.

- ***Anovulation*** (NCIt) "The absence of ovulation" (C34388)

- ***anovulation***$_{n:1}$ (PWN) "the absence of ovulation due to immaturity or post-maturity or pregnancy or oral contraceptive pills or dysfunction of the ovary" (`i107333`)

Fourth, we found that the assumption that all concepts are nouns is not true. Entries such as ***unfavorable*** are clearly adjectives. Fortunately, the UMLS Semantic Type 'Qualitative Concept' and the wordnet coarse domain `adj.all` both give an indication that it should be an adjective, so we should be able to tell this largely automatically. There are about 1,000 candidate adjectives, and even a few tens of verbs (such as ***mutate***), whose definitions tend to start with infinitive ***to*** in NCIt.

- ***Unfavorable*** (NCIt) "Expressing something as negative, undesired or adverse" (C102561)

- ***unfavorable***$_{a:1}$ (PWN) "not encouraging or approving or pleasing" (`i5455`)

- ***Mutate*** (NCIt) "To undergo or cause genetic mutation" (C28031)
- ***mutate***$_{v:1}$ (PWN) "undergo mutation" (i22358)

Finally, we need to decide how to handle multiword expressions that have been annotated with '2' such as ***Breast Cancer Prognostic Factor*** (C19601). One approach is to create a new concept in the CILI. However, further consideration needs to be given to which concepts are too domain specific to be included in the CILI. Another approach is to map these to the CILI by way of the head word using the hyponym relation. For example, ***Breast Cancer Prognostic Factor*** (C19601) would be mapped to i75200 by way of PWN ***factor***. However, as the number of concepts in the CILI grows, we anticipate that concepts that are not lexicalized in Princeton WordNet will appear in the CILI. For example, the concept prognostic factor may be added to the CILI in the future. In the long term, strategies for detecting and properly aligning such concepts will need to be developed.

We used UMLS Semantic Types, which were created to help disambiguate and cluster senses (McCray et al., 2001), to improve our automatic alignment to PWN coarse domains. PWN contains more detailed domain-category links such as ***tobacco***$_{n:1}$ is in the domain of ***pharmacy***$_{n:1}$. We could exploit both them and the hypernyms to improve the automatic mapping. Finally, if all the UMLS Semantic Types can be mapped to synsets, we can link them using domain-category. This will enrich the overall graph in ncitWN and facilitate mapping other UMLS terminologies to the CILI.

Acknowledgements

This research was supported in part by the MOE Tier 1 grant *Semi-automatic implementation of clinical practice guidelines in Singapore hospitals* and by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362. Association for Computational Linguistics, Sofia, Bulgaria. URL http://aclweb.org/anthology/P13-1133.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Global WordNet Conference (GWC2016)*, pages 50–57. Bucharest, Romania.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF) for NLP multilingual resources. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 1–8. Association for Computational Linguistics. URL http://aclanthology.coli.uni-saarland.de/pdf/W/W06/W06-1001.pdf.

John McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing 2014 (LDL-2014)*. Association for Computational Linguistics, Reykjavik, Iceland.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

Luís Morgado da Costa and Francis Bond. 2015. OMWEdit - The Integrated Open Multilingual Wordnet Editing System. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 73–78. Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Beijing, China. URL http://www.aclweb.org/anthology/P15-4013.

National Library of Medicine. 2009. *UMLS Reference Manual*, chapter Chapter 5 - Semantic

Network. U.S. National Library of Medicine, National Institutes of Health.

Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, et al. 2007. Advancing translational research with the Semantic Web. *BMC bioinformatics*, 8(Suppl 3):S2.

Nadine Schuurman and Agnieszka Leszczynski. 2008. Ontologies for bioinformatics. *Bioinformatics and biology insights*, 2:187.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. 2007. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.

Barry Smith and Christiane Fellbaum. 2004. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, page 371. Association for Computational Linguistics.

Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. Singapore.

Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference (GWC2016)*, pages 419–26.

Piek Vossen, Claudia Soria, and Monica Monachini. 2013. *Wordnet-LMF: standard representation for multilingual wordnets*, chapter 4, pages 51–66. John Wiley & Sons, Inc, Hoboken, NJ USA.