MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220267

Bottom-Up Natural Language Processing Based Evaluation of the Fitness of UMLS as a Semantic Source for a Computer Interpretable Guidelines Ontology

George Despotou[a], Ioannis Korkontzelos[b] and Theodoros N. Arvanitis[a]

a Institute of Digital Healthcare, WMG, University of Warwick, UK
b Department of Computer Science, Edge Hill University, UK

Abstract

Background: CIGs languages consist of approach specific concepts. More widely used concepts, such as those in UMLS are not typically used. Objective: An evaluation of UMLS concept sufficiency for CIG definition. Method: A popular guideline is mapped to UMLS concepts with NLP. Results are reviewed to evaluate gaps, and appropriateness. Results: A significant number of the guideline text mapped to UMLS concepts. Conclusions: The approach has shown promise and highlighted further challenges.

Keywords:
natural language processing, knowledge representation, practice guideline, computer interpretable guidelines

Introduction

Clinical Practice Guidelines are seen as encapsulating the best available evidence regarding management of a condition, constituting a widely accepted standard. CIGs will allow guidelines to be run by decision support systems [1]. Numerous approaches convert clinical guidelines into computer interpretable guidelines [2, 3]. One common aspect is that use languages with project specific concepts and semantics. All these languages aim to represent a common artefact (i.e., clinical guidelines). However, except for a few obvious concepts (e.g., patient), there is variation in the name, definition as well as structure of many of them. There are numerous coding systems such as UMLS and SNOMED CT, which provide unambiguous definition of healthcare related concepts. Furthermore, there is a high degree of traceability amongst most of these coding systems. Natural Language Processing (NLP) allows the computational analysis and processing of data in natural languages (i.e., text), increasingly applied in healthcare [4]. The objective of this paper is to understand, using NLP, whether the expressive prowess of UMLS is sufficient, to effectively provide semantic definitions for a CIG ontology.

Methods

The paper adopts a bottom-up approach by examining how published clinical guidelines can be mapped on UMLS concepts. MetaMap [5] analyses free text, and maps its contents to UMLS concepts. An utterance that is a block of text, is split into phrases. The tool then looks for words in the phrase that can be matched to UMLS concepts, with maximum confidence of 1000. When MetaMap cannot map words to UMLS concepts it produces a 'NOT FOUND' result. Each matched word may be mapped to multiple concepts, depending on the context of the sentence. A local server of MetaMap 2020 was used, in its default mode without additional parameters. The top results were reviewed manually to examine whether the MetaMap mapping is accurate and unambiguous in a specific context. For this, the work analyzed the NICE NG28 guideline. The following tests were performed to the MetaMap output: a) Descriptive statistical analysis of the number of mappings of each matched word; b) Examination of the confidence variation in the mappings of each word; c) Review of the words that gave a NOT FOUND result from MetaMap to identify potential semantic gaps in UMLS.

Results

The analysis found a total of 3527 unique mapping interpretations, on 1188 concepts. These 1188 concepts were associated with 85 semantic types. Figure 1 presents the distribution of the frequencies of concepts (CUIs). For example, 1 concept was mapped 130 times, whereas the majority of concepts were mapped fewer than 20 times, with 665 concepts being identified only once.

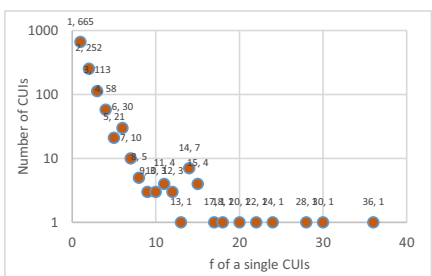

Figure 1 – Frequency of number of concepts appearing in mappings of words (e.g. 1 CUI appeared 30 times)

Table 1– Most common concepts in NG28

CUI	(f)	%	Description
C0027361	36	3.0%	person
C0001675	30	2.5%	adult
C0011860	24	2.4%	diabetes type 2
C0150600	22	2.0%	recommendation to
C0087111	20	1.9%	therapy
C0039798	18	1.7%	therapy (MeSH)
C2347489	17	1.5%	person observer
C0021641	15	1.4%	insulin
C0013227	15	1.3%	medicines

Table 1 presents the 10 most common concepts that were mapped from NG28. Figure 2 shows the number of mappings

for each instance of matched word (i.e., excluding those not found). It can be seen that the majority of the words had fewer than 10 mappings.

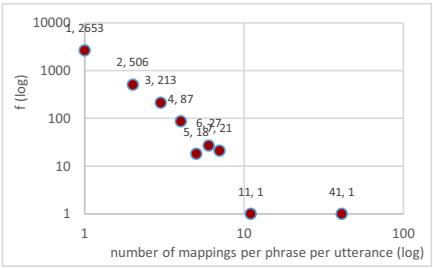

Figure 2 – Number of mappings per word per utterance

Table 2 – UMLS semantic types

SemType	(f)	%	Description
ftcn	333	10.2	functional concept
qlco	296	9.1	qualitative concept
fndg	231	7.1	finding
inpr	202	6.2	intellectual product
dsyn	181	5.6	disease or syndrome
topp	167	5.1	therapeutic/preventive procedure
idcn	151	4.6	idea or concept
qnco	150	4.6	quantitative concept
aapp	131	4.0	amino acid peptide or protein
hlca	119	3.7	health care activity

In total the analysis found 85 semantic types. Table 2 shows the 10 most common semantic types associated with mappings. The top 20 semantic types were associated with 85.6% of the mappings. Figure 3 shows the frequency of the standard deviations of the mappings for each matched word (where a word had more than one mappings). It can be seen that the majority of mappings had the same confidence scored.

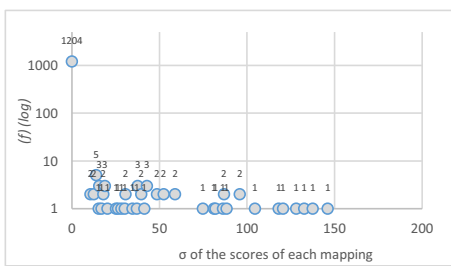

Figure 3 – Standard deviation distribution of the scores of mappings for each phrase.

Most mapping scores were the same, without prevailing interpretation.

MetaMap was not able to match 1135 phrases. This is a significant number (approx 39% of all phrases) indicating a potential gap in being able to capture the guideline. Table 3 shows the most frequent phrases that were not matched, and the most frequent unmatched phrases that were considered to constitute semantic gaps.

Table 3 – Top NOT FOUND words considered to be semantically relevant (left column) and top NOT FOUND words in general

Missing Semantics	(f)	Most Frequently Missing Phrases	(f)
Offer	12	and	198
Encourage	8	or	102
achieve	6	:	94
Tolerated	6	is	66
Reduce	5	If	61
Explain	4	who	45
HbA1c to	4	that	44
Caution	3	(	38
Reasses	3	and	198
the HbA1c	3	or	102
Advise	2	is	66
Individualise	2	If	61
Prescribe	2	is	66

Conclusions

UMLS does provide a rich semantic basis that we can use to model guidelines as CIG. An extract of the NICE NG28 guideline was used as a proof of concept. There is little information that cannot be mapped supporting the position that CIG can converge to existing semantic standards. The main challenge in UMLS to represent CIG seems to be logical operators. A number of issues were revealed in need of further research focusing including accuracy of MetaMap, detection and representation of 'not found' concepts, and analysis of more syntactically complex phrases, which may incorporate logical operators applied to the concepts.

Acknowledgements

Parts of this work were performed as part of Health Data Research (HDR) UK. The authors have not declared any conflicts for this work.

References

[1] E. Bilici, G. Despotou, T.N. Arvanitis. The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: A review. *DIGITAL HEALTH*. January 2018. doi:10.1177/2055207618804927

[2] A.A. Boxwala, M. Peleg, S. Tu,. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *J Biomed Inform* 2004; 37(3): 147–161

[3] D.R. Sutton, J. Fox,. The syntax and semantics of the PROforma guideline modeling language. *J Am Med Inform Assoc* 2003; 10(5): 433–443.

[4] Y. Liu, C. Whitfield, T. Zhang, A. Hauser, T. Reynolds, M. Anwar. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. Health Inf Sci Syst. 2021;9(1):25. Published 2021 Jun 25. doi:10.1007/s13755-021-00158-4

[5] A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17-21. PMID: 11825149; PMCID: PMC2243666.

Address for correspondence

Contact: Dr George Despotou. The preferred method of contact is email at: g.despotou@warwick.ac.uk