Reconciliation of SNOMED CT and domain clinical model for interoperable medical knowledge creation

Taqdir Ali and Sungyoung Lee*

***Abstract*— Use of heterogeneous data models in hospital information systems (HIS), obstructs the integration of clinical decision support system (CDSS) with clinical workflows. The diverse concepts diminish the interoperability level among the CDSS knowledge bases and data models of HIS. Standard terminology utilization in knowledge acquisition and its reconciliation with HIS data models are the candidate solution to overcome the interoperability barrier. We propose a reconciliation model to map concepts of diverse domain clinical models (DCM) with the standard terminology. In the proposed model, the implicit and explicit semantics are complemented to the word set of the targeted DCM concepts. The inclusion of semantics, mapped the DCM concepts to the SNOMED CT concepts with high accuracy. The results showed that the system correctly mapped 95% of concepts of DCM with standard terminology SNOMED CT concepts.**

I. INTRODUCTION

Clinical decision support system (CDSS) is an effective assistant to physicians in diagnosis and treatment decision process during the patient care [1]. The intelligence of decision making systems depend on the latest and up-to-date knowledge [2]. The evolutionary and up-to-date knowledge base is highly influenced by creating, enhancing, managing, and disseminating the knowledge [3, 4]. However, the most prominent barrier of the CDSS adoption is knowledge integration and interoperability with workflows of hospital information systems (HIS) [5]. The standard data model virtual medical record (vMR) is a recommended unified model to achieve the objective of knowledge interoperability [6, 7]. However, the standard data model requires to bind its classes and attributes with one of the standard terminologies to enhance the level of interoperability. The standard terminologies (i.e. SNOMED CT, LOINC, and ICD10) can be aligned with vMR data model [8].

SNOMED CT is a well-known and comprehensive terminology, which is used by physicians worldwide [9]. It contains more than 0.3 million clinical concepts, which are easily understandable in medical institutions with multi-lingual support [10]. However, physicians are comfortable with use of local concepts instead of standard terminologies during knowledge creation due to high recall of concepts. A mapping system is needed to provide semantically interoperable mappings between the standard terminologies and localized concepts of domain clinical model (DCM). Therefore, we proposed a semantic reconciliation model, which provides semantic mappings between standard terminology SNOMED CT and localized concepts of DCM. The proposed semantic reconciliation model guarantees the semantic interoperability mappings with high accuracy.

The current mapping systems reflect the internal semantics of the source and target ontologies only and lack the external semantics during execution of mapping algorithms [11]. In the current context, the internal semantics mean that the relationships among the concepts of terminologies specified by the owner institutions based on their requirements, while the external semantics are the more generic relationships based on the standard terminologies. Therefore, in addition to the internal semantics, we inset the external semantics to the source and target concepts before the execution of matching algorithms. We transmute the short form of used acronyms, if exist, into their extended forms using acronyms library before the insertion of external semantics. The acronyms transmutation helps in enhancing the mapping accuracy up to some extent. In proposed mapping model, we inset the explicit semantics with the insertion of synonyms, hypernyms, hyponyms, and meronyms of the source and target concepts vectors. We used the existing two libraries, WordNet [12] and ConceptsNet5 [13], to inset the explicit semantics.

The proposed model allows to define the strategy for algorithms execution. The strategy may allow to execute all matching algorithms in sequential manner, or it may execute some of algorithms in defined sequence. The system provides different matching algorithms, such as string matching, label matching, child matching (means internal Childs of source and target), and property matching, these are well-known algorithms in ontological mappings. The proposed system is evaluated to map localized concepts of a DCM of our collaborative hospital with standard terminology SNOMED CT. The localized DCM covered the concepts of head and neck cancer. The system mapped the concepts of DCM with SNOMED CT with 95% of precision, 86.9% of recall, and 90% F-measure. The proposed system only focuses on high precision and recall of the mapping.

II. RELATED WORK

In the existing mapping systems, a number of matching algorithms are used, such as string matching, label matching, child matching as a linguistic and structural matching with different strategies. Peigang Xu et al. [14] have used differentor-based similarity matrix creation technique to incorporate multiple similarity measures. In this technique, weights are assigned to different entities of the source and target ontologies for accumulation tasks after finding the similarity measures. Falcon [11] provides fundamental

T.Ali, and S.Y.Lee {taqdir.ali, sylee}@oslab.khu.ac.kr
Department of Computer Science and Engineering, Kyung Hee University, Seocheon dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea, 446-701,
* Corresponding Author

technologies for finding, aligning and matching ontologies using a divide-and-conquer approach. This ontology matching tool has proved best results in Ontology Alignment Evaluation Initiative (OAEI) campaign. Although this system is still progressive in improvement of ontologies mappings using its matching techniques with help of user interface. The system lacks extendibility and reusability with respect to adoption of new matching techniques and algorithms.

Another ontology matching tool, called Agreement Maker [15, 16] proposed resolving the extendibility issue by displaying the ontologies and presenting the generated mappings for producing the ontology alignments. The developed system provides a flexible and extensible framework but it not scalable for matching the large scale ontologies. However, the authors extended the systems to Agreement Maker Light for resolving the scalability issue. GOMMA [17] provides a framework to manage the matching and evolution of ontologies and its influence on mappings. The system demonstrates high accuracy as compared to other systems, but it lacks expressive mapping representation. An ontology matching system, called YAM++ [18, 19] is developed to supports self-configuration, extensibility and combining multiple matchers. It discovers mappings using information-retrieval techniques and also deals with multilingual ontology-matching problems.

To summarize, existing terminology matching tools and approaches are unable to imitate a comprehensive system that exploits the explicit and implicit semantics during the execution of matching algorithms. The semantics inclusion into matching concepts enhances the accuracy of matching algorithms. Therefore, our proposed semantic reconciliation model overcomes the limitation of the existing matching approaches and insets semantics to enhance the matching accuracy up to better level.

III. Proposed Methodology

A. Methodology and architecture

The significant feature of semantic reconciliation methodology is the inclusion of semantics into the vectors of the source and target terminologies. The existing algorithms only focused on the internal semantics as child, siblings, and parents' similarity matching within the source and target terminologies. In addition to the external semantics, the proposed system insets the external semantics of the matching concepts in the form of their synonyms, hypernyms, hyponyms, and meronyms from semantically rich libraries. We used well-known libraries WordNet and ConceptsNet5 for including the external semantics. The system transmutes the short form of acronyms into their extended form using acronyms library. The DCM, developed for our collaborative hospital, contains acronyms for some concepts; acronyms highly effect the accuracy of matching algorithms. Therefore, the proposed system included the acronyms inset using the All-Acronyms library.

We designed and implemented a library to orchestrate multiple matching algorithms based on the selected algorithm execution strategy. The strategy can be selected to execute all matching algorithms sequentially or to execute some selected algorithms in a specific manner. Figure 1 demonstrates the reconciliation model to map the standard terminology and DCM. Following are the detail description of the reconciliation model.

Execution Control fetches the concepts of DCM and standard terminology SNOMED CT. When concepts exist in both of the terminologies then forwards to the *Explicit Semantic Inset* for insertion the external semantics to the matching concepts' vectors. In Explicit Semantic Inset, the concepts are preprocessed for the basic operations such as tokenize the concepts and remove the stop words, if exist; using the *Concepts Preprocessing* component. In second step, the *Acronym Inset* ($\forall$ *Acronym* $\exists$ *in* D_i || S) transmutes the short form of acronyms, if exist; to the corresponding extended forms using *All-Acronyms* library [20]. For instance, the concept "EB Virus" is transformed to "Epstein-Barr Virus". In third step, *Stem words Inset* transforms the concepts to their stem words. We focused on the stem transformation of nouns, verb, and adjectives because these parts of speech are mostly used in the terminology concepts.

In fourth step, we include the synonyms of the concepts using *Synonym Inset* component. We limited the synonym recursion depth up to first level due to performance with respect to handling the multi-dimensionality of the concepts. In fifth step, the semantics with respect to hypernyms are included into the concepts vector using *Hypernyms Inset* component. In case of hypernyms insertion, we extended the recursion depth up to two, because the multi-dimensionality cannot be effected by two hypernym concepts. In sixth step, the external semantics with respect hyponyms are inserted into the concepts by *Hyponyms Inset*. There is a chance of the existence of many Childs for the matching concepts. Therefore, the recursion depth is selected up to first level of child concepts and it regulates the multi-dimensionality to some extent. In final step of *Explicit Semantics Inset*, we included the meronyms of the matching concepts using *Meronyms Inset.* Meronyms may not be available for each matching concept; in unavailability case, the system discards the meronyms insertion. We selected the recursion depth as first level for the meronyms insertion.

After the insertion of explicit semantics, a strategy is built for the execution of matching algorithms. The system has two options to execute algorithms; a) execute all algorithms sequentially, b) execute only selected algorithms; from the *Matching Algorithms Library*. In the proposed system, we have String Matching, Label Matching, Child Matching, Parent Matching, and Property Matching algorithms. Based on the selected strategy, the system executes the algorithms. Each algorithm calculates the similarity score and provides it to the *Generate Mappings* component. This component evaluates and compares the similarity score with the defined threshold value "0.8", which is a recommended threshold value for concept matching in medical domain [21]. When the similarity score is greater than or equal to the threshold value then the concepts are considered as matched concepts, otherwise concepts are unmatched and system is executed for another iteration. The *Verify Mappings* component verifies the mapped concepts, when a single concept is mapped with multiple concepts of SNOMED CT then *Verify Mappings* gives alert to physicians for verification and final decision. Based on verification criteria, the mapped concepts are stored into the

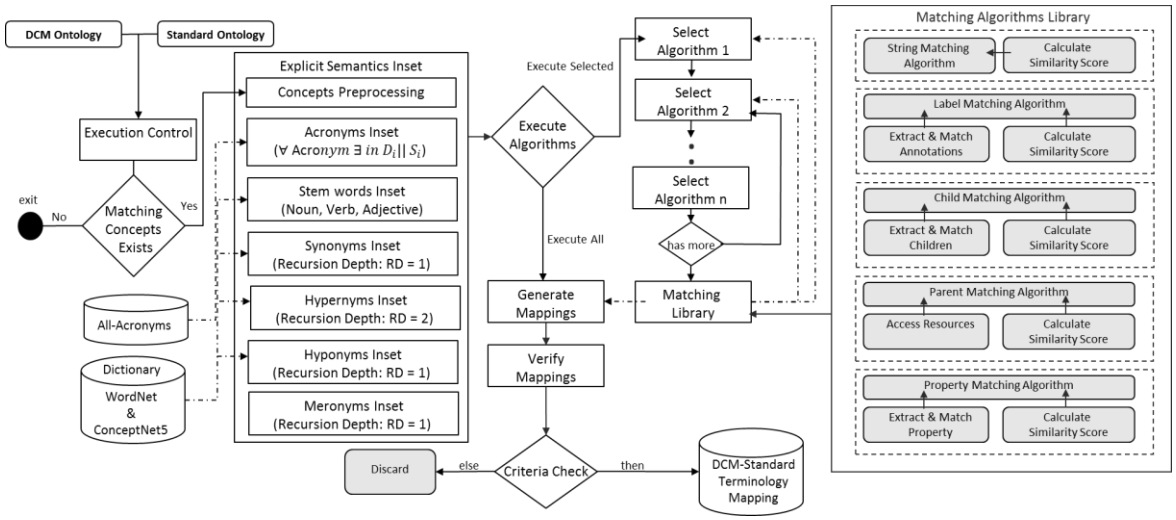

Figure 1: Semantic reconciliation model for mapping of terminologies

DCM-Standard Terminology Mapping repository otherwise the concepts are discarded.

B. *Scenario: Finding similarity score*

The proposed system insets the external semantics before the execution of matching algorithms and finding the similarity score. This approach enhanced the mapping accuracy and resolved the issue of multiple matched concepts. We extended the standard cosine similarity formula (Equation (1)), based on the semantics insertion.

$$\text{Cosine Similarity} = \frac{A \,.\, B}{|A|\,|B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad \ldots\ldots\ldots (1)$$

We added the union of stems, synonyms, hypernyms, hyponyms, and meronyms into Equation (1) and extended to Equation (2). As an example, one DCM concept "Smoking status" is mapped with three different SNOMED CT concepts "Smoking status at 4 week", "Smoking status at 52 week", and "Smoking monitoring status" according to the standard cosine similarity in Equation (1) with similarity score 0.816. The same similarity score for three concepts creates misperception in matching the exact concept. The proposed system resolved this misperception with the insertion of explicit semantics using Equation (2). Our approach calculated the similarity score 0.739 for "Smoking status at 4 week" and "Smoking status at 52 week", while 0.926 is calculated for "Smoking monitoring status". Therefore, it is easily distinguishable and it is considered as matched concepts. The inserted semantics are shown Figure 2.

IV. Results and Discussion

We evaluated the system using two datasets SNOMED CT ontology downloaded from the website of International Health Terminology Standard Development Organization (IHTSDO) [10] and DCM local terminology developed for our collaborative hospital. The SNOMED CT contains more than 0.3 million concepts and DCM comprises 199 concepts of head and neck cancer domain. The results' statistics are shown in Table 1. We calculated the precision, recall, and F-measure using corresponding standard formulas based on the values described in Table 1. Figure 3 showed the precision, recall, and F-measure.

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n}(A_i \cup Stem(A_i) \cup Syn(A_i) \cup Hyper(A_i) \cup Hypo(A_i) \cup Mero(A_i).(B_i \cup Stem(B_i) \cup Syn(B_i) \cup Hyper(B_i) \cup Hypo(B_i) \cup Mero(B_i)).}{\sqrt{\sum_{i=1}^{n}({A_i}^2 \cup Stem(A_i)^2 \cup Syn(A_i)^2 \cup Hyper(A_i)^2 \cup Hypo(A_i)^2 \cup Mero(A_i)^2)}\sqrt{\sum_{i=1}^{n}({B_i}^2 \cup Stem(B_i)^2 \cup Syn(B_i)^2 \cup Hyper(B_i)^2 \cup Hypo(B_i)^2 Mero(B_i)^2)}} \ldots\ldots (2)$$

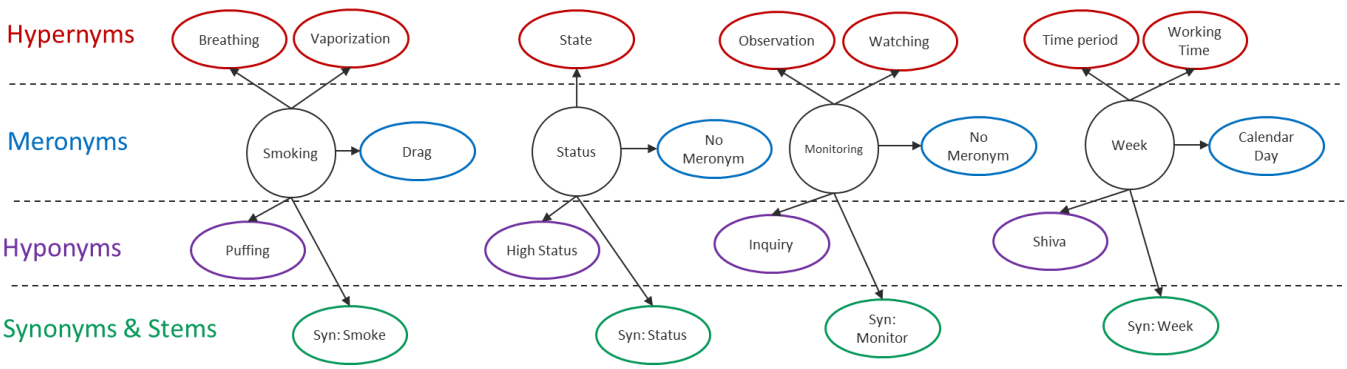

Figure 2: Example of external semantics insertion

TABLE 1: MAPPING STATISTICS OF SNOMED CT AND DCM

Description	Concepts
SNOMED CT concepts	More than 0.3 million
DCM concepts	199
Correctly mapped concepts	182
Incorrectly mapped concepts	9

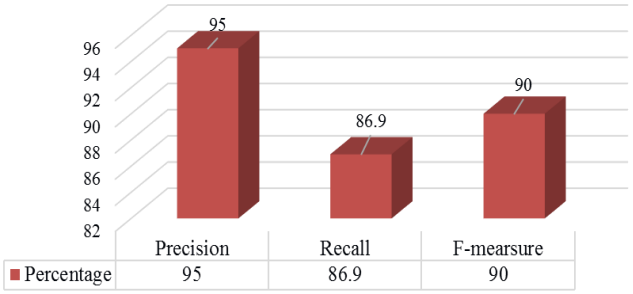

Figure 3: System evaluation results

The objective of our study is to achieve high precision and recall. The precision is highly effected by the regional concepts and some non-standard acronyms used in DCM. Some regional concepts related to drugs such as "naswar" and "paan", which only use in the specific region of our collaborative hospital. Similarly, some non-standard acronyms such as "S Proc 1" and "C S RT", are used in DCM by local physicians. Therefore, the non-standard acronyms and regional concepts do not exist in standard SNOMED CT terminology and it effected the precision and recall of the system.

V. CONCLUSION AND FUTURE WORK

In this study, we proposed a reconciliation model to align and reconcile the standard terminology concepts with localized concepts of domain clinical model (DCM). We observed that explicit and external semantics inclusion enhanced the accuracy of semantic reconciliation model. Therefore, we included the external semantics into the matching concepts vectors using existing concepts libraries. Similarly, the extended forms of acronyms also enhances the mapping accuracy.

In future, we will focus on mapping of regional and non-standard acronyms to increase the precision and recall of matching terminologies. We will evaluate the proposed system on a larger DCM dataset, compared to one employed in the current study.

ACKNOWLEDGEMENT

This work was supported by the Industrial Core Technology Development Program (10049079, Development of Mining Core Technology Exploiting Personal Big Data) funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea)

References

[1] A.Soumeya, D.Michel, R.Claire, B.Philippe and L.Eric. *"A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development."* Journal of the American Medical Informatics Association , vol 8, no. 4, pp. 351-360, 2012.

[2] K. Fehre, K.-P. Adlassnig, "*Service-oriented arden-syntax-based clinical decision support"*, Proceedings of eHealth2011. Vienna: Austrian Computer Society (2011) 123-8.

[3] D. F. Sittig, A. Wright, J. A. Oshero , B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, D. W. Bates, "*Grand challenges in clinical decision suppor*t", Journal of biomedical informatics 41 (2) (2008) 387-392

[4] D. Dustin, D. Jemery, B. Christopher, F. Simon and A. J. Mark. *"A Knowledge Authoring Tool for Clinical Decision Support"* Journal of Clinical Monitoring and Computing, vol 22, no. 3, pp. 189-198, 2008

[5] A. Wright, D. F. Sittig, A four-phase model of the evolution of clinical decision support architectures, International journal of medical informatics 77 (10) (2008) 641-649.

[6] K. Kawamoto, G. Del Fiol, H. R. Strasberg, N. Hulse, C. Curtis, J. J. Cimino, B. H. Rocha, S. Maviglia, E. Fry, H. J. Scherpbier, et al., "*Multinational, multi-institutional analysis of clinical decision support data needs to inform development of the hl7 virtual medical record standard*", in: AMIA Annual Symposium Proceedings, Vol. 2010, American Medical Informatics Association, 2010, p. 377.

[7] P. D. Johnson, S. W. Tu, M. Musen, I. Purves, *A virtual medical record for guideline-based decision support*., in: Proceedings of the AMIA symposium, American Medical Informatics Association, 2001, p. 294

[8] J. A. Oshero, J. M. Teich, B. Middleton, E. B. Steen, A. Wright, D. E. Detmer, "*A roadmap for national action on clinical decision support"*, Journal of the American medical informatics association 14 (2) (2007) 141-145.

[9] Z. Aleksovski, M. Sevenster, "*Identifying breast cancer concepts in snomed-ct using large text corpus"*, in: Electronic Healthcare, Springer, 2012, pp. 27-34

[10] Summary of SNOMED CT Benefits, *http://www.ihtsdo.org/snomedct/whysnomedct/benefits/* [Visited on January, 2017]

[11] Hu, W., & Qu, Y. (2008). "*Falcon-AO: A practical ontology matching system"*. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), 237-239.

[12] Miller, George A. "*WordNet: a lexical database for English*." Communications of the ACM 38.11 (1995): 39-41.

[13] Liu, Hugo, and Push Singh. "*ConceptNet—a practical commonsense reasoning tool-kit*." BT technology journal 22.4 (2004): 211-226.

[14] Xu P, Wang Y and Liu B. "*A differentor-based adaptive ontology-matching approach"*. Journal of Information Science 2012; 38(5): 459–475

[15] Cruz IF, Sunna W, Makar N and Bathal S. "*A visual tool for ontology alignment to enable geospatial interoperability"*. Journal of Visual Languages and Computing 2007; 18(3): 230–254

[16] Cruz IF, Antonelli FP and Stroe C. AgreementMaker: "*Efficient matching for large real-world schemas and ontologies"*. Proceedings of the VLDB Endowment 2009; 2(2): 1586–1589.

[17] Kirsten T, Gross A, Hartung M and Rahm E. "*GOMMA: A component-based infrastructure for managing and analyzing life science ontologies and their evolution"*. Journal of Biomedical Semantics 2011; 2: 6

[18] Ngo D, Bellahsene Z and Coletta R. "*A flexible system for ontology matching"*. In: IS Olympics: Information systems in a diverse world. Berlin: Springer, 2012, pp. 79–94

[19] Ngo D and Bellahsene Z. "*YAM++: A multi-strategy based approach for ontology matching task"*. In: Knowledge engineering and knowledge management. Berlin: Springer, 2012, pp. 421–425

[20] AllAcronyms 3,134,000 Acronyms and Abbreviations. 16,440 Topics, https://www.allacronyms.com/ [Visited on January, 2017]

[21] Christen, V., Groß, A., Varghese, J., Dugas, M., & Rahm, E. (2015, July). "*Annotating medical forms using UMLS"*. In International Conference on Data Integration in the Life Sciences (pp. 55-69). Springer International Publishing.