doi:10.3233/978-1-61499-808-2-190

Analysis of Annotated Data Models for Improving Data Quality

Hannes ULRICH [a,b,1], Ann-Kristin KOCK-SCHOPPENHAUER [a],
Björn ANDERSEN [b], Josef INGENERF [a,b]
[a] *IT for Clinical Research, Lübeck (ITCR-L), University of Lübeck, Germany*
[b] *Institute of Medical Informatics, University of Lübeck, Germany*

Abstract. The public Medical Data Models (MDM) portal with more than 9.000 annotated forms from clinical trials and other sources provides many research opportunities for the medical informatics community. It is mainly used to address the problem of heterogeneity by searching, mediating, reusing, and assessing data models, e. g. the semi-interactive curation of core data records in a special domain. Furthermore, it can be used as a benchmark for evaluating algorithms that create, transform, annotate, and analyse structured patient data. Using CDISC ODM for syntactically representing all data models in the MDM portal, there are semi-automatically added UMLS CUIs at several ODM levels like *ItemGroupDef, ItemDef,* or *CodeList item*. This can improve the interpretability and processability of the received information, but only if the coded information is correct and reliable. This raises the question how to assure that semantically similar datasets are also processed and classified similarly. In this work, a (semi-)automatic approach to analyse and assess items, questions, and data elements in clinical studies is described. The approach uses a hybrid evaluation process to rate and propose semantic annotations for under-specified trial items. The evaluation algorithm operates with the commonly used NLM *MetaMap* to provide UMLS support and corpus-based proposal algorithms to link datasets from the provided CDISC ODM item pool.

Keywords. CDISC ODM, UMLS, Semantic Interoperability, Natural Language Processing

1. Introduction

To prove effectiveness and efficiency of medication and medical therapies, clinical studies are performed. The results are documented in Case Report Forms (CRFs), which can easily consist of hundreds of documentation items, e. g. weight in kg. The Meta Data Models (MDM) Portal [1] developed by the Institute of Medical Informatics at the University of Muenster provides a huge collection of documentation items out of CRFs and routine care documentation with the aim to improve sharing and reuse of items. All forms and items are provided in the Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM) format [2], can be downloaded in various formats, and are publicly available. For the majority of data items, semantic annotations from the Unified Medical Language System (UMLS) were added [3].

[1] Corresponding Author: **Hannes.Ulrich@itcr.uni-luebeck.de**

Sharing and reuse of documentation items is deeply dependent on the understanding of the meaning of such an item. This meaning can be expressed using coding systems such as UMLS or SNOMED Clinical Terms (CT). Our approach aims for a method to check whether semantically similar datasets are processed and classified similarly. The MDM Portal therein serves as the source system for documentation items.

2. Methods

This work describes an automatic approach for the analysis and assessment of the semantic codes assigned to items, questions, and data elements. It uses a hybrid evaluation process to rate and propose semantic annotations for under-specified, non-annotated trial items. The implemented evaluation algorithm utilises the commonly used National Library of Medicine's (NLM) MetaMap API [4] to provide UMLS support and extends it by corpus-based proposal algorithms to link datasets from the provided CDISC ODM item pool.

2.1. Text and String Similarity

Information about the pairwise similarity of terms and sentences is required for further processing of text and plays an important role in the field of information retrieval and text classification. The given terms can either be lexically or semantically similar. The former denotes that the given inputs share a common sequence within their respective string representations, whereas the latter signifies whether the inputs represent the same cognitive concept and is more difficult to determine. The field of string similarity and the measurement thereof is well explored and is divided into three major types of algorithms: string-based, corpus-based, and knowledge-based approaches [5].
This approach proposes a similarity algorithm that combines two string-based measurements: the *five-gram* algorithm and the *metric Longest Common Subsequence* (mLCS) [6]. The first splits the given terms into subsequences of five characters and measures their similarity based on correspondence. The mLCS is based on the longest subsequence that both terms share. The subsequence differs therein from the substring that for the subsequence it is not mandatory to have the same position in the terms. In order to minimise false positives, a quality threshold rejects proposals with a low score.

2.2. Unified Medical Language System and MetaMap

UMLS is a linked collection of the majority of biomedical vocabularies. The NLM initiated the project in 1996 to help researchers to retrieve and integrate electronic biomedical information. One of the most well-known programmes for natural language processing of biomedical texts is *MetaMap*, which is also developed and published by the NLM. It provides access to UMLS by analysing a given biomedical text and mapping it to the corresponding concepts. Hence, a link between unstructured free text and the rich knowledge of UMLS is established, including all synonymy relationships and further references from other medical knowledge systems.

2.3. Processing Pipeline

In order to analyse given items, a processing pipeline was designed and implemented as a modular Java project, containing eight submodules with distinct functionality to minimise overlapping source code and to support reusability, see Figure 1.
In addition, a data pool was generated that contains 250 trial forms, provided by the MDM portal. The fields of the clinical forms are heterogeneous to broaden the applicability and reliability. The forms contain 4240 items and are annotated with 3291 different UMLS concepts. The most frequently used item describes the patients' age and the most frequent concept is "Date in time" as a temporal concept.

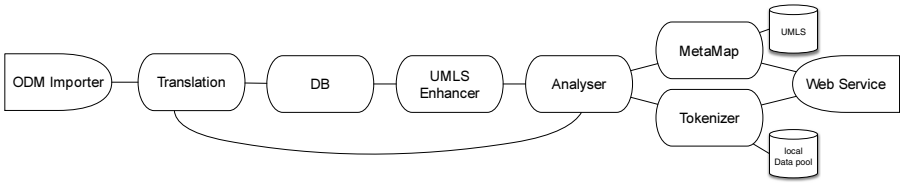

Figure 1 Central modules are the database access, ODM processing, UMLS querying, *Tokenizer*, and the web service. Special, customised services are the MetaMap access, *Translation*, and *Importer*.

The given ODM-XML file is read and split into sections of items grouped by their corresponding *ItemGroup.* The new trial items can enter the pipeline in two different manners: (1) in order to expand the data pool or (2) to be analysed. If they are to expand the information base, it is reviewed whether they are already known and can be linked. The UMLS querying module then enhances the item by adding UMLS semantic concept descriptors in addition to the corresponding Medical Subject Headings (MeSH) and SNOMED CT reference, if existing. The entered item is also checked for the languages it provides and is translated if no English translation is already given. The English version is necessary to use the *MetaMap* module and to be able to find further UMLS concepts to enhance the given item and therefore the quality of the data pool. The translation module benefits from Google's translation API. If the given item is to be analysed, two major criteria are determined: (1) Is the item or a related one already in the data pool and (2) does the item contain any UMLS references. Based on these criteria, the item is either reviewed or new UMLS concepts are proposed. Upon review, for the item and its included references it is tested as to whether *MetaMap* can confirm the connection or whether a related item in the curated pool contains the same references. For the concept proposal, the algorithm uses *MetaMap* to link the term and the previously described similarity measurement to determine related items in the database, see Figure 2. The entire data pool is examined by the analyser component and shown next to the pipeline health status, e. g. connection to the database, *MetaMap,* and the Google Cloud services.

```
PROPOSE to Allergies (text)

Tokenizer
Allergens( text )
[5289] C0002092 - Allergens
Allergies( text )
[5698] C0020517 - Allergies

MetaMap
[-1000] C0020517 - Allergies
```

Figure 2 Example output for the under-specified trial item *Allergies* of datatype *text:* The tokenizer finds similar items in the data pool and *MetaMap* generates a corresponding proposal.

3. Results

The herein described environment was successfully implemented and tested to measure the overall benefit of the approach. To evaluate the implementation, 25 trial forms containing 639 items were randomly chosen from the MDM portal and assessed. The system exhibits good performance at proposing UMLS concepts based on the similarity results of the tokenizer and *MetaMap*, as shown in Table 1.

Table 1 Proposal of under-specified trial items based on the pipeline.

	proposal		no proposal	total
	correct proposal	wrong proposal		
abs. number	64	0	17	81
ratio	79%	0%	21%	100%

The good hit ratio of 79% demonstrates that the system provides reliable proposals for the 81 under-specified trial items. After expert review, it was stated that the algorithm did not suggest a single wrong proposal, but for 17 items the system could not provide any proposal. The reasons are two-fold: the system could either not identify any concepts by processing the term or the results were not satisfying the quality threshold. The quality check of given 764 UMLS concepts in relation to the *item name* provided a 62% correspondence ratio, see Table 2.

Table 2 Evaluation of the accuracy of UMLS code and *item name*.

	correct match	incorrect match	total
CUIs	474	291	764
ratio	62%	38%	100%

The mismatch of 291 items has various reasons: additional domain-specific knowledge that could not be determined by the name, e. g. an item named “Date” has the additional concept “Date of Death”, the composition of long and rich item names, e. g. “maintenance treatment tablets Injection Infusion”, or names like “Unequivocalprogressivediseaseinnontargetlesionsisbasedon: (pleasedecribe)”.

4. Discussion and Conclusion

The system meets the specified requirements and provides good results in proposing and analysing the given trial items. However, during implementation and evaluation some limitations were discovered. The most challenging is the need for a curated and well-structured item pool to provide good and reliable results. But the acquisition of high-quality data sources is difficult: Whereas the MDM portal provides a reliably good data pool, additional resources – ideally covering a broad range of clinical disciplines – are needed for the similarity algorithm to process the majority of items.

The combination of mLCS, nGram, and quality thresholds is an improvement over previous approaches: It achieves better results than simple annotation comparison and Levenshtein distance [7]. Due to the introduction of a quality threshold, the results could be optimised and the output refined. The use of *MetaMap* to identify UMLS concepts corresponding to the given trial item increases the data quality significantly. The gained link between term and UMLS is an enabler for linking further medical

concept knowledge, e. g. relations to anatomy or drugs from *Snomed CT* or *MeSH.* However, MetaMap does not always provide suitable results for a given term.
Regarding UMLS, another problem was discovered: Older forms are annotated with concept codes that are not included in the current UMLS version, which raised errors processing them. As the given ODM forms do not contain any version information regarding UMLS, solving this conflict remains difficult [8]. The processing of non-English forms could be realised using the Google Translation API, but due to occurring inaccuracies an alternative approach would be desirable [9]. Nevertheless, the prototypical pipeline already supports the identification of candidates for an independent review of the quality and consistency of annotations.
Dugas et al. [3] reused known concepts from their repository to minimise the concept variability in large terminologies in their ODM data pool. But sometimes the concept choice appears to be arbitrary. For coding "Height" they prefer "Patient height" (C0005890) instead of the more general "Height" (C0489786). However, temperature items are most often annotated with "Temperature" (C0039476, i. e. in the sense of bio-specimen characteristics), and "Body Temperature" (C000590). In general, the pipeline based on two different knowledge bases yields good and reliable results. The modular software design enables the integration into existing study and metadata repositories as well as the reuse of the developed tools. The semi-automatic approach can accelerate the process of curating item annotations, e. g. by externally and independently enhancing the repository provided by the MDM portal [10]. A predicted mismatch shall not perturb, but encourage the user to review the items and either modify or rather improve the actual items and annotated concepts.

5. Conflict of Interest/Acknowledgment

The authors state that they have no conflict of interests. The authors express their gratitude to Martin Dugas and the MDM portal team for their kind support!

References

[1] Dugas M: „Metadata Repository for Medical Forms Portal". *Medical-data-models.org*. Retrieved from http://www.medical-data-models.org (17. 03. 2017).
[2] Operational Data Model (ODM)-XML". Retrieved from https://www.cdisc.org/standards/foundational/odm. (17. 03. 2017)
[3] Dugas M, Neuhaus P, Meidt A, Doods J, et al. (2016): „Portal of medical data models: information infrastructure for medical research and healthcare". *Database (Oxford)*. 2016 10.1093/database/bav121.
[4] MetaMap - A Tool For Recognizing UMLS Concepts in Text". *Metamap.nlm.nih.gov*. Retrieved from https://metamap.nlm.nih.gov/. (17. 03. 2017)
[5] Gomaa WH, Fahmy AA: „A Survey of Text Similarity Approaches". In: *International Journal of Computer Applications*. 68 (13), 2013), 13-18.
[6] Bakkelund D. "An LCS-based string metric" *University of Oslo* (2009).
[7] Ulrich, H., et al. "Metadata Repository for Improved Data Sharing and Reuse Based on HL7 FHIR." S*tud Health Technol Inform.* 2016, 228: 162-6.
[8] Seerainer C, Sabutsch SW: "eHealth Terminology Management in Austria" in *Studies in Health Technology and Informatics 2016, 228: 426-30*
[9] Schlegel DR, Crowner C, Elkin PL: "Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia". S*tud Health Technol Inform.* 2015; 216:619-23.
[10] Varghese J, Dugas M: Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. *Methods Inf Med.* 2015; 54(1):83-92.