Approaches for Annotating Medical Documents

Victor Christen, Anika Groß, Erhard Rahm

Department of Computer Science, University of Leipzig, Germany
{christen,gross,rahm}@informatik.uni-leipzig.de

Abstract. Annotations are useful to semantically enrich documents and other datasets with concepts of ontologies. In the medical domain, many documents are not annotated at all and manual annotation is a difficult process making automatic annotation methods highly desirable to support human annotators. We propose a linguistic-based and a reuse-based approach annotating medical documents by concepts from an ontology. The reuse-based approach utilizes previous annotations to annotate similar medical documents. The approach clusters items in documents such as medical forms according to previous ontology-based annotations and uses these clusters to determine candidate annotations for new items.

1 Introduction

The annotation of data with concepts of standardized vocabularies and ontologies has gained increasing significance due to the huge number and size of available datasets as well as the need to deal with the resulting data heterogeneity. Annotations of medical documents such as Electronic Health Records (EHR) that are used to document the history of patients can also support advanced analyses and searches. For instance, they can be used to identify significant co-occurrences between the use of certain drugs and negative side effects in terms of occurring diseases [5]. Moreover, case report forms are used for examining clinical trials, e.g. to ask for the medical history of probands. To enable an efficient search for medical documents, annotations can be used to semantically look for a certain set of forms, e.g., in the MDM repository of medical data models [2] and to design new forms with a similar topic.

To improve the value of medical documents for analysis, reuse and data integration it is thus crucial to annotate them with concepts of ontologies. Since the number, size and complexity of medical documents and ontologies can be very large, a manual annotation process is time-consuming or even infeasible. Hence, automatic annotation methods become necessary to support human annotators with recommendations for manual verification. The goal of an annotation method is the identification of annotations for a collection of medical documents . An annotation is an assocoitation between a document and a concept from an ontology, where the concept covers the semantics of the document. Therefore, a document might be annotated with more than one concept to precisely describe the content of the document. The use of annotations enables a standardized representation, since an ontology is a unified set of concepts and a set of relationship interrelating the ontology concepts by certain relationship types, e.g. $is-a$, $part-of$ or domain-specific relationships such as $is-located-in$. The annotation of documents by using concepts of an ontology is related to the entity-linking problem that is a well studied field [6]. Moreover, there exist different annotation methods such as MetaMap [1] that annotates medical documents with concepts of UMLS by applying a linguistic-based approach.

In our recent work, we realized different annotation methods to identify annotations for medical forms based on concepts of UMLS. We initally start with a linguistic-based annotation approach [4]. A crucial part of an annotation method is the identification of annotation candidates in terms of effectivness and efficiency. In general, a medical document or a collection of medical documents cover topically a subset of an ontology. Moreover, the quality of annotation candidates depends on the quality of synonyms and labels for a concept. We overcome such issues by creating a reuse repository for utilizing verified annotated documents [3]. We are able to build more compact and preciser representatives for a concept based on the verified documents than the synonyms and labels for a concept. Morover, the reuse of the genenerated representatives to annotate a set of medical documents is more efficient than using the whole ontology.

2 Linguistic-Based Annotation Approach

The workflow consists of a preprocessing, a candidate identification and a selection step (see Fig. 1). The input of the workflow is a set of forms $\mathcal{F}$, an ontology $\mathcal{O}$, and a similarity threshold δ. This kind of documents consists of a set of question that we want to annotate with a set of concepts. In our case, we use concepts from the Unified Medical Language System (UMLS) that is an integrated knowledge system including several biomedical ontologies. First, we normalize the labels and synonyms of ontology concepts by removing stop words, transforming all string values to lower case and removing delimiters. The same preprocessing steps are applied for each form F_i. We identify an intermediate annotation mapping $\mathcal{M}'_{F_i,\mathcal{O}}$ by lexicographically comparing each question with the labels and synonyms of ontology concepts. For this purpose, we apply three string similarity measures, namely trigram, TF/IDF as well as a longest common sequence string similarity approach. We keep an annotation (q, c, sim) for a question q and a concept c, if the maximal similarity sim of the three string similarity approaches exceeds the threshold δ. Finally, we select annotations from the intermediate result by not only choosing the concepts with the highest similarity but also by considering the similarity among the concepts. For

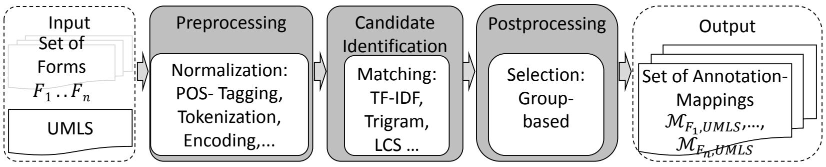

Input
Set of Forms
$F_1 .. F_n$
UMLS
Preprocessing
Normalization: POS- Tagging, Tokenization, Encoding,...
Candidate Identification
Matching: TF-IDF, Trigram, LCS ...
Postprocessing
Selection: Group-based
Output
Set of Annotation-Mappings
$\mathcal{M}_{F_1,UMLS},...,$
$\mathcal{M}_{F_n,UMLS}$

Fig. 1. Workflow of the linguistic-based annotation approach

this purpose, we group the concepts associated with a question based on their mutual similarity and only choose the concept with the highest similarity per group in order to avoid the redundant selection of highly similar concepts. This group-based selection proved to be quite effective in [4] albeit it only considers the string-based (linguistic) similarity between questions and concepts, and among concepts.

3 Reuse-based Annotation Approach

The workflow for the reuse-based annotation approach is shown in Figure 2. Its input includes a set of verified annotation mappings containing the annotations for reuse. The result is a set of annotation mappings $\mathcal{M}_{\mathcal{F},\mathcal{O}}$ for the unannotated input forms $\mathcal{F}$ w.r.t. ontology $\mathcal{O}$. In the first step, we use the verified annotations to determine a set of *annotation clusters* $\mathcal{AC} = \{ac_{c_1}, ac_{c_2}, ..., ac_{c_m}\}$. For each concept c_i used in the verified annotations, we have an annotation cluster ac_{c_i} containing all questions that are associated to this concept. To calculate the similarity between an unannotated question and the questions of an annotation cluster we determine for each cluster a *representative* (feature set) $ac_{c_i}^{fs}$ consisting of relevant term groups in this cluster. A relevant term group is either a frequently co-occuring term group in the questions of the cluster or the maximized overlap between the terms of a question and the synonyms or the label of a concept, i.e., we do not use term groups that build a subset of another frequently occurring term group. As an example, Figure 3 shows the resulting annotation cluster $ac_{C0023467}$ for UMLS concept *C0023467* about the disease *Acute Myeloid Leukaemia*. In the UMLS ontology, this concept is described by a set of 32 synonyms (Figure 3 left). The annotation cluster also contains 25 questions associated to this concept in the verified annotation mappings. Most questions only relate to some of the synonym terms of the concept while other synonyms remain unused. So the abbreviation 'AML' that is a part of some synonyms is often used but the abbreviation 'ANLL' does not occur in the medical forms used to build the annotation clusters. For this example, we generate only 9 relevant term groups, i.e., the representative feature set of the cluster is much more compact than the free text questions and large synonym set.

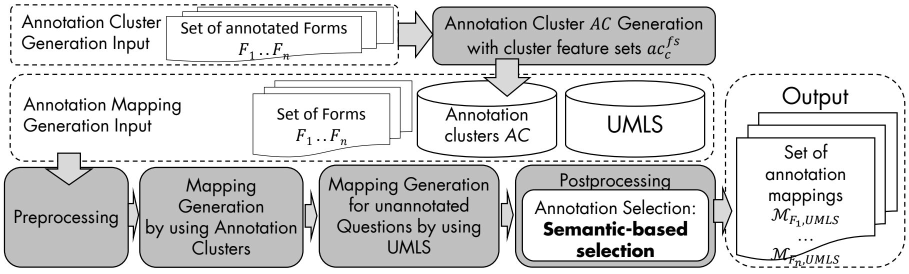
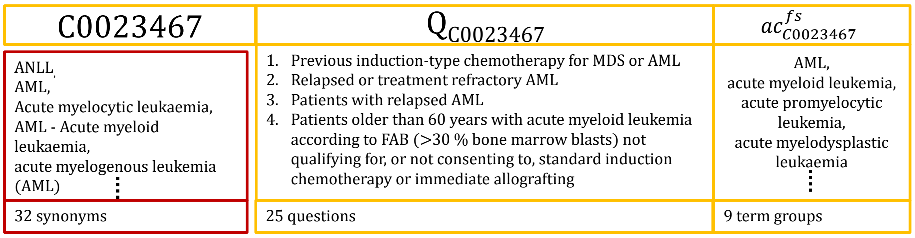

Annotation Cluster Generation Input
Set of annotated Forms $F_1..F_n$
Annotation Cluster AC Generation with cluster feature sets ac_c^{fs}
Annotation Mapping Generation Input
Set of Forms $F_1..F_n$
Annotation clusters AC
UMLS
Output
Set of annotation mappings $\mathcal{M}_{F_1,UMLS}$ … $\mathcal{M}_{F_n,UMLS}$
Preprocessing
Mapping Generation by using Annotation Clusters
Mapping Generation for unannotated Questions by using UMLS
Postprocessing
Annotation Selection: **Semantic-based selection**

Fig. 2. Workflow of the reuse-based annotation approach



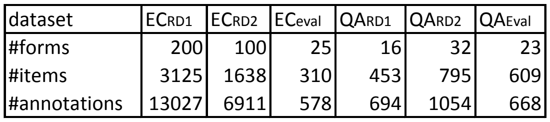
C0023467	$Q_{C0023467}$	$ac^{fs}_{C0023467}$
ANLL, AML, Acute myelocytic leukaemia, AML - Acute myeloid leukaemia, acute myelogenous leukemia (AML) ⋮	1. Previous induction-type chemotherapy for MDS or AML 2. Relapsed or treatment refractory AML 3. Patients with relapsed AML 4. Patients older than 60 years with acute myeloid leukemia according to FAB (>30 % bone marrow blasts) not qualifying for, or not consenting to, standard induction chemotherapy or immediate allografting	AML, acute myeloid leukemia, acute promyelocytic leukemia, acute myelodysplastic leukaemia ⋮
32 synonyms	25 questions	9 term groups

Fig. 3. Sample annotation cluster $ac_{C0023467}$ for UMLS concept *C0023467* with its set of associated questions $Q_{C0023467}$ and feature set $ac^{fs}_{C0023467}$.

After these initial steps we determine the annotation mapping for each unannotated input form F_i. We first preprocess a form and the ontology as in the base approach (see Fig. 1). Then we determine an annotation mapping $\mathcal{M}^{Reuse}_{F_i,\mathcal{O}}$ for the form based on the annotation clusters. Depending on the degree of reusable annotations the determined mapping is likely to be incomplete. We thus identify all questions that are not yet covered by the first mapping. For these questions we apply the base algorithm to match them to the whole ontology and obtain a second annotation mapping. We then take the union of the two partial mappings to obtain the intermediate mapping $\mathcal{M}'_{F_i,\mathcal{O}}$. Finally, we apply a context-based selection strategy to determine the annotations for the final mapping $\mathcal{M}_{\mathcal{F},\mathcal{O}}$. The input for the selection of annotations is a set of grouped candidate concepts for each question in the medical forms $\mathcal{F}$. To determine the final annotations per question, we rank the candidate concepts within each group based on a combination of both linguistic and context-based similarity among the candidate concepts. For this purpose, we consider two criteria for a set of candidate concepts of a certain question: first, the degree to which concepts co-occurred in the annotations for the same question within the verified annotation mapping, and second, the degree of semantic (contextual) relatedness of the concepts w.r.t. the ontological structure. The goal is to give a high contextual similarity (and thus a high chance of being selected) to frequently co-occurring concepts and to semantically close concepts. To determine a context-based similarity, we construct a *context graph* $G_q = (V_q, E_q)$ for each question q. The vertices V_q represent candidate concepts that are interconnected by two kinds of edges in E_q to express that concepts have co-occurred in previous annotations or that concepts are semantically related within the ontology. In both cases we assign distance scores to the edges that will be used to calculate the context similarity between concepts.

4 Evaluation

We evaluate the proposed annotation approaches for medical forms and compare it with the MetaMap tool. Our evaluation uses medical forms about eligibility criteria (EC) and about quality assurance (QA) w.r.t cardiovascular procedures from the MDM platform [2]. To evaluate the quality of automatically generated annotations, we use manually created reference mappings from the MDM portal. These reference mappings might not be perfect (”a silver standard”) since the huge size of UMLS makes it hard to manually identify the most suitable concepts for each item. To analyze the quality of the resulting annotation mappings, we compute precision, recall and F-measure using the union of all annotated form items in the evaluation dataset. Table 4 shows the number of forms, items and verified annotations for the reuse and evaluation datasets.



dataset	ECRD1	ECRD2	ECeval	QARD1	QARD2	QAEval
#forms	200	100	25	16	32	23
#items	3125	1638	310	453	795	609
#annotations	13027	6911	578	694	1054	668

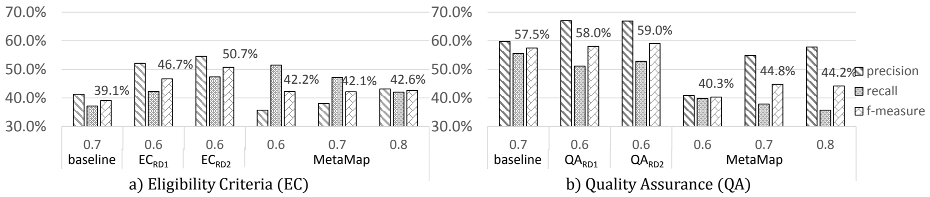
Table 4. Statistics on the reuse and evaluation datasets for EC and QA

Figure 5 shows the results for the two datasets and different configurations. Our reuse-based approach outperforms MetaMap in terms of mapping quality for each dataset. For the EC dataset, F-Measure is improved by ∼4%(EC_{RD1}) and ∼8.6% (EC_{RD2}) indicating that the the computed annotation clusters allow a more effective identification of annotations than with the original concept definition. In addition, our approach benefits from using the ontological relationships for selecting annotations resulting in a much better precision than using MetaMap (54.5% for EC_{RD2} than compared to 43.1%). While MetaMap achieved a better F-Measure than the baseline approach for the EC dataset it performed poorly for the QA dataset where its best F-Measure of 44.8% was much lower for the baseline approach and reuse-based approaches (57.5 and 59%), mainly because of a very low recall for Metamap.

A positive side of MetaMap is its high performance due to the use of an indexed database for finding annotations. Its runtimes were up to 13 times faster than for the baseline approach and it was also faster than the reuse-based approach. In future work we will study whether the use of MetaMap in combination with the reuse approach, either as an alternative or in addition to the baseline approach, can further improve the annotation quality.

5 Conclusion

We proposed a linguistic-based and a reuse-based approach to semantically annotate medical documents such as EHRs with concepts of an ontology. The linguistic-based approach identifies an annotation mapping between a form and an ontology by comparing each question of the form with the synonyms or labels of each concept from an ontology. The reuse-based approach avoids the comparison of each concept by utilizing already found and verified annotations for similar CRFs. It builds so-called annotation clusters combining all previously annotated questions related to the same medical concept. New questions are matched with the identified cluster representatives to find candidates for annotating concepts.



a) Eligibility Criteria (EC)

b) Quality Assurance (QA)

Fig. 5. Comparison of the quality for the resulting annotation mappings from the baseline approach, reuse-based approach and MetaMap.

To identify the most promising annotations, we proposed a context-based selection strategy based on the semantic relatedness of concept candidates as well as known co-occurrences from previous annotations. We compared our approaches with MetaMap and showed that the reuse-based approach outperforms the annotation method of MetaMap in terms of quality. However, the efficiency is lower than MetaMap, since it uses an indexed database.

For future work, we plan to use different annotation frameworks for generating more candidates and to get more evidence for correctness. We also plan to build a reuse repository covering annotation clusters and their feature sets for different medical subdomains. Such a repository can be used to identify annotations for new medical documents. It further enables a semantic search for existing medical document annotations. This can be useful to define new medical forms by finding and reusing suitable annotated items instead of creating new forms from scratch.

References

1. A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proc. AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
2. B. Breil, J. Kenneweg, F. Fritz, et al. Multilingual medical data models in ODM format–a novel form-based approach to semantic interoperability between routine health-care and clinical research. *Appl Clin Inf*, 3:276–289, 2012.
3. V. Christen, A. Groß, and E. Rahm. A reuse-based annotation approach for medical documents. In *Submmited for: International Semantic Web Conference(ISWC)*, 2016.
4. V. Christen, A. Groß, J. Varghese, M. Dugas, and E. Rahm. Annotating medical forms using UMLS. In *Data Integration in the Life Sciences (DILS)*, volume 9162 of *LNCS*, pages 55–69. 2015.
5. P. LePendu, S. Iyer, C. Fairon, N. H. Shah, et al. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics*, 3(S-1):S5, 2012.
6. W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.