# A Reuse-based Annotation Approach for 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 standardized vocabularies and 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 reuse-based annotation approach that 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. The final annotations are selected according to a new context-based strategy that considers the co-occurrence and semantic relatedness of annotating concepts. The evaluation based on previous UMLS annotations of medical forms shows that the new approaches outperform a baseline approach as well as the use of the MetaMap tool for finding UMLS concepts in medical documents.
**Keywords:** Semantic annotation, medical documents, ontology, UMLS.

## 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. In the biomedical domain, gene or protein functions are thus often described by concepts of the Gene Ontology(GO) [2], scientific publications can be annotated with Medical Subject Headings (MESH) [14], and electronic health records (EHRs) can be semantically classified by concepts of SNOMED CT [7]. Annotations of medical documents such as EHRs can also support advanced analyses, e.g. significant co-occurrences between the use of certain drugs and negative side effects in terms of occurring diseases [12]. Still many medical documents are not annotated at all, impeding data analysis and data integration. For instance, more than 200.000 trials are registered on `http://clinicaltrials.gov` and every study requires a set of so-called case report forms (CRFs), e.g. to ask for the medical history of probands. For every new clinical trial, CRFs are usually built from scratch, although previous forms might already cover similar topics. CRF annotations are helpful to search for existing form collections, e.g., in the MDM repository of medical data models [4].

**Question**

*Confirmed*(1) *diagnosis*(2) of *AML*(3) according to the WHO definition (*except(4)* for acute *promyelocytic leukaemia, APL*(5))

○ **yes** ○ **no**

**Annotations**

| | | |
|---|---|---|
| **1** | C0750484 | label:confirmation<br>synonyms: confirmatory, confirm |
| **2** | C0011900 | label: diagnosis (observable entity)<br>synonyms: diagnostic, diagnosis (DX) ; DX ;... |
| **3** | C0023467 | label: AML - acute myeloid leukaemia<br>synonyms: acute myeloid leukaemia ; acute granulocytic leukaemia ;ANLL; ... |
| **4** | C1554961 | label: exception |
| **5** | C0023487 | label: acute promyelocytic Leukemia<br>synonyms: APL; acute myeloid leukaemia, PML/RAR-alpha;... |

**Fig. 1.** Example medical form item and associated annotations to UMLS concepts.

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. Figure 1 shows an exemplary annotation for one item in a medical form (CRF) on eligibility criteria for a clinical trial on acute myeloid leukaemia (AML). Such an item comprises a question as well as a response field or a list of answer options. The shown question has been manually annotated based on a reference mapping with five concepts of the Unified Medical Language System (UMLS) [3], a comprehensive knowledge base integrating many biomedical ontologies. The associated UMLS concepts relate to different terms of the item text (italicized) as indicated by the numbers (1) to (5).

The automatic annotation of medical documents is challenging for several reasons. In particular, it is difficult to correctly identify relevant terms and medical concepts within natural language sentences such as the items (questions) occurring in medical forms. This is because concepts typically have several synonyms that may occur in sentences in different variations. Furthermore, concepts are often described by labels or synonyms consisting of several words, e.g., *AML-Acute myeloid leukaemia* (*C0023467*), that can match many irrelevant terms in the items to be annotated. We might further need to identify complex many-to-many mappings between items and ontology concepts without knowing a priori how many medical concepts should be associated per item. Moreover, UMLS is very large (2.8 mio. concepts) making it difficult to identify the best fitting concepts for annotation.

We recently proposed already an initial approach to annotate medical forms with UMLS concepts by extracting terms from items and matching these terms to UMLS concepts based on linguistic ontology matching techniques [5]. The study revealed the mentioned challenges and showed the difficulty of automatically achieving high quality annotations especially for long natural language sentences. Moreover, we observed frequent errors due to the high number of available concept synonyms and misleading terms in synonyms. In this study we aim at improving the quality of annotations and reducing the manual anno-

tation effort by reusing already determined and manually verified annotations. This assumes that there are similar questions in different medical forms of a domain of interest so that previous annotations can be reapplied. For this purpose, we propose and evaluate a new reuse-based annotation approach for annotating medical forms and documents.

Specifically, we make the following contributions:

- To enable annotation reuse, we propose to cluster all previously annotated items that are annotated with the same medical concept. For such annotation clusters, we identify representative features that are more compact than the large set of terms in concept labels and synonyms. We use these clusters and their features to find likely annotations for new items that are similar to already annotated ones.
- We propose a new context-based strategy to select the most promising annotations from a set of previously determined candidates. The strategy considers both the semantic relatedness of the annotating concepts as well as their co-occurrence in previously annotated items.
- We evaluate the proposed approaches based on reference mappings between a set of medical forms and UMLS and compare them with a baseline annotation approach as well as with using the MetaMap tool [1] to identify UMLS concepts within medical documents.

The remainder of this paper is organized as follows. We first provide a more formal problem definition and introduce a base workflow for determining annotations (Sec. 2). We then propose our new reuse-based annotation approach and the context-based selection strategy (Sec. 3). Sec. 4 presents evaluation results for the new approaches. Finally, we discuss related work in Sec. 5 and conclude in Sec. 6.

## 2 Preliminaries

We first present the formal definition of the annotation problem we address. Next we present a base workflow to determine annotation mappings for medical forms. This workflow has already been proposed in [5] and serves as a basis for our new approach that can reuse previous annotations (Sec. 3).

### 2.1 Problem Definition

We are given a set of medical forms $\mathcal{F}$ and an ontology $\mathcal{O}$. Each form $F \in \mathcal{F}$ consists of a set of items $\{i_1, i_2, ...i_k\}$ where each item has a question $q$ and a response part. The response may be provided as free text or by selecting an answer from a list of possible values (as in Figure 1). While the list of possible answers may include valuable information for the annotation of items, in this work we concentrate on using the question parts for finding suitable annotations. An ontology $\mathcal{O}$ consists of a set of concepts $C_\mathcal{O} = \{c_1, c_2, ..., c_l\}$ and a set of relations $R_\mathcal{O} = \{(c_1, c_2, rel\_type_1), ...(c_i, c_j, rel\_type_k)\}$ interrelating the ontology

concepts by certain relationship types, e.g. $is-a$, $part-of$ or domain-specific relationships such as $is-located-in$. The concepts in $\mathcal{O}$ are typically described by an id, a label and several synonyms as shown on the right side of Figure 1. The goal is to annotate each question (item) with one or several concepts from the given ontology $\mathcal{O}$. More specifically, we aim to determine an annotation mapping $\mathcal{M}_{F,\mathcal{O}} = \{(q, c, sim) | q \in F, c \in \mathcal{O}, sim[0,1]\}$ for each form $F$. An annotation $(q, c, sim)$ in these mappings indicates that question $q$ is semantically described by concept $c$; the similarity value $sim$ indicates the strength of the association according to the underlying method to compute the annotations.

Note that a question may be annotated by several concepts and that a concept may describe several questions. The challenge is to develop automatic methods that can determine annotation mappings of good quality (recall, precision). Ideally, all questions are correctly annotated, i.e. they are annotated with the ontology concepts that provide the best semantic description for the questions. A secondary goal is to efficiently determine the annotation mappings in a short time, even for large form collections and large ontologies.

**Algorithm 1:** annotation method $\mathcal{A}$

**Input**: Set of forms $\mathcal{F}$, ontology $\mathcal{O}$= $(C_\mathcal{O}, R_\mathcal{O})$, threshold $\delta$
**Output**: Annotation mapping $\mathcal{M}_{\mathcal{F},\mathcal{O}}$

```
1 O←preprocess (O);
2 M_{F,O}← ∅ ;
3 foreach F_i ∈ F do
4     F_i ← preprocess (F_i);
5     M'_{F_i,O} ←identifyCandidates (F_i, C_O, δ);
6     M_{F_i,O} ← selectAnnotations (M'_{F_i,O});
7     M_{F,O}←M_{F,O}∪M_{F_i,O};
8 return M_{F,O};
```

### 2.2 Base Workflow

In our previous work [5] we used the basic workflow shown in Algorithm 1 to determine annotation mappings for medical forms. The input of the workflow is a set of forms $\mathcal{F}$, an ontology $\mathcal{O}$, and a similarity threshold $\delta$. First, we normalize the label 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 label 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 $(q, c, sim)$, if the maximal similarity of the three string similarity approaches exceeds the threshold $\delta$. 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 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 [5] albeit it only considers the string-based (linguistic) similarity between questions and concepts, and among concepts.

## 3 Reuse-based Annotation Approach

In this section we outline an extended workflow to determine annotation mappings that reuses previously found annotations for similar questions. The goal is to reduce the complexity of the annotation problem by avoiding to search a very large ontology for finding concepts that describe or match terms of a question to annotate. By reusing verified annotations we also hope to achieve a good annotation quality since the previous annotations may include concepts that are difficult to find by common match techniques based on linguistic similarity. The reuse approach is also motivated by the existence of a high number of related forms in a specific domain, e.g. dealing with a specific disease. It would thus be desirable to reuse the annotations of a subset of these forms to more quickly and effectively annotate the remaining ones. The proposed approach is not limited to the annotation of medical forms but could be generalized for other medical documents such as electronic health records (EHRs) where we would associate medical concepts from an ontology to specific sentences or sections of the document rather than to questions.

We will first outline the new workflow for reuse-based annotation and then provide more details about its main steps, i.e., the generation of so-called annotation clusters (Sec. 3.2), determination of candidate annotations (Sec. 3.3) and a context-based strategy for selecting the final annotations (Sec. 3.4).

### 3.1 Workflow for Reuse-based Annotation

The workflow for the reuse-based annotation approach is shown in Algorithm 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 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. These term groups are identified based on common terms between the questions $q \in ac_{c_i}$ and the description (label and synonyms) of the corresponding concept of $ac_i$.

After these initial steps we determine the annotation mapping for each unannotated input form $F_i$ (lines 3-7 in Algorithm 2). We first preprocess a form as in the base approach of Algorithm 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 (line 7). 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 new strategy to select the annotations for the final mapping $\mathcal{M}_{\mathcal{F},\mathcal{O}}$. This selection strategy considers the context of concepts, their linguistic similarity as well as their co-occurrences in previous annotations.

**Algorithm 2:** Extended annotation method $\mathcal{A}^{reuse}$

**Input**: Set of unknown forms $\mathcal{F}$, ontology $\mathcal{O}$= $(C_\mathcal{O}, R_\mathcal{O})$, set of verified annotation mappings $\mathcal{M}^{verified}_{\mathcal{F},\mathcal{O}}$, threshold $\delta$
**Output**: Annotation mapping $\mathcal{M}_{\mathcal{F},\mathcal{O}}$

1 $\mathcal{AC} \leftarrow$`determineAnnotationCluster` $(\mathcal{M}^{verified}_{\mathcal{F},\mathcal{O}})$ ;
2 $\mathcal{AC} \leftarrow$ `determineFeatureSets` $(\mathcal{AC}, \mathcal{O})$;
3 $\mathcal{O} \leftarrow$`preprocess` $(\mathcal{O})$;
4 **foreach** $F_i \in \mathcal{F}$ **do**
5 $\quad F_i \leftarrow$ `preprocess` $(F_i)$;
6 $\quad \mathcal{M}^{Reuse}_{F_i,\mathcal{O}} \leftarrow$ `identifyCandidatesByReuse` $(F_i, \mathcal{AC}, \delta)$;
7 $\quad F'_i \leftarrow$ `findUnannotatedQuestions` $(F_i, \mathcal{M}^{Reuse}_{F_i,\mathcal{O}})$;
8 $\quad \mathcal{M}^{reduced}_{F'_i,\mathcal{O}} \leftarrow$ `identifyCandidates` $(F'_i, \mathcal{O}, \delta)$;
9 $\quad \mathcal{M}'_{F_i,\mathcal{O}} \leftarrow \mathcal{M}^{reduced}_{F'_i,\mathcal{O}} \cup \mathcal{M}^{Reuse}_{F_i,\mathcal{O}}$;
10 $\quad \mathcal{M}_{F_i,\mathcal{O}} \leftarrow$ `selectAnnotationsByContext` $(\mathcal{M}'_{F_i,\mathcal{O}})$;
11 $\quad \mathcal{M}_{\mathcal{F},\mathcal{O}} \leftarrow \mathcal{M}_{\mathcal{F},\mathcal{O}} \cup \mathcal{M}_{F_i,\mathcal{O}}$;
12 **return** $\mathcal{M}_{\mathcal{F},\mathcal{O}}$;

### 3.2 Generation of Annotation Clusters and Representatives

We build annotation clusters from verified annotation mappings by creating a cluster for each applied ontology concept $c_k$ and associating to it all input questions that are assigned to this concept. Formally, an annotation cluster $ac_{c_k}$ is represented as triple:

$$ac_{c_k} := (c_k, Q_{c_k}, ac^{fs}_{c_k}).$$

It includes the concept $c_k$, the set of questions $Q_{c_k}$ annotated with $c_k$, as well as a cluster representative or feature set $ac^{fs}_{c_k}$. The purpose of the cluster representative is to provide a compact cluster description that is more suitable for finding further annotations than the free text questions or the label and synonym terms of the ontology concept.

A feature set is formed by terms or groups of terms that frequently co-occur in the questions of the cluster and that are similar to the synonym description of the corresponding concept. To identify frequently co-occurring terms, we use a frequent itemset mining algorithm where the frequency of term groups has to exceed a given *min_support*. Moreover, we only keep term groups that maximize

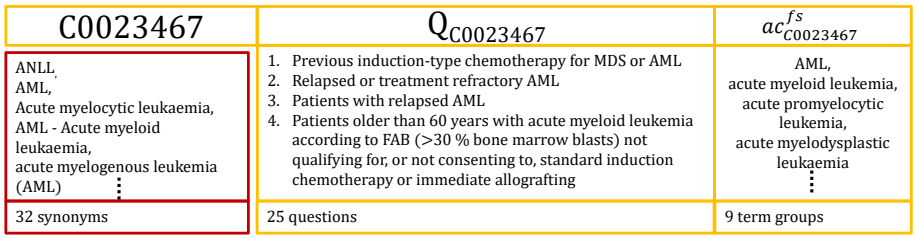


**Fig. 2.** Sample annotation cluster $ac_{C0023467}$ for UMLS concept *C0023467* with its set of associated questions $Q_{C0023467}$ and feature set $ac^{fs}_{C0023467}$.

the 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. The resulting feature sets build representatives for the annotation clusters that will be used to identify new annotations by matching unannotated forms to cluster representatives.

As an example, Figure 2 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 2 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.

### 3.3 Identification of Annotation Candidates

To reuse the confirmed annotations for unannotated forms we have to determine the annotation clusters (and thus their concepts) that match best the new questions to be annotated. One difficulty is that we need to find several annotations per question, i.e., we aim at identifying several annotation clusters. Since we may find too many related annotation clusters it is also important to select the most promising ones from the set of candidates.

We first describe how we determine the set of candidate annotation clusters. The example in Figure 1 showed that annotating concepts typically refer to some portion, i.e., succeeding terms, of the question text. Our approach to find matching annotation clusters thus uses a sliding window with a specified size *wnd_size* that partitions a given question into smaller portions according to the order of words in the question. Every text portion is compared with the feature set of every existing annotation cluster using a linguistic similarity measure. For this linguistic comparison we apply a soft TF/IDF string similarity function. TF/IDF weights the different terms based on their significance in all considered

documents. A soft variant of TF/IDF is more robust than TF/IDF w.r.t. different word forms. An annotation cluster and thus its concept is an annotation candidate for a given question, if the linguistic similarity exceeds a threshold $\delta$ for one portion of the question.

In the final selection of annotations, we want to avoid choosing similar annotations referring to the same medical concept. We therefore group the annotation candidates per question that relate to the same tokens and text portions of a question. For selecting the best matching concept per candidate group we apply the context-based selection strategy to be described next.

### 3.4 Context-based Selection of Annotations

The input for the final 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 calculate an aggregated similarity ($aggSim$) for each question and candidate concept based on weighted linguistic ($lsim$) and context ($csim$) similarity scores:

$$aggSim_{q,Candidates}(c_k) = \omega_{lsim} \cdot lsim(q, c_k) + \omega_{csim} \cdot csim(c_k, Candidates)$$

The linguistic similarity between candidate concepts is determined by the linguistic similarity of their concept descriptions, similarly as in the selection strategy of the base approach (Sec. 2.2). The calculation of the context-based similarity is more involved and will be described below. For each question in the set of input forms, we select the concepts with the highest $aggSim$ value per candidate group to obtain the final set of annotations.

For the context-based similarity between candidate concepts we consider two criteria: 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. These concepts are more likely to fit the context of a question which is typically about one subject, e.g. different medical aspects such as medications for a specific disease.

For the context-based selection of candidate concepts, 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. Figure 3 a shows the sample input for annotation selection consisting of a question and the set of grouped candidate concepts. In the context graph of the question (Figure 3 b), green edges interconnect concepts that have co-occurred before and red edges interconnect semantically related concepts.

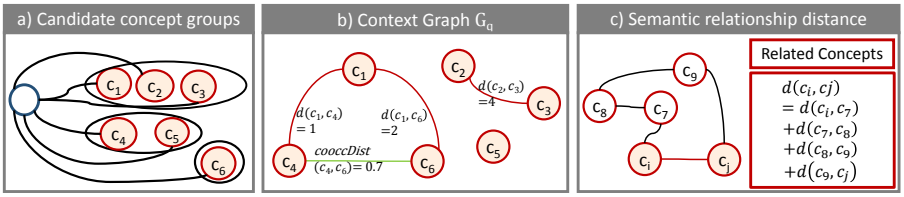


**Fig. 3.** Context-based similarity computation. **a)** candidate concept groups for one question; **b)** context graph with different edges for concept co-occurrence (green edges) and semantic relatedness (red edges); **c)** computation of semantic relatedness between concepts with related concepts from UMLS.

To determine the co-occurrence score between concepts $c_1$ and $c_2$ we count how often the two concepts have been annotated to the same question and compute the following normalized overlap of their annotation clusters:

$cooccDist(c_1, c_2) = 1 - \frac{|ac_{c_1} \cap ac_{c_2}|}{|ac_{c_1}|}$.

Concepts that often co-occur thus have a small distance score.

We further assign a semantic distance between concepts in the context graph based on the shortest path between two considered concepts in the ontological structure (see Figure 3 c), similarly to common techniques [18]. The ontological structure consists of the $is-a$, $part-of$ relationships and further domain specific relationships. We determine the semantic distance between two candidate concepts by summarizing the weighted distances of each relationship within the shortest path. We currently use the same distance 1 for each relationship type. Hence the semantic distance between two concepts corresponds to the path length, e.g., distance 4 for the concept pair in the example of Figure 3 c.

Based on the context graph and its distance scores we compute a context-based similarity for each concept by computing the distance to all other concepts in the candidate set of a question. Thereby, we favor concepts that often co-occur and those with a close semantic relatedness for our selection, i.e. selected concepts should have a small distance to other annotated concepts. We use the closeness centrality measure $cc$ that computes the reciprocal of the sum of all distances $d$ between a vertex $v$ and all other vertices $w$ in the graph $G$:

$$cc(v) = \frac{1}{\sum_{w \in G} d(v,w)}$$

We adopt a modified version of the closeness centrality to compute the context-based similarity score as follows. In our graph concepts can be isolated in case they do not co-occur with any other concepts and have a very different semantic context (e.g., concept $c_5$ in the context graph of Figure 3 b). Such isolated concepts should get a lower similarity score than concepts in a larger subgraph of $G_q$. However, isolated concepts have infinite distances $d$ to all other vertices such that $cc(v)$ would often converge to zero. To compute a normalized context-based similarity score $csim(c_i) \in [0, 1]$ for each concept $c_i$ in the set of vertices $V$ of the context-graph $G_q$, we sum up single reciprocal values of distances and normalize it by the number of concepts in the context-graph:

$$csim(c_i, V) = \frac{\sum_{c_j \in V \setminus \{c_i\}} \frac{1}{d(c_i, c_j)}}{|V|-1}$$

Concepts with a small distance to every other concept in the graph have high *csim* values meaning they are highly related to the other candidate concepts due to annotation co-occurrences and relationships from UMLS.

For instance, the context similarity for the concept $c_4$ is computed by the semantic distance $d(c_4, c_1) = 1$ and the co-occurrence distance $cooccDist(c_4, c_6) = 0.7$. The distances to the other concepts in the context graph are infinite. Therefore, we get the following context-similarity $csim(c_4) = \frac{\frac{1}{1} + \frac{1}{0.7} + \frac{1}{\infty} + \frac{1}{\infty} + \frac{1}{\infty}}{6-1} \approx 0.49$.

## 4 Evaluation

We now evaluate the proposed reuse-based annotation approach for medical forms and compare it with the baseline approach and the MetaMap tool. In the next subsection we introduce the used datasets and workflow configurations. We then evaluate the annotation quality compared to the baseline approach (Sec. 4.2) and analyze the effectiveness of the context-based selection strategy (Sec. 4.3). Finally, we provide the comparison with MetaMap (Sec. 4.4).

### 4.1 Evaluation Setting

Our evaluation uses medical forms about eligibility criteria EC and about quality assurance QA w.r.t cardiovascular procedures from the MDM platform [4]. The forms in the first dataset are used to recruit patients in clinical trials. Most questions in this dataset are long natural language sentences since the recruitment of clinical trial participants requires a precise definition of inclusion and exclusion criteria. The sentences contain $\sim 8$ tokens on average and often mention several medical concepts. The QA forms are used by health service providers in Germany since 2000 to document the quality of their services. The questions of the QA forms are shorter than the eligibility criteria ($\sim 3$ tokens on average), therefore a question is probably annotated with only one concept. The forms will be annotated with concepts of a reduced version of UMLS [3] covering all UMLS concepts that possess at least one preferred label or synonym ($\sim$1 Mio. concepts with $\sim 7$ Mio. labels/synonyms). Moreover, we do not consider general concepts ($\sim$ 12000 concepts) that are associated with one of the following semantic types: *Qualitative Concept, Quantitative Concept, Functional Concept, Conceptual Entity*.

To evaluate the quality of automatically generated annotations, we use manually created reference mappings from the MDM portal [4]. 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. We divide the set of input forms into disjoint reuse and evaluation datasets. For both use cases, EC and QA, we consider two reuse datasets of different sizes to study the impact of the amount of reusable annotations. Table 4 shows the number of forms, items and verified annotations for the reuse and evaluation datasets.

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.

| dataset | $EC_{RD1}$ | $EC_{RD2}$ | $EC_{eval}$ | $QA_{RD1}$ | $QA_{RD2}$ | $QA_{Eval}$ |
|---|---|---|---|---|---|---|
| #forms | 100 | 200 | 25 | 16 | 32 | 23 |
| #items | 1638 | 3125 | 310 | 453 | 795 | 609 |
| #annotations | 6911 | 13027 | 578 | 694 | 1054 | 668 |

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

For our reuse-based annotation workflow, we set a fixed window size *wnd_size* of five tokens for the *Candidate Identification* and fixed weights $\omega\_lsim/\omega\_csim$ to 0.5 for the *Context-based Selection*. In our experiments, we observe that these parameters only slightly affected the results for the considered datasets. We evaluate different thresholds $\delta = \{0.5, 0.6, 0.7\}$ to present the recall and precision trends. For the selection strategy we consider both the previously proposed group-based strategy [5] as well as the new context-based strategy. Note that we can use the group-based strategy not only for the base workflow but also in the reuse-based approach by setting the weight $\omega_{csim}$ for the context similarity to 0.

### 4.2 Reuse-based Annotation

Figure 5 shows evaluation results w.r.t. the mapping quality (precision, recall, F-measure) for the baseline approach and the different configurations of the reuse-based approach for the two datasets. For the baseline approach we only show the results for the best threshold of $\delta = 0.7$ for both datasets. The reuse-based approaches uses the context-based selection strategy. We observe that the reuse-based approach can significantly improve the annotation quality and that the improvement grows with the amount of annotations that we can reuse. Compared to the baseline approach, the reuse of existing annotations increases the F-measure from 39.1% to 50.7% for the EC dataset and from 57.5% to 59% for the QA dataset for the best threshold setting of $\delta = 0.6$. Using more existing annotations ($EC_{RD2}$ and $QA_{RD2}$) improves the mapping quality - and especially

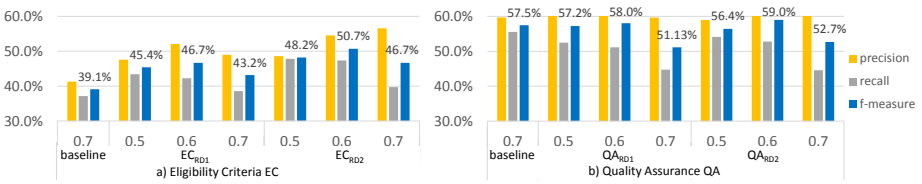


**Fig. 5.** Results on the quality of annotation results for the baseline and reuse-based annotation using the *EC* dataset and the *QA* dataset with both configurations.

recall - compared to the smaller reuse datasets ($EC_{RD1}$ and $QA_{RD1}$) since annotation clusters and their feature sets become more accurate and are thus more valuable to match to unannotated questions. The reuse-based approach is especially effective for the EC dataset where we could apply more annotations (Table 4) to build the annotation clusters compared to the QA dataset. The results confirm that matching questions to the feature sets of annotation clusters (*reuse-based*) helps to identify more correct annotations than trying to find the best matches in the UMLS (*baseline*). At the same time, the reuse-based approaches with the context-based selection strategy usually improve precision compared to the baseline approach.

An added benefit is that the execution time of the reuse-based approaches is lower than for the baseline approaches since matching questions with the compact annotation clusters is much faster than matching with the large UMLS ontology. Overall, runtimes could be reduced by half for our experiments compared to the baseline. Moreover, the execution time depends on the number of reused forms and the coverage of reused annotation clusters.

### 4.3 Context-based Selection

To analyze the effectiveness of the proposed context-based selection strategy ($CS$), we now compare its use with the group-based selection strategy ($GS$) that was used in the baseline approach but can also be applied for the reuse-based approaches. Table 6 shows the resulting mapping quality for the two selection strategies for the different EC and QA reuse configurations and threshold 0.6 that led to the best mapping quality for the reuse-based approach. The results show that the context-based selection strategy improves F-measure in all cases (up to 2.2%) compared to the simpler group-based approach. While recall is generally reduced this is more than outweighed by an increase in precision by up to almost ∼7% ($EC_{RD2}$). This indicates that considering the context eliminates many false candidates.

| dataset$_{configuration}$ | EC$_{RD1}$ | | EC$_{RD2}$ | | QA$_{RD1}$ | | QA$_{RD2}$ | |
|---|---|---|---|---|---|---|---|---|
| selection-strategy | gs | cs | gs | cs | gs | cs | gs | cs |
| precision | 45.9% | **52.1%** | 47.9% | **54.5%** | 61.9% | **67.0%** | 60.4% | **66.9%** |
| recall | **43.6%** | 42.2% | **49.2%** | 47.3% | 51.0% | **51.2%** | **54.6%** | 52.8% |
| f-measure | 44.7% | **46.7%** | 48.5% | **50.7%** | 55.9% | **58.0%** | 57.4% | **59.0%** |

**Table 6.** Results on the quality of annotation results for the group-based ($GS$) and context-based ($CS$) selection strategies for both datasets

### 4.4 Comparing reuse-based annotation approach with MetaMap

We finally compare our reuse-based annotation method with the MetaMap tool that is commonly used for annotating medical documents (see Sec. 5). We gen-

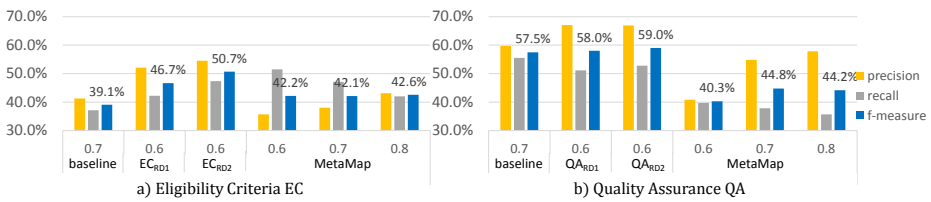


**Fig. 7.** Comparison of the quality for the resulting annotation mappings from the baseline approach, reuse-based approach and MetaMap.

erate the annotations with a local installation of a MetaMap server and the MetaMapAPI and use the provided word sense disambiguation service and the configuration considering several variants for a concept. We select annotations based on the generated MetaMap score. This score ranges from 0 to 1000 and is computed by applying several ranking functions for each identified term. If MetaMap generates more than one annotation per question, we select the annotations with an aggregated score above a threshold. We normalize the scores by dividing by 1000 for comparing with our approach and evaluate different thresholds $\delta = \{0.6, 0.7, 0.8\}$ for selecting the candidates.

Figure 7 shows the results for the two datasets and different configurations. Our reuse-based approach outperforms Meta-Map in terms of mapping quality for each dataset. For the EC dataset, F-Measure is improved by $\sim$4%($EC_{RD1}$) and $\sim$8.6% ($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 Related Work

The automatic annotation of medical forms and documents with concepts of standardized vocabularies is related to the well-studied fields of ontology matching [20,8] and entity linking [22]. Both research domains provide useful generic methods to identify concepts or names in full-text documents and match them to concepts or entities of a knowledge base or standardized vocabulary. These

techniques can also be applied to the medical domain. In fact, our base workflow proposed in [5] uses linguistic ontology matching techniques to map terms of medical forms to the concepts and their synonyms of the UMLS ontology. Entity linking approaches focus on the identification of named entities in text documents and their linking to corresponding entities of a knowledge base for enrichment. Many approaches (e.g. [6,16,24]) use a dictionary-based strategy to identify entity occurrences by searching the whole knowledge base.

Moreover, there are many approaches to select the correct entities from a set of candidates (e.g. [6,11,9]). For instance, in [9] co-occurrences of entities in Wikipedia articles are transformed into a graph model to consider the global interdependence between different candidate entities in a document.

There is also some research focusing on the manual or automatic annotation of certain kinds of medical documents. The MetaMap tool [1] considered in our evaluation applies information retrieval methods such as tokenization, lexical lookup and term-based ranking methods to retrieve UMLS concepts within medical documents. There is evidence in the literature that MetaMap results are not fine-grained enough [15], contain many spurious annotations [19] and do not cover mappings to longer medical terms [21]. These observations confirm that a correct annotation of medical documents with UMLS concepts is challenging. Our reuse-based approach could significantly outperform MetaMap due to its use of annotation clusters derived from verified annotations and due to its context-based approach to select and disambiguate concept candidates.

In the medical domain, the standardization of eligibility criteria has become an active field of research and datasets from this subdomain are often used for method evaluation (e.g. [10,13,23,17]). For instance, the study in [23] identified the most frequent ECs in clinical trial forms and performed a manual annotation of eligibility criteria top terms. In [10], similar clinical trials have been clustered by performing nearest neighbor search using annotated eligibility criteria, and the application of a dictionary-based pre-annotation method [13] showed to improve the speed of manual annotation for clinical trial announcements. In [17], a set of eligibility criteria in the context of clinical trials on breast cancer is formalized by defining eligibility criteria specific patterns in order to improve their comparability.

In contrast to previous research we propose a novel reuse-based annotation approach for medical documents. Our method is especially valuable to annotate documents from different biomedical domains with ontology concepts, i.e. it is not restricted to a specific medical subdomain. The proposed use of annotation clusters and their feature sets has not been explored before. Furthermore, we apply a novel context-based selection of annotations considering both, the co-occurrences of verified annotations as well as the semantic relatedness of concepts. Our comparative evaluation showed that the new approaches outperform previous annotation schemes including tools like MetaMap.

## 6 Conclusion

We proposed and evaluated a new reuse-based approach to semantically annotate medical documents such as CRFs with concepts of an ontology. The approach utilizes 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. Clusters are represented by features covering meaningful term groups from the annotated questions and concept description. New questions are matched with these cluster representatives to find candidates for annotating concepts. We further presented a context-based selection strategy to identify the most promising annotations based on the semantic relatedness of concept candidates and well as known co-occurrences from previous annotations. In a real-world evaluation, our methods showed to be effective and we could generate valuable recommendations to reduce the manual annotation effort. Moreover, reusing annotation clusters is more efficient than searching a large knowledge base such as UMLS for suitable annotation candidates.

For future work, we plan to evaluate further annotation approaches, in particular the combined use of several reuse-based and other techniques. For example, the MetaMap tool alone was inferior to the reuse-based scheme but it could be used in a combined scheme to find further annotation candidates. 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 efficiently and effectively 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.

## Acknowledgment

This work is funded by the German Research Foundation (DFG) (grant RA 497/22-1, "ELISA - Evolution of Semantic Annotations").

## References

1. A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 2000.
3. O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.
4. 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.
5. 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.

6. S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, 2007.
7. K. Donnelly. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. *Studies in Health Technology and Informatics–Medical and Care Compunetics 3*, 121:279–290, 2006.
8. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, 2007.
9. X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *Proc. 34th Int. ACM SIGIR Conf.*, pages 765–774, 2011.
10. T. Hao, A. Rusanov, M. R. Boland, et al. Clustering clinical trials with similar eligibility criteria features. *Journal of Biomedical Informatics*, 52:112–120, 2014.
11. S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of Wikipedia entities in web text. In *Proc. 15th ACM SIGKDD Conf.*, pages 457–466, 2009.
12. 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.
13. T. Lingren, L. Deleger, K. Molnar, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, 21(3):406–413, 2014.
14. H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association (JAMA)*, 271(14):1103–1108, 1994.
15. Z. Luo, R. Duffy, S. Johnson, and C. Weng. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. *AMIA Summits on Translational Science Proceedings*, 2010:26, 2010.
16. R. Mihalcea and A. Csomai. Wikify! Linking Documents to Encyclopedic Knowledge. In *Proc. 16th ACM CIKM*, pages 233–242, 2007.
17. K. Milian, R. Hoekstra, A. Bucur, A. Ten Teije, F. Van Harmelen, and J. Paulissen. Enhancing Reuse of Structured Eligibility Criteria and Supporting their Relaxation. *Journal of Biomedical Informatics*, 2015.
18. C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology*, 5(7), 2009.
19. G. S. Philip Ogren and C. Chute. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In *Proc. (LREC) Conf.*, pages 3143–50, 2008.
20. E. Rahm. Towards Large-Scale Schema and Ontology Matching. In *Schema Matching and Mapping*, pages 3–27. Springer, 2011.
21. K. Ren, A. M. Lai, A. Mukhopadhyay, et al. Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. *BMC Medical Genomics*, 7(Suppl 1), 2014.
22. 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.
23. J. Varghese and M. Dugas. Frequency Analysis of Medical Concepts in Clinical Trials and their Coverage in MeSH and SNOMED-CT. *Methods of Information in Medicine*, 53(6), 2014.
24. W. Zhang, C. L. Tan, Y. C. Sim, and J. Su. NUS-I2R: Learning a Combined System for Entity Linking. In *Proc. 3rd Text Analysis Conf. (TAC)*. NIST, 2010.