Evaluating and Improving Annotation Tools for Medical Forms

Ying-Chi Lin[1(✉)], Victor Christen[1], Anika Groß[1], Silvio Domingos Cardoso[2,3], Cédric Pruski[2], Marcos Da Silveira[2], and Erhard Rahm[1]

[1] Department of Computer Science, Universität Leipzig, Leipzig, Germany
{lin,christen,gross,rahm}@informatik.uni-leipzig.de
[2] LIST, Luxembourg Institute of Science and Technology, Esch-sur-Alzette, Luxembourg
{silvio.cardoso,cedric.pruski,marcos.dasilveira}@list.lu
[3] LRI, University of Paris-Sud XI, Orsay, France

Abstract. The annotation of entities with concepts from standardized terminologies and ontologies is of high importance in the life sciences to enhance semantic interoperability, information retrieval and meta-analysis. Unfortunately, medical documents such as clinical forms or electronic health records are still rarely annotated despite the availability of some tools to automatically determine possible annotations. In this study, we comparatively evaluate the quality of two such tools, cTAKES and MetaMap, as well as of a recently proposed annotation approach from our group for annotating medical forms. We also investigate how to improve the match quality of the tools by post-filtering computed annotations as well as by combining several annotation approaches.

Keywords: Annotation · Medical documents · Ontology · UMLS

1 Introduction

The interest for annotating datasets with concepts of controlled vocabularies, terminologies or ontologies is increasing, especially in the biomedical domain. Semantic annotations help to overcome typical data heterogeneity issues and thus improve interoperability for different data providers and applications. For instance, exchanging and analyzing the results from different clinical trials can lead to new insights for diagnosis or treatment of diseases. Semantic annotations of electronic health records (EHRs) showed to be valuable to identify adverse effects of drugs and thus for developing better drugs [11,13]. NCI Metathesaurus has been used to annotate semantically related entities in clinical documents to achieve enhanced document retrieval [22]. Furthermore, annotations of publications help to better deal with the huge volume of research literature by enhancing systems for automatically generating hypotheses from documents about relevant factors, phenotypes, or biological processes [1].

In the healthcare sector there is a high and increasing number of documents such as research publications, EHRs or case report forms (CRFs). For instance,

© Springer International Publishing AG 2017
M. Da Silveira et al. (Eds.): DILS 2017, LNBI 10649, pp. 1–16, 2017.
https://doi.org/10.1007/978-3-319-69751-2_1

there are almost 250,000 clinical studies registered on `ClinicalTrials.gov`. Dugas et al. estimate that more than 10 million different CRFs have been used so far [8], e.g., to document the medical history of patients or to evaluate eligibility criteria of probands of a study. Unfortunately, the vast majority of medical documents is still not annotated at all. For example, from the 11,000 forms and their 700,000 questions in the MDM portal[1], only about 1/7 have currently been annotated with concepts of the Unified Medical Language System (UMLS) Metathesaurus [14], the most widely used integrated vocabulary for clinical annotations [7]. The Metathesaurus currently contains more than 3.4 million concepts from over 200 controlled vocabularies and ontologies, such as ICD-10, SNOMED CT and MeSH. The huge number of documents, the use of natural language within the documents as well as the large complexity of biomedical ontologies such as UMLS make it challenging to find correct annotations for both automatic approaches as well as human experts. The most promising approach is thus to first apply a tool to automatically determine annotation candidates. A human expert can then select the final annotations from these candidates.

There exist several tools and approaches for such a semi-automatic annotation as well as a few initial evaluations of them [10,15,17,21]. In [21], the tools MetaMap, MGrep, ConceptMapper, cTAKES Dictionary Lookup Annotator and NOBLE Coder have been evaluated for annotating medical documents from the ShARe corpus[2] (containing clinical free-text notes from electrocardiogram and radiology reports) with concepts from the UMLS SNOMED-CT ontology. While the reported findings seem to indicate the usability of the tools the results cannot be generalized to different kinds of medical documents, such as other EHRs or CRFs.

In this study, we focus on the comparative evaluation of three tools/approaches for annotating CRFs and whether we can improve annotation quality by post-processing the tool results or by combining different approaches. We selected the tools *MetaMap* [2] and *cTAKES* [16] as well as our previous research approach [5] to which we refer here as *AnnoMap*. MetaMap is a well established tool and has been applied in many different types of tasks such as text mining, classification and question answering [2]. We chose cTAKES as it performed best in the mentioned evaluation study [21]. Specifically, we make the following contributions:

- We comparatively evaluate the three annotation tools based on the annotation of two kinds of English medical forms with the UMLS.
- We investigate to which degree the annotation results of cTAKES and MetaMap can be improved by additionally applying the group-based selection of annotation candidates from AnnoMap [5].
- We propose and evaluate annotation approaches combining the results generated by different tools in order to improve overall annotation quality.

[1] https://medical-data-models.org.
[2] https://sites.google.com/site/shareclefehealth/.

We first introduce the considered annotation tools and their combination in Sect. 2. We then describe the evaluation methodologies and analyze the results in Sect. 3. Finally, we summarize the findings and conclude.

2 Annotation Tools

The task of *annotation* or *concept recognition* has as input a set of documents $D = \{d_1, d_2, ..., d_n\}$, e.g., publications, EHRs, or CRFs, to annotate as well as the ontology ON from which the concepts for annotation are to be found. The goal is to determine for each relevant document fragment *df* such as sentences or questions in medical forms the set of its most precisely describing ontology concepts. The annotation result is a set of so-called *annotation mappings* $\mathcal{AM}_{d_i,ON} = \{(df_j, \{c_1, ..., c_m\}) | df_j \in d_i, c_k \in ON\}$ where each mapping refers to one document d_i and consists of the associations between the document fragments and its set of annotating concepts.

Several tools for the automatic annotation of documents in the life sciences have been developed in the last years. Such *annotators* can be generally categorized into dictionary-based and machine learning-based approaches [3]. The learning-based approaches typically require a training corpus which is rarely available for a new set of documents to annotate. As a result, the more general-purpose dictionary-based approaches are mostly favored. To speedup the annotation process, they typically create a dictionary for the ontology (e.g., UMLS) to be used for finding annotating concepts. Examples of such tools include MetaMap [2], NCBO Annotator [6], IndexFinder [24], MedLEE [9], ConceptMapper [20], NOBLE Coder [21], cTAKES [16] as well as our own AnnoMap approach [5]. We further developed an extension of AnnoMap utilizing previous annotations which can be seen as a special kind of training data [4].

In this study, we evaluate three annotation tools and their combination: MetaMap, cTAKES and AnnoMap. Table 1 summarizes the main features of these annotators w.r.t. three phases: preprocessing, candidate generation and postprocessing. The preprocessing phase is divided into an offline and an online step. The offline step is devoted to generating the dictionary for the ontology with indexed entries for the concepts to support fast lookup. The online step is used to preprocess the input documents by using NLP approaches. In the candidate generation phase, the annotation candidates for each text fragment are identified by using a dictionary lookup strategy or a fuzzy matching based on similarity functions. Finally, the postprocessing phase selects the annotations from the annotation candidates.

In the following, we discuss the three tools in more detail. At the end, we discuss possible combinations of the individual tools aiming at improving the annotation quality compared to the use of only one approach.

Table 1. Components and functions of MetaMap, cTAKES and AnnoMap. POS: Part of Speech, LCS: Longest Common Substring

Tool	Ontology prepocessing	Form preprocessing	Candidate generation	Post-processing
MetaMap	dictionary construction (UMLS, SPECIALIST lexicon)	sentence detector, tokenizer, POS tagger/filter, shallow parser, variant generation (static/dynamic), abbreviation identifier	dictionary lookup (first word)	word sense disambiguation, score-based filtering
cTAKES	dictionary construction (UMLS)	sentence detector, tokenizer, POS tagger/filter, shallow parser, variantgeneration (dynamic)	dictionary lookup (rare word)	-
AnnoMap	-	tokenizer, POS tagger/filter, TF/IDF computation	fuzzy match (TF/IDF, Trigram, LCS)	threshold-based, group-based

2.1 MetaMap

MetaMap was originally developed to improve the retrieval of bibliographic documents such as MEDLINE citations [2]. It is designed to map biomedical mentions to concepts in UMLS Metathesaurus. MetaMap is based on a dictionary-lookup by using several sources such as UMLS itself as well as SPECIALIST lexicon. The SPECIALIST lexicon contains syntactic, morphological, and spelling variations of commonly occurring English words and biomedical terms of UMLS [14]. The input text is first split into sentences and further parsed into phrases. These phrases are the basic units for the variant generation and candidate retrieval. MetaMap provides several configurations for the lookup of annotation candidates per phrase such as *gap* allowance, *ignore* word order, and *dynamic* as well as *static* variant generation. For each annotation candidate MetaMap computes a complex score function considering linguistic metrics [2] for each phrase of a sentence. The final result is determined by the combination of candidates maximizing the aggregated score. MetaMap also provides an optional postprocessing step, word sense disambiguation (WSD), for cases when the final result has several Metathesaurus concepts with similar scores. WSD selects the concept that is semantically most consistent with the surrounding text [12].

2.2 cTAKES

cTAKES[3] is built on the Apache UIMA framework[4] providing a standardized architecture for processing unstructured data. To annotate medical documents, cTAKES provides several components for specifying preprocessing and lookup

[3] Clinical Text Analysis and Knowledge Extraction System http://ctakes.apache.org.
[4] Unstructured Information Management Architecture [16] https://uima.apache.org.

strategies. The components are used to define customized annotation pipelines where each component uses the intermediate output of the previous component as input. In addition to general components used in a default pipeline, cTAKES offers domain-specific components such as for the classification of smoking status [19], the extraction of drug side effects [18], and coreference resolution [23].

In the following, we describe the default pipeline with its components. During (offline) preprocessing, an ontology dictionary is built where each property of a concept becomes an entry in the dictionary. The rarest word of an entry is used to index it for fast lookup. The rareness of a word is based on the global occurrence frequency in the ontology. For the (online) preprocessing of the input documents, cTAKES uses the following components: sentence boundary detector, customized part of speech (POS) tagger and a lexical variant generator. The model of the POS tagger is trained for medical entities based on clinical data since general POS taggers do not cover domain-specific characteristics such as abbreviations. In general, medical entity mentions within documents can be different according to the name and synonyms of concepts. Therefore, cTAKES applies a lexical variant generator (LVG) to transform differently inflected forms, conjugations or alphabetic cases to a canonical form for improved comparability. While cTAKES permits the addition of customized postprocessing steps to the pipeline such strategies are not part of the cTAKES core project.

2.3 AnnoMap

AnnoMap implements a general approach for annotating documents with concepts of arbitrary ontologies. In the current version, it does not create a dictionary of the ontology during preprocessing for fast lookup but directly searches in the ontology for finding suitable annotations. In the preprocessing step, the concept entries and the documents are normalized by applying several text transformation functions such as lower case, stop word elimination, POS filtering or removing characters that are not alpha-numeric. For candidate generation, AnnoMap loads the ontology into main memory and applies a general match approach by comparing each document fragment with each ontology concept. Matching is based on the combined similarity score from different string similarity functions, in particular TF/IDF, Trigram and LCS (longest common substring) similarity. AnnoMap retains all annotation candidates with a score above a given threshold δ. This corresponds to a fuzzy matching that tolerates name variations and typos which might not be the case for the lookup techniques of MetaMap and cTAKES.

During postprocessing, AnnoMap filters the candidates of a document fragment with a *group-based selection strategy* that aims at keeping only the best annotations for groups of similar annotation candidates. Figure 1 illustrates this selection approach for the candidates of two document fragments (e.g. CRF questions) df_1 and df_2. The candidates of a document fragment are first grouped or clustered based on the mutual (string) similarity of the annotating concepts. For groups of highly similar candidates, the approach then only retains the one

with the highest annotation score. In the example, concepts c_1 and c_2 represent a group for df_1 and c_3 and c_4 form a group for df_2. From these groups, c_1 and c_4 have the highest score and are retained while candidates c_2 and c_3 are removed from the result.

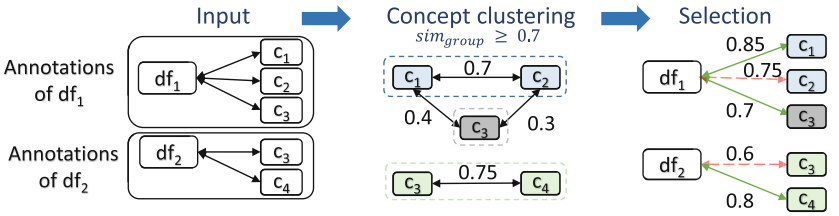


Fig. 1. Example for the group-based selection strategy

The described group-based selection can also be applied to postprocess the results of the tools MetaMap and cTAKES and we will evaluate the effectiveness of such a combined strategy. Furthermore, we can combine the three tools as described in the next subsection.

2.4 Combining Different Tools

The considered tools follow different approaches for finding annotations that may complement each other. Hence, it is promising to combine the results of the individual approaches to hopefully improve overall annotation quality, e.g., to improve recall (find more correct annotations) or/and precision (eliminate less likely annotations that are not confirmed by two or more tools). For combining the annotation results of two or more tools we follow three simple approaches: *union*, *intersection* and *majority*. The *union* approach includes the annotations from any tool to improve recall while *intersection* only preserves annotations found by all tools for improved precision. The *majority* approach includes the annotations found by a majority of tools, e.g., by at least two of three tools. There are further variations for combining the approaches by differentiating whether the proposed group-based selection is applied before or after the combination (aggregation) of the individual tool results. The two resulting workflows, wf_1 and wf_2, are illustrated in Fig. 2. In the first case (wf_1) we combine postprocessed annotation results after we have applied group-based selection to the results of the respective tools. For wf_2, we aggregate the results without individual postprocessing but apply group-based selection only on the combined annotation result.

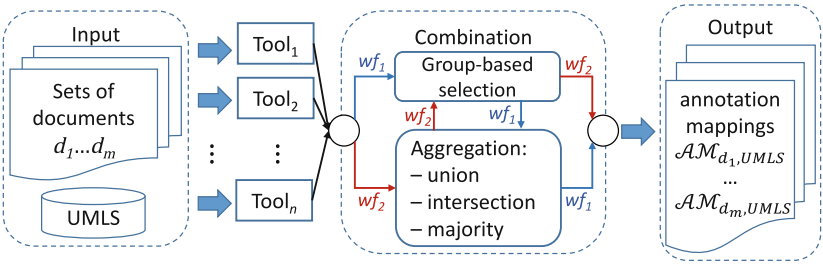


Fig. 2. Workflows for combining the annotation results of different tools. Workflow wf_1 first applies group-based selection for each tool and then combines the selection results. Workflow wf_2 first combines the annotation results of each tool and then selects the final annotations on the combined result.

3 Evaluation and Results

We now comparatively evaluate the three annotation tools MetaMap, cTAKES and AnnoMap and their combinations using two sets of medical forms and the UMLS Metathesaurus. We first describe our experimental setup including the datasets and tool configurations. We then evaluate the annotation results for single tools and the additional use of group-based selection (Sect. 3.2) as well as for the combination of two or three tools (Sect. 3.3). Sect. 3.4 summarizes the results. The evaluation focuses on the standard metrics *recall*, *precision* and their harmonic mean *F-measure* as the main indicator for annotation quality.

3.1 Experimental Setup

Document Sets and Ontologies: We use two datasets with medical forms (CRFs) from the MDM portal that have also been used in previous annotation evaluations [4,5] and for which a reference mapping exists: a dataset with forms on *eligibility criteria* (EC) and a dataset with *quality assurance* (QA) forms.

The EC dataset contains 25 forms with 310 manually annotated questions. These forms are used to recruit patients in clinical trials for diseases such as epilepsy or hemophilia. The QA dataset has 24 standardized forms with 543 annotated questions used in cardio-vascular procedures. The number of annotations in the reference mappings is 541 for EC and 589 for QA. The previous evaluations [4,5] showed that it is very challenging to correctly identify all annotations for these datasets.

To annotate we use UMLS version 2014AB that was used for the manual annotation. We include five vocabularies: UMLS Metathesaurus, NCI (National Cancer Institute) Thesaurus, MedDRA[5], OAC-CHV[6], and SNOMED-CT_US[7], covering most annotations in the manually determined reference mappings.

[5] Medical Dictionary for Regulatory Activities.
[6] Open-access and Collaborative (OAC) Consumer Health Vocabulary (CHV).
[7] US Extension to Systematized Nomenclature of Medicine-Clinical Terms.

Since we use different subsets of UMLS in this paper and in the previous studies [5], the results are not directly comparable.

Tool Configuration and Parameter Settings: The considered tools provide a large spectrum of possible configurations making it difficult to find suitable parameter settings. To limit the scope of the comparative evaluation and still allow a fair comparison we analyzed the influence of different parameters in a preparatory evaluation to arrive at default configurations achieving reasonable annotation quality per tool.

Table 2 lists the considered parameters for cTAKES, MetaMap and AnnoMap. For MetaMap, we found that the best *scoreFilter* values are (700/800/900) for EC and also 1000 for QA. WSD delivered significant better F-measures than $default$ only for the EC dataset, for which $dynVar$ does not provide noticeable improvements. For QA, $dynVar$ as well as *gaps* could produce better results than $default$ but the results were inferior to the use of WSD when we combine several tools. Hence, we omit the results of *gaps* and $dynVar$ and focus on MetaMap results for $default$ and WSD with different *scoreFilter* values.

For cTAKES, using *longestMatch* results in improved precision and F-measure. While *overlap* is supposed to increase recall, this is not the case for our datasets so that we exclude experiments using this parameter. For AnnoMap,

Table 2. Tested parameters in MetaMap, cTAKES and AnnoMap

Parameter	Description
MetaMap	
gaps	allows gaps between tokens
$dynVar$	generates variants dynamically rather than only lookup table
WSD	enables word sense disambiguation
scoreFilter	sets the threshold to filter out mapping candidates. MetaMap score values range between 0–1000 (tested values: 700/800/900 for EC and 700/800/900/1000 for QA)
wordOrder	matches also terms in different orders
derivVar	specifies which type of derivational variations to be used (tested settings: default/none/all). Default uses only derivational variations between adjectives and nouns
cTAKES	
overlap	allows matches on discontiguous spans
longestMatch	returns only the concept with the longest matched span
AnnoMap	
threshold δ	sets the minimum similarity for filtering annotation candidates (tested values: 0.6–0.8 with 0.5 interval)

we tested the thresholds δ ranging from 0.6 and 0.8 based on our previous investigation in [5]. We apply the best-performing results in the experiments, i.e., $\delta = 0.7$ for EC and $\delta = 0.75$ for QA.

To use the group-based selection strategy (Sect. 2.3) for the tools cTAKES and MetaMap, we need a score per annotation candidate to select the one with the highest score from a group of similar candidates. For MetaMap, we use the generated scores divided by 1000 (to obtain a value between 0 to 1) for this purpose. Since cTAKES does not determine any score, we calculate a linguistic similarity between each question and its matched concept using Soft TF/IDF as the annotation score.

3.2 Evaluation of Single Tools and Use of Group-Based Selection

We first assess the annotation quality for the single approaches cTAKES, MetaMap and AnnoMap without and with the additional use of group-based selection. Figure 3 presents the results for the datasets (a) EC and (b) QA with different parameter settings. AnnoMap with group-based selection achieves the highest F-measure among all tools/parameter settings for both EC (39.5%) and QA (56.1%). Group-based selection is an integral part of AnnoMap but for comparison we also show AnnoMap results without this selection method. We observe

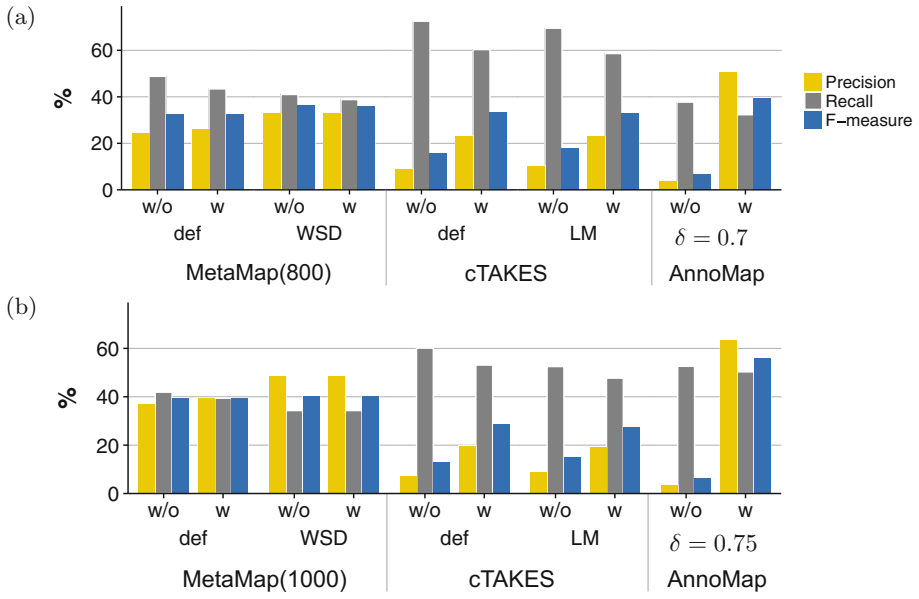


Fig. 3. Annotation quality of MetaMap, cTAKES and AnnoMap without (w/o) and with (w) group-based selection in datasets (a) EC and (b) QA. The MetaMap results refer to the best-performing *scoreFilter* setting.
def: *default* setting, WSD: Word Sense Disambiguation, LM: *longestMatch*

that group-based selection dramatically increases precision by filtering out a large amount of false positives after initial candidate generation in AnnoMap (e.g., from 8,573 to only 170 for QA). Overall, AnnoMap achieves the best precision among all tools in both datasets (50.6% for EC and 63.5% for QA).

The highest recall values are obtained by cTAKES default (def) without group-based selection (72.5% for EC and 60.1% for QA), at the expense of very low precision values (less than 9%) and thus poor F-measure. Applying the $longestMatch$ (LM) function decreases the number of false positives by about 1/4 (e.g., from 4,407 to 3,164 for QA) and slightly improves precision to about 10%. Significantly more effective is the proposed extension of cTAKES with group-based selection which improves precision as well as F-measure by about a factor of two for both datasets. As a result, the best cTAKES F-measure results, 33.4% for EC and 28.8% for QA, are achieved with cTAKES(def) with group-based selection.

MetaMap achieves better F-measure results than cTAKES especially when applying WSD with a maximum of 36.4% for EC (with $scoreFilter = 800$) and 40.2% for QA ($scoreFilter = 1000$). In contrast to cTAKES, the use of group-based selection did not improve annotation quality since MetaMap itself already filters similar annotation candidates based on their scores within phrases (Sect. 2.1). Applying WSD improved F-measure over the default strategy of MetaMap by up to 4% by further filtering the annotation candidates.

3.3 Results of Combining Tools

We first compare the effectiveness of the two workflows wf_1 and wf_2 for combining the annotation results. We then analyze combinations of two and three tools for a union, intersection or majority aggregation of annotations.

Influence of Combination Workflow: As described in Sect. 2.4, we consider two workflows differing in whether group-based selection is applied before (wf_1) or after (wf_2) the combination of the individual tool results. The motivation for wf_2 is that we may improve recall if we do not filter already the individual tool results but postpone the filter step until after we have combined the annotation candidates from different tools. Our evaluation, however, showed that wf_1 outperforms wf_2 in almost all cases, i.e., it is beneficial to first apply group-based selection per tool and then combine filtered results. For a *union* aggregation, wf_2 results in a large number of annotation candidates as input to the final group-based selection. Many of these candidates share common tokens and are thus grouped into the same group from which only one candidate is finally selected. Hence, wf_2 leads to fewer true positives and more false negatives than wf_1. For the *intersection* or *majority* combinations, wf_2 suffered from more false positives and such a reduced precision compared to wf_1 which can not be outweighed by a slightly higher recall. Given the superiority of wf_1 we will only present results for this approach in the following.

Combining Two Tools: For two tools, we support a union or intersection of the individual results (the majority approach corresponds to intersection here). We have three possible tool combinations for which the average results on annotation quality are shown in Fig. 4. The averages are taken over all configurations of a tool while the vertical bars (*variance bars*) denote the spectrum between the minimal and maximal result per combination. As expected, we see that the *union* combinations achieve high recall values while *intersection* leads to increased precision over the single tools. More importantly, we note that *intersection* consistently leads to improved F-measure compared to the *union* combination indicating that the improvements on precision are more decisive than the recall increases. The large variance bars for some combinations reflect a substantial influence of some parameter settings such as the *scoreFilter* value of MetaMap.

For the EC dataset (Fig. 4a), the best F-measure of 42.1% is achieved for the (intersection) combination of MetaMap and cTAKES. This combination also outperforms all single tools including AnnoMap (39.5%). The combinations AM-CT and AM-MM cannot reach the F-measure of AnnoMap but outperform the single tools cTAKES and MetaMap, respectively, mainly due to an improved precision (ranging from 62.8% to 81.6%).

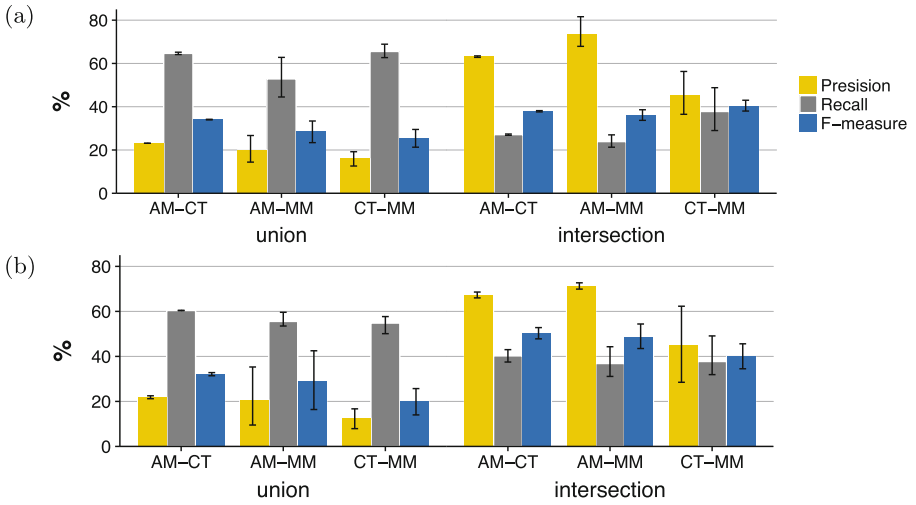
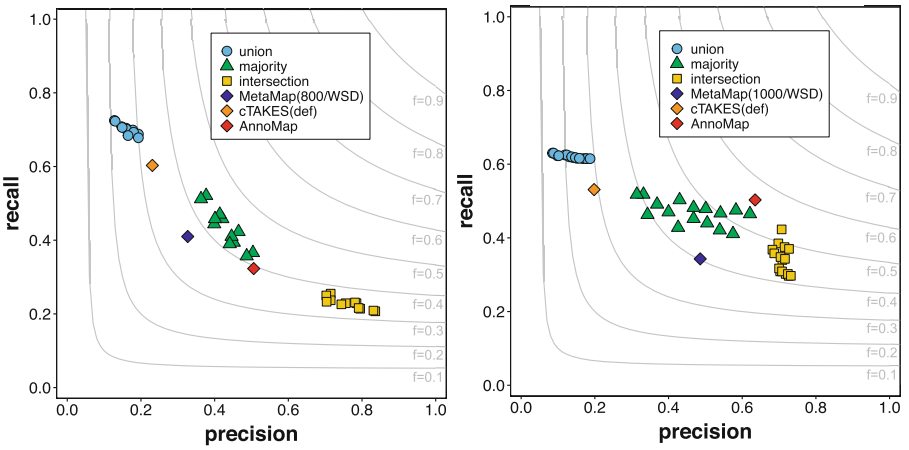


Fig. 4. Average annotation quality of combining two tools for different parameter settings with datasets (a) EC and (b) QA. Variance bars show the maximum and minimum quality values. The results are obtained using wf_1.

For the QA dataset (Fig. 4b), the highest F-measure (54.4%) for combining two tools is obtained by intersecting the results of AnnoMap and MetaMap (700/def). While this result is slightly lower than for AnnoMap alone (56.1%) it substantially outperforms the F-measure of MetaMap alone (40.2%). Similarly, the combination AM-CT leads to a strong F-measure improvement compared to

cTAKES alone. By contrast, the combination CT-MM is less effective than for EC but still improves on the single tools.

Combining Three Tools: For three tools, we can apply three aggregation approaches (union, intersection, majority) and have many combinations depending on which configuration per tool we select. We therefore use now precision-recall plots in Fig. 5 to present the results for the (a) EC and (b) QA datasets. The curves inside these plots refer to different F-measure values (f). Both plots show that the results of different aggregation methods form three distinctive clusters. The combinations based on a *union* aggregation have the best recall but the lowest precision while the *intersection* combinations have opposite characteristics. The *majority* combinations are able to better balance recall and precision and lie therefore in-between the two other approaches and achieve mostly the best F-measure values.



(a) EC (b) QA

Fig. 5. Annotation quality of combining three tools using wf_1. Each point refers to a union, intersection or majority combination for a specific cTAKES and MetaMap configuration, as well as the best single tool results.

For EC, all *majority* combinations outperformed each single tool in terms of F-measure (Fig. 5a). This is because the combinations improved recall over MetaMap and AnnoMap and precision over cTAKES. The best F-measure (44.3%) is obtained by the *majority* of AnnoMap, cTAKES (def) and MetaMap (800/WSD), i.e., with the configurations for cTAKES and MetaMap that performed best when using these tools alone. As for two tool combinations, the *union* approach achieves always lower F-measure than with *intersection*.

For the QA dataset (Fig. 5b), the best F-measure (53.2%) is achieved by the *majority* aggregation of the combination AnnoMap, cTAKES (def) and

MetaMap (1000/WSD). Again, these are the best performing QA configurations of the single tools. The single tool results for both cTAKES and MetaMap are outperformed by all combinations of three tools using either *majority* or *intersection*. However, different from the EC dataset the F-measure of AnnoMap alone can not be topped by the combined schemes. This is because recall decreased compared to AnnoMap alone indicating that AnnoMap can determine many valid annotation candidates that are not found by another tool to build a majority. The precision for *majority* also differs over a large range (31.4%–62.1%) mainly due to a strong dependency on the *scoreFilter* of MetaMap.

3.4 Result Summary

The presented evaluation showed that the annotation quality of existing tools such as cTAKES and AnnoMap can be substantially improved by the proposed combinations such as adding a group-based selection of annotation candidates and aggregating the results of different tools. Most effective is the use of both optimizations, i.e., the use of group-based selection and the aggregation of results from two or more tools. In this case, it is better to first apply group-based selection per tool before aggregating the results (combination workflow wf_1). From the considered aggregation strategies, *intersection* performs best for two tools and *majority* for three tools. For the EC and QA datasets, the single tool performance of cTAKES is lower than for MetaMap and the research approach AnnoMap. However, by applying the combination strategies these differences can be reduced to a large degree.

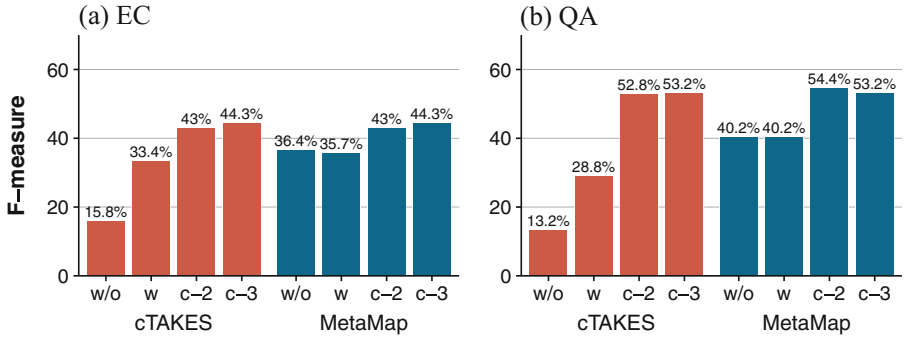


Fig. 6. Summarizing F-measure results for cTAKES and MetaMap and the proposed combinations for the (a) EC and (b) QA datasets. The combinations are w/o: without group-based selection, w: with group-based selection, c-2: best results from combining two tools, c-3: best results from combining three tools. For EC, the c-2 configuration is cTAKES (def) with MetaMap (700/def), and for c-3: AnnoMap with cTAKES (def) and MetaMap (800/WSD). For QA, the c-2 configurations are cTAKES (def) with AnnoMap and MetaMap (700/def) with AnnoMap. For c-3: AnnoMap with cTAKES (def) and MetaMap (1000/WSD).

Figure 6 summarizes the impact of the proposed combination strategies on F-measure for cTAKES and MetaMap. We observe that the F-measure of cTAKES can be dramatically improved (about a factor 3–4) for both datasets. Adding group-based selection alone already doubles F-measure. Combining cTAKES with MetaMap or AnnoMap further improves F-measure noticeably. For MetaMap with the WSD option, the additional use of group-based selection is not useful but the aggregation with other tools also improved F-measure substantially. Interestingly, most improvements can already be achieved by combining only two tools. AnnoMap is the best-performing single tool and its combination with other tools generally improves annotation quality for these tools. The quality for AnnoMap itself can be topped for the EC dataset by a majority combination of all three tools but not for the QA dataset. We therefore see a need to investigate strategies to further improve annotation quality for tools such as AnnoMap.

4 Conclusions

The large-scale annotation of documents in healthcare such as medical forms or EHRs is of high benefit but still in an early stage. In this paper, we comprehensively evaluated the quality of three existing annotation tools (MetaMap, cTAKES and AnnoMap) for real-world medical forms and proposed several combination approaches to improve their effectiveness and thus their practical applicability. We showed that post-processing the annotation results with group-based selection of annotation candidates as well as the aggregation of annotation results from two or more tools can substantially increase F-measure, for one of the tools even by a factor 3–4. In future work, we plan to investigate more sophisticated, e.g., supervised combination strategies that are tailored to the specific document corpus to annotate, and that are able to apply different weights when aggregating the results of different tools.

Acknowledgment. This work is funded by the German Research Foundation (DFG) (grant RA 497/22-1, "ELISA - Evolution of Semantic Annotations"), German Federal Ministry of Education and Research (BMBF) (grant 031L0026, "Leipzig Health Atlas") and National Research Fund Luxembourg (FNR) (grant C13/IS/5809134).

References

1. Abedi, V., Zand, R., Yeasin, M., Faisal, F.E.: An automated framework for hypotheses generation using literature. BioData Min. **5**(1), 13 (2012)
2. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. **17**(3), 229–236 (2010)
3. Campos, D., Matos, S., Oliveira, J.: Current methodologies for biomedical named entity recognition. In: Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, pp. 839–868 (2013)

4. Christen, V., Groß, A., Rahm, E.: A reuse-based annotation approach for medical documents. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 135–150. Springer, Cham (2016). doi:10.1007/978-3-319-46523-4_9
5. Christen, V., Groß, A., Varghese, J., Dugas, M., Rahm, E.: Annotating medical forms using UMLS. In: Ashish, N., Ambite, J.-L. (eds.) DILS 2015. LNCS, vol. 9162, pp. 55–69. Springer, Cham (2015). doi:10.1007/978-3-319-21843-4_5
6. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F., et al.: An efficient solution for mapping free text to ontology terms. In: AMIA Summit on Translational Bioinformatics 21 (2008)
7. Doan, S., Conway, M., Phuong, T.M., Ohno-Machado, L.: Natural language processing in biomedicine: a unified system architecture overview. In: Trent, R. (ed.) Clinical Bioinformatics. Methods in Molecular Biology (Methods and Protocols), vol 1168, pp. 275–294. Humana Press, New York (2014)
8. Dugas, M., Neuhaus, P., Meidt, A., Doods, J., Storck, M., Bruland, P., Varghese, J.: Portal of medical data models: information infrastructure for medical research and healthcare. Database: The Journal of Biological Databases and Curation p. bav121 (2016)
9. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. **11**(5), 392–402 (2004)
10. Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., Verspoor, K.: Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinform. **15**(1), 1–29 (2014)
11. Heinemann, F., Huber, T., Meisel, C., Bundschus, M., Leser, U.: Reflection of successful anticancer drug development processes in the literature. Drug Discovery Today **21**(11), 1740–1744 (2016)
12. Humphrey, S.M., Rogers, W.J., Kilicoglu, H., Demner-Fushman, D., Rindflesch, T.C.: Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. J. Am. Soc. Inform. Sci. Technol. **57**(1), 96–113 (2006)
13. LePendu, P., Iyer, S., Fairon, C., Shah, N.H., et al.: Annotation analysis for testing drug safety signals using unstructured clinical notes. J. Biomed. Semant. **3**(S-1), S5 (2012)
14. McCray, A.T., Srinivasan, S., Browne, A.C.: Lexical methods for managing variation in biomedical terminologies. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 235–239 (1994)
15. Oellrich, A., Collier, N., Smedley, D., Groza, T.: Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. PLoS ONE **10**(1), e0116040 (2015)
16. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**(5), 507–513 (2010)
17. Shah, N.H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A.P., Musen, M.A.: Comparison of concept recognizers for building the open biomedical annotator. BMC Bioinform. **10**(Suppl. 9), S14–S14 (2009)
18. Sohn, S., Kocher, J.P.A., Chute, C.G., Savova, G.K.: Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J. Am. Med. Inform. Assoc. **18**(Suppl. 1), i144–i149 (2011)

19. Sohn, S., Savova, G.K.: Mayo clinic smoking status classification system: extensions and improvements. In: AMIA Annual Symposium Proceedings, pp. 619–623 (2009)
20. Tanenblatt, M.A., Coden, A., Sominsky, I.L.: The ConceptMapper approach to named entity recognition. In: Proceedings of 7th Language Resources and Evaluation Conference (LREC), pp. 546–551 (2010)
21. Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., Jacobson, R.S.: NOBLE-Flexible concept recognition for large-scale biomedical natural language processing. BMC Bioinform. **17**(1), 32 (2016)
22. University of Pittsburgh: TIES-Text Information Extraction System (2017). http://ties.dbmi.pitt.edu/
23. Zheng, J., Chapman, W.W., Miller, T.A., Lin, C., Crowley, R.S., Savova, G.K.: A system for coreference resolution for the clinical narrative. J. Am. Med. Inform. Assoc. **19**(4), 660 (2012)
24. Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangarloo, H.: Indexfinder: a knowledge-based method for indexing clinical texts. In: AMIA Annual Symposium Proceedings, pp. 763–767 (2003)