Health and Technology (2018) 8:361–376
https://doi.org/10.1007/s12553-018-0261-3

ORIGINAL PAPER

Evolving semantic annotations through multiple versions of controlled medical terminologies

Silvio Cardoso[1,2] · Chantal Reynaud-Delaître[2] · Marcos Silveira[1] · Ying-Chi Lin[3] · Anika Groß[3] · Erhard Rahm[3] · Cédric Pruski[1]

Received: 14 September 2017 / Accepted: 13 September 2018 / Published online: 28 September 2018

Abstract
The extensive use of semantic annotations in the medical domain to enhance information retrieval or encode clinical notes to improving information reuse and sharing demands for high quality annotations generation and services for guaranteeing their validity over time. In this paper we present the extension of an existing framework supporting the (semi-)automatic maintenance of semantic annotations rendered outdated by the evolution of the knowledge organization system they are extracted from. This is done by the adding new rules and improving existing ones to overcome some shortcomings of the framework. We also propose an experimental evaluation of the extension using seven successive versions of four standard controlled terminologies within the domain.

Keywords Semantic annotations · Ontology evolution · Life sciences · Controlled terminologies

1 Introduction

The use of Knowledge Organization Systems (KOS) [22], such as classification schemes, controlled terminologies, thesauri or ontologies in the medical field has been gaining interest over the last years [15, 17, 41]. Usually, KOS entities are associated with medical documents such as clinical reports or medical images in order to make their semantics explicit for humans and software applications. The association between KOS entities (concepts, relationships, attributes, etc.) are called semantic annotations [10]. This process is usually carried out by domain experts or automatic annotators and the metadata produced bring many benefits for end users. Actually, MeSH annotations are used by the MEDLINE application to index scientific publications which enhances the retrieval of relevant documents. Moreover, the combined usage of semantic annotations and ontology mappings improves semantic interoperability between systems [40].

However, the dynamic nature of medical knowledge forces KOS content to be continuously revised. Thus, semantic annotations based on previous versions of the KOS can be impacted and lose their validity. For instance, a concept can be removed from a given KOS making its associated annotations obsolete. Therefore, concept and tools to adapt those impacted annotations to the new version of the KOS are urgently needed by virtue of the amount of annotated documents. In our previous work [6], we have shown a strong correlation between the modification of KOS elements and the modification of semantic annotations. We also manage to categorize the various evolution that can affect KOS and associate these changes with modifications

✉ Silvio Cardoso
silvio.cardoso@list.lu

Chantal Reynaud-Delaître
chantal.reynaud@lri.fr

Marcos Da Silveira
marcos.dasilveira@list.lu

Ying-Chi Lin
lin@informatik.uni-leipzig.de

Anika Groß
gross@informatik.uni-leipzig.de

Erhard Rahm
rahm@informatik.uni-leipzig.de

Cédric Pruski
cedric.pruski@list.lu

[1] LIST, Luxembourg Institute of Science and Technology, 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

[2] LRI, University of Paris-Sud XI, Orsay, France

[3] Department of Computer Science, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Germany

of elements defining annotations. We have also drawn the contours of a generic framework for maintaining annotations [7]. It is a multi-layer approach to manage semantic annotations when their underlying KOS evolve over time without re-annotating the documents. It implements rules, external background knowledge [33] and change patterns [12] to incrementally modify outdated semantic annotations. However, we observed some shortcomings with the current version of the framework [5]:

- the implemented rules do not considering the plural form of medical terms,
- the use of the termino-ontological resources composing the background knowledge is made without distinguishing the versions of the ontology used,
- change patterns are only considering the neighborhood of an evolving concept.

In this paper, we propose an extension of this framework. We introduce new rules and improve existing ones to overcome the gaps mentioned. Moreover, we experimentally assess the relevance of the extension of the framework using seven successive versions of four standard controlled medical terminologies: SNOMED CT, MeSH, NCIt and ICD-9-CM.

We structure the remainder of this paper as follows: Section 2 presents the related work of the field. Section 3 we introduce the basic notions used throughout the paper. Section 4 introduces the extension of our framework for (semi-)automatic maintenance of semantic annotations. Section 5 deals with the experimental evaluation of our approach. We present the obtained results and the discussion in Section 6 and 7. Finally, we wrap up with concluding remarks and outline future work in Section 8.

2 Related work

The previous work in annotation maintenance can be categorized into three families. The first one addresses the automatic detection of inconsistent annotations [13, 24, 34, 42]. This is mainly done by the combined identification of concepts that have changed from one KOS version to the next and their associated set of annotations. However, mechanisms to support the correction of impacted annotations are not proposed.

The second family of approaches focus on the automatic detection and manual correction of invalid annotations [1, 2, 4, 27]. These approaches only consider basic ontology changes, e.g., the deletion and addition of concepts in KOS while more complex changes are also important and need to be considered. Moreover, the requirement of human intervention to perform the maintenance is hardly applicable in the medical domain by virtue of the huge amount of annotations to adapt.

Lastly, the most advanced work implements an automatic correction of the annotation [14, 26, 30, 38]. This is mostly done based on reasoning techniques which rely on the logic formalism of the KOS. However, as medical KOS are mostly expressed and incorporate only lightweight description logics, these techniques must be adapted.

The literature review highlights that there is no annotation maintenance framework able to cope with the specificity of the medical domain e.g., size of the KOS, amount of annotations. Therefore, in this paper, we aim at further to improve our framework in [5, 7].

3 Background

In this section, we introduce the key notions that will be used throughout the paper. We start with a description of the annotation model used to formalize the semantic annotations of this work. We then present the initial version of the framework we aim to extend.

3.1 Model for formalizing semantic annotations

In our previous study [6] we proposed an annotation model that takes into account of the evolution and quality aspects for annotations. Herewith we give a brief overview of its main aspects.

A single annotation is defined as $= (i \;\; c \;\; \{q\}$ where an instance item $\in I_u$ is annotated with an ontology concept $c \in ON_v$, and a set of quality indicators $\{q\} \in Q$. An instance might be a text segment of an electronic health record (EHR), a question from a case report form (CRFs) as used within clinical trials. In general, a concept can be used to annotate many items and an item might be annotated with several concepts. Instance data might undergo modifications and resulting different versions, for instance, in the case of CRFs. Similarly, ontologies might be altered due to e.g., newly discovered knowledge. Hence, we included the subscripts u and v to denote the versions of instance data I and ontology ON. Different quality indicators q can be used to retain quality, reliability and provenance information for each annotation, e.g. by attaching numerical confidence values, categorical ratings or evidence codes [18].

The model also contains several elements for tracing the ontology changes and are used for the maintenance of the annotations. Firstly, we include $of \;\; set$ in the model. Due to ontology changes the range of the annotated text segment, i.e., the $of \;\; set$, might change. For instance, in the new ontology version a more precise concept is introduced. Such concepts generally consist of more words as opposed to more general concepts in the old version. With $of \;\; set$

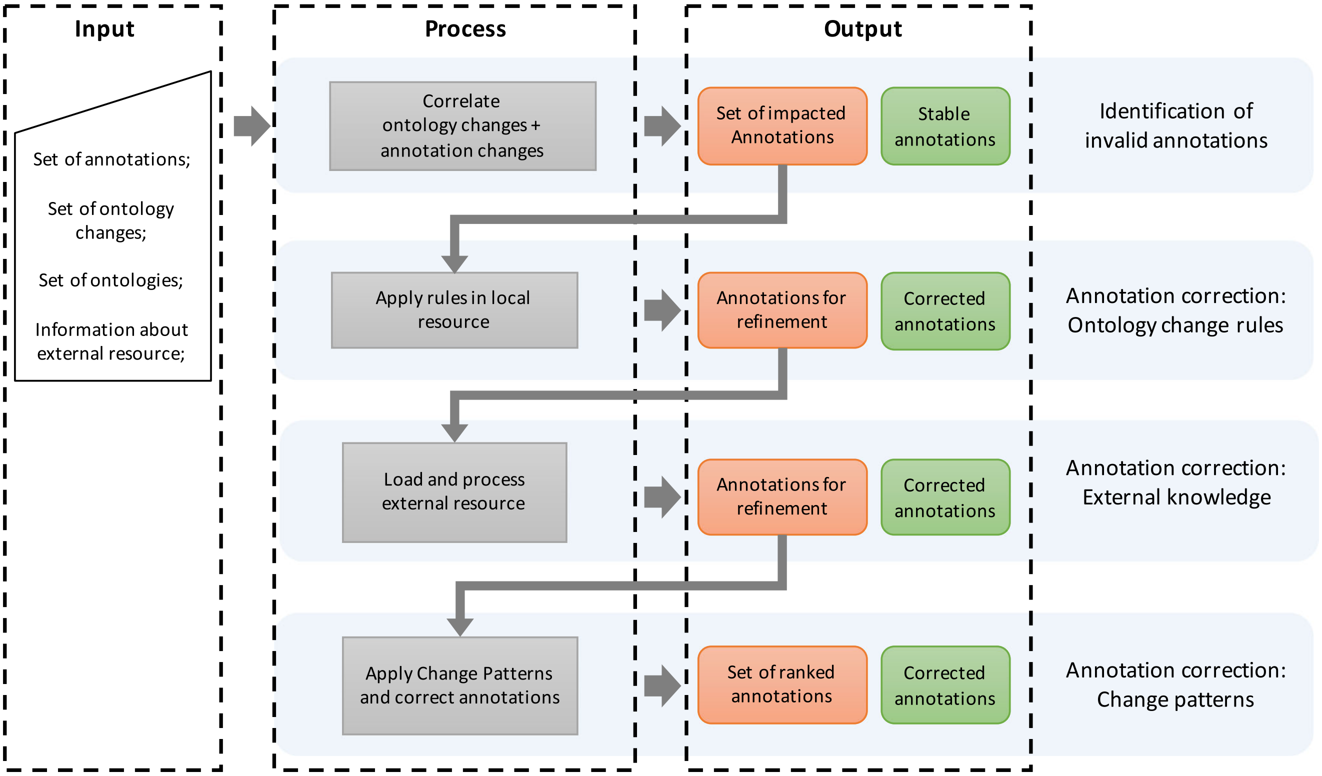
we can trace the two concepts annotating the same part of the text segment. We also incorporate an element indicating which concept attribute (e.g. title, synonym, preferred terms, etc.) was used to produce an annotation. This information can be used, for example, to determine if the used attribute of the concept is impacted by the ontology evolution and that further triggers the corresponding annotation modification. Moreover, we denote the *sem ntic type* of an annotation in the model to indicate the relationship between a concept and a text segment. For instance, instead of removing an impacted annotation after concept deletion, we preserve the annotated segment by linking it to the superclass of the removed concept and changing its semantic type to "less specific". We also included other relationships such as equivalent, more/less specific, partial match, and other ontology region.

3.2 Initial version of the framework

The framework we aim to extend has been largely presented in [5, 7]. As depicted in Fig. 1, it is a multi-level approach that implements rules, background knowledge and change patterns to maintain semantic annotations impacted by the evolution of the underlying KOS.

In this framework, the tasks carried out at each level can be summarized as follows:

1. **Identification of invalid annotations**. This consists of identifying invalid annotations by analyzing the evolution of the associated KOS using the COnto-Diff tool [21].
2. **Annotation correction using ontology change rules**. This consists of using information derived from the set of annotations itself as well as the data of the Diff between the two KOS versions coming from the previous level to adapt the invalid annotations identified. The annotation maintenance is governed by seven rules (*MergeAnnot, IncreaseAnnot, ResurrectAnnot, PluralAnnot, ChangeConceptAnnot, SplitAnnot, SuperClassAnnot*)), which are described in [7].
3. **Annotation correction using background knowledge**. This consists of using information inferred from external knowledge sources to maintain the annotations that could not be corrected using the local resources of the previous level.
4. **Annotation correction using change patterns**. This relies on the analysis of the morphosyntactic form of the attribute values used to generate the annotation and its evolution to decide which new attribute values can be used to keep the annotation valid. It is only applied when the previous approaches do not produce outputs [7]. The Change patterns are classified into lexical and semantic change patterns (LCP, SCP), respectively. The


Input
Process
Output
Set of annotations;
Set of ontology changes;
Set of ontologies;
Information about external resource;
Correlate ontology changes + annotation changes
Set of impacted Annotations
Stable annotations
Identification of invalid annotations
Apply rules in local resource
Annotations for refinement
Corrected annotations
Annotation correction: Ontology change rules
Load and process external resource
Annotations for refinement
Corrected annotations
Annotation correction: External knowledge
Apply Change Patterns and correct annotations
Set of ranked annotations
Corrected annotations
Annotation correction: Change patterns

ig. 1 The framework for supporting annotation maintenance. Source: [7]

LCP algorithm identifies four categories of the attribute value changes, i.e. Total Copy, Total Transfer, Partial Copy, and Partial Transfer. On the other hand, the SCP algorithm recognizes if the evolved attribute value is equal to, has become more or less specific or is a partial match of the original attribute value [12].

3.3 Ontology-based semantic similarity measures

Ontology-based semantic similarity measures (SSM) aim to estimate the likeness of two concepts considering the taxonomical knowledge modelled in ontologies [20]. It is used in a wide range of applications: to validate automatic annotations in Gene Ontology [9], information retrieval algorithms [37], Linked Data paradigms [28], etc.

There are numerous SSM in literature [9, 20, 28, 35]; in our framework, we focus on the pairwise measures which computes the semantic similarity between a pair of concepts and are commonly divided into four groups, described below:

i *Edge-based* measures which estimate the similarity of two concepts as a function of the distance which separates the two concepts in the ontology.

ii *Feature-based* rely on the taxonomic interpretation of the feature model proposed in Tversky [39]; generally, the representation of a concept corresponds to a set of neighbor concepts or instances. Feature-based strategies root semantic similarity in the context of classical binary or distance measures (e.g. set-based measures, vector-based measures).

iii *IC-based* measures assess the similarity of concepts as a function of the Information Content (IC) from their Most Informative Common Ancestor (MICA), e.g. the deepest concept which subsumes two verified concepts. [20, 35]

iv *Hybrid* measures which combines previous approaches.

These measures have been extensively evaluated across multiple benchmark and KOS [8, 16, 20, 31]. As result the IC-based measures in general outperform the edge-based. One of the main drawbacks of Feature-based measures is that they consider dimensions as mutually orthogonal and do not exploit concept relationships. Finally, the hybrid-measures require specifics parameters making a generalization across multiple KOS difficult. Therefore, in our framework, we focus on the use of IC-based measures.

3.4 Lexical Measures

In the biomedical domain, various Lexical Similarity Methods (LSM) have been used in order to improve the information retrieval of biomedical documents [36], support the mapping adaption process [12], improve the semantic relatedness between terms in named entity recognition process, e.g., “ammonium” ↔ “ammonium ion”, etc.

These LSM have been extensively evaluated [12, 36]. The results show that they are capable of improving the relatedness between the terms. However, different thresholds must be considered. Therefore, each domain must be carefully studied in order to apply such techniques.

In our framework we utilize LSM such as Levenshtein, TF-IDF, Jaro-Winkler, etc, during the annotation maintenance task, specifically in Change Patterns and `PartialMatch` rule, in order to improve the adaptation process to correctly evolve the annotations.

4 An extension of the framework

As discussed in Section 1, one of our objectives is to design a (semi-)automatic approach for maintaining semantic annotations valid over time even if the underlying KOS is evolving. This must be done without a complete re-annotation of the document and by guaranteeing a high quality in the annotation after maintenance. Although efficient for maintaining the validity of a semantic annotation, the initial framework described in Section 3 can be improved. Indeed, several limitations can be highlighted:

1. The existing `Rules` that exploit the morphosyntactic form of terms denoting attribute values, especially MergeAnnot, IncreaseAnnot and SplitAnnot, do not take the plural form of the terms into account. Depending on the complexity of the plural form, some maintenance decisions may recommend irrelevant concepts for annotation evolution, impacting the overall quality of the annotations.
2. The termino-ontological resources contained in the background knowledge are not exploited inline with the version of the KOS used to produce the annotations. As a result, the content of the KOS and the background knowledge are hardly comparable leading to bad maintenance decisions. For instance, in the Bioportal application, only the last version of an ontology is available. Thus, if the annotation version is not the same one, `BK` can provide the wrong evolution of the annotation.
3. The definition of the change patterns limits their scope to the neighborhood of a concept `Rules`, its direct super, sub-concepts and siblings. However, ontological change may result in moving a concept to another part of the new version of the ontology preventing the use of change patterns.

In order to overcome the above mentioned limitations, we have decided to extend the initial framework. In the following subsections we detail the proposed extensions.

4.1 Extension of existing rules to deal with plural

Since medical terms have various origins like ancient Greek or Latin, their plural form is a derived form in these languages. To this end, we have implemented the following well-accepted rules, in addition to the common English rules for plural, in order to take the evolution of medical terms from their singular to plural form and vice versa.

- Change the “a” ending term to “ae”
- Change the “um” ending term to “a”
- Change the “us” ending term to “i”
- Change the “is” ending term to “es”
- Change the “ma” or “oma” ending term to “mata”
- When a medical term ends in “yx”, “ax” or “ix” change the “x” to “c” and add “es”
- When a medical term ends in “nx”, change the “x” to “g” and add “es”
- For medical terms that have Latin roots and that are composed of a noun and adjective, both terms must include their plural form.

In order to integrate these rules, we had to modify the definition of MergeAnnot, IncreaseAnnot and SplitAnnot.

4.2 Adaptation of the background knowledge

In our approach, the use of background knowledge consists in reusing information inferred from external termino-ontological resources to maintain the annotations that could not be corrected using the `Rules` implemented at phase 2. Actually, in many cases, the evolution of ontological concepts can be characterized only by considering the semantic relationships provided by other ontologies [33]. Often labels of concept are completely different, from the syntactic point of view, before and after evolution. Therefore, considering only local resources does not allow the characterization of their evolution and, in turn, cannot be reused for annotation maintenance purpose. Nevertheless, the nature of the external knowledge sources can vary. Whether RDF datasets like BIO2RDF [3] or expressive OWL ontologies contained in Bioportal [29] are considered, the inferred information can be of different quality and can affect the quality of the maintenance process.

In our previous implementation of BK [5], relying on Bioportal, provided complementary results for the above `Rules`, the AUC for ICD-9-CM and MeSH increased from 0.899 to 0.915 and 0.850 to 0.863 respectively. However, it also provided unaligned mappings to past KOS versions leading to the development of an additional phase to filter inconsistent result. We also observed that the concept labels in SNOMED CT and ICD-9-CM are not unique *e.g.*, in SNOMED CT concepts 31113003, 397881000, and 68047000 have the same label “diverticulosis”. Therefore, this ambiguity can lead the system to select the inappropriate concept to replace the impacted annotations needing a disambiguation phase.

To this end, we have improved our framework to overcome these two limitations. Considering the problem of concept version, we are still using the last version of the KOS provided by the BK. But, we consider only the mapping provided by Algorithm 1 as candidate mapping. In this case, we filter the mappings retrieved by the BK (lines 1 to 3) keeping only those which exists in the new ontology version KOS_{v1}. The next step (line 4) retrieves all stable ancestors of a source concept $_s$ within a specified period, e.g., (2009/2010). From all candidates that satisfy the previous conditions, we compute the similarity between the ancestors and the source concept (lines 5 to 6). Then, we take the most similar ancestor MS (line 7). Finally we select the best candidate to maintain our annotation (lines 8-10). It is the mapping which presents the highest similarity to the MS .

To overcome the ambiguity problem, we use an adequate similarity measure able to correctly determine which concept is the most similar to the one used to annotate it before its evolution. In the first version of our framework we used the Tversky similarity measure [39]. However, this metric exploits the intersection between the features of two concepts, e.g. siblings, sub and super classes to determine their similarity. This metric failed when using a flat ontology *e.g.*, ICD-9-CM. For this reason, we replaced the Tversky similarity measure by the Jiang-Conrath (JC) [23] one (see Section 4.3).

Algorithm 1 Similarity between mappings and ontology source in the Background Knowledge. MSA: the most similar ancestor.

```
Input: Concept source $C_s$; Mappings $MappingSet$ ;
       Ontology v0 $KOS_{v0}$; Ontology v1 $KOS_{v1}$
Output: Concept Target $C_t$
forall the $obj \in MappingSet$ do
    if $(obj \in KOS_{v1}) == TRUE$ then
        $ValidMappings \leftarrow obj$
$Set\_Sup\_Classes \leftarrow$
$getAllStableAncestor(C_s, KOS_{v0}, KOS_{v1})$
forall the $obj \in Set\_Sup\_Classes$ do
    $calSemanticDistance(obj, C_s, KOS_{v0})$
$MSA \leftarrow$
$getMostSimilarAncestor(Set\_Sup\_Classes)$
forall the $obj \in ValidMapping$ do
    $calSemanticDistance(MSA, obj, KOS_{v1})$
$C_t \leftarrow getHighSimilarity(ValidMappings)$
return $C_t$
```

4.3 artial Match and Change atterns

The major limitation of the framework is due to the limited scope of the change patterns to correct the annotations that are still invalid after the application of the `Rules` and of the `BK`. As change patterns consider the local evolution of concept, we add a new rule, called `PartialMatch`, able to deal with global evolution of a concept.

This rule changes the term and/or the concept ID of an annotation considering Lexical Similarity Measure (LSM) and Semantic Similarity Measure (SSM). We defined it based on the results we had using Semantic Change Patterns (`SCP`) in [5]. The analysis of `SCP` showed good precision and low recall. The reason here is that `SCP` considers only changes between concepts that are in the same neighbourhood `Rules`, siblings, super-, and sub-concepts. Therefore, new methods which also consider other ontology regions are needed. To combine both measures and compute this rule we utilized an arithmetic mean, where LSM and SSMs represents the similarity values in the interval of [0,1]. Therefore, we calculate the `PartialMatch` as:

$$sim(c_1,c_2 = \frac{LSM + SSMs}{2} \tag{1}$$

In our framework, we choose Jiang Conrath 1997 (JC) [23] as SSMs and AnnoMap [25] as LSM. The choice of JC was based on its usage and good performance in multiple domains [8, 16, 20, 23, 31]. Regarding the use of AnnoMap, it showed the best performance in our previous work annotating clinical forms [25].

The similarity computed by AnnoMap, see Eq. 2, is based on the combined similarity score from different string similarity functions, in particular TF/IDF, Trigram and LCS (longest common substring).

$$sim_{AnnoM\ \ p}(t_1,t_2 \ = MA\ \ (\ \ \ Id\ \ \ \ riGram\ \ L\ \ S \tag{2}$$

The Jiang Conrath utilized (see Eq. 3) is the adaptation proposed in [19]. It calculates the MICA between two concepts c_1,c_2 and the IC of each one considering structural information extracted from the ontology, i.e., Intrinsic Information Content (iIC) which avoids the dependency of a corpus to calculate the concept usage. Furthermore, they prevent errors related to bias on concept usage, e.g., many annotations associated to two different concepts do not mean they are similar.

In our framework, we utilized the Depth Max Linear [19] to compute the iIC. This approach considers that the depth of a concept w.r.t in a Graph G is proportional to its degree of expressiveness. Therefore, the MICA in Eq. 3 is calculated in function of the depth from Most Specific Common Abstraction (MSCA) that subsumes both concepts c_1 and c_2, expressed as $dept\ \ (MS\ \ \ \ (c_1,c_2\ \ $ [43].

$$sim_{JC}(c_1,c_2 \ = 1 - \frac{IC(c_1\ +IC(c_2\ -2 \cdot MI\ \ \ \ (c_1,c_2}{2} \tag{3}$$

Thus, the new proposed `PartialMatch` rule is capable to maintain semantic annotations by changing the term/concept even if it is in a different ontology region as the following example: *167696007:feces examination* in SNOMED CT 2009 → *167592004:examination of feces* in SNOMED CT 2010.

5 Methodology

This section describes the data and method used to evaluate our approach. The main points are: i) the terminologies ii) the extension of our evaluation dataset (silver standard) [5], iii) the method used to evolve impacted annotations through seven successive versions of four standard KOS and iv) the metrics for evaluation.

5.1 erminologies

As described in Section 4 our maintenance method utilizes consecutive KOS versions in order to detect and adapt the impacted annotations. In our experiments we have used: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), Medical Subject Headings (MeSH), National Cancer Institute Thesaurus (NCIt) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT). We are using the versions 2009AA to 2016AA (excluding the AB versions), downloaded from the UMLS and we transformed into OWL files.[1] To compute the difference between the terminologies which are strongly correlated with the validity of annotations [6], we are using COnto-Diff [21].

2 S lver Standard

Since no annotation baseline generated with successive ontologies versions for the above terminologies exists in the literature, we had to improve our silver standard described in [5] including the reference for 2016AA.

Table 1 shows an illustrative example related to our silver standard. It shows one annotation produced with the MeSH:2009AA using the PubMed document 232[2] and the concept D009133. The annotated text is "muscular atrophy", and it can be found at position [5561,5577] of

[1] https://git.list.lu/ELISA/ParserUMLS

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638829/

Table 1 Example of an evolving annotation, extracted from our silver standard

KOS	Doc.	Concept	Annotation	Start	End	Prefix	Suffix
2009AA	232	D009133	muscular atrophy	5561	5577	(HD), spinal and bulbar	, drpla and various forms
2010AA	232	D055534	spinal and bulbar muscular atrophy	5543	5577	(HD),	, drpla and various forms
2016AA	232	D020966	spinal and bulbar muscular atrophy	5543	5577	(HD),	, drpla and various forms

The red color indicates the changes that occurred in the annotation at KOS evolution time

the document. We customized our system to have four words as a prefix *"(HD), spinal and bulbar"* and *", drpla and various"* as suffixes. It can be observed that the concept label and ID used to annotate the text increased and changed respectively, from 2009AA to 2010AA. Furthermore, in 2016AA the concept ID changed from D055534 to D020966. Therefore, we have an annotation impacted multiple times by the evolution of the MeSH.

The silver standard, can be downloaded from http://www.elisa-project.lu/, under menu publications/downloads. We adopt the term "silver" to indicate that our reference is based on only one viewpoint, i.e., each expert validated a set of annotations and no discussions between them were organized.

5.3 Experimental etup

In order to evaluate the capacity of our framework to adapt impacted annotations into consistent ones, we utilized the three different configurations described below:

- The **first Setup** verify if the extensions of `Rules` and `BK` (see Sections 4.1 and 4.2, respectively) improve our framework, we compare the new implementation to the results presented in [5]. The comparisons include not only single component of the framework (i.e., `Rules`, `BK`, `SCP`, `LCP`) but also the combinations of them (e.g., `Rules/BK`, `Rules/SCP`, Combination of all). Note that these evaluations do not include the extension of `PartialMatch` as whose effectiveness is examined in the next Setup.
- The **second Setup** utilizes only basic pipelines, i.e., `Rules`, `SCP` and `LCP` without their combinations. Furthermore, the `Rules` only contain the `PartialMatch`. Thus, we will be able to verify our second proposed objective of Section 1, i.e., if our newly proposed rule is able to outperform the other techniques.
- The **third Setup** we first determine the position for the `PartialMatch` rule. Since we did not find any annotation guideline related to it, we tested two possibilities: i) Before the *SuperClassAnnot* ii) After the *SuperClassAnnot*. After the determination of the `PartialMatch` positioning, we evaluate the framework with all the extensions proposed in Section 4 and present the best pipelines for each terminology.

For the first Setup we used the silver standard version 2009AA/2010AA in order to compare the results with the previous study in [5]. For the other two Setups we also evaluated the successive evolution from 2009AA to 2016AA. For all the Setups, we tested the effectiveness of methods in two aspects: i) the capacity of our framework to detect impacted annotations after changing a KOS concept; and ii) the ability to correctly evolve the impacted annotations into consistent ones. In this case, consistency means adequacy with the silver standard.

Metrics

To evaluate the effectiveness of our method, i.e., whether our predictions are similar to the silver standard we used classic well-known metrics, such as, *Precision*, *Recall*, *F1-score*, *Area Under the Curve* (AUC) and *Accuracy* [32].

6 Results

Setup 1 The first results for this set-up concern the capacity of our framework to detect impacted annotations (cf. Table 2). The first column of this table shows the pipelines used, i.e., `BK`, `Rules`, `SCP`, `LCP` and their combinations.

Table 2 Precision (P), Recall (R) and F1-Score (F1) of adaptation of the impacted annotations computed using three different methods (BK, SCP, Rules) and the combination of them in *Setup 1*

	ICD-9-CM			MeSH			NCIt			SNOMED CT		
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BK in [3]	1	0.161	0.278	1	0.025	0.049	1	0.135	0.237	0.963	0.542	0.693
BK	1	0.129	0.229	1	0.050	0.094	1	0.115	0.207	1	0.625	0.769
Rules in [3]	1	0,984	0.992	0.982	0.933	0.957	0.979	0.885	0.929	1	0.792	0.884
Rules	1	0,982	0.991	0.991	0.941	0.966	0.980	0.942	0.961	0.929	0.812	0.867
SCP in [3]	1	0.081	0.149	1	0.008	0.017	0.833	0.096	0.172	0	0	0
SCP	1	0.041	0.078	1	0.091	0.167	0	0	0	0	0	0
LCP	1	0.048	0.092	1	0.099	0.180	1	0.019	0.038	0	0	0
BK/Rules in [3]	1	0.984	0.992	0.982	0.933	0.957	0.979	0.885	0.929	0.975	0.813	0.886
BK/Rules	1	0.982	0.991	0.991	0.941	0.966	0.980	0.942	0.961	0.929	0.812	0.867
BK/SCP in [3]	1	0.194	0.324	1	0.025	0.049	0.909	0.192	0.317	0.963	0.542	0.693
BK/SCP	1	0.161	0.278	1	0.140	0.246	0.857	0.115	0.203	1	0.625	0.769
BK/LCP	1	0.161	0.278	1	0.149	0.259	1	0.135	0.237	1	0.625	0.769
Rules/SCP in [3]	1	0.984	0.992	0.982	0.933	0.957	0.979	0.885	0.929	1	0.792	0.884
Rules/SCP	1	0.982	0.991	0.991	0.941	0.966	0.980	0.942	0.961	0.929	0.812	0.867
Rules/LCP	1	0.982	0.991	0.991	0.941	0.966	0.980	0.942	0.961	0.929	0.812	0.867
CombineAll in [3]	1	0.984	0.992	0.982	0.933	0.957	0.979	0.885	0.929	0.975	0.812	0.886
CombineAll	1	0.982	0.991	0.991	0.941	0.966	0.979	0.940	0.959	0.929	0.812	0.867

The red and blue colours indicate decrease and improvement of recall, respectively

The first line shows the KOS used, ICD-9-CM, MeSH, NCIt, SNOMED-CT and for each column in these KOS, the Precision (P), Recall (R) and F1-score (F1) values are demonstrated. We verified that there was an improvement in all of the methods when compared to [5].

Regarding the MeSH terminology, all values associated with recall and F1-score increased for all configurations. In contrast, results for ICD-9-CM did not improved, but this is a minor difference of 0.2%. Regarding SNOMED-CT, except for the BK method, it can be noticed that the values obtained are similar to those in [5] or had a minor gain of 2.5%. Results for NCIt showed a significant improvement. The Rules were capable of reaching 0.942 for recall and 0.961 for F1-Score. A gain of 6% is demonstrated in the recall and 3% in F1-Score when compared to our previous values of 0.885 and 0.926. Finally, the changes for precision only demonstrated an expressive difference for the SCP method in NCIt. The newly proposed configurations were not capable of finding impacted annotations in NCIt. The reason for this is discussed in Section 7. Nevertheless, we obtained acceptable results for precision with a minimal value of 0.929 considering all the other pipelines.

The second evaluation consists in the application of the pipelines, BK, Rules, etc, to find an adaptation for the annotations impacted. The results obtained (cf. Table 3) showed that the AUC and F1-score values increased for MeSH, NCIt and SNOMED CT. The changes in the Rules method were capable of smoothly improving the AUC results to 3.76% for MeSH, 1,39% for NCIt and 1,32% for SNOMED CT. On the other hand, the method applied to ICD-9-CM loss 4,12% of its capacity to propose correct adaptations.

Regarding the BK method, we verified that for SNOMED CT, it improved to 5.93% of the AUC and 13.44% for the F1-Score. The other terminologies showed the same results or a maximal loss of 2.61%. Furthermore, the SCP and LCP still not produced results for SNOMED CT. We also verified that the LCP technique was capable of providing better results than the SCP. It occurs when we used this method alone or in combination with other methods, e.g., BK/LCP, Rules/LCP. The AUC results show a difference in favor of LCP around 2,2% for ICD-9-CM and 1,23% for NCIt.

Finally, the combination of all methods showed some variations when compared with our previous work. The terminologies SNOMED CT and MeSH had an improvement in the AUC of 1.3% and 3% respectively. While NCIt demonstrated the same values and ICD-9-CM showed a

Table 3 Accuracy (ACC), Area Under the Curve (AUC) and F1-Score (F1) values of developed heuristics used to maintain annotations in *Setup 1*

	ICD-9-CM			MeSH			NCIt			SNOMED CT		
Method	AC	AUC	F1	AC	AUC	F1	AC	AUC	F1	AC	AUC	F1
`BK` in [3]	0.518	0.613	0.368	0.457	0.554	0.195	0.611	0.663	0.493	0.699	0.708	0.588
`BK`	0.497	0.597	0.324	0.442	0.545	0.167	0.611	0.663	0.493	0.742	0.75	0.667
`Rules` in [3]	0.874	0.899	0.888	0.817	0.850	0.824	0.678	0.721	0.613	0.828	0.833	0.800
`Rules`	0.834	0.862	0.839	0.856	0.882	0.867	0.689	0.731	0.632	0.839	0.844	0.815
`SCP` in [3]	0.492	0.593	0.313	0.452	0.550	0.182	0.544	0.606	0.349	0.484	0.500	0
`SCP`	0.49	0.589	0.303	0.472	0.57	0.246	0.556	0.615	0.375	0.484	0.500	0
`LCP`	0.487	0.589	0.301	0.472	0.57	0.246	0.567	0.625	0.4	0.484	0.500	0
`BK/Rules` in [3]	0.894	0.915	0.907	0.832	0.862	0.841	0.678	0.721	0.613	0.828	0.833	0.80
`BK/Rules`	0.797	0.83	0.796	0.841	0.87	0.85	0.689	0.731	0.632	0.839	0.844	0.815
`BK/SCP` in [3]	0.503	0.601	0.336	0.457	0.554	0.195	0.611	0.663	0.493	0.699	0.708	0.588
`BK/SCP`	0.487	0.589	0.301	0.482	0.579	0.271	0.611	0.663	0.493	0.742	0.75	0.667
`BK/LCP`	0.487	0.589	0.301	0.482	0.579	0.271	0.622	0.673	0.514	0.742	0.75	0.667
`Rules/SCP` in [3]	0.869	0.895	0.883	0.802	0.838	0.806	0.689	0.731	0.632	0.828	0.833	0.80
`Rules/SCP`	0.786	0.821	0.783	0.851	0.878	0.861	0.689	0.731	0.632	0.828	0.833	0.800
`Rules/LCP`	0.807	0.839	0.809	0.851	0.878	0.861	0.7	0.74	0.649	0.839	0.844	0.815
CombineAll in [3]	0.905	0.923	0.917	0.832	0.862	0.841	0.689	0.731	0.632	0.828	0.833	0.80
CombineAll	0.786	0.821	0.783	0.862	0.887	0.872	0.693	0.73	0.63	0.839	0.844	0.815

The red and blue color highlight the lower and higher values for each dataset, respectively

loss of 11,05%. These variations are further discussed in Section 7.

Setup 2 The results for this Setup are demonstrated in Figs. 2 and 3. Figure 2 shows the results concerning the ability of these methods to detect impacted annotations, while Fig. 3 shows the ability to propose correct adaptations for the impacted annotations. We are utilizing the references (2009/2010) and (2009/2016) of our silver standard.

As observed in Fig. 2, the precision of all methods is high for both cases (2009/2010) or (2009/2016). However, the recall varies according to the terminology used and the year. The terminologies SNOMED CT and NCIt in (2009/2010) show the highest recall for the `Rules`, while `SCP` and `LCP` show null values or close to 0.

In the following years (2009/2016), SNOMED CT showed an improvement of 18% when compared to the previous version (2009/2010), while the other terminologies have a smooth variation. In short, the proposed rule outperforms the `SCP` and `LCP` in all terminologies and years in detecting impacted annotations. This result is very clear when we observe SNOMED CT in Fig. 2.

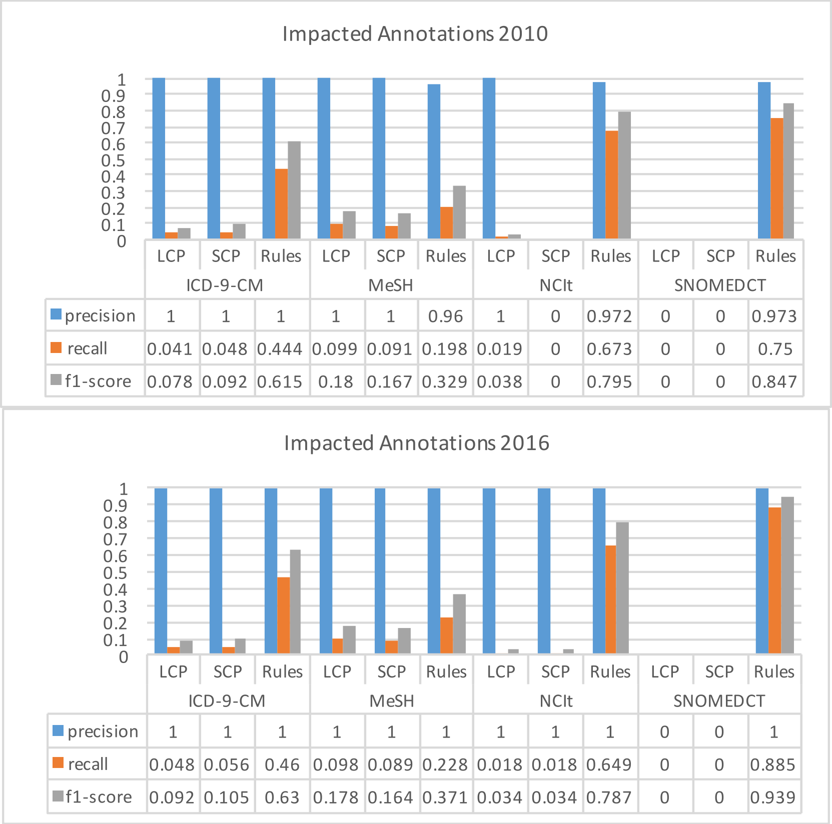
Considering the ability to provide correct adaptations in our Setup 2. The `Rules` also demonstrated good results for all KOS versions (2009/2010) and (2009/2016), see Fig. 3. The AUC values for NCIt in 2009/2010 are 14% and 15.77% greater than `LCP` and `SCP` respectively. This difference increases when we observe the F1-Score reaching 33% for `LCP` and 37% for `SCP`.

The other terminologies ICD-9-CM and MeSH also show better results for the `PartialMatch` rule. The observed difference in ICD-9-CM is 19.72% of F1-Score and 23.62% for AUC. While in MeSH this difference is less expressive reaching 2% of AUC and 9% of F1-Score.

Setup 3 The results concerning these experiments are showed in Table 4. We verified that only ICD-9-CM 2009/2010 and SNOMED-CT 2010/2016 showed best performance when the `PartialMatch` is placed before the *SuperClassAnnot*. It also has a huge impact in NCIt, since the F1-Score shows a difference of 11% in 2009/2010 and 13% in 2009/2016. Therefore, our next results concern only the usage of `PartialMatch` **before** the *SuperClassAnnot* to adapt the annotations.

Table 5 shows the best pipelines for the period: 2009/2010. In the first column we have the KOS used, i.e., ICD, MeSH, NCIt and SNOMED CT. The configuration described in Section 5.3 is mentioned in the second column and the third column shows the pipelines followed by the metrics.

Fig. 2 Performance of methods in *Setup 2* to detect impacted annotations



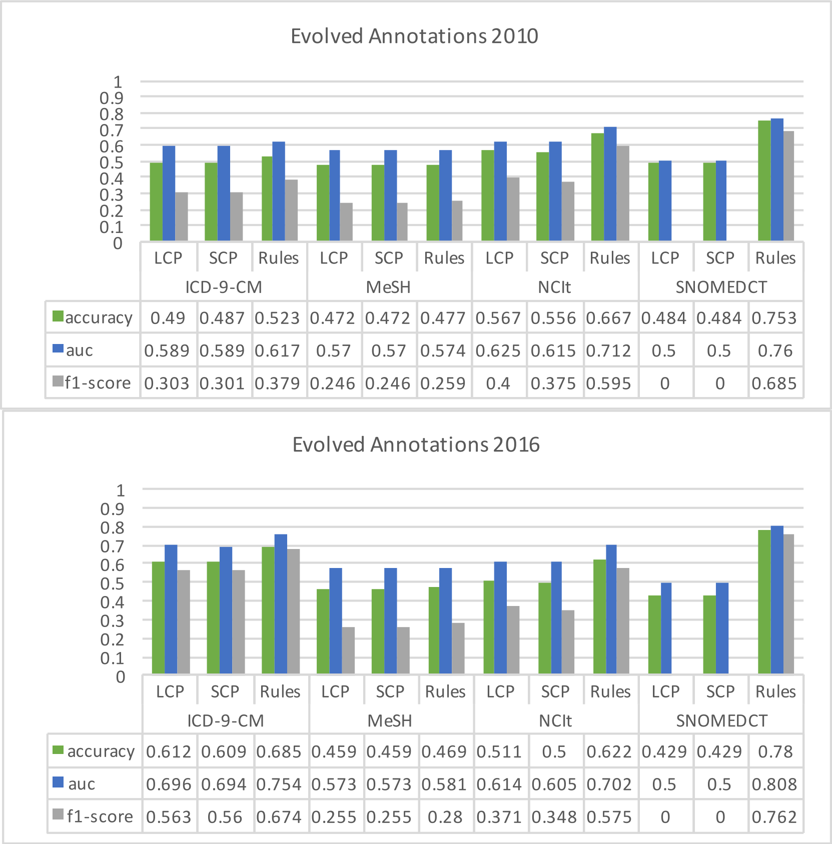
Impacted Annotations 2010

	ICD-9-CM			MeSH			NCIt			SNOMEDCT		
	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules
🟦 precision	1	1	1	1	1	0.96	1	0	0.972	0	0	0.973
🟧 recall	0.041	0.048	0.444	0.099	0.091	0.198	0.019	0	0.673	0	0	0.75
⬜ f1-score	0.078	0.092	0.615	0.18	0.167	0.329	0.038	0	0.795	0	0	0.847

Impacted Annotations 2016

	ICD-9-CM			MeSH			NCIt			SNOMEDCT		
	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules
🟦 precision	1	1	1	1	1	1	1	1	1	0	0	1
🟧 recall	0.048	0.056	0.46	0.098	0.089	0.228	0.018	0.018	0.649	0	0	0.885
⬜ f1-score	0.092	0.105	0.63	0.178	0.164	0.371	0.034	0.034	0.787	0	0	0.939

Fig. 3 Performance of methods in *Setup 2* to adapt impacted annotations



Evolved Annotations 2010

	ICD-9-CM			MeSH			NCIt			SNOMEDCT		
	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules
🟩 accuracy	0.49	0.487	0.523	0.472	0.472	0.477	0.567	0.556	0.667	0.484	0.484	0.753
🟦 auc	0.589	0.589	0.617	0.57	0.57	0.574	0.625	0.615	0.712	0.5	0.5	0.76
⬜ f1-score	0.303	0.301	0.379	0.246	0.246	0.259	0.4	0.375	0.595	0	0	0.685

Evolved Annotations 2016

	ICD-9-CM			MeSH			NCIt			SNOMEDCT		
	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules	LCP	SCP	Rules
🟩 accuracy	0.612	0.609	0.685	0.459	0.459	0.469	0.511	0.5	0.622	0.429	0.429	0.78
🟦 auc	0.696	0.694	0.754	0.573	0.573	0.581	0.614	0.605	0.702	0.5	0.5	0.808
⬜ f1-score	0.563	0.56	0.674	0.255	0.255	0.28	0.371	0.348	0.575	0	0	0.762

Table 4 Experiments to determine the place of PartialMatch

KOS	Year	Config	Accuracy	AUC	F1-Score
ICD-9-CM	2009/2010	Before	0.834	0.862	0.839
		After	0.845	0.871	0.851
	2009/2016	Before	0.757	0.803	0.754
		After	0.741	0.789	0.733
Mesh	2009/2010	Before	0.867	0.891	0.877
		After	0.851	0.878	0.861
	2009/2016	Before	0.88	0.905	0.895
		After	0.859	0.888	0.874
NCIt	2009/2010	Before	0.767	0.798	0.747
		After	0.697	0.735	0.64
	2009/2016	Before	0.753	0.804	0.756
		After	0.685	0.75	0.667
SNOMED-CT	2009/2010	Before	0.849	0.854	0.829
		After	0.839	0.844	0.815
	2009/2016	Before	0.912	0.923	0.917
		After	0.923	0.933	0.928

The blue values are related to the best performance

As first result we observed that only the new Rules are capable to reach the same results than its combination with BK method. Furthermore, the use of BK for annotations generated with ICD-9-CM and SNOMED CT leaded to smoothly decrease the values.

It also noticed that the new Rules were capable to outperform (or have the same results) the best pipelines of *Setup 1* and the previous study in [5]. The major difference is related to NCIt showing, 7.84% of improvement in AUC and 15% for F1-Score. While in the previous study we have 10% of improvement.

Besides, we compared these results using Sign Test [11]. It is a non-parametric test used to verify whether or not two groups are equally sized, i.e., the amount of success cases remains the same, before and after a procedure, see Table 6. In the first column we have the used terminologies. It is followed by the configuration to maintain the annotations in columns two and three, respectively. In the fourth column we have the amount of annotations correctly maintained, where the labels *before* and *after* are related to the second column (*Config*). Finally, we have the amount of impacted annotations in our silver standard, followed by the

Table 5 Results regarding the adaptation of annotations of *Setup 3* during the period 2009/2010

KOS	Config	Method	ACC	AUC	F1
	3	BK/Rules PM	0.856	0.879	0.863
		Rules PM	0.834	0.862	0.839
ICD9CM	1	Rules	0.834	0.862	0.839
	3	BK/Rules PartialMatch	0.867	0.891	0.877
		Rules PartialMatch	0.867	0.891	0.877
MeSH	1	CombineAll	0.862	0.887	0.872
	3	BK/Rules PartialMatch	0.767	0.798	0.747
		Rules PartialMatch	0.767	0.798	0.747
NCIT	1	LCP/Rules	0.7	0.74	0.649
	3	BK/Rules PartialMatch	0.828	0.833	0.8
		Rules PartialMatch	0.849	0.854	0.829
SNOMEDCT	1	BK/Rules	0.839	0.844	0.815

The blue values indicate the best pipelines

Table 6 Results regarding the Sign Test

			Corrected			
Kos	Config	Method	Before	After	Annots	p-value
	Before: Previous in [3] After: Setup 3	Rules	99	81	124	0,000011
	Before: Previous in [3] After: Setup 3	BK/Rules	103	85	124	0,000011
	Before: Previous in [3] After: Setup 3	CombineAll	105	85	124	0,0000039
	Before: Setup 1 After: Setup 3	Rules	81	81	124	0,5
	Before: Setup 1 After: Setup 3	BK/Rules	74	85	124	0,000455
ICD-9-CM	Before: Setup 1 After: Setup 3	CombineAll	72	85	124	0,000155
	Before: Previous in [3] After: Setup 3	Rules	84	93	120	0,0013499
	Before: Previous in [3] After: Setup 3	BK/Rules	87	93	120	0,0071529
	Before: Previous in [3] After: Setup 3	CombineAll	87	93	120	0,0071529
	Before: Setup 1 After: Setup 3	Rules	91	93	120	0,07864
	Before: Setup 1 After: Setup 3	BK/Rules	88	93	120	0,01267
MeSH	Before: Setup 1 After: Setup 3	CombineAll	92	93	120	0,1586
	Before: Previous in [3] After: Setup 3	Rules	23	31	52	0,0023389
	Before: Previous in [3] After: Setup 3	BK/Rules	23	31	52	0,0023389
	Before: Previous in [3] After: Setup 3	CombineAll	24	31	52	0,0040755
	Before: Setup 1 After: Setup 3	Rules	24	31	52	0,0040755
	Before: Setup 1 After: Setup 3	BK/Rules	24	31	52	0,0040755
NCIt	Before: Setup 1 After: Setup 3	CombineAll	23	31	52	0,0023389
	Before: Previous in [3] After: Setup 3	Rules	32	34	48	0,0786496
	Before: Previous in [3] After: Setup 3	BK/Rules	32	32	48	0,5
	Before: Previous in [3] After: Setup 3	CombineAll	32	32	48	0,5
	Before: Setup 1 After: Setup 3	Rules	33	34	48	0,1586553
	Before: Setup 1 After: Setup 3	BK/Rules	33	32	48	0,1586553
SNOMED CT	Before: Setup 1 After: Setup 3	CombineAll	33	32	48	0,1586553

The values in blue refers to p 0.05 and red values p = 0.05

p-value that indicates whether the Sing Test was capable or not to refuse the null hypothesis: H_0 : *Population median difference = 0.*

As result, we verified that MeSH and NCIt have significant differences between the methods. In NCIt all the tests refused the null hypothesis, it means that the new methods are capable to maintain more annotations than the previous ones. In a real world scenario with thousands of annotations our contribution is even more evident. Consider for instance, the impacted annotations related to MeSH in [6], around 367400 annotations. Using the configuration: *Config: Before: Previous in* [5] *After: Setup 3; Method: CombineAll.* We will be able to maintain up to annotations 266367 using the old method and 284738 using the new methods. It is a difference of 6.66%, around 18370 annotations, that will have huge impact in the day-to-day work of the healthcare facility.

The good performance of the new Rules are also verified in the period 2009/2016. The results in Table 7

Table 7 Results regarding the adaptation of annotations of *Setup 3* during the period 2009/2016

KOS	Config	Method	ACC	AUC	F1
		BK/Rules PartialMatch	0.73	0.781	0.719
	3	Rules PartialMatch	0.757	0.803	0.754
ICD9CM	1	SCP/Rules	0.676	0.737	0.643
		BK/Rules PartialMatch	0.875	0.901	0.89
	3	Rules PartialMatch	0.88	0.905	0.895
		SCP/Rules	0.859	0.888	0.874
MeSH	1	CombineAll	0.859	0.888	0.874
		BK/Rules PartialMatch	0.75	0.8	0.75
	3	Rules PartialMatch	0.753	0.804	0.756
		BK/Rules	0.685	0.75	0.667
	1	SCP/Rules	0.685	0.75	0.667
		LCP/Rules	0.685	0.75	0.667
NCIT		CombineAll	0.685	0.75	0.667
		BK/Rules PartialMatch	0.901	0.913	0.905
	3	Rules PartialMatch	0.912	0.923	0.917
SNOMEDCT	1	Rules	0.923	0.933	0.928

The blue values indicate the best pipelines

shows that, except for SNOMED CT, it is capable to outperform all the other techniques applied to the other KOS. Furthermore, it shows expressive differences when we compare with its application to NCIt and ICD-9-CM. The F1-Score is 13.34% bigger for NCIt and 17.26% for ICD-9-CM. Detailed explanations for these observations are provided in the next section.

7 Discussion

The evaluation showed that the evolution of annotations can be substantially supported by our framework, either by using one consecutive year or through multiple successive versions of KOS. The outcomes presented in Section 6 demonstrated that we were capable to obtain high F1-Score and AUC for the whole maintenance process.

When analyzing the results provided by each method, we verified that the BK method had significant changes. The implemented filter solved the problem of finding unaligned mappings and increased the F1-Score and AUC. However, our modifications also impact the specificity of the adapted annotations. For instance, in ICD-9-CM the mappings related to *"acute renal failure"*, concepts 584 and 584.9. When computed by our method, the chosen candidate was 584, because it is the most similar to the stable ancestor 580-589.99. In our silver standard the valid reference to adapt the annotation *"acute renal failure" 584.9*, is *"acute kidney failure" 584.9*, i.e., the same concept using the new term. However, this mapping can also be considered as valid because both refers to the term "acute kidney failure". Therefore, our method is able to provide right adaptations and our silver standard can be improved for the next usage.

The change in the Rules also provided better results. For instance, the inclusion of plural forms in *IncreaseAnnot, MergeAnnot* and *SplitAnnot* increase the recall during the detection of impacted annotations. Moreover, it was capable to provide the right adaptions. As example *['PMC2642994'; 'D002875'; 'chromosome'; 36563; 36573]* in MeSH 2009 was correctly changed to *['PMC2642994', 'D056905', 'chromosome breakpoints', 36563, 36585]* in MeSH 2010.

It smoothly contributed to improve the results showed in Section 6 for all terminologies. However, we also found some limitations. For example, in SNOMED CT the annotation *[PMC2633322; '31113003'; 'diverticulum'; 6412; 6424; 'association with the meckel'; ', the appendix,']* had to evolve to *"Meckel diverticulum"* ConceptID *37373007*, in order to be similar to our silver standard. In spite of that, *IncreaseAnnot* provided random results, adapting this annotation sometimes using the ConceptID *37373007* or *127962001*. It is sharp to observe in the method *Rules/SCP* since the change patterns do not provide results for SNOMED CT. The main reason for this miss adaptation, is that SNOMED CT includes those two different concepts in version 2010 using the same term *"Meckel diverticulum"*. The concept *37373007* has as super class: *"Congenital anomaly of small intestine"* and *"Diverticulosis of small intestine"*, while the concept *127962001* is children of *"Persistent embryonic structure"*, *"Structure of yolk stalk"*, *"Structure of distal portion of ileum"* and *"Diverticulum"*. Therefore, extensions of annotations also have to consider the semantic similarity between concepts in the KOS to disambiguate terms to annotate.

Regarding the SCP and LCP, we observed that LSM has a major role to explain the results produced by these techniques. In our previous work [5], we computed the SCP using Levenshtein distance while in the current version we used AnnoMap. This change produced positive impacts for some terminologies. For example, in MeSH the LCP and SCP were capable to improve all results for both phases, i.e., i) to detect impacted annotations and ii) to correctly adapt annotations. In NCIt, the LCP improved the second phase. The main reason here is the sensitivity of string matching methods to a specific domain. Therefore, our experiments highlight that the use of string based techniques to adapt annotations has to be carefully analyzed.

Another aspect to mention regarding the SCP lies in the computed annotation *"granulocyte-macrophage colony stimulating factor"* as evidenced in the F1-Score of Fig. 3 and Recall of Fig. 2. We verified that the concept associated to this annotation, C1287, changed in 2009/2010 since some terms changed from *is_pref_label* to *synonym*. In our silver standard, we did not include it as an impacted annotation since the concept definition remains the same. The main point here is that the ontology diff tool considers all types of change including those related to a class, data or object properties. However, all ontological changes are not necessary affecting existing annotations therefore, methods to detect impacted annotations must also identify if such ontology changes impact the annotations.

Besides, the inclusion of another layer (LCP) after the SCP in *CombineAll* method increased the adaptation of annotations. It can be observed for annotations generated with MeSH since we obtain the best results for version 2009/2010 and 2009/2016. The same does not occur for NCIt as showed by the results. In NCIt the Rules provided random results and it is associated to *"glycerol kinase gene"* which has different CUI (C1415082, C2700225) and codes (C75498, C75499).

Regarding the *Setup 2*, we verified that the use of SSMs to find candidate concepts in other ontology regions produces relevant associations. It is sharply observed in Fig. 3 when we compare LCP, SCP and PartialMatch. The main reason is that PartialMatch covers more situations than only the neighborhood utilized in Change Patterns. Therefore, it can also be extended to LCP and SCP in future versions to increase the definition of the context of the concept.

The positioning of PartialMatch also demonstrated significant improvements in our framework, see Table 4. However, this rule is not 100% accurate. For instance, the adaptation of *C11197:"folfox"* to *C11197:"folfox regimen"* is not aligned to our silver standard. It should evolve to *C63590:"FOLFOX-4 Regimen"* which also considers the suffix of this annotation. It forced us to verify if the inclusion of weights for LSM and SSMs or a threshold in future versions aim at overcoming this limitation. Moreover, PartialMatch is capable to provide adaptations which BK method cannot provide. This is for instance the case of the annotation *"postoperative myocardial infarction"*. Actually, no mappings contained in BioPortal with the used terminologies exist.

Regarding the results of *Setup 3*, we observed that any annotation was adapted by the BK method. It occurred in both utilized versions, 2009/2010 and 2009/2016. The variation noticed refers only the random results provoked by the *IncreaseAnnot* rule. In Fact the BK technique did not compute any annotations because all of them were adapted at previous layer. As mentioned before and observed in [5] the adaptation proposed by PartialMatch and *SuperClassAnnot* are not 100% precise. Therefore, in future versions of the framework, the BK method will be extended to re-adapt those annotations.

Finally we highlight that our framework can adapt annotations over several years. In our framework the lower AUC is 0.803 using the Setup 3, which is higher than the one for all previous configurations (see Table 7) and the results of our previous work in [5] using only one version 2009/2010.

The analyzes of these adaptations also demonstrated that the way these terminologies are changed as well as their internal structure have remarkable influence on the adaptations. For instance, in MeSH the reuse of CODEs and synonyms aids the adaptation method, in SNOMED CT the generation of new IDs which move the entire concept to another region of the terminology or add new ones also have a positive impact.

We also observed that few annotations remain invalid and marked as unsolved by our framework. What we have seen in such cases is the extension of the framework to adapt these annotations is very complex. Basically the concepts are in different ontology region and have different terms from the past annotation. It leads us to work with more sophisticated methods of string matching combined to semantic similarity.

Besides, the way these terminologies are structured is important and clearly observed in adaptations of ICD-9-CM. This terminology has a basic structure of a tree with maximum depth of 3. Furthermore, it does not have many synonyms. The drawback here is that the application of semantic techniques or string similarity methods do not aid in the maintenance task as verified with other KOS. For instance, the annotation 854.00*"brain injury"* is extended in 2010 to V80.01:*"traumatic brain injury"* which is located in a different region. Then in 2011 it evolves again, because the concept V80.01 became more specific *"screening for traumatic brain injury"*.

In our silver standard, the domain specialists decided to reduce the expressivity of this annotation returning to the first concept V80.01 and decreasing the annotated text. We verified that when applying the current `Rules` of our framework we cannot provide a good adaptation for this annotation. The `PartialMatch` Rule was not able to find a reasonable result since it produced feebly results for semantic similarity between concepts V80.01 and 854.00. Furthermore, the string similarity value of "screening for traumatic brain injury" is higher than *"brain injury"* when compared to *"traumatic brain injury"*. Therefore, future versions also have to deal with the reduction of expressivity in annotations through multiple versions.

Conclusion

We have presented in this paper an extension of a general framework for the semi-automatic maintenance of semantic annotations affected by the evolution of KOS. Moreover, our experimental analyzes demonstrated it is capable to reaches good results to adapt annotations using one or multiple successive versions. We observed that the use of semantic similarity approaches is important to determine the relatedness during the evolution process. As a result, we proposed a new rule, *Partial Match*, designed to support lexical and semantic measures. In future work, we plan to apply more sophisticated NLP techniques and combine them to semantic approaches in order to select better maintenance strategies.

Acknowledgments This work is funded by The National Research Fund (FNR) of Luxembourg ((grant C13/IS/5809134) and German Research Foundation (DFG) (RA 497/22-1).

Compliance with Ethical tandards

Conflict of interests Authors Silvio D. Cardoso and Cédric Pruski has received research grants from National Research Fund (FNR) of Luxembourg (grant C13/IS/5809134). The author Ying-Chi Lin has received grants from Deutsche Forschungsgemeinschaft (DFG) (RA 497/22-1). The authors Marcos Da Silveira, Chantal Reynaud-Delaître, Anika Groß and Erhard Rahm have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Abgaz YM. Change impact analysis for evolving ontology-based content management. Ph.D. thesis: Dublin City University; 2013.
2. Auer S, Herre H. A versioning and evolution framework for rdf knowledge bases. Berlin: Springer; 2007, pp. 55–69.
3. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008;41(5):706–716.
4. Burger T, Morozova O, Zaihrayeu I, Andrews P, Pane J. Report on methods and algorithms for linking user-generated semantic annotations to semantic web and supporting their evolution in time. 2010.
5. Cardoso SD, Reynaud-Delaître C, Da Silveira M, Pruski C. Combining rules, background knowledge and change patterns to maintain semantic annotations. AMIA Annu Symp Proc. 2017;2017:505–514. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977713/. 2017.
6. Cardoso SD, Pruski C, Da Silveira M, Lin Y-C, Groß A, Rahm E, Reynaud-Delaître C. Leveraging the impact of ontology evolution on semantic annotations. In: Blomqvist E, Ciancarini P, Poggi, F, Vitali, F, editors. Knowledge engineering and knowledge management. Cham: Springer International Publishing; 2016. P. 69–82. ISBN:978-3-319-49004-5.
7. Cardoso SD, Reynaud-Delaître C, Da Silveira M, Lin Y-C, Groß A, Rahm E, Pruski C. Towards a multi-level approach for the maintenance of semantic annotations. In: Proceedings of the 10th international joint conference on biomedical engineering systems and technologies (BIOSTEC 2017). HEALTHINF, Porto, Portugal, February 21-23. 2017.
8. Costa T, Leal JP. Semantic measures: How similar? how related?. Cham: Springer International Publishing; 2016, pp. 431–438.
9. Couto FM, Silva MJ, Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. Goannotator: linking protein go annotations to evidence text. J Biomed Discov Collab. 2006;1(1):19.
10. Da Silveira M, Dos Reis JC, Pruski C. Management of dynamic biomedical terminologies: Current status and future challenges. Yearb Med Inform. 2015;10(1):125–133.
11. Dixon WJ, Mood AM. The statistical sign test. J Am Stat Assoc. 1946;41(236):557–566.
12. Dos Reis JC, Dinh D, Da Silveira M, Pruski C, Reynaud-Delaître C. Recognizing lexical and semantic change patterns in evolving life science ontologies to inform mapping adaptation. Artif Intell Med. 2015;63(3):153–170.
13. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinforma. 2009;10(1):67.
14. Frost HR, Moore JH. Optimization of gene set annotations via entropy minimization over variable clusters (emvc). Bioinformatics (Oxford England). 2014;30(12):1698–1706.
15. Funk C, Baumgartner W, Garcia B, Roeder C, Bada M, Cohen KB, Hunter LE, Verspoor K. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinforma. 2014;15(1):1–29. https://doi.org/10.1186/1471-2105-15-59.
16. Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinforma. 2012;13(1):261.
17. Gimenez F, Xu J, Liu Y, Liu TT, Beaulieu CF, Rubin DL, Napel S. Automatic annotation of radiological observations in liver CT images. In: AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, Illinois, USA, November 3-7, 2012. 2012.

18. Gross A, Hartung M, Kirsten T, Rahm E. Estimating the quality of ontology-based annotations by considering evolutionary changes. In: International Workshop on Data Integration in the Life Sciences, pp. 71–87. Springer. 2009.
19. Harispe S, Ranwez S, Janaqi S, Montmain J. Semantic similarity from natural language and ontology analysis. Synthesis Lectures on Human Language Technologies Morgan &Claypool Publishers. 2015.
20. Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform. 2014;48:38–53.
21. Hartung M, Gross A, Rahm E. Conto-diff: Generation of complex evolution mappings for life science ontologies. J Biomed Inform. 2013;46:15–32.
22. Hodge G. Systems of knowledge organization for digital libraries: Beyond traditional authority files. Reports - Descriptive. 2000.
23. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. 1997. arXiv:9709008.
24. Köpke J, Eder J. Semantic invalidation of annotations due to ontology evolution. On the move to meaningful internet systems: OTM 2011, Lecture Notes in Computer Science. In: Meersman R, Dillon T, Herrero P, Kumar A, Reichert M, Qing L, Ooi BC, Damiani E, Schmidt D, White J, Hauswirth M, Hitzler P, and Mohania M, editors. Berlin: Springer; 2011. p. 763–780.
25. Lin Y-C, Christen V, Groß A, Cardoso SD, Pruski C, Da Silveira M, Rahm E. Evaluating and improving annotation tools for medical forms. In: Da Silveira M, Pruski C, Schneider R, editors. Data integration in the life sciences. Cham: Springer International Publishing; 2017. P. 1–16. ISBN:978-3-319-69751-2.
26. Luong PH, Dieng-Kuntz R. A rule-based approach for semantic annotation evolution in the coswem system. Canadian semantic web, semantic web and beyond. In: Koné M and Lemire D, editors. US: Springer; 2006. p. 103–120. https://doi.org/10.1007/978-0-387-34347-1_7.
27. Maynard D, Peters W, Sabou M. Change management for metadata evolution. 2007.
28. Meymandpour R, Davis JG. A semantic similarity measure for linked data: An information content-based approach. Knowledge-Based Systems. 2016;109:276–293.
29. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research p gkp440. 2009.
30. Park YR, Kim J, Lee HW, Yoon YJ, Kim JH. Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products. BMC Bioinforma. 2011;12(1): 1–7.
31. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5(7):1–12.
32. Powers DM. Evaluation: from precision, recall and f-measure to roc, informedness markedness and correlation. 2011.
33. Pruski C, Dos Reis JC, Da Silveira M. Capturing the relationship between evolving biomedical concepts via background knowledge. In: the 9th Semantic Web Applications and Tools for Life Sciences International Conference. 2016.
34. Qin L, Atluri V. Evaluating the validity of data instances against ontology evolution over the semantic web. Inf Softw Technol. 2009;51(1):83–97.
35. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence - Volume 1, IJCAI'95. San Francisco: Morgan Kaufmann Publishers Inc.; 1995. p. 448–453.
36. Soualmia LF, Prieur-Gaston E, Moalla Z, Lecroq T, Darmoni SJ. Matching health information seekers' queries to medical terms. BMC Bioinforma. 2012;13(14):S11.
37. Sy MF, Ranwez S, Montmain J, Regnault A, Crampes M, Ranwez V. User centered and ontology based information retrieval system for life sciences. BMC Bioinforma. 2012;13(1):S4.
38. Tissaoui A, Aussenac-Gilles N, Hernandez N, Laublet P. EVONTO - Joint evolution of ontologies and semantic annotations. (short paper). In: Dietz, J, editor. International conference on knowledge engineering and ontology development (KEOD), Paris, 26/10/2011-29/10/2011, pp. 226–231. 2011.
39. Tversky A. Features of similarity. Psychol Rev. 1977;84(4):327–352.
40. Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: Science Services and Agents on the World Wide Web. 2006;4(1):14–28.
41. Yimam SM, Biemann C, Majnaric L, Šabanović Š, Holzinger A. An adaptive annotation approach for biomedical entity and relation recognition. Brain Inf. 2016;3(3):157–168.
42. Zavalina OL, Kizhakkethil P, Alemneh DG, Phillips ME, Tarver H. Building a framework of metadata change to support knowledge management. J Inf Knowl Manag. 2015;14(01):1550,005.
43. Zhang X, Sun S, Zhang K. A novel comprehensive approach for estimating concept semantic similarity in wordnet. 2017. arXiv:1703.01726.