A Novel Atomic Annotator for Quality Assurance of Biomedical Ontologies

Rashmi Burse, Michela Bertolotto and Gavin Mcardle
University College Dublin, Belfield, Dublin 4, Ireland

Keywords: Biomedical Named Entity Recognition, Lexical Auditing, Semantic Analysis, Quality Assurance, Biomedical Ontologies, SNOMED.

Abstract: Existing lexical auditing techniques for Quality Assurance (QA) of biomedical ontologies exclusively consider lexical patterns of concept names and do not take semantic domains associated with the tokens constituting those patterns into consideration. For many similar lexical patterns the corresponding semantic domains may not be similar. Therefore, not considering the semantic aspect of similar lexical patterns can lead to poor QA of biomedical ontologies. Semantic domain association can be accomplished by using a Biomedical Named Entity Recognition (Bio-NER) system. However, the existing Bio-NER systems are developed with the goal of extracting information from natural language text, like discharge summaries, and as a result do not annotate individual tokens of a clinical concept. Annotating individual tokens of a clinical concept with their semantic domains is important from a QA perspective, since these annotations can be leveraged to gain insight into the type of attributes that should be associated with the concept. In this paper we present an annotator that atomically annotates the tokens of a clinical concept by crafting atomic dictionaries from the sub-hierarchies of Systematized Nomenclature of Medicine (SNOMED). Semantic analysis of lexically similar concepts by atomically annotating semantic domains to the tokens will ensure improved QA of biomedical ontologies.

1 INTRODUCTION

Incomplete and inconsistent representations of biomedical ontologies reduce their expressiveness and represent real-world facts inaccurately. Clinical concepts represented in biomedical ontologies are referred to by various Electronic Health Record (EHR) systems, discharge summaries and several Health Information Systems (HIS). Therefore, it is crucial to ensure that the data represented in biomedical ontologies is accurate and impeccable. Over the years, several auditing techniques have been developed to identify inconsistent/incorrect representations, missing relationships, and incomplete definitions of clinical concepts. Based on their approach to identify inconsistencies, they can be roughly classified into lexical, structural, and ontological techniques. Structural techniques focus on the graphical structure of biomedical ontologies to identify missing edges (relationships), ontological methods evaluate the soundness of an ontology based on ontological principals, and lexical techniques exploit the names of biomedical concepts to refine their definitions. Each of these have their own strengths and can be used to identify a particular set of inconsistencies. However, a robust technique that can be employed to seamlessly identify all inconsistencies is still an area of ongoing research. Given the large variety of inconsistencies that can be identified by lexical auditing techniques (Rector et al., 2011), this paper focuses on the lexical auditing techniques and aims to improve their quality by analyzing semantic aspects along with the lexical aspects. Lexical aspects include consideration of common words and the sequence in which they appear in the Fully Specified Name (FSN) whereas semantic aspects also consider the meaning of the word by associating a semantic domain to it. The paper presents a method to associate semantic domains to the individual lexical tokens of a concept name, which will ensure better identification of inconsistencies in an ontology and therefore suggest more appropriate attribute relationships to correct those inconsistencies, as compared to purely lexical auditing techniques. The remainder of the paper is organized as follows: Section 2 describes the state of the art and several gaps identified with the existing lexical auditing techniques and Bio-NER systems. Section 3 discusses the proposed method

Burse, R., Bertolotto, M. and Mcardle, G.
A Novel Atomic Annotator for Quality Assurance of Biomedical Ontologies.
DOI: 10.5220/0010782500003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 5: HEALTHINF*, pages 35-44
ISBN: 978-989-758-552-4; ISSN: 2184-4305

to atomically annotate the individual tokens of a concept name with their semantic domains. Section 4 presents the obtained results along with a detailed discussion. Finally, Section 5 concludes the paper and discusses some directions for future work.

2 RELATED WORK

2.1 Lexical Auditing Techniques

The existing lexical auditing techniques employ a variety of approaches to exploit several lexical features to identify inconsistencies in clinical concepts. For example, (Bodenreider et al., 2002) identified missing concepts in SNOMED by targeting concepts containing binary antonymous adjectives such as (acute , chronic), (unilateral , bilateral), (primary, secondary) etc. The method suggested new concepts by creating combinations of adjectives and nouns. (Bodenreider et al., 2001) identified missing hyponomic relationships by intuitively assuming that concepts conforming to a “modifier+noun” form should be hyponyms of the “noun” form concepts: e.g., acute appendicitis should be a child of appendicitis. (Pacheco et al., 2009) developed a method by eliminating the common sub-words appearing in both parent and child concept’s Fully Specified Name (FSN) to suggest attribute relationships. (Agrawal and Elhanan, 2014) created similarity sets containing concepts whose FSNs were lexically similar and identified 5 types of inconsistencies by comparing the concepts within a similarity set. (Bodenreider, 2016) re-created logical definitions from the lexical features of a concept name and inferred hierarchical relationships among these newly defined concepts. The newly obtained hierarchy was then compared with the original SNOMED hierarchy to detect differences. (Schulz et al., 2017) detected ambiguities in hierarchy tags, attribute relationships, and IS-A relationships based on the lexical features of SNOMED concepts. (Rector and Iannone, 2012) focused on finding concepts from the findings and diseases sub-hierarchies of SNOMED that should be classified as chronic or acute according to CORE problem list but currently are not and studied the effect of this misclassification on post-coordination queries. (Ceusters et al., 2007) scrutinized concepts containing negation words like absence, negation, and not and misclassification caused due to these words. The author introduced a new ”lacks” relationship to correctly classify such negative concepts. (Agrawal et al., 2013) presented the results of a study that statistically concluded that the complexity and thereby the chances of identifying errors increases with the length (number of words) of a concept name and the number of parents of a concept. (Agrawal, 2018; Agrawal and Qazi, 2020) proposed an auditing method based on the hypothesis that if two concepts are lexically similar then their structural and logical modeling should also be similar. (Cui et al., 2017) proposed a hybrid method combining the structural and lexical aspects of a CT system and identified four lexical patterns in non-lattice subgraphs that suggested potential missing hierarchical relationships and potential missing concepts. (Damme et al., 2018) suggested OWL axioms to be added in a concept definition by analyzing concept FSNs having a similar lexical structure.

2.1.1 Discussion

Based on this literature survey, we observed that the auditing techniques focusing on the lexical features of concept names fail to integrate the semantic meaning of the tokens constituting the concept name, leading to many false positives in the identification of inconsistencies and subsequently suggestion of inappropriate relationships to rectify those inconsistencies. For example, the axioms suggested by (Damme et al., 2018), based purely on lexical analysis, suggest the attribute “finding site” to be present in all disorders containing “of aorta”. However, this axiom only holds true if the identified lexical pattern contains a body structure after ”of”. If the token following ”of” were to belong to another hierarchy, the method would not work. Figure 1 illustrates the shortcomings of a purely lexical approach with an example and how the proposed work would improve on the existing approach, by including a semantic perspective.

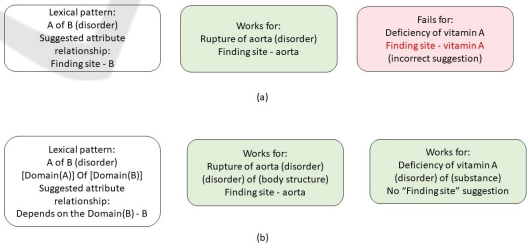

Figure 1: (a) Shortcomings of a purely lexical approach (b) Atomic annotation solution to address the shortcoming.

In Figure 1 (a), both “Rupture of aorta (disorder)” and “Deficiency of vitamin A (disorder)” have similar lexical patterns, “A of B (disorder)”. However, the attribute suggestion “finding site” while suitable for “Rupture of aorta (disorder)” turns out to be an inappropriate attribute suggestion for “Deficiency of vitamin A (disorder)”, since vitamin A is a substance and not a body structure. For many similar lexical patterns the corresponding semantic domains associ-

ated with the tokens in the lexical pattern may not be similar, which is the primary cause of false positives in the identification of inconsistencies in such methods. In cases where both the part of speech (POS) and lexical pattern are similar, the semantic domain is the deciding factor for suggesting an appropriate attribute relationship for the concept. We hypothesize that the rate of incorrect attribute suggestions can be reduced by basing suggestions on lexical similarities and also taking the semantic domains associated with the tokens in the lexical pattern into consideration. In order to improve the performance of purely lexical auditing techniques, we propose to add a semantic layer in the analysis (Figure 1 (b)). We propose to associate a semantic tag (semantic domain) to each individual token in an FSN and then suggest attributes to rectify the identified missing relationships based on both lexical and semantic aspects associated with individual tokens. The next section discusses the state of the art in named entity recognition for annotating biomedical entities.

2.2 Biomedical Named Entity Recognition

Named entity recognition is a subtask of Information Extraction (IE) that includes locating and classifying named entities from unstructured text into certain predefined categories. With huge amounts of biomedical text being generated, domain specific Named Entity Recognizers (NERs) were developed that could identify named entities, like disorders, treatment, diagnosis, etc. The growth of Bio-NERs led to a demand for biomedical lexicons that could aid the process of identifying clinical entities in free, unstructured text. As a result the major biomedical ontologies like UMLS (Bodenreider, 2004), SNOMED (IHTSDO, 2021), Gene Ontology (Consortium, 2003) etc. were used as dictionaries to identify biomedical entities in free text. Along with these dictionaries a number of manually annotated corpora including NCBI disease corpus (Dogan et al., 2014), GENIA corpus (Ohta et al., 2002), MCN corpus (Luo et al., 2019), corpus of manually annotated clinical notes (Pakhomov et al., 2006), (Ogren et al., 2008), BioScope corpus (Vincze et al., 2008), Distributional Semantics Resources corpus (Pyysalo et al., 2013), CLEF corpus (Roberts et al., 2007) were developed to train NLP algorithms. Genia (Ohta et al., 2002) is to date one of the most widely used corpora and specialises in gene and protein identification. With dictionaries and corpora in order, various Bio-NER methods were employed to extract biomedical concepts from clinical narratives like discharge summaries. Commonly used approaches for Bio-NER (Allahyari et al., 2017) can be classified into:

- Dictionary-based approaches (Friedman et al., 2004; Long, 2005), that extract biomedical entities by searching for a match in the biomedical dictionaries.
- Rule-based approaches (Hina et al., 2010; Ina et al., 2013), that aid the dictionary-based approaches by defining specific rules for matching biomedical concept patterns.
- Statistical approaches (Kulick et al., 2004) that employ various statistical models for extracting named entities.
- Machine Learning (ML) based approaches that employ artificial neural network models like SVM (Ju et al., 2011), KNN classifiers (Keretna et al., 2015), Conditional Random Field (CRF) (Skeppstedt et al., 2014; Lee et al., 2018), biLSTM-CRF (Lample et al., 2016), LSTM-CRF (Habibi et al., 2017), Recurrent Neural Networks (RNNs) (Unanue et al., 2017), unsupervised models (Zhang and Elhadad, 2013; Pérez et al., 2017), deep learning models (Liu et al., 2019), and the latest of all BERT (Lee et al., 2020). Although ML based techniques automated the process of Bio-NER to a large extent, a major requirement was the provision of huge volumes of manually annotated corpora to train the ML models. Manual annotation of huge volumes of biomedical data was both laborious and time intensive. This led to the development of methods that attempted to eliminate the need for creation of manually annotated datasets (Chen et al., 2015; Ghiasvand and Kate, 2018; Tulkens et al., 2019; Usami et al., 2011).
- Finally, Hybrid models that combine the strengths of aforementioned approaches for improved Bio-NER (Sasaki et al., 2008; Wang et al., 2019).

To make Bio-NERs reachable to non-expert users, easy-to-use pipelines (Dernoncourt et al., 2017), online web services (Jonquet et al., 2009a), and tools like CONANN (Reeve and Han, 2007), Cimind (Cabot et al., 2019), ABNER(Settles, 2005), CLAMP(Soysal et al., 2018), CliNER (Boag et al., 2018), MetaMap (Aronson and Lang, 2010) and CTakes (Savova et al., 2010) were also developed. (Rais et al., 2014) has presented a comparative study of seven Machine Learning methods and (Sniegula et al., 2019) has systematically reviewed all NER methods employed in the biomedical domain and highlighted specific areas that need to be attended in order to improve the performance of Bio-NERs.

2.2.1 Discussion

The existing Bio-NERs were developed with the goal of extracting biomedical entities from free unstructured text like discharge summaries, medical abstracts, and other clinical narratives. As a result, these Bio-NERs use existing biomedical ontologies as a reference source / dictionary in order to locate and annotate biomedical concepts appearing in free text. These methods try to club as many sequential words together until they can match a concept name in one of the biomedical ontologies. However, when annotating clinical concepts from the perspective of Quality Assurance (QA) of biomedical ontologies, concept names tagged with their respective semantic categories are already available in the biomedical ontology. On the contrary, the annotation process from a QA perspective requires tokenizing a concept name and annotating semantic domains to individual tokens in the name in order to gain insight into the type of attributes that can be associated with the concept. This makes the end goal and therefore the annotation approach applied for QA different from the existing Bio-NER systems. After extensively reviewing the literature and assessing the functionality of some of the existing Bio-NERs (Aronson and Lang, 2010; Jonquet et al., 2009b; Mahi, 2019; Kocaman and Talby, 2021), we came to the conclusion that none of the existing tools could be satisfactorily used to serve the requirement of atomic annotation. For example, when the input "injury of knee" was fed to National Center for Biomedical Ontology (NCBO) Bio-ontology Annotator (Jonquet et al., 2009b), the annotator either tagged "Injury of knee" as a "clinical finding" or "injury" was individually tagged as a "traumatic abnormality" and "knee" was left unannotated. Interactive MetaMap tool (Aronson and Lang, 2010) tagged "Injury of knee" as "knee injuries" / "injury or poisoning". While an implementation of Scispacy (Mahi, 2019) using the *en_ner_bionlp13cg_md* model in python only tagged "knee" as an organ and left "injury of" unannotated. Based on a cursory examination of the demo provided by SparkNLP (Kocaman and Talby, 2021), a licensed software tool, we noted that the entire disorder "malignant neoplasm of thyroid gland" was tagged as a "Problem". Some of the limitations that do not allow us to use existing Bio-NER tools for the purpose of atomic annotation are listed below.

- Non-uniformity: The tags used by each of the tools were different and there was no uniformity among them. For example, NLTK tagged knee as an "organ" instead of "body structure" which is a more general term and used widely in SNOMED. SparkNLP (Kocaman and Talby, 2021) followed a more generic approach and tagged the entire disorder as a "problem" instead of a more specific "traumatic abnormality" as tagged by bio-portal's annotator (Jonquet et al., 2009b).
- Dataset Variation: Firstly, we cannot use the existing trained ML models for atomic annotation as the datasets employed in their training are different. For example, if an ML model is trained on a dataset that has manually annotated "injury of knee" as a disorder, it will not tag the individual elements atomically. Manually annotating new datasets to train a model for atomic annotation is a very laborious and time intensive task. Automated creation of annotated datasets discussed earlier is also not of much use since they use existing biomedical ontologies as dictionaries and this would again tag the entire disorder instead of tagging its individual elements. Secondly, given the highly constrained lexical structure of clinical concept FSNs, the computing requirement for context detection is very limited and employing the time and memory intensive ML models for this purpose seems unnecessary. The task of analyzing such highly constrained lexical clinical concept FSNs can be easily accomplished by using a simple dictionary and rule based approach.
- Partial annotation: While some bio-NERs partially annotated the FSN, this is not sufficient. We require uniform atomic annotation of all tokens in the FSN, in order to create semantic patterns for analysis in the future. For example, NLTK annotated knee as an organ in "injury of knee" but left "injury of" unannotated. We need to annotate both injury as a disorder and knee as a body structure in order to classify a concept in a semantic pattern "disorder of body structure".

After reviewing the existing lexical auditing techniques we proposed combining semantic aspects along with lexical aspects to improve the QA of biomedical ontologies. Based on the literature survey of the state-of-the-art Bio-NERs we found that none of the existing Bio-NERs could be satisfactorily used to serve the purpose of atomic annotation in order to add a semantic aspect into lexical auditing techniques. In this paper, we present a method to develop an atomic annotator that uniformly annotates individual tokens in concept FSNs using the SNOMED hierarchy / semantic tags. The next section discusses the method in detail.

3 METHOD

As discussed in Section 2.2, based on an analysis of our requirements we adopted a hybrid approach that combined the strengths of dictionary-based and rule-based approaches to develop the atomic annotator.

3.1 Dataset

RF2 files of the SNOMED International Edition released in January 2021 were used to test and validate our approach. Based on a preliminary manual inspection of the *Disorder* subhierarchy of SNOMED, we found that a lexical pattern "A of B" had at least 3 different semantic patterns associated with it. Therefore, *Disorder* subhierarchy was chosen to evaluate the performance of our atomic annotator. We created a subset of the *Disorder* subhierarchy to only include disorders containing the stop word "of". The dataset was further simplified by limiting the number of tokens in the FSN to a maximum of 3 words, excluding the semantic tag "(disorder)". This reduced the complexity of the dataset without hindering the performance evaluation process of the atomic annotator. Finally, the dataset consisted of 2777 three-word disorder concepts containing the stop-word "of".

3.2 Atomic Dictionary Creation

In order to atomically annotate each token in the lexical pattern "A of B", 3 atomic dictionaries were created. Based on the semantic patterns observed, these included:

- Atomic disorder dictionary (tag "DIS")
- Atomic body structure dictionary (tag "BOD")
- Atomic substance dictionary (tag "SUB")

Each of these dictionaries was created by processing the respective SNOMED subhierarchies. The token "of" was annotated using the tag GEN (for general English word) followed by the token itself, i.e., "GEN-of". Tokens that could not be annotated using any of the aforementioned tags, were annotated as unknown (tag "UNK"). The detailed process employed in the creation of each of the atomic dictionaries is described next.

3.2.1 Atomic Disorder Dictionary

All single-word disorders extracted from the *Disorder* subhierarchy were stripped of their semantic tag "(disorder)" and added to the atomic disorder dictionary. Since the atomic disorder dictionary only contained disorders, the semantic tag "(disorder)" was removed to avoid redundancy (this step was repeated for all atomic dictionaries). To ensure high annotation performance for disorders, we logically assumed that if "A of B" is a disorder then "A" must be a disorder. For example, consider if "Carcinoma" was not present as a single-word concept in the *Disorder* subhierarchy of SNOMED, then it would not have been added to the atomic disorder dictionary. As a result, the atomic annotator would have tagged "Carcinoma" as "UNK" in the concept "Carcinoma of breast", in spite of a disorder being clearly present in the FSN. Based on this assumption, we added all tokens appearing before "of" to the dictionary. A few exceptions were eliminated from the dictionary after consulting a medical expert. A medical expert manually examined all tokens appearing before "of" and checked if they could be atomically added to the disorder dictionary based on his medical knowledge. For example, in the concept "Band of Ladd", "Band" was not included in the atomic disorder dictionary but in the concept "Edema of pharynx", "Edema" was included in the atomic disorder dictionary. For future references, we define this as refinement 1.

- Refinement 1: All tokens appearing before "of" were included in the atomic disorder dictionary, except a few that were removed based on the opinion of a medical expert.

3.2.2 Atomic Body Structure Dictionary

The approach followed for atomic disorder dictionary creation would not work well for a body structure dictionary because SNOMED follows a Structure Entire Part (SEP) model to represent body structures, in which a body structure is preceded by words like entire, part of, etc. For example, "knee" is present as "entire knee joint" or "entire left knee" in the *Body Structure* sub-hierarchy of SNOMED. As a result, the body structure "knee" would not have been added to the atomic body structure dictionary, if only single-word concepts from the *Body Structure* sub-hierarchy of SNOMED were included. Therefore, in case of a concept like "Injury of knee", the atomic annotator would have annotated "knee" as "UNK", in spite of a body structure being present in the FSN. After taking SNOMED's SEP model into consideration, we refined the atomic body structure dictionary. For future references, we define this as refinement 2.

- Refinement 2: The atomic body structure dictionary was recreated by extracting two-word concepts (excluding the (body structure) tag) from the *Body Structure* subhierarchy of SNOMED.

3.2.3 Atomic Substance Dictionary

An atomic substance dictionary was created by extracting single-word concepts from the *Substance* subhierarchy of SNOMED. To ensure high annotation performance for substances we added a few more entries that included (a) substances missing from the *Substance* subhierarchy of SNOMED, which were identified after an analysis of the common "UNK" tags and (b) substances that were usually referred to in their plural forms in the concept FSNs. In most of the cases, substances listed as proper nouns appeared in their singular form whereas a general reference to substances was always represented in the plural form. For example, "Deficiency of vitamin A" vs "Deficiency of vitamins". We did not find it accurate to process the disorder FSNs to eliminate plurals as we are not medical experts and instead refined the atomic substance dictionary to include such entries. Let us define this as refinement 3.

- Refinement 3: The substance hierarchy was refined to include plural forms of certain substances and a few substances that were missing in the *Substance* subhierarchy of SNOMED.

3.3 Atomic Annotation

After creating and refining the atomic dictionaries, the dataset of 2777 concepts was passed as input for atomic annotation of tokens. In cases where a token could be annotated using multiple tags, rules were defined to ensure that the tokens were annotated accurately. The rules were defined after a manual inspection and taking medical expert opinion into consideration. The highest priority was given to disorder followed by body structure and lastly substance. Indeed, based on the position of the token, if the token appeared before "of", it was always tagged as a disorder. For example, in "Dehiscence of fascia (disorder)", Dehiscence could be tagged as both a disorder and a body structure but higher priority was given to disorder based on the position. In the case of "Necrosis of flap (disorder)", flap could be tagged as both a substance and a body structure but higher priority was given to body structure, as suggested by the medical expert. While searching for a token in the body structure dictionary a partial match logic was applied since the body structure present in the concept FSN was always a substring of the body structure modelled using the SEP model in the atomic body structure dictionary. Minute considerations like adding trailing spaces to tokens before applying partial match logic were made to ensure accurate atomic annotation. For example, not adding trailing spaces to a token would have resulted in inaccurate results like "liver" being tagged as a "DIS" instead of a "BOD" due to the presence of "Hyperbiliverdinemia" in the atomic disorder dictionary. So trailing spaces were appended to both the entries in the atomic dictionary and the tokens before matching to avoid such mishaps. For disorder and substance tagging, a complete match was considered while searching through atomic dictionaries. A concept was considered to be correctly annotated if

- All individual tokens of the FSN were annotated.
- The semantic pattern formed after atomically annotating the FSN belonged to one of these patterns : "DIS GEN-of BOD", "DIS GEN-of DIS", or "DIS GEN-of SUB".
- The medical expert found the annotated concept to be semantically sound after manually inspecting the tags

The next section discusses the results obtained by the atomic annotator.

4 RESULTS & DISCUSSION

The atomic annotator correctly annotated 2653 out of the 2777 concepts (95.53%) passed to it. Table 1 describes the gradual improvement in annotation results after applying each of the refinements defined in section 3.2. In table 1, column 1 describes the number of dictionary refinements that were applied before annotating the dataset. Column 2 displays the number of concepts that were correctly annotated by the atomic annotator, as per our definition. Column 3 displays the respective percentages calculated out of a total 2777 concepts fed to the atomic annotator. Initially when all atomic dictionaries were created by adding single-word concepts from the respective SNOMED sub-hierarchies, only 12.32 % of the concepts were correctly annotated. After applying refinement 1 to the atomic disorder dictionary, this percentage increased from 12.32 % to 12.99 %. After taking into consideration the SEP model of SNOMED and refining the atomic body structure dictionary 94.59 %

Table 1: Improvement in annotation results after applying each refinement.

Refinements applied	# correct annotations	% correct annotations
None	342	12.32
1	361	12.99
1,2	2627	94.59
1,2,3	2653	95.53

of the concepts were correctly annotated. The reason for this drastic improvement is the fact that the majority of the disorder concepts in our input data set belonged to the semantic pattern "DIS GEN-of BOD" and after refinement 2 the majority of them were annotated correctly. Finally, after applying refinement 3 to the atomic substance dictionary, the results further improved to 95.53 %. An additional output was the identification of missing concepts in the SNOMED *Substance* sub-hierarchy, which were added to the atomic substance dictionary as a part of refinement 3. This list of identified missing substances will be submitted to SNOMED authors for review. The 124 concepts that could not be annotated correctly by the atomic annotator include rare tokens that could not be found in any of the aforementioned atomic dictionaries. Table 2 displays some of the patterns that were not correctly annotated by the atomic annotator.

Table 2: Examples of concepts incorrectly annotated by the atomic annotator.

Concept FSN (excluding (disorder) tag)	Annotations by atomic annotator
Caries of infancy	DIS GEN-of UNK
Disorder of fluency	DIS GEN-of UNK
Gangrene of newborn	DIS GEN-of UNK
Vegetation of heart	UNK GEN-of BOD
Barotrauma of ascent	DIS GEN-of UNK
Barotrauma of descent	DIS GEN-of UNK
Fibroepithelioma of Pinkus	DIS GEN-of UNK

The majority of the cases where concepts were incorrectly annotated belonged to the semantic pattern "DIS GEN-of UNK". In a few of the cases, the token appearing after "of" was a qualifier value (e.g. infancy), a social concept (e.g. newborn), an observable entity (e.g. fluency). In "Barotrauma of ascent" and "Barotrauma of descent", "ascent" was listed as an *Event* but "descent" was listed as a *Situation*, *Finding* and an *Event* in SNOMED. Analyzing "UNK" tags will provide interesting insights into the lexical modelling of SNOMED FSNs and highlight additional erroneous and inconsistent regions of SNOMED. In rare cases like "Vegetation of heart", the token before "of" was not annotated as a disorder. This happened because of the manual elimination of a few concepts appearing before "of" from the atomic disorder dictionary, based on the opinion of a medical expert. In a few cases, for example, "Fibroepithelioma of Pinkus", The token after "of", i.e., "Pinkus" represents the name of the person who discovered the disorder. It would be more ideal to model such disorders using "Pinkus's Fibroepithelioma" instead.

To conduct a comparative evaluation of our atomic annotator with the state-of-the-art, a set of three concepts belonging to each semantic pattern, which were picked randomly, was passed to two of the most widely used Bio-NERs , i.e., Bioportal (Jonquet et al., 2009b) and MetaMap (Aronson and Lang, 2010). Furthermore, the aforementioned concepts, represented in table 2, that could not be annotated by our atomic annotator were also passed to bioportal (Jonquet et al., 2009b) and metaMap (Aronson and Lang, 2010), to check if they could annotate them individually and extract semantic patterns. Table 3 presents a comparative evaluation of the annotation results obtained by our atomic annotator vs Bioportal (Jonquet et al., 2009b) and MetaMap (Aronson and Lang, 2010).

Table 3: Comparative evaluation of the atomic annotator with Bioportal and MetaMap.

Concept FSN	Atomic annotator	Bioportal annotator	MetaMap annotator
Calcification of lung	DIS GEN-of BOD	calcification of lung **structure**	Disease or Syndrome
Tuberculosis of bronchus	DIS GEN-of BOD	Tuberculosis of bronchus	Disease or Syndrome
Lipoma of hip	DIS GEN-of BOD	hip	Neoplastic Process
Overdose of metformin	DIS GEN-of SUB	Overdose of metformin	Injury or Poisoning
Abuse of laxatives	DIS GEN-of SUB	Abuse of laxatives	Mental or Behavioral Dysfunction
Extravasation of urine	DIS GEN-of SUB	Extravasation of urine	Pathologic Function
Sequela of trachoma	DIS GEN-of DIS	Sequela of trachoma	Pathologic Function
Rupture of neoplasm	DIS GEN-of DIS	Rupture of neoplasm	Neoplastic Process
Hyperkeratosis of pinta	DIS GEN-of DIS	Hyperkeratosis of pinta	Disease or Syndrome
Caries of infancy	DIS GEN-of UNK	Caries of infancy	Disease or Syndrome
Disorder of fluency	DIS GEN-of UNK	Disorder of fluency	Disease or Syndrome
Fibroepithelioma of Pinkus	DIS GEN-of UNK	Fibroepithelioma of Pinkus	Neoplastic Process

Based on a comparative evaluation, a notable feature of bioportal (Jonquet et al., 2009b) was that it was able to tokenize the disorder phrase. However, despite tokenizing, the atomic annotations rendered were of no use since they were only mapped to the same term in a reference biomedical ontology, rather than being mapped to its semantic domain. Figure 2 displays a partial view of the bioportal annotator results illustrating a few annotations for each of the tokens of the disorder "Tuberculosis of bronchus". Comparatively, metaMap (Aronson and Lang, 2010), did annotate the entire phrase with its semantic domain but could not tokenize the phrase and annotate its individual elements. Figure 3 displays the annotation results of metaMap for the disorder "Tuberculosis of bronchus". It is clearly evident from the comparative evaluation results that our atomic annotator outperforms both bioportal and metaMap annotators as far as atomic annotation is concerned. The tags annotated by bioportal and metaMap, although useful in IE from discharge summaries, cannot be used to extract semantic patterns for QA of biomedical ontologies.

CLASS	ONTOLOGY	TYPE	MAPPING SEMANTIC TYPE	CONTEXT	MATCHED CLASS	MATCHED ONTOLOGY
Tuberculosis of bronchus	SNOMED CT	direct		**Tuberculosis of bronchus**	Tuberculosis of bronchus	SNOMED CT
Tuberculosis of bronchus	Read Codes, Clinical Terms Version 3 (CTV3)	direct		**Tuberculosis of bronchus**	Tuberculosis of bronchus	Read Codes, Clinical Terms Version 3 (CTV3)
Tuberculosis of bronchus	International Classification of Diseases, Version 9 - Clinical Modification	direct		**Tuberculosis of bronchus**	Tuberculosis of bronchus	International Classification of Diseases, Version 9 - Clinical Modification
Tuberculosis	Logical Observation Identifier Names and Codes	direct		**Tuberculosis** of bronchus	Tuberculosis	Logical Observation Identifier Names and Codes
Tuberculosis	SNOMED CT	direct		**Tuberculosis** of bronchus	Tuberculosis	SNOMED CT
Tuberculosis	Logical Observation Identifier Names and Codes	direct		**Tuberculosis** of bronchus	Tuberculosis	Logical Observation Identifier Names and Codes
Bronchus	Read Codes, Clinical Terms Version 3 (CTV3)	direct		Tuberculosis of **bronchus**	Bronchus	Read Codes, Clinical Terms Version 3 (CTV3)
bronchus	Computer Retrieval of Information on Scientific Projects Thesaurus	direct		Tuberculosis of **bronchus**	bronchus	Computer Retrieval of Information on Scientific Projects Thesaurus

Figure 2: Bioportal annotator results.

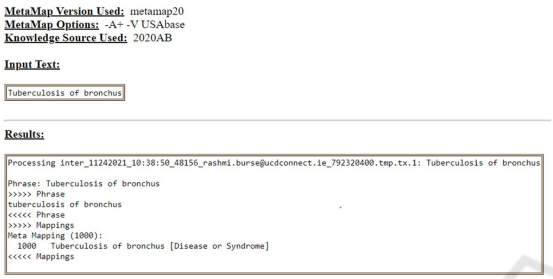

Figure 3: MetaMap annotator results.

5 CONCLUSION

In this work, we highlighted the importance of taking into account a semantic aspect along with the lexical aspect in order to reduce the number of false positives in the identification of inconsistencies in biomedical concepts. In order to introduce a semantic perspective to the exclusive lexical auditing techniques, currently employed in the QA of biomedical ontologies, we presented an atomic annotator that annotates individual tokens of a concept FSN with their semantic domains to provide insight into the nature of attributes that should be suggested for an inconsistent or incompletely defined concept. The atomic annotator has shown a promising potential by correctly annotating 95.53% of the concepts from the input dataset. The atomic annotator presented in this paper is a part of ongoing work. In the future we plan on validating the semantic patterns identified by our atomic annotator and using the atomically annotated output to extract specific attribute suggestions for inconsistent biomedical concepts based on the semantic pattern to which the concept belongs.

ACKNOWLEDGEMENTS

We would like to thank Dr. K.S. Burse (M.S. (ENT)) for providing his expert medical opinion which tremendously helped us in refining the atomic dictionaries.

REFERENCES

Agrawal, A. (2018). Evaluating lexical similarity and modeling discrepancies in the procedure hierarchy of snomed ct. *BMC Medical Informatics and Decision Making*, 18.

Agrawal, A. and Elhanan, G. (2014). Contrasting lexical similarity and formal definitions in snomed ct: Consistency and implications. *Journal of biomedical informatics*, 47:192–8.

Agrawal, A., Perl, Y., Chen, Y., Elhanan, G., and Liu, M. (2013). Identifying inconsistencies in snomed ct problem lists using structural indicators. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2013:17–26.

Agrawal, A. and Qazi, K. (2020). Detecting modeling inconsistencies in snomed ct using a machine learning technique. *Methods*.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutiérrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv*, abs/1707.02919.

Aronson, A. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17 3:229–36.

Boag, W., Sergeeva, E., Kulshreshtha, S., Szolovits, P., Rumshisky, A., and Naumann, T. (2018). Cliner 2.0: Accessible and accurate clinical concept extraction. *ArXiv*, abs/1803.02245.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Bodenreider, O. (2016). Identifying missing hierarchical relations in snomed ct from logical definitions based on the lexical features of concept names. In *ICBO/BioCreative*.

Bodenreider, O., Burgun, A., and Rindflesch, T. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the umls.

Bodenreider, O., Burgun-Parenthoine, A., and Rindflesch, T. (2002). Assessing the consistency of a biomedical terminology through lexical knowledge. *International journal of medical informatics*, 67 1-3:85–95.

Cabot, C., Darmoni, S., and Soualmia, L. (2019). Cimind: A phonetic-based tool for multilingual named entity recognition in biomedical texts. *Journal of biomedical informatics*, 94:103176.

Ceusters, W., Elkin, P., and Smith, B. (2007). Negative findings in electronic health records and biomedical ontologies: A realist approach. *International journal of medical informatics*, 76 Suppl 3:S326–33.

Chen, Y., Lasko, T., Mei, Q., Denny, J., and Xu, H. (2015). A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Consortium, G. O. (2003). The gene ontology (go) database and informatics resource gene ontology consortium.

Cui, L., Zhu, W., Tao, S., Case, J., Bodenreider, O., and Zhang, G.-Q. (2017). Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in snomed ct. *Journal of the American Medical Informatics Association : JAMIA*, 24:788 – 798.

Damme, P., Quesada-Martínez, M., Cornet, R., and Fernández-breis, J. (2018). From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies. *Journal of biomedical informatics*, 84:59–74.

Dernoncourt, F., Lee, J. Y., and Szolovits, P. (2017). Neuroner: an easy-to-use program for named-entity recognition based on neural networks. In *EMNLP*.

Dogan, R., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Friedman, C., Shagina, L., Lussier, Y., and Hripcsak, G. (2004). Research paper: Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association : JAMIA*, 11 5:392–402.

Ghiasvand, O. and Kate, R. J. (2018). Learning for clinical named entity recognition without manual annotations. *Informatics in Medicine Unlocked*, 13:122–127.

Habibi, M., Weber, L., Neves, M., Wiegandt, D., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33:i37 – i48.

Hina, S., Atwell, E., and Johnson, O. (2010). Secure information extraction from clinical documents using snomed ct gazetteer and natural language processing. *2010 International Conference for Internet Technology and Secured Transactions*, pages 1–5.

IHTSDO (2021). *SNOMED International*. last accessed 24/08/2021.

Ina, S., Twell, E. R. A., and Ohnson, O. W. J. (2013). Snomedtagger : A semantic tagger for medical narratives.

Jonquet, C., Shah, N., and Musen, M. (2009a). The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009:56 – 60.

Jonquet, C., Shah, N., Youn, C., Musen, M., Callendar, C., and Storey, M. (2009b). Ncbo annotator : Semantic annotation of biomedical data.

Ju, Z., Wang, J., and Zhu, F. (2011). Named entity recognition from biomedical text using svm. *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4.

Keretna, S., Lim, C., Creighton, D., and Shaban, K. (2015). Enhancing medical named entity recognition with an extended segment representation technique. *Computer methods and programs in biomedicine*, 119 2:88–100.

Kocaman, V. and Talby, D. (2021). Spark nlp: Natural language understanding at scale. *Softw. Impacts*, 8:100058.

Kulick, S., Bies, A., Liberman, M., Mandel, M. A., McDonald, R. T., Palmer, M., Schein, A., Ungar, L., Winters, S., and White, P. S. (2004). Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *NAACL*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Lee, W., Kim, K., Lee, E. Y., and Choi, J. (2018). Conditional random fields for clinical named entity recognition: A comparative study using korean clinical texts. *Computers in biology and medicine*, 101:7–14.

Liu, X., Zhou, Y., and Wang, Z. (2019). Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. *J. Vis. Commun. Image Represent.*, 60:1–15.

Long, W. (2005). Extracting diagnoses from discharge summaries. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 470–4.

Luo, Y.-F., Sun, W., and Rumshisky, A. (2019). Mcn: A comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92:103132.

Mahi, M. (2019). *scispaCy for Bio-medical Named Entity Recognition*. last accessed 30/08/2021.

Ogren, P. V., Savova, G., and Chute, C. (2008). Constructing evaluation corpora for automated clinical named entity recognition. In *LREC*.

Ohta, T., Tateisi, Y., and Kim, J.-D. (2002). The genia corpus: an annotated research abstract corpus in molecular biology domain.

Pacheco, E., Stenzhorn, H., Nohama, P., Paetzold, J., and Schulz, S. (2009). Detecting underspecification in snomed ct concept definitions through natural language processing. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:492–6.

Pakhomov, S. V. S., Coden, A., and Chute, C. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *International journal of medical informatics*, 75 6:418–29.

Pérez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M., and Dalianis, H. (2017). Semi-supervised medical entity recognition: A study on spanish and swedish clinical corpora. *Journal of biomedical informatics*, 71:16–30.

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing.

Rais, M., Lachkar, A., Lachkar, A., and Ouatik, S. A. (2014). A comparative study of biomedical named entity recognition methods based machine learning approach. *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*, pages 329–334.

Rector, A. and Iannone, L. (2012). Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in snomed ct. *Journal of biomedical informatics*, 45 2:199–209.

Rector, A. L., Iannone, L., and Stevens, R. (2011). Quality assurance of the content of a large dl-based terminol-

ogy using mixed lexical and semantic criteria: experience with snomed ct. In *K-CAP '11*.

Reeve, L. H. and Han, H. (2007). Conann: An online biomedical concept annotator. In *DILS*.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., and Wheeldin, B. (2007). The clef corpus: Semantic annotation of clinical text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 625–9.

Sasaki, Y., Tsuruoka, Y., McNaught, J., and Ananiadou, S. (2008). How to make the most of ne dictionaries in statistical ner. *BMC Bioinformatics*, 9:S5 – S5.

Savova, G., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Schuler, K., and Chute, C. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17 5:507–13.

Schulz, S., Martínez-Costa, C., and Miñarro-Giménez, J. A. (2017). Lexical ambiguity in snomed ct. In *JOWO*.

Settles, B. (2005). Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21:3191–3192.

Skeppstedt, M., Kvist, M., Nilsson, G., and Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–58.

Sniegula, A., Poniszewska-Marańda, A., and Chomatek, L. (2019). Study of named entity recognition methods in biomedical field. In *EUSPN/ICTH*.

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S. V. S., Liu, H., and Xu, H. (2018). Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association : JAMIA*, 25:331 – 336.

Tulkens, S., Suster, S., and Daelemans, W. (2019). Unsupervised concept extraction from clinical text through semantic composition. *Journal of biomedical informatics*, 91:103120.

Unanue, I. J., Borzeshi, E. Z., and Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of biomedical informatics*, 76:102–109.

Usami, Y., Cho, H.-C., Okazaki, N., and Tsujii, J. (2011). Automatic acquisition of huge training data for biomedical named entity recognition. In *BioNLP@ACL*.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9:S9 – S9.

Wang, Q., Xia, Y., Zhou, Y., Ruan, T., Gao, D., and He, P. (2019). Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of biomedical informatics*, 92:103133.

Zhang, S. and Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46 6:1088–98.