Scoring semantic annotations returned by the NCBO Annotator

Soumia Melzi, Clement Jonquet

Scoring semantic annotations returned by the NCBO Annotator

Soumia Melzi and Clement Jonquet

Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
& Computational Biology Institute (IBC) of Montpellier
University of Montpellier, France
soumia.melzi@lirmm.fr, jonquet@lirmm.fr

Abstract. Semantic annotation using biomedical ontologies is required to enable data integration, interoperability, indexing and mining of biomedical data. When used to support semantic indexing the scoring and ranking of annotations become as important as provenance and metadata on the annotations themselves. In the biomedical domain, one broadly used service for annotations is the NCBO Annotator Web service, offered within the BioPortal platform and giving access to more than 350+ ontologies or terminologies. This paper presents a new scoring method for the NCBO Annotator allowing to rank the annotation results and enabling to use such scores for better indexing of the annotated data. By using a natural language processing-based term extraction measure, C-Value, we are able to enhance the original scoring algorithm which uses basic frequencies of the matches and in addition to positively discriminate multi-words term annotations. We show results obtained by comparing three different methods with a reference corpus of PubMed-MeSH manual annotations.

Keywords: semantic web, biomedical ontologies, semantic annotation, semantic indexing, NCBO Annotator, term extraction, scoring, C-Value, Bio-NLP.

1 Introduction

The large quantity of biomedical data available today requires the provision of efficient tools to search, process, explore and query the data. These data are often unstructured and are presented in different formats (database, documents, etc.), which prevents an efficient integration and interoperability. To address these problems, the biomedical community has turned to ontologies and terminologies to describe their data and turn them into structured and formalized knowledge [2,20]. Ontologies help to address the data integration problem by playing the role of common denominator. One way of using ontologies is by means of creating semantic annotations. An annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the term [7,25]. When doing ontology-based indexing, one might use these annotations to 'bring together' the data elements from different resources. The knowledge formalized within the ontologies enables then semantic search [12,8,13]. Previous work has encouraged and exalted the use

of ontologies for annotation at various levels [19,4,22,8]. Manual annotation, although highly desirable [11], is very expensive as it requires both an important manual work and an excellent domain knowledge, which becomes harder and harder when scaling to the huge amount of data and ontologies available today [14]. To overcome these limitations, several tools for automatic annotation have been proposed [1,18,6,10,21,23,16]. However, when annotations are generated automatically, different systems will return a very large resultset among which it is hard to distinguish the most relevant annotations. Therefore, the scoring and ranking of the result annotations become crucial to use them in real world scenarios.

In this paper, we repport on the use of one of the mostly used biomedical annotation tool: the NCBO Annotator offered offered within the BioPortal platform [15] and giving access to more than 350+ ontologies or terminologies[10]. We propose to improve the annotation results (while not changing the service implementation) by ranking the produced annotations according to their relevance by taking into account their frequencies (as originally proposed) and a term extraction measure, called C-Value, used to positively discriminate annotations generated from matches with multi-word terms. In the following, we propose two new scoring methods allowing to score and rank annotations by their importance in the given input data. We present each method and compare them on the same text example along the paper. In addition, we analyse the results obtained by each methods when compared to a corpus of 1250 PubMed citations manually annotated with MeSH.

2 Background - The NCBO Annotator & annotation scoring

The NCBO Annotator Web service [10] (http://bioportal.bioontology.org/annotator), provides a mechanism to employ ontology-based annotation in curation, data integration, and indexing workflows using any of the several hundred public ontologies or terminologies in the BioPortal repository. In a first step the user submitted text is given as input to a concept recognition tool along with a dictionary. The dictionary (or lexicon) consists of a list of strings that identify ontology classes. The dictionary is constructed by pooling all concept names and other lexical identifiers, such as synonyms or alternative labels that identify concepts. The Annotator uses Mgrep [3], a concept recognizer developed by the University of Michigan that enables fast and efficient matching of text against a set of dictionary terms to recognize concepts and generate direct annotations. In a second step, semantic expansion components use the knowledge within the ontologies or between them, to create additional annotations. For example, the is-a transitive closure component traverses an ontology parent-child hierarchy to create additional annotations with parent concepts. The ontology-mapping component creates additional annotations based on existing mappings between ontology terms. The direct annotations and the set of semantically expanded annotations are then returned to the user. Figures 1 and 2 show the Annotator user interface.

The NCBO Annotator Web service was released in 2009 and used by an increasing numbers of users since then (approx. 400 GB of data per year) and is also embedded in commercial platforms. The Annotator service returns annotations in JSON or XML and uses URIs as defined within BioPortal. Our contribution consists in ranking the an-

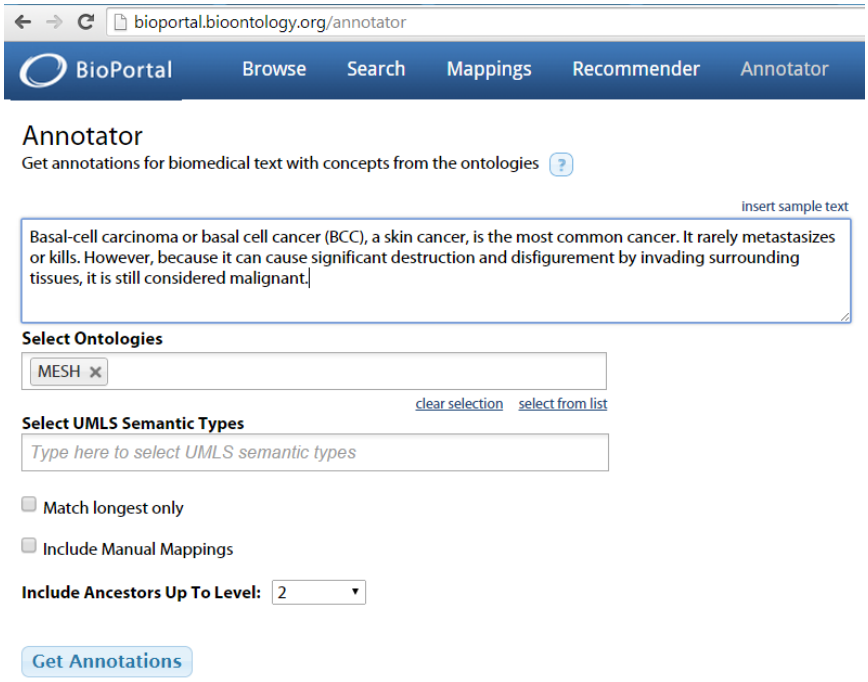

Fig. 1. NCBO Annotator user interface within BioPortal

Annotations

total results **35** (direct **12** / ancestor **23** / ma

CLASS filter	ONTOLOGY filter	TYPE filter	CONTEXT	MATCHED CLASS filter	MATCHED ONTOLOGY
Cells	Medical Subject Headings	direct	Basal-**cell** carcinoma or basal ...	Cells	Medical Subject Headings
Anatomy (MeSH Category)	Medical Subject Headings	ancestor	Basal-**cell** carcinoma or basal ...	Cells	Medical Subject Headings
Topical Descriptor	Medical Subject Headings	ancestor	Basal-**cell** carcinoma or basal ...	Cells	Medical Subject Headings
Cells	Medical Subject Headings	direct	... or basal **cell** cancer (BCC), a ...	Cells	Medical Subject Headings
Anatomy (MeSH Category)	Medical Subject Headings	ancestor	... or basal **cell** cancer (BCC), a ...	Cells	Medical Subject Headings
Topical Descriptor	Medical Subject Headings	ancestor	... or basal **cell** cancer (BCC), a ...	Cells	Medical Subject Headings
Cell	Medical Subject Headings	direct	Basal-**cell** carcinoma or basal ...	Cell	Medical Subject Headings
Cell	Medical Subject Headings	ancestor	Basal-**cell** carcinoma or basal ...	Cell	Medical Subject Headings
Cell	Medical Subject Headings	direct	... or basal **cell** cancer (BCC), a ...	Cell	Medical Subject Headings
Cell	Medical Subject Headings	ancestor	... or basal **cell** cancer (BCC), a ...	Cell	Medical Subject Headings
Carcinoma	Medical Subject Headings	direct	Basal-cell **carcinoma** or basal cell ...	Carcinoma	Medical Subject Headings
Neoplasms, Glandular and Epithelial	Medical Subject Headings	ancestor	Basal-cell **carcinoma** or basal cell ...	Carcinoma	Medical Subject Headings
Neoplasms by Histologic Type	Medical Subject Headings	ancestor	Basal-cell **carcinoma** or basal cell ...	Carcinoma	Medical Subject Headings
Neoplasms, Basal Cell	Medical Subject Headings	direct	... carcinoma or **basal cell cancer** (BCC), a skin ...	Neoplasms, Basal Cell	Medical Subject Headings
Neoplasms, Glandular and Epithelial	Medical Subject Headings	ancestor	... carcinoma or **basal cell cancer** (BCC), a skin ...	Neoplasms, Basal Cell	Medical Subject Headings
Neoplasms by Histologic Type	Medical Subject Headings	ancestor	... carcinoma or **basal cell cancer** (BCC), a skin ...	Neoplasms, Basal Cell	Medical Subject Headings
Neoplasms	Medical Subject Headings	direct	... basal cell **cancer** (BCC), a skin ...	Neoplasms	Medical Subject Headings
Diseases (MeSH Category)	Medical Subject Headings	ancestor	... basal cell **cancer** (BCC), a skin ...	Neoplasms	Medical Subject Headings
Topical Descriptor	Medical Subject Headings	ancestor	... basal cell **cancer** (BCC), a skin ...	Neoplasms	Medical Subject Headings
Neoplasms	Medical Subject Headings	direct	... a skin **cancer**, is the most ...	Neoplasms	Medical Subject Headings

Fig. 2. NCBO Annotator results within BioPortal. The 35 annotations are not ordered by importance within the given text.

notations by post-processing them, without changing either the annotation themselves or the implementation of the service.

When scoring annotations, a number is assigned to an annotation to indicate its importance. Higher scores reflect more important or relevant annotations. Typically, methods will use context information [13,24] (i.e., part of the document where the annotation was generated), frequencies (i.e., number of occurrence of the same annotating concepts), or user feedbacks [26] to score annotations. However, only a few tools in the literature mentioned above offer a scoring feature although it is mandatory when considering to use the annotations for sematic indexing of the data [17].

3 Methods

3.1 Old Annotator scoring method

In the latest versions of the Annotator, since BioPortal 4.0, end of 2013, the scoring method has been removed from the implementation, thus transferring the task of scoring to the users of the service, when all the annotations have been retrieved. In a previous version of the web service, as released in 2009 [10], the annotations were scored based on the type of match as well as the matching term, as described in Table 1: the importance of each annotation was measured with a numerical value (weight). For instance, an annotation done by matching a concept's preferred name would get a higher weight than one done by matching a concept's synonym or one done with a parent-level-3 (ancestor) concept obtained by traversing the ontology is-a hierarchy. As another example, an annotation done with a concept obtained thanks to a mapping from a another concept directly matched would get a slightly smaller weight. Finally, the global score of an annotation would be obtained by summing the weights of all the annotations made with the same concept. This method was therefore mostly based on type of match and frequency. In the following, we have re-implemented this scoring method to be able to compare it to our new methods; in the following, it is refered as *"OldScoreNcbo"*.

Table 1. Annotation weights based on the provenance (as of [9]).

Type of match (context & matching terms)	**Weights**	**Noted**
Direct annotation done with a concept's preferred name	10	pref
Direct annotation done with a concept's synonym	8	syn
Expanded annotation done with a mapping	7	map
Expanded annotation done with a parent level n (e.g., 9 for n=1; 7 for n=2; 4 for n=5; 3 for n=8; 1 for n>12)	$1 + 10.e^{-0.2*n}$	exp

As an example, Table 2 describes the annotations obtained with the Annotator using MesH and the *"OldScoreNcbo"* scoring method for the text:

"Basal Cell Carcinoma ... Basal Cell Carcinoma ... Basal Cell Carcinoma ... Basal Cell Carcinoma ... Basal Cell ...Basal Cell..."

For each annotating concept, we show: the code, preferred name, matched terms, the matching type, the frequency (F) of this match and finally the annotation scores (S1). The results show that the concept T025 (*cell*) is the most represented in the text with a score of 96 (twice nested in the expression *basal cell* and four times nested in the expression *Basal cell carcinoma*). However, this concept is never present "alone" in the text and one would like to see the concept D002280 (*basal cell carcinoma*) higher in the ranking as it is one of the most representative concept of the text. Minimally, D002280 should be ranked above D002277 as it is an hyponym (more precise) of carcinoma mentioned in the text. Therefore, the results obtained with the *OldScoreNcbo* scoring are not very relevant to the needs and expectations of the user although it is true than the text is about the notion of *cell*.

Table 2. Annotations obtained with the *OldScoreNcbo* method and ranked by score S1.

Code	Preferred name	Matched terms	Type	F	S1	
T025	Cell	Cell	pref	6	60	**96**
		Cell	exp n=1	6	36	
D002277	Carcinoma	Basal cell carcinoma	exp n=1	4	36	**76**
		Carcinoma	pref	4	40	
D009375	Neoplasms, Glandular	Basal cell carcinoma	exp n=2	4	28	**64**
		Carcinoma	exp n=1	4	36	
U000002	Anatomy (MeSH Category)	Cell	exp n=1	6	**54**	
D009370	Neoplasms by histologic type	Basal cell carcinoma	exp n=3	4	24	**52**
		Carcinoma	exp n=2	4	28	
D002477	Cells	Cell	syn	6	**48**	
U000019	Topical Descriptor	Cell	exp n=2	6	**42**	
D018295	Neoplasms, Basal Cell	Basal cell carcinoma	exp n=1	4	**36**	
U000017	MeSH Descriptors	Cell	exp n=3	6	**36**	
D002280	Carcinoma, Basal Cell	Basal cell carcinoma	syn	4	**32**	
D009369	Neoplasms	Carcinoma	exp n=3	4	**24**	

During our study, we have more extensively experimented the *OldScoreNcbo* scoring method with different texts and ontologies and varying the hierarchy level. This led us to conclude that this method does not penalize enough the single-word matches within multi-word terms because, as the scoring method considers both full term matches and single word matches, it increases the weight of single word terms, which are also nested in longer terms.[1] To address these limitation, we turned ourselves towards ap-

[1] A few months after our study, in Sept. 2014, the Annotator was enriched by a new 'longest-only' parameter which makes the service ignore shortest matches if a longer match (within

proaches proposed in the natural language processing community when doing term extraction from text corpora.

3.2 Old Annotator method + C-value

Frequently, biomedical terminologies and ontologies contains several concepts whose names are nested in names of other concepts. For instance, the word *disease*, preferred name of the concept D004194 in MeSH is nested in the term *periodontal disease*, preferred name of D010510. During the annotation process, in order to penalize the nested concepts and to favor non-nested concepts we proposed to use the *C-value* measure [5].

C-value is an automatic term recognition method well known in the literature that combines statistical and linguistic information for the extraction of multi-word and nested terms. It often gets best precision results when used for automatic term extraction especially in biomedical studies [27]. The *C-value* method combines linguistic and statistical information; the linguistic information is the use of a general regular expression as linguistic patterns, and the statistical information is the value assigned with the *C-value* measure (hereafter) based on term frequencies but discriminating nested terms.

$$C\text{-}value(a) = \begin{cases} \log_2 |a| \times f(a) & \text{if a is not nested} \\ \log_2 |a| \times \left(f(a) - \frac{1}{|P(T_a)|} \times \sum_{b \in T_a} f(b) \right) & \text{otherwise} \end{cases} \quad (1)$$

Where a is the candidate string, $f(a)$ the frequency of occurrence of a in the corpus (the text to annotate in our case), $|a|$ the number of words in a, T_a the set of terms that contain a and $P(T_a)$ the number of terms in this set. In a nutshell, *C-value* either uses frequency of the term if the term is not included in other terms (first line), or decrease this frequency if the term appears in other terms, by using the frequency of those other terms (second line). For example, for a given text containing exactly four times the expression 'breast cancer' and twice the expression 'cancer': the C-value of the term *breast cancer* will be 6.33 which is higher than the C-value of *cancer* will be 2, while the frequency of terms are respectively 4 and 6.

In order to use the *C-value* measure to score annotations we had to overcome the problem that *C-value* operates at the term level where as annotations are related to concepts. We proposed to deal with this by applying the *C-value* measure on all the matched terms and summing the values of all the terms matched for a given concept in order to obtain the C-Value of that concept. In addition, to keep the advantages of the previous measure, taking into account the type of match, we used the log to weaken the effect of the *OldScoreNcbo* when calculating the new score. This new scoring method is referred after as *ScoreNcboCvalue*.

---

the same ontology) exists e.g., *carcinoma* matches would be ignored in front of *basal cell carcinoma* ones. Although, this new parameter fixes partially the problem by removing some matches it does not replace a proper scoring methods that allows to rank all the annotations. In addition, the problem remains when someone wants to get the maximum number of annotations for a given text and thus does not use the new longest-only parameter.

$$ScoreNcboCvalue(c) = \begin{cases} \log(OldScoreNcbo(c)) \times \sum_{t \in T_c} C\text{-}value(t) & \text{if } \sum_{t \in T_c} C\text{-}value(t) \neq 0 \\ \log(OldScoreNcbo(c)) & \text{otherwise} \end{cases} \quad (2)$$

Where c is an annotating concept, $OldScoreNcbo(c)$ is the score obtained with the previous method for that concept, T_c is the set of matched terms with the concept c and $C\text{-}value(t)$ is the C-value score of the term t in the annotated text. As an example, for the same previous text, Table 3 describes the annotations obtained with the *ScoreNcboCvalue* scoring method. The results show that all annotations generated from matches with the term *basal cell carcinoma* are now ranked first, followed by the annotations generated from matches with the term *cell*. Therefore, this new scoring method addresses the first problem identified in previous section by giving better score to the annotations obtained with non-nested matching terms.

However, Table 3 also illustrates than the annotation with T025 (*cell*) has been pushed after all the annotations generated with the term *basal cell carcinoma* including the annotations with high level concepts such as D009375 (*neoplasms, glandular*) or D009370 (*neoplasms by histologic type*) obtained thanks to the is-a hierarchy expansion. This could be a problem as those terms are more general than the term *cell*, which had indeed generated direct annotations with preferred name. Therefore, to improve the results of the *ScoreNcboCvalue* scoring method, we need to decrease the effect of the hierarchy semantic expansion.

3.3 Old Annotator method + C-value + H

To penalize the score of expanded annotations, we propose to ignore when calculating the C-value of a concept the C-value of hierarchical annotations. A new formula is proposed with *ScoreNcboCvalue*, where the set T_c refers then to the set of terms annotated with the concept c when ignoring hierarchical annotations (the C-value of hierarchical annotations is consider zero). This new scoring method is designated by *ScoreNcboCvalueH*.

The scores of annotations generated using the *ScoreNcboCvalueH* method are presented in Table 4. This scoring method allows us to score and rank efficiently the annotations of the example text. The most appropriate annotation, done with concept D002280, is now ranked first with a very big score compared to the rest of the annotations bellow. In the second pool of annotations, with score between 1 and 2, both the ranking and the score are satisfying as they favor first precise concepts with direct matches such as *cell* or *carcinoma*, while keeping after the annotations with more general concepts.

Table 3. Annotations obtained with the *ScoreNcboCvalue* method and ranked by score S2.

Code	Preferred name	Matched terms	Type	F	S1		C-value		S2
D002277	Carcinoma	Basal cell carcinoma	exp n=1	4	36	76	6.33	6.33	**11.92**
		Carcinoma	pref	4	40		0		
D009375	Neoplasms, Glandular	Basal cell carcinoma	exp n=2	4	28	64	6.33	6.33	**11.45**
		Carcinoma	exp n=1	4	36		0		
D009370	Neoplasms by histologic type	Basal cell carcinoma	exp n=3	4	24	52	6.33	6.33	**10.87**
		Carcinoma	exp n=2	4	28		0		
D018295	Neoplasms, Basal Cell	Basal cell carcinoma	exp n=1	4	36		6.33		**9.86**
D002280	Carcinoma, Basal Cell	Basal cell carcinoma	syn	4	32		6.33		**9.54**
T025	Cell	Cell	pref	6	60	96	0	0	**2.05**
		Cell	exp n=1	6	36		0		
U000002	Anatomy (MeSH Category)	Cell	exp n=1	6	54		0		**1.73**
D002477	Cells	Cell	syn	6	48		0		**1.68**
U000019	Topical Descriptor	Cell	exp n=2	6	42		0		**1.62**
U000017	MeSH Descriptors	Cell	exp n=3	6	36		0		**1.55**
D009369	Neoplasms	Carcinoma	exp n=3	4	24		0		**1.38**

Table 4. Annotations obtained with the *ScoreNcboCvalueH* method and ranked by score S3.

Code	Preferred name	Matched terms	Type	F	S1		C-value		S2	S3
D002280	Carcinoma, Basal Cell	Basal cell carcinoma	syn	4	32		6.33		9.54	**9.54**
T025	Cell	Cell	pref	6	60	96	0	0	2.05	**2.05**
		Cell	exp n=1	6	36		0			
D002277	Carcinoma	Basal cell carcinoma	exp n=1	4	36	76	6.33	6.33	11.92	**1.88**
		Carcinoma	pref	4	40		0			
D009375	Neoplasms, Glandular	Basal cell carcinoma	exp n=2	4	28	64	6.33	6.33	11.45	**1.80**
		Carcinoma	exp n=1	4	36		0			
U000002	Anatomy (MeSH Category)	Cell	exp n=1	6	54		0		1.73	**1.73**
D009370	Neoplasms by histologic type	Basal cell carcinoma	exp n=3	4	24	52	6.33	6.33	10.87	**1.71**
		Carcinoma	exp n=2	4	28		0			
D002477	Cells	Cell	syn	6	48		0		1.68	**1.68**
U000019	Topical Descriptor	Cell	exp n=2	6	42		0		1.62	**1.62**
D018295	Neoplasms, Basal Cell	Basal cell carcinoma	exp n=1	4	36		6.33		9.86	**1.55**
U000017	MeSH Descriptors	Cell	exp n=3	6	36		0		1.55	**1.55**
D009369	Neoplasms	Carcinoma	exp n=3	4	24		0		1.38	**1.38**

4 Evaluation & Results

4.1 Evaluation using manual MeSH annotations as reference

In this section, we evaluate the relevance of the three annotation-scoring methods described previously using a manually annotated corpus: PubMed citations manually annotated with MeSH terms by experts from the US National Library of Medicine. Our hypothesis is that such a MeSH terms should count among the first automatic annotations obtained with the automated method when processing the related title and abstract. In the following, we describe our general evaluation procedure (illustrated in Figure 3):

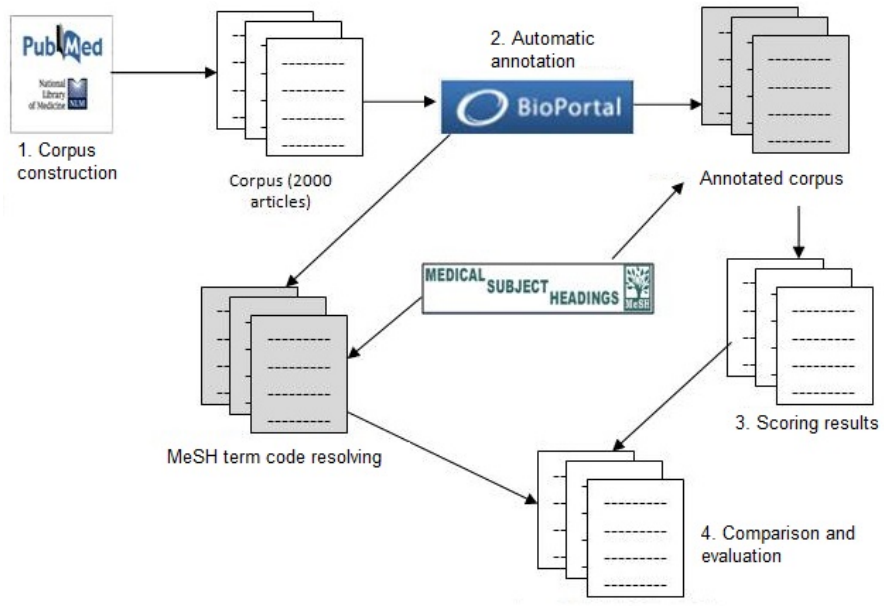

Fig. 3. Evaluation procedure

Corpus building: We have first collected 1,250 citations in English from the PubMed[2] database, previously annotated with the Medical Subject Heading (MeSH) terminology. For each article, we have a title, an abstract, and the referenced annotations (cf. Figure 4).

Automatic annotation: We have annotated the title and the abstract of each citation of the corpus using the Annotatorwith MeSH. During that step we have generated two sets of annotations for each citation according to the hierarchy level parameter of the Annotator either fixed to 0 (no is-a expansion) or 10 (expansion up to parent-level-10).

Mesh code mapping: Because PubMed citations annotations do not use the Mesh codes (identifiers) but the term itself in their format (cf. Figure 4), we had to reconcile the term with its code using the NCBO BioPortal search service which allows for a given term to retrieve the different identifiers of a concept, including the ones also returned by the Annotator.

Scoring of annotations: During this step, we have scored the annotations retrieved from the previous step using the three scoring methods described in Section 3.

Ranking of annotations: After scoring the annotations, we ranked them in ascending order. We grouped annotations with the same scores and assigned each group a sequential position starting with the group with the highest annotation score (cf. Figure 5).

Normalization: Because the rank value of a given annotation (or group of annotations) will be different with each scoring methods, we had to normalize the ranking to compare them.

[2] http://www.ncbi.nlm.nih.gov/pubmed

Comparison with MeSH annotations: At this step we look up the manual annotations retrieved among all the automatic annotations returned by the Annotator and keep only the citations with a recall above 30% (i.e., citations for which at least 30% of manual annotations have been found by the Annotator). For each subset of manual annotations we then compute above two evaluation measures: (i) Average reference annotation rank by citation; (ii) Average reference annotation rank of the whole corpus.

$$\text{Average\_citation\_rank}_M(\text{citation}_i) = \frac{100 * \sum_{c \in C} score_M(c)}{|C|} \quad (3)$$

Where C is the set of the manual annotations of the PubMed $citation_i$ found among the automatic annotations of this article with scoring method M.

$$\text{Average\_corpus\_rank}_M = \frac{\sum_{i=1}^{N} \text{Average\_citation\_rank}_M(\text{citation}_i)}{N} \quad (4)$$

Where N is the total number of citations in the corpus.

```
<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="MEDLINE">
        <PMID Version="1">15785422</PMID>
        <Article PubModel="Print">
            <ArticleTitle>[Intraosseous meningioma of the skull: radiologic pathologic correlation].
            </ArticleTitle>
            <Abstract>
                <AbstractText>Intraosseous meningiomas are rare ectopic meningiomas. The authors report
                the case of a hyperostotic intraosseous meningioma of the parietal bone without dural
                extension. The preoperative imaging findings, as well as imaging features of the
                surgical specimen and pathologic findings are discussed.</AbstractText>
            </Abstract>
        </Article>
        <MeshHeadingList>
            <MeshHeading> <Annotation>Female</Annotation> </MeshHeading>
            <MeshHeading> <Annotation>Humans</Annotation> </MeshHeading>
            <MeshHeading> <Annotation>Meningioma</Annotation> </MeshHeading>
            <MeshHeading> <Annotation>Middle Aged</Annotation>   </MeshHeading>
            <MeshHeading> <Annotation>Skull Neoplasms</Annotation> </MeshHeading>
        </MeshHeadingList>
    </MedlineCitation>
</PubmedArticle>
```

Fig. 4. Example of PubMed citation, including title, abstract and Mesh heading annotations.

4.2 Results

We compare the three scoring methods by calculating the average rank of the corpus for each scoring method. The lower the average rank is, the better, as it means that the reference annotations obtained better scores. Table 5 shows the results obtained with or without hierarchy expansion for each scoring methods.

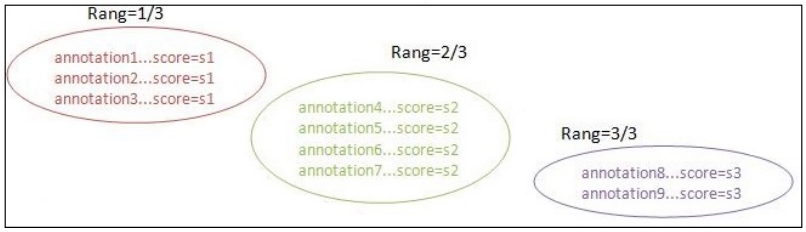

Fig. 5. Ranking of annotations by groups with same score.

We first note that the average corpus rank is lower (on results obtained with is-a hierarchy semantic expansion, independently of the scoring method. We also note that without hierarchy expansion *ScoreNcboCvalue* and *ScoreNcboCvalueH* obtained the same value, which is normal because the second measure was only affected hierarchy annotations. Finally, we note that with or without hierarchy expansion the *ScoreNcboCvalueH* method gets the best (lowest) rank which confirms the improvement of the scoring and ranking method. In addition, during our studies, we experimented with smaller recall values and obtained similar results.

Table 5. Average corpus rank obtained by each method over the PubMed corpus

Scoring method	Without hierarchy	With hierarchy
OldScoreNcbo	68.55	61.02
ScoreNcboCvalue	61.53	55.35
ScoreNcboCvalueH	61.53	48.68

This experiment has shown that our new scoring methods are more efficient than the old method used by the Annotator in terms of average corpus rank. Now, it is interesting to know in which cases this conclusion is true i.e., what types of annotations are ranked better. With the use of C-value, created to favour multi-term extraction, our assumption was that annotations made with multi-word terms will be ranked better. Therefore, we have performed basic statistics on the three following sets of annotations:

Improved annotations: when the two new methods provide the best rank;
Equal annotations: when the three methods give the same normalized rank;
Not improved annotations: when the *OldScoreNcbo* method gives the best rank.

For each set, we estimate the percentage of annotations done directly with a term made of one, two, three, four or five words or indirectly with is-a semantic expansion. Table 6 shows the results obtained for each annotation sets. We notice that in both cases, more than 60% of annotations that are equal or not improved by our new methods are annotations done with a single word term match. In addition, between 70% and 90% of improved annotations are done with a multi-word term match. This confirms us the improvements brought by our new methods was mainly on annotations done with multi-word terms. Also, it is important to notice that single word matches are not explicitly penalized by our method, but just pushed after multi-word matches. This means that

in the case of ontologies with mostly single world terms, our method will still behave properly.

Table 6. Statistics (i.e., number of annotations) in each set over the PubMed corpus and percentage of each annotation type

Annotations set	Without hierarchy			With hierarchy		
Equal	**48**	Hierarchy	0	12	Hierarchy	13
		1 word	**59**		1 word	**74**
		2 words	36		2 words	22
		3 words	3		3 words	1
		4 words	0		4 words	1
		5 words	0		5 words	0
Not improved	31	Hierarchy	0	**51**	Hierarchy	12
		1 word	**81**		1 word	**66**
		2 words	16		2 words	28
		3 words	1		3 words	4
		4 words	0		4 words	0
		5 words	0		5 words	0
Improved	21	Hierarchy	0	26	Hierarchy	0
		1 word	10		1 word	28
		2 words	**49**		2 words	**43**
		3 words	36		3 words	24
		4 words	3		4 words	2
		5 words	0		5 words	0

5 Discussion & Conclusion

In this paper, we have presented our approach to improve the results of the NCBO Annotator by ranking annotations according to their relevance. We have proposed two new methods and compared them one another as well as with the original annotation scoring provided in the first version of the Annotator (not available anymore since BioPortal 4.0). The introduction of the C-value measure has led to significant improvements in the scoring and ranking allowing to discriminate positively annotations made with multi-word terms and penalizing hierarchical annotations. We have evaluated the performance of these scoring methods using PubMed-MeSH annotations as a reference corpus. The evaluation demonstrated that the two new methods are more efficient than the old method of the Annotator in terms of average corpus rank. One limit of our approach is due to the fact that the Annotator service cannot handle structured data for

which more context information (e.g., title, abstract, preconditions, etc.) will improve the scoring of the annotations as explained in [13].

We are currently working to offer our new scoring methods to the community of users as an add-on when calling the NCBO Annotator. Indeed, an important advantage is that the scores can be computed only by processing the Annotator results and with no requirement to change the service implementation. As of today, the Java implementation of the methods is available as a JAR file offering the scoring functions taking as input the exact Annotator outputs and generating exactly the same XML or JSON outputs but completed and ranked with the scores. This JAR library is available on request. Our long term perspective, within the Semantic Indexing of French Biomedical Data Resources (SIFR) project (http://www.lirmm.fr/sifr) is to offer a service endpoint implementing several improvements (negation, disambiguation, new semantic expansion, new outputs formats, etc.) of the NCBO Annotator done with pre and post processing while still calling the Annotator service.

6 Acknowledgements

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University of Montpellier, CNRS and the Computational Biology Institute (IBC) of Montpellier. We thank the National Center for Biomedical Ontology (NCBO) for latest information about the Annotator.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, AMIA'01. pp. 17–21. Washington, DC, USA (November 2001)
2. Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 7(3), 256–274 (August 2006)
3. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F.: An Efficient Solution for Mapping Free Text to Ontology Terms. In: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'08. San Francisco, CA, USA (March 2008)
4. Dwinell, M.R., Worthey, E.A., Shimoyama, M., Bakir-Gungor, B., DePons, J., Laulederkind, S., Lowry, T., Nigram, R., Petri, V., Smith, J., Stoddard, A., Twigger, S.N., Jacob, H.J.: The Rat Genome Database 2009: variation, ontologies and pathways. Nucleic Acids Research 37((database)), 744–749 (Jan 2009)
5. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value Method. Digital Libraries 3(2), 115–130 (August 2000)
6. Hancock, D., Morrison, N., Velarde, G., Field, D.: Terminizer – Assisting Mark-Up of Text Using Ontological Terms. In: 3rd International Biocuration Conference. Berlin, Germany (April 2009)
7. Handschuh, S., Staab, S. (eds.): Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications, vol. 96. IOS Press (2003)
8. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: A Concept-based Search Engine for Structured Biomedical Text. American Medical Informatics Association 14(3), 253–263 (May-June 2007)

9. Jonquet, C., Musen, M.A., Shah, N.H.: Building a Biomedical Ontology Recommender Web Service. Biomedical Semantics 1(S1) (June 2010)
10. Jonquet, C., Shah, N.H., Musen, M.A.: The Open Biomedical Annotator. In: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09. pp. 56–60. San Francisco, CA, USA (March 2009)
11. Jr, W.A.B., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.A.: Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 23(13), 41–48 (2007)
12. McCool, R.G.R., Miller, E.: Semantic Search. In: 12th International Conference on World Wide Web, WWW'03. pp. 700–709. ACM, Budapest, Hungary (2003)
13. Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. American Medical Informatics Association 14(2), 164–174 (March 2007)
14. Murdoch, T.B., Detsky, A.S.: The Inevitable Application of Big Data to Health Care. Journal of the American Medical Association 309(13), 1351–1352 (2013)
15. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37((web server)), 170–173 (May 2009)
16. Pereira, S., Névéol, A., Kerdelhué, G., Serrot, E., Joubert, M., Darmoni, S.J.: Using multiterminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. In: American Medical Informatics Association Annual Symposium, AMIA'08. pp. 586–590. Washington DC, USA (November 2008)
17. Popov, B., Kiryakov, A., Kitchukov, I., Angelov, K., Kozhuharov, D.:
18. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A.: Text processing through Web services: Calling Whatizit. Bioinformatics 24(2), 296–298 (2008)
19. Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S.: Use and misuse of the gene ontology annotations. Nature Reviews Genetics 9, 509–515 (July 2008)
20. Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. Briefings in Bioinformatics 9(1), 75–90 (2008)
21. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. American Medical Informtics Association 17, 507–513 (June 2010)
22. Shah, N.H., Jonquet, C., Chiang, A.P., Butte, A.J., Chen, R., Musen, M.A.: Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. BMC Bioinformatics 10(2:S1) (February 2009)
23. Song, D., Chute, C.G., Taoa, C.: Semantator: Annotating Clinical Narratives with Semantic Web Ontologies. In: AMIA Joint Summits on Translational Science. pp. 20–29. San Francisco, USA (March 2012)
24. Syarifah, B.R., Shahrul, A.N., Wardhana, A.: Ranking and scoring semantic document annotation. In: International Conference on Science and Social Research, CSSR'10. pp. 691–694. IEE, Kuala Lumpur, Malaysia. (December 2010)
25. 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 4(1), 14–28 (January 2006)
26. Wu, A.: Ranking Biomedical Annotations with AnnotatorŠs Semantic Relevancy. Computational and Mathematical Methods in Medicine p. 11 (May 2014)
27. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms. In: 6th International Conference on Language Resources and Evaluation, LREC'08. pp. 2108–2113. Marrakech, Morocco (June 2008)