Improving term extraction with linguistic analysis in the biomedical domain

Wiktoria Golik[1], Robert Bossy[1], Zorana Ratkovic[1,2] and Claire Nédellec[1]

[1]MIG INRA UR1077, Domaine de Vilvert, F-850 Jouy-en-Josas, France 2
[2]LaTTiCe UMR 8094 CNRS, 1 rue Maurice Arnoux, F-92120 Montrouge, France
Wiktoria.Golik@jouy.inra.fr, Robert.Bossy@jouy.inra.fr, Zorana.Ratkovic@jouy.inra.fr, Claire.Nedellec@jouy.inra.fr

Abstract.
This paper presents a linguistic-based approach to term extraction from corpora in the biomedical domain. The method is based on an analysis of terms and their context that verify linguistic constraints. It focuses on participles and prepositional complements. The purpose of our approach is to obtain terms that are relevant for knowledge acquisition applications, such as the creation and enrichment of terminologies and ontologies. We report on the evaluations we conducted by applying two complementary strategies, using a reference terminology and a manual validation. They were applied to two corpora of differing genres and Life Science domains, namely pharmacology patents and animal physiology scientific articles. Our work shows that the linguistic analysis-based developments significantly improve the extraction results. The method is especially efficient when dealing with gerunds and *to* prepositional modifiers.

Keywords: term extraction, biomedical corpora, linguistic approach

1 Introduction

The amount of biomedical information is growing exponentially. Most of this information is made available through domain literature and is expressed in natural language (Jensen *et al.*, 2006). The need to automatically process this large amount of data has led to advancement in the field of biomedical text mining in the past several years. Most of the work has focused on Information Extraction (IE). The use of terminological resources and ontologies has been found to be necessary for high-quality IE (Nenadic *et al.*, 2006; Bodenreider, 2006). Recent developments in NLP and term extraction methods offer a powerful and efficient way to design terminological resources and facilitate access to scientific information.

In recent years, several approaches have been proposed for the acquisition of terms from text. Reviews by (Pazienza *et al.*, 2005; Zhang *et al.*, 2008) have described and compared the most popular techniques. Traditionally, term extraction methods are divided into linguistic, statistical and hybrid ones. Linguistic methods attempt to identify terms by their linguistic properties, while the statistical methods are based on frequency, association and the distribution of terms in documents. By handling the particularities of language structure, linguistic tools usually provide a large number of

well-formed and diverse candidate terms (CTs). Conversely, statistical approaches are knowledge-poor, but are fully automatic and are language and domain independent. Also, they provide noisier results due to their reliance on the frequency and length of CTs. Hybrid systems combine the advantages of both approaches, using linguistic techniques for term acquisition and statistics for term ranking (Sclano & Velardi, 2007; Wang *et al.*, 2007).

In this paper, we focus on the linguistic approach as a critical step for term extraction. The paper describes our method, which aims to improve the quality of extracted terms from biomedical corpora. The purpose is to obtain relevant terms for knowledge acquisition applications, such as the building of terminologies and ontologies. Unlike other applications (*e.g.* document ranking), the well-formedness of terms is crucial. Our method is based on the use of appropriate filtering and the processing of important linguistic structures that are commonly found in biomedical texts and beyond. In particular, they include structures containing prepositional noun phrases (NN PREP NN) and participles (past participles and gerunds). The paper details two different evaluations and discusses their complementarities.

The term extraction experiments were performed on two different types of corpora from two biomedical subdomains: patents and scientific papers from the pharmacology domain, and scientific journals from the animal physiology domain. In the next section we describe previous work in this field. Section 3 details the motivation of our work. The method is described in Section 4. Section 5 details the experiments. The results are reported in section 6, while in section 7 we discuss the results and conclude.

2 Related Work

Statistical methods have been found to be very effective at ranking CTs (Zhang *et al.*, 2010) and research in this field has been very active in recent years. Linguistic-based methods are founded on a deep analysis of different linguistic phenomena that can be observed in the data. Traditionally, most of the linguistic and hybrid approaches focus on noun phrase (NP) extraction, since NPs usually contain domain relevant semantic information (Justeson and Katz, 1995). The extraction generally targets multi-word units, since it has been found that 85% of domain-specific terms are multi-words (Nakagawa & Mori, 2002).

A typical extraction process includes POS tagging, tokenization, chunking and the use of linguistic patterns (Sclano & Velardi, 2007, Wermter & Hahn, 2005). POS tagging has a strong impact on the extraction of CTs, most notably on chunking, which is based on POS tags. POS tagging is particularly important for the correct extraction of phrases containing participles, since participles can play the role of verbs, adjectives or nouns. Despite the high performance of existing taggers, some errors do occur. The problem of participle tagging has been addressed in tagging annotation guidelines (Santorini, 1990) and syntactic parsing (Hara, Miyao & Tsuji, 2009). However, it still remains problematic for term extraction systems whose results strongly depend on the quality of the tagging.

The use of well-defined linguistic patterns improves the extraction accuracy and correctness. As described in (Frantzi, Ananiadou & Mina, 2000), they are built in order to extract frequent and simple structures (NN NN) as well as more complex ones, such as those including prepositions. The more a pattern is complex, the more variable and unpredictable the results may be, including incorrect terms. Simple patterns are less error prone, but they are too restrictive and less productive. Some systems use syntax-based endogenous disambiguation to handle noisy data. For example, Syntex (Bourigault, 2007) learns and compares syntactic contexts of CTs in order to distinguish relevant and irrelevant forms. This comparison results in a better extraction of prepositional phrases that typically depend on the context.

The quality of extraction also depends on the capability to extract various forms of terms. Variant extractors, such as FastR (Jaquemin, 1999) improve the completeness of extracted NPs. A large number of variants, including prepositional ones (*e.g.* NN at NN, NN in NN) are identified using different meta-rules (*i.e.* permutation, insertion, coordination). The abundance and accuracy of the produced variants depend on terms that are already recognized and accepted as valid (attested terms), as well as certified resources that are used as a starting point.

The use of existing domain-specific resources is another valuable way to enhance term extraction. POS-tagged resources help to deal with the POS tagging quality problem, as well as the lack of extraction patterns (Aubin & Hamon, 2006; Roberts et al., 2008).

Finally, the quality of CTs depends on their well-formedness and domain relevance. As already mentioned, this task is often handled by statistical approaches or the use of limited but efficient techniques such as stop lists or simple linguistic filters. TermExtractor (Sclano & Velardi, 2007) is a hybrid tool that relies on both. Some systems additionally use contextual information to capture the domain relevance of terms. In (Frantzi et al. 2000) the linguistic and statistical analysis of nested terms serves to identify domain-specific term markers. More recently, a similar method has also been applied to filter out incomplete phrases (Gojun *et al.*, 2012).

3 Motivation

There are many resources that have been developed for the biomedical domain, such as the UMLS metathesaurus (Bodenreider, 2004) and Gene Ontology (Harris *et al.*, 2004). Due to the fast evolution of the field there is a constant need to produce new resources and to complete existing ones. Data from different biomedical subdomains cannot be semantically processed using a single resource. The diversity of the goals further stresses the importance of having different and new resources for different datasets and different tasks. There is a need for efficient and automatic systems that can be used to create and guide the creation of such resources (Nédellec *et al.*, 2010).

Our work focuses on the improvement of term extraction of biomedical corpora that produces grammatically well-formed and application relevant terms. Most term extractors focus on NPs without taking into account possible prepositional phrases. This is due to the high attachment ambiguity. However, NPs with prepositional modi-

fiers such as *NN in NN*, *NN for NN*, *NN at NN* and *NN to NN* are interesting and useful to consider. The exploration of biomedical corpora, as well as the consultation of domain experts, have led us to conclude that these structures can generate highly relevant domain terms. They can considerably enrich specific domain resources where prepositional phrases are infrequent.

Next, we consider that past and present participles, when correctly POS tagged and extracted, can improve the extraction comprehensiveness, in particular in the biomedical domain where they are frequently employed.

We also observed that the extracted terms often contain terms that are referential or too general. These terms are irrelevant to the application domain and should be filtered out. We propose to improve term extraction in three steps: (i) extension of extracted terms by considering prepositional phrases (ii) supervision of the extraction through an enhanced processing of *–ing* and *–ed* forms (iii) elimination of irrelevant terms through the use of filters.

4 Methods

Our method aims to improve term extraction of biomedical corpora, for the creation of lexical resources, such as terminologies or ontologies. The method is based on a linguistic analysis of biomedical texts. We make no assumptions as to the nature and exact usage of the resource. Rather, we aim to extract terms which are well-formed, complete from a syntactic and semantic point of view and could be useful for a given application. We leave the term relevance decision to experts. Our approach captures the linguistic phenomena that are found in biomedical corpora differing in style and genre. For this reason it will be broadly applicable. We did not use a machine learning (ML) approach since this requires annotated data, which is not available and is costly to produce. Moreover, ML approaches always introduce bias with respect to the training corpus that is used. The extraction is based on linguistics patterns reinforced by additional context-based rules in order to handle specific prepositional phrases and participles. We are unaware of any previous work that combines both in order to enhance the extraction of such structures.

4.1 Extraction of Structures Containing Prepositions

Prepositional attachment resolution is a well-known problem in NLP, especially for syntactic parsing (Ratnaparkhi *et al.*, 1994; Nakov & Hearst, 2005). Due to their high degree of ambiguity, PPs are often not taken into account in term extraction. In the case of shallow parsing the prepositions are usually treated as boundaries of chunks (Ash & Daelemans, 2009), except for the frequent prepositional structure *NN of NN*. PP ambiguity is strongly related to the nature of predicate-argument structures, and more precisely to the difference between an argument and an adjunct (Grimshaw, 1992). According to the context, PPs act as arguments or adjuncts. This ambiguity is difficult to resolve automatically since it depends on both syntax and semantics. Our corpus analysis revealed terms containing PPs, which were both well-formed and

domain relevant. A closer look at the most frequent prepositions showed that most frequently they are either arguments of verbs (such as *by* and *in*) or are adjuncts and parts of NPs (such as *of*, *to* and *at*). Our work focuses on the latter case. Since the *of* preposition is treated in previous works, we focus on *at* and *to*. For instance, NPs with the preposition *at* often contain information about level, condition or period (*e.g.*, age at parturition, body weight at birth). NPs containing *to* denote reactions to different stimuli and situations (*e.g.*, susceptibility to mastitis, response to fish oil supplementation).

The extraction of terms containing *at* and *to* is done in two steps: (i) the application of extraction patterns that include the prepositions (*e.g.*, NN *to* NN or NN *at* JJ NN) (ii) the filtering of irrelevant attachments by a set of five context based rules (see Table 1). Their role is either to trigger the extraction of relevant PPs or to prevent the extraction of irrelevant ones. For instance, for CTs containing *to*, the first rule in the table checks if the structure NN *to* NN is preceded by *from* or *by* (*e.g.*, from mother to young), in which case the CT is not extracted because *to* is directly related to *from* and not to the NP. The proposed rules are generic in order to be applicable to different corpora.

Table 1. Context-based rules for the extraction of *to* and *at*.

Context-based rules	**Relevant POS tag**
[from|by][not SENT][**to**]	Reject
[not NN] [not V][**to**]	Reject
[not V|VVN][**to**]	Reject
[NN|VVN and not stop-list][**to**]	Reject
[stop-list][not SENT][**at**] e.g. *weight at birth*	AT

4.2 Candidate Term Refinement by Filtering

Filtering means the automatic removal of forms considered to be non-terms, similarly to Pazienza *et al.* (2005). We use basic linguistic filters that are efficient and easy to build and maintain. The method aims to improve the extraction results by filtering out two kinds of irrelevant terms. First, terms which are structurally incoherent (*i.e.* invalid). The filter is simple, yet very efficient. Second, terms which are structurally coherent, but are referential or too vague (*i.e.* semantically poor). The aim of the filter is to remove terms that are grammatically well-formed, but that are not useful for any domain application. Most often, they are referential expressions where the context is needed to interpret the term, or terms that reflect the writing style and do not convey domain knowledge.

Filtering of Invalid Forms

The quality of the extraction depends on the quality of preceding tokenization and POS tagging steps. The filtering handles incorrectly tokenized or POS-tagged CTs. The filters capture: surface forms that start or end with invalid characters (*e.g.*, +, ~, *, \, .); surface forms that start with coordination marks (*e.g.*, and, or), contain only parenthesis or square bracket (*e.g.*, B2(lipid source), start or end with a unit of measurement (*e.g.*, kg clozapine, 9 mm), or contain only numbers (*e.g.*, 1666-1673).

Filtering is also used to handle function words that are traditionally filtered out due to their high frequency and their lack of semantic information, such as definite and indefinite articles, demonstratives and *wh*-determiners (*e.g.*, the, these, which, each).

Filtering of Semantically Poor Terms

The filtering performs a preliminary semantic refinement of CTs. It identifies correctly extracted NPs that cannot be considered as true terms from a semantic point of view. They can be divided into four main types:

- NPs usually containing non discriminative modifiers (*e.g.*, important, particular, useful, various, certain, amount of)
- NPs that depend on the context in order to be properly interpreted (*e.g.*, day 33, position 1978*)*; they often include comparatives (*e.g.*, greater DMI, higher number of assays)
- NPs directly linked to the nature and style of the corpora (*e.g.*, embodiment, point of view, above-mentioned feature, present experiment)
- Named entities related to the references present in the documents, dates (*e.g.*, Smith et al., November 1986)

4.3 The Extraction of Gerunds and Past Participles

The last part of our method focuses on the POS-tagging of participles, which is a common problem in NLP, notably in the biomedical domain (Teteisi & Tsuji, 2006). According to the context, the participles play either the role of verbs, adjectives or nouns (*i.e.*, binding). These three POS tags are particularly difficult to distinguish and the context is usually discriminant. The erroneous tagging of participles usually leads to the omission of relevant NPs. The number of NPs with *–ed* and *-ing* is high in most biomedical corpora. While there are POS taggers that have been adapted to the biomedical domain, the problem still persists. We do not consider retraining a tagger for two main reasons. First, such a process requires manually annotated data. Secondly, the tagger will be influenced by the corpora that it is trained on.

To improve the completeness of the extraction, we propose to supervise the tagging stage using five context-based rules. These rules take into account the words surrounding an *–ing* or *–ed* form and their POS tags. Additionally, the rules use a stop list of forms that are always verbs (*i.e.* being, using, getting). The list was collected from the corpus and tested using the criterion proposed in (Santorini, 1999). For instance, an *–ing* form (not in the stop list) preceded by *of* and not followed by a verb or punctuation mark will be tagged as NN (*e.g.* day of calving, role of farming).

Table 2. Context-based rules for the disambiguation of participle POS tags. [ing*= ing and not stop-list]

Rule	Relevant POS tag
[DT|JJ|SENT] [–**ing***)][NN|NNS|NP|,] e.g. *eating quality*; *a training period*	NN
[DT|JJ][–**ing***][JJ] e.g. *increasing perinatal mortality*	JJ
[of][-***ing****][not V|,|SENT] e.g. day *of calving* ; role *of farming*	NN
[of][-***ing****][DT|JJ|PP|WDT] e.g. accuracy *of predicting the* percentage	VVG
[-***ed***][NN|NP|JJ|NNS] e.g. autumn *saved* pasture; *immunized animals*	JJ

5 Experiments

5.1 The Dataset

Biomedical literature is very rich and diverse. The variation in document language significantly varies with respect to the scientific field and the document genre (Lippincott *et al.*, 2010). In our experiments, we use two corpora of different genres: patents and scientific papers. They also belong to two very distinct biomedical subdomains: pharmacology and animal physiology. They are representative of the heterogeneity found in biomedical texts. The patents belong to the legal literature and are characterized by a highly controlled structure and vocabulary, while the scientific papers express different structural constraints and language. Further, scientific papers describe experimental hypotheses, procedures and results. By dealing with these very different corpora, we demonstrate our method to be general and applicable to different biomedical domain texts.

Pharmacology Domain Corpora

For the first evaluation experiment we used the pharmacology corpora used for the term extraction challenge of the *Quaero*[1] project organized in 2010 and 2011 in which we participated (Mondary *et al.*, 2012). It consists of four corpora: three are made up of patents (C1, C2, and C3) and one of scientific papers (CA). The patents are the European patents from the A61K class of the ECLA classification on preparations for medical, dental or toilet purposes. For the evaluation, we reused the largest patent corpus (C3) and CA. C3 contains 157 patents (2,500,000 words). CA is made up of

[1] http://www.quaero.org/modules/movie/scenes/home/index.php?FUSEBOX_LANG=2

7,030 scientific paper abstracts (1,500,000 words) from the PASCAL database[2]. The extraction results were evaluated against a reference terminology (see section 5.3).

Animal Physiology Domain Corpus

The animal physiology corpus is made up of full-text papers from the *Animal* journal (Cambridge University Press) published until 2011. The corpus contains 697 scientific communication papers that cover a large number of subjects from the animal physiology domain. Scientific papers use descriptive language and are characterized by a high variability of expressions and by the frequent presence of specific linguistic forms such as gerunds. In this corpus these forms are used to describe the states or activities of animals (*i.e.* abnormal calving, adequate laying space, grazing behavior).

5.2 BioYaTeA

For our experiments we used the BioYaTeA[3] term extractor, an extended version of YaTeA (Aubin & Hamon, 2006). YaTeA's extraction method includes the detection of morpho-syntactic boundaries and the matching of parsing patterns. It also comprises exogenous (supervised) and endogenous (unsupervised) disambiguation. For the experiment, we used YaTeA version v.0.6.

We extended YaTeA with new syntactic patterns, context-based rules and the post-processing filtering described in section 4. This new version is called BioYaTeA. BioYaTeA is integrated into a generic NLP platform developed at INRA-MIG, namely the AlvisNLP pipeline (Nédellec *et al.*, 2008). It takes as input the results of the AlvisNLP tokenizer and the POS tagger TreeTagger (Schmid, 1994). BioYaTeA was designed to build domain specific ontologies for automatic fine-grained indexing of biomedical texts (Nédellec *et al.*, 2010) for semantic search engine applications. Here, we use it to measure the added-value of our method. However, our approach is universal and it could be implemented with any term extractor, whether it is a statistical, linguistic or hybrid one.

5.3 Evaluation Methods

The evaluation of extracted CTs is a difficult and costly task. It requires the definition of what a term is and how to determine it (Pazienza, 2005; Vivaldi, 2007). Early works aimed to define termhood (Kageura & Umino, 1996) and the different ways of measuring it (Ananiadou *et al.*, 1998). Difficulties remain due to the complex nature of terms and the lack of a general consensus. Also, the evaluation of corpus term extraction results depends on the target application, the strategy being used and the nature of the corpora (Zhang *et al.*, 2008).

Traditionally, the results are evaluated using (i) a comparison to a reference terminology (ii) expert judgment with respect to a target end-user application (*i.e.* infor-

[2] The multidisciplinary bibliographical database produced by INIST-CNRS.PASCAL can be found on INIST's official website: http://inist.fr/spip.php?article11

[3] BioYaTeA is available at http://search.cpan.org/~bibliome/Lingua-BioYaTeA/

mation extraction). The advantage of the reference-based approach is that it is fast, fully automatic, and it allows for the use of standard metrics such as recall, precision and F-measure. However, relevant corpus terms might be counted as irrelevant because they are missing from the reference. To be useful, the recall should be measured with respect to what is effectively extractable from the corpus. This requires a manual annotation of corpus terms. The reference-based approach is not an absolute measure, but is used to measure the relative performance of different tools.

Manual validation is time consuming and requires the participation of domain experts or knowledge engineers. It relies on human judgment that can vary from one person to another. However, this type of evaluation is the most efficient and popular since it evaluates the extraction quality for a targeted purpose.

Our experiments include both types of evaluation. They are fundamentally different in their goal; the reference-based evaluation allows us to check the domain relevance of CTs, while the manual validation aims to estimate the value of CTs according to the ontology building task.

Automatic Evaluation of Pharmacology Term Candidates

In order to evaluate the CTs extracted from the pharmacology corpora, we used the same evaluation method as proposed for the *Quaero* term extraction challenge (Mondary *et al.*, 2012). The aim of the original evaluation was to compare the results of different term extractors using pharmacology domain corpora. The extractions were compared to a gold standard reference terminology containing 76,466 terms. According to the protocol proposed in (Nazarenko & Zargayouna, 2009) the standard precision and recall metrics were adapted to terminological result evaluations by taking into account partial matches. The results of the *Quaero* evaluation showed that filtering played an essential role in the quality of the results. We repeated the same experiments in order to better characterize the impact of each improvement described in Section 4. We applied the same protocol with the same corpora, as well as the same gold reference (see section 6.1).

Manual Evaluation of Animal Physiology Candidate Terms

The second evaluation was done using the animal physiology domain corpus. The evaluation was manual and involved ten annotators who were not animal physiology experts, but were familiar with the biomedical domain and with the development of ontology driven IR and IE applications. To validate terms, they referred to detailed guidelines[4] as defined by a knowledge engineer.

A sample of terms was judged according to their correctness in general, as well as according to their potential relevance for the design of an ontology. The annotators pointed out CTs that were too general or rhetorical. This allowed us to measure the impact of the extraction improvements. Since the annotators relied on the guidelines and not their expertise, this suggests that their validation should be more consistent compared to a domain expert validation. The evaluation was performed using the TyDI interface (Nédellec *et al.*, 2010) that gives access to the context of the terms,

[4] http://bibliome.jouy.inra.fr/GuideEvaluateur.pdf

which can be critical in some cases. The validation of terms was double-blind, with two annotators assigned to each term. We measured the Cohen kappa inter-annotator agreement (Cohen, 1960).

6 Results

The evaluation compares three versions of the term extractor, from the basic to the most enhanced version: YaTeA, YaTeA with filter and BioYaTeA (*i.e.* YaTeA with filter and rules). BioYaTeA does not perform any additional processing of single-word CTs. However, we decided to include single-word CTs in the reference-based evaluation due to their presence in the reference terminology. For the manual validation we only kept multi-word CTs.

6.1 Pharmacology Corpora Extraction: Evaluation and Result Analysis

The three term extractors were applied to the two pharmacology corpora (C3 and CA). The extracted CTs were evaluated and compared to the reference terminology. The terminological precision (t-precision), the terminological recall (t-recall) and the terminological F-measure (t-F-measure) were calculated for each corpus and each extraction as displayed in Table 3.

Corpus		YaTeA	YaTeA+filter	BioYaTeA (YaTeA+filter+rules)
C3	TP	34.2	48.0 (+13.8)	52.9 (+4.9)
	TR	33.1	29.4 (-3.7)	29.1 (-0.3)
	TF	33.7	36.4 (+2.7)	37.5 (+1.1)
CA	TP	46.2	56.7 (+10.5)	55.5 (-1.2)
	TR	37.3	33.9 (-3.4)	33.9 (+0)
	TF	41.3	42.4 (+1.1)	42.1 (-0.3)

Table 3**.** Pharmacology corpora extraction evaluation results.

There are several general trends to be observed from Table 3. BioYaTeA increases the precision, while decreasing the recall. However, there is an overall improvement, as shown by the increase in F-measure. For both of the corpora, the filtering results in a global improvement. These results show that the filtering is efficient for both corpora, even thought they differ in size and genre. In particular, the filters avoided an overly greedy extraction, making the extraction more precise by keeping a large number of true positives.

The effect of the rules is not the same for the two corpora. Overall, the effect of the rules on C3 results is positive, as shown by the last column in Table 3. This is mainly due to an increase in precision. Surprisingly, the effect of the rules on the CA corpus is less pronounced. Their impact on the t-recall is negligible while the t-precision slightly decreased. Given that the rules target only specific structures (participles and prepositional NPs) they had a less significant impact on the results than the filters.

The analysis of the effects of the rules is complex because of possible interactions between rules and filters. Our hypothesis is that it may be due to a different coverage of the targeted syntactic structures. For instance, the rules that limit the extraction of unwanted prepositional and gerund NPs are responsible for the t-precision increase in C3. The more permissive rules handling the participial NP extraction may decrease the t-recall in CA, where these structures are particularly abundant.

6.2 Animal Corpus Extraction: Evaluation and Result Analysis

We performed two manual validations on the *Animal* corpus in order to evaluate the filtering and the rules.

Validation of the Filter

The extractor comparison sample was built using the two sets of CTs extracted by YaTeA: CTs rejected by the filter (within BioYaTeA) and CTs accepted by the filter. The manual evaluation focused on a subset of 1,125 candidate terms (0.5% of the total) that were randomly selected from the two CT sets. The proportion of CTs from each sample as related to the total had been kept during the sample selection. Each term was validated by two annotators in a double blind-mode and each annotator validated 225 terms. The annotators did not know which terms were rejected or accepted by the filter. The Cohen kappa inter-validator agreement was 0.92, which indicates a very high agreement. The results are displayed in Table 4. Most of the CTs retained by the filters were manually validated as correct by the human annotators (68%). Similarly, most of the terms rejected were validated as incorrect (77%). As with the pharmacology evaluation, the filter significantly improved the extraction result. A detailed analysis of CTs that were wrongly omitted by the filter led us to observe that 30% of them were incomplete forms, due to POS tagging or attachment errors. The remaining 70% were well-formed according to the annotators but were either irrelevant to the target application, or too general for a domain ontology (*e.g.*, models of work, subset of data, progressive increase), which is outside the scope of the term extractor.

Table 4. Validation of terms rejected or accepted by the filter.

	TOTAL (1125)	Rejected (124)	Passed (1001)
Correct terms	713 (63%)	27 (22%)	686 (68%, +5)
Incorrect terms	376 (33%)	95 (77%)	281 (28%, -5)
Controversial terms	36 (3%)	2 (1.6%)	34 (4%)

Finally, the analysis of false negatives showed that a large number of them were filtered out because they end with numbers (*e.g.*, FIL 2001) or start with a single capital letter. It is not within a general-purpose term extractor's ability to distinguish forms such as *N basis* (correct) from *S lambs* (incorrect). This suggests that the filter should

be enriched in order to take into account additional semantic domain-specific knowledge, when possible.

Validation of the Rules

In order to evaluate the effect of the context-based rules, we manually validated two sets of candidate terms: those extracted by YaTeA and those extracted by BioYaTeA. The filter was applied to both extractions. As with the previous validation we selected a random sample. We took care to retain the relative frequency of the syntactic structure distributions (Table 5).

Table 5. Number of terms extracted by two term extractors.

Type of structure	YaTeA + filter	BioYaTeA
-ing	21	81
-ed	15	324
to	41	12
at	0	17
Other	455	159
Total sampled	532	593
Total extracted	27589	31206

The prepositional structures with *at* were not extracted by YaTeA, but only by BioYaTeA. The number of expressions containing the *to* preposition was lower in the second extraction because of more restrictive rules. Conversely, the number of *–ing* and *–ed* forms was higher in the second extraction. This was due to the rules being more inclusive.

The annotators validated the same number of terms as for the previous validation and they referred to the same guidelines. The Cohen kappa was 0.78, which indicates a reasonably high agreement. The results in Table 6 show that for the specific syntactic structures that were tackled, the rules improve the results. In particular, for the *–ing* and *to* structures there is an increase in the number of correct CTs that were extracted and a decrease in the extraction of incorrect CTs. For the *at* CTs, the rules permit the extraction of such structures, which for the most part are correct (76%). The results for the *-ed* structures are less pronounced. There is an increase in both the number of correct and incorrect CTs. This suggests that the context-based rules concerning the *-ed* forms were too permissive. After a deeper analysis of the *-ed* false positives, we noticed that 38% of the CTs were validated as incorrect because they played the role of verbs, 45% were viewed as irrelevant for the domain (*e.g.*, detailed description of the dissection, improved likelihood, tested oils), 8% were incomplete or incorrectly extracted and 10% were validated as incorrect but no reason was specified.

The results confirm that the extraction of *–ed* NPs is a complex task. The extraction should be improved using an exhaustive stop list. Furthermore, an additional

deeper analysis of the syntactic context should be completed in order to better define the context of correct terms.

Table 6**.** Validation results of the terms extracted with and without rules.

	Correct		**Incorrect**		**Conflicts**	
	BioYaTeA	BioYaTeA + rules	BioYaTeA	BioYaTeA + rules	BioYaTeA	BioYaTeA + rules
-ed	53%	55%	20%	36%	27%	8%
-ing	48%	62%	38%	30%	14%	9%
to	41%	83%	51%	8%	7%	8%
at	-	76%	-	24%	0	0
Other	55%	50%	32%	41%	13%	8%
Total	53%	56%	34%	36%	13%	8%

Finally, the quality of the extraction of *other* CTs is surprisingly lower when using the rules. We noticed that most of these CTs (46%), despite their valid form, were incorrect because they were irrelevant as ontology concepts. The rest of the rejected NPs were incomplete or they were erroneously extracted. This shows that although the new context-based rules cannot capture semantic information, they are efficient at dealing with syntactic criteria. We need to extend the method by taking into account the semantic information, together with the syntactic information that the filters and rules handle.

7 Conclusion

Both experiments show promising and interesting results. First, the positive impact of the filtering is clear in both of the evaluations. The role of the context-based rules turned out to be more difficult to assess, especially using the gold standard approach. However, the more detailed analysis of the manual validation showed powerful rules (*to*, *at*, *-ing*), as well as concerns that need to be addressed more thoroughly (*–ed*). Finally, the results confirm that the relevance of CTs for a given application is an important criterion. Even thought well-formed, a part of the CTs were evaluated as incorrect because they were inappropriate for the application. To better address this problem, we plan on adding complementary techniques that examine the semantics of the CTs, such as distributional analysis (Harris, 1954) in order to group CTs into semantic clusters. Distributional semantics could bring us new relevant extraction patterns and help us to built appropriate stop lists. Building well-limited semantic clusters could be an efficient way to better distinguish domain and application relevant CTs.

In this paper, we presented the improvements designed to increase the quality, completeness and accuracy of extracted terms from biomedical corpora, with respect to the design of domain resources. They consist of linguistic based filtering of unwanted candidate terms and a rule-controlled extraction of NPs containing prepositions and participles. Two different evaluation strategies were applied, confirming the positive impact of such improvements for two different biomedical corpora. Further work should be conducted to confirm the results in other domains.

Acknowledgements

The authors thank the following persons for having validated the terms: Philippe Bessières, Tristan Bitard-Feildel, Paul Bui-Quang, Julien Jourde, Dialekti Valsamou and Pierre Warnier. This work was partially supported by the Quareo Programme funded by OSEO (French agency for innovation).

References

1. Aubin S. and Hamon T. (2006). Improving Term Extraction with Terminological Resources, in T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (eds.), *Proc. of the Advances in Natural Language Processing*, FinTAL'06, LNAI 4139, Springer, p. 380-387, 2006.
2. Bodenreider, O. (2006) Lexical, terminological and ontological resources for biological text mining, in Ananiadou S., McNaught J., (eds.) *Text mining for biology and biomedicine*, Artech House, p. 43-66.
3. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Researcher*, 32: supplement 1.
4. Bossy, R., Kotoujansky, A., et al. (2008). *Close Integration of ML and NLP Tools in BioAlvis for Semantic Search in Bacteriology.* In: Burger, A. et al. (eds.) Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences. UK (2008).
5. Bourigault, D. (2007). Un analyseur syntaxique opérationnel: SYNTEX. *Mémoire d'Habilitation*, Université de Toulouse-le-Mirail.
6. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
7. Frantzi, K., Ananiadou, S. and Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal of Digital Libraries*, 3(2):117–132.
8. Frantzi, K. T., Ananiadou, S., and Tsujii, J. (1998). The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. In Goos G., Hartmanis J. & van Leeuwen J. (eds.), *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference*, ECDL'98 (Vol. 1513, p. 585–604). Lecture Notes in Computer Science. Berlin/ Heidelberg: Springer.
9. Gojun, A., Heid, U., Weissbach, B., Loth, C. and Mingers, I. (2012). *Adapting and evaluating a generic term extraction tool.* Proceeding of LREC-8.
10. Grimshaw, J. (1992). *Argument structure.* MIT Press, Cambridge, MA.
11. Harris, MA., et al. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acid Research, 32 (Database issue).
12. Harris, Z. (1954). Distributional structure. *Word*, 10: 146-162.

13. Jacquemin C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL'99*, p. 341-348.
14. Jensen, L. J., Saric, J. and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7: 119-129.
15. Justeson, J. S. and Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.
16. Kageura K. and Umino B. (1996). Methods for Automatic Term Recognition: A Review. *Terminology* 3(2), p. 259-289.
17. Lippincott, T., O Seaghdha, D., et al. (2010). *Exploring variation across biomedical subdomains*. In Proceedings of Coling. Beijing, China.
18. Mondary, T., Nazarenko, A., et al. (2012). The Quæro Evaluation Initiative on Term Extraction. In *Proceedings of LREC-8*, p. 663–669. Istambul, Turkey.
19. Nakagawa, H. and Mori, T. (2002). A simple but powerful automatic term extraction method. In *COMPUTERM 2002 – Proceedings of the 2nd International Workshop on Computational Terminology*, p. 29-35. Taipei, Taiwan.
20. Nakov, P. and Hearst, M. (2005). Using the web as an implicit training set: Application to structural ambiguity resolution. In *Proceedings of HLT-EMNLP*, p. 17–24.
21. Nédellec, C, Golik, W., et al. (2010). Building Large Lexicalized Ontologies from Text: a Use Case in Indexing Biotechnology Patents, *EKAW 2010.* Lisbon, Portugal, Oct. 11-15, 2010.
22. Nenadic, G., Okazaki, N., and Ananiadou, S. (2006). *Towards a terminological resource for biomedical text mining.* In *Proceedings of LREC-5*, Genoa, Italy, May.
23. Pazienza, M. T., Pennacchiotti, M. and Zanzotto, F. M. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. In *S. Sirmakessis (eds.), Knowledge mining: Proceedings of the NEMIS 2004 final conference*, p. 255–279, Berlin Heidelberg. Springer.
24. Ratnaparkhi, A., Reynar, J. and Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, p. 250 – 252.
25. Sant, P.M. (2004). Levenshtein distance. In *Dictionary of Algorithms and Data Structures* [online], Black P.E., (ed.), U.S. National Institute of Standards and Technology, 2004.
26. Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.
27. Sclano, F. and Velardi, P. (2007). TermExtractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of I-ESA 2007*.
28. Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK
29. Van Asch, V., & Daelemans, W. (2009). *Prepositional phrase attachment in shallow parsing*. In Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (RANLP), p. 12-17. Borovets, Bulgaria: Association for Computational Linguistics
30. Vivaldi J. & Rodríguez H. (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13:2, p. 225–248.
31. Wang, X., McCallum, A. and Wei, X. (2007). Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE international conference on data mining*, p. 697–702. Washington, DC. IEEE Computer Society.

32. Wermter, J. and Hahn, U. (2005). *Paradigmatic modifiability statistics for the extraction of complex multi-word terms*. In Proc. of HLT-EMNLP'05, 843–850.
33. Zargayouna, H. and Nazarenko, A. (2010). *Evaluation of Textual Knowledge Acquisition Tools: a Challenging Task.* In Proceedings of LREC 2010, p. 435–440, Valletta, Malta.
34. Z. Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). *A comparative evaluation of term recognition algorithms.* In Proceedings of LREC 2008.
35. Zhang, Z., Iria, J. and Ciravegna, F. (2010). Improving Domain-specific Entity Recognition with Automatic Term Recognition and Feature Extraction. In *Proceedings of LREC 2010*, p. 2606–2613. Valletta, Malta.