Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary

Hammad Afzal, Robert Stevens, Goran Nenadic
School of Computer Science, University of Manchester, Manchester, UK
{H.Afzal@postgrad., R.Stevens@, G.Nenadic@}manchester.ac.uk

Abstract

Most bio text-mining efforts so far have focused on identification of biological, molecular and chemical entities from the literature to support knowledge acquisition and discovery in the life sciences. There are also a growing number of bioinformatics services and tools available. This raises the challenging problem of semi-automated annotation, documentation and discovery of services suitable for a specific data analysis and/or integration into workflows. The first step in this process would be to build a controlled vocabulary to describe bioinformatics services, which can then be used for service retrieval and discovery. In this paper we present a methodology that combines lexical and contextual profiles of candidate terms to suggest terms for the bioinformatics vocabulary. The method achieved an estimated precision in the range 70-90% with recall between 20 and 90%. After processing the whole of BMC Bioinformatics, almost 80% of the top 300 terms were deemed as conceptual terms relevant for describing the major concepts in bioinformatics. In addition to this, the method has also extracted a number of service and tool names. The controlled vocabulary is freely available at: http://gnode1.mib.man.ac.uk/bioinf/CV.

1 Introduction

Along with the huge amount of experimental data, both raw and curated, and together with the literature being published in the biomedical domain, various bioinformatics data sources and tools have exposed programmatic interfaces as services. Resource sharing has already been established as a common policy within the community, and many groups have dedicated significant efforts to organise both internal and public repositories of bioinformatics tools, typically classifying them in broad categories (e.g. EBI Web services[1] are organised into data retrieval, analytics, similarity searches, multiple alignment, literature processing, etc.). Several projects and initiatives (e.g. myGrid[2] and myExperiment[3]) are annotating functional capabilities and semantically describing resources in a way which would make them discoverable and usable both by bioinformaticians and machines. Service descriptions typically include both textual explanations and ontological annotations. For example, EBI's *emma* service[4] is represented by the following (textual) description in the myGrid repository:[5]

> *Performs a multiple alignment of nucleic acid or protein sequences using ClustalW program*

along with a set of myGrid ontology[6] tags describing its operation (*multiple local aligning*), type (*Soaplab service*), parameters (including name, semantic type (e.g. *biological sequence*) and format (e.g. *single sequence format*), etc.).

Currently, most of the frameworks cataloguing bioinformatics services and workflows (e.g. myGrid/Taverna (Oinn et al, 2007)) describe resources manually, which – like any curation task – requires a lot of time and effort. As the number of services is increasing, manual annotation is becoming a bottleneck for discovering and using relevant services and tools (Cannata et al, 2005). Therefore, (semi)automatic methodologies to describe services are becoming inevitable, including automatic extraction of functional descriptions of services from available documents (articles, blogs, documentation, user manuals, etc.). Furthermore, since the domain is extremely dynamic, controlled vocabularies and/or ontologies that are (or can be) used for annotations need to be regularly updated and adjusted to include emerging methods, functionalities, data formats, etc. For example, the myGrid ontology

[1] http://www.ebi.ac.uk/Tools/webservices/

[2] http://www.mygrid.org.uk/

[3] http://www.myexperiment.org/

[4] http://www.ebi.ac.uk/soaplab/emboss4/services/alignment_multiple.emma

[5] http://www.mygrid.org.uk/feta/mygrid/descriptions/Soaplab_EBI/alignment_multiple/emma.xml

[6] http://www.mygrid.ac.uk/ontology

(Wolstencroft et al, 2007) contains around 440 bioinformatics terms that are currently used to describe services; still, for many potentially useful services, there may not be a set of adequate ontological terms or keywords for their description. For example, it would be difficult to precisely describe *GeneSom* service (Yan, 2002) for clustering-based microarray data analysis, as term *clustering* is not included in the current set of the ontology terms (the closest related term would be *grouping*).

In this paper we present a methodology and results in building a controlled vocabulary (CV) of bioinformatics terms that can be used for semantic annotation and description of services. By CV, we mean a set of key terms that are used to convey relevant information in a given domain or task (Kageura and Umino, 1996). Our main hypothesis is that new potential descriptors are likely to appear in documents that report on service design or utilisation. Therefore, our method for identification of terms related to bioinformatics services is based on processing full text articles from relevant journals. We have combined an automatic term recognition technique with a term classification approach based on lexical and contextual properties of candidate terms. Since not all terms that appear in a given corpus are of interest for a given task (e.g. specific protein/gene names, drugs, etc. may not be of interest to service descriptions), the method aims to filter out candidate terms that are not relevant for the task. The results obtained are very encouraging, showing that 70-90% of terms obtained are relevant for service descriptions, making the CV generation a first step towards facilitating automated annotation of services.

The paper is organised as follows. In Section 2 we present the overall methodology. The results and discussions are presented in sections 3 and 4 respectively, while related work is examined in Section 5. Finally, Section 6 concludes the paper and gives an outline of topics for future work.

2 Methodology

We have designed the following general methodology (see Figure 1): we start with the candidate term recognition process from a corpus and apply a classification method that rearranges the candidate terms according to their relevance to the task and/or domain of interest (in our case bioinformatics tools/services**)**. Term classification is based on a hybrid approach combining terms' lexical and contextual properties, represented as term profiles. Task/domain relevance is then assessed by comparing profiles of candidate terms with profiles of seed terms and ontological concepts that portray the task/domain. These steps are described below in detail.

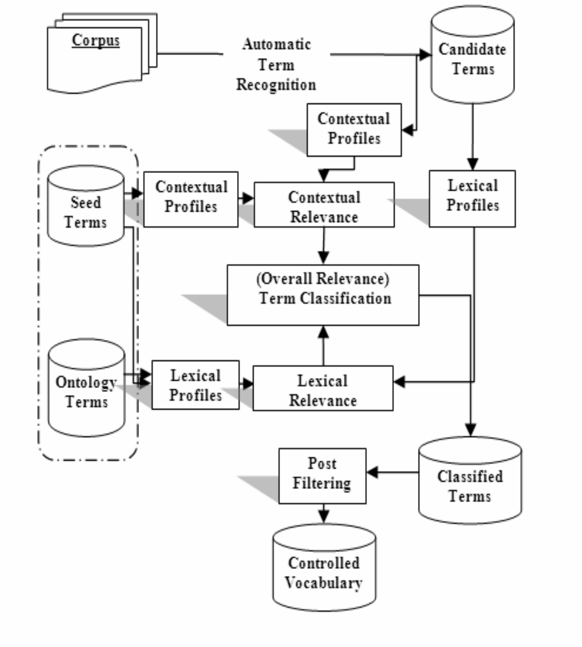

Figure 1: System architecture

2.1 Collecting candidate terms

To support the task, we have collected a corpus of full text articles from a bioinformatics journal. The corpus was processed by an Automatic Term Recognition (ATR**)** service (TerMine[7], based on the C-value method (Frantzi et al, 2000)), in order to obtain the candidate terms to be considered for the CV. The C-value method is, however, a generic ATR approach that considers only statistical information (frequency of occurrence, string nestedness, etc) and recognises terms that are relevant for the whole collection, irrespective of sub-domains/tasks of interest. Therefore, the candidate terms would typically include gene and protein names, drugs, organisms, chemical terms, various procedures, tools, etc. Our aim is to identify only terms related to bioinformatics by assessing correlation between the candidate terms and a set of pre-prepared concepts representing the task/domain of interest.

2.2 Knowledge resources

To represent the domain, we have created a knowledge base that comprises two resources: a list of seed terms and a list of ontological terms. Both resources are used to provide the lexical profile of the domain, with the seed terms also used to "illustrate" textual behaviour in docu-

[7] http://www.nactem.ac.uk/software/termine/

ments (i.e. pragmatics) of the domain terms, providing positive “use cases”. Obviously, ontological terms may not appear in the literature since they describe concepts and are used for domain modelling. For the bioinformatics CV task, the seed terms (ST) have been collected from existing Web service descriptions provided by various sources (e.g. EBI Web services) and from the relevant literature cited at the myGrid website[8]. These terms have been collected automatically using TerMine and then manually pruned on the basis of their relevance to our domain. A total of 250 terms have been identified: these are “real” terms used for service descriptions in the literature. Ontological terms (OT) are extracted from 440 concepts of the bioinformatics ontology prepared by the myGrid team. The ontology includes informatics concepts (the key concepts of data, data structures, databases and metadata); bioinformatics concepts (domain-specific data sources e.g. model organism sequencing databases, and domain-specific algorithms for searching and analysing data e.g. the sequence alignment algorithm); molecular biology concepts (higher level concepts used to describe bioinformatics data types, used as inputs and outputs in services e.g. protein sequence, nucleic acid sequence); task concepts (generic tasks a service operation can perform e.g. retrieving, displaying and aligning).

For each of the seed and ontological terms, we have generated (as explained below) lexical profiles that will be used to identify potential bioinformatics terms. For the seed terms only, we have also generated contextual profiles to provide a case-base with typical contexts in which the seed terms have appeared.

2.3 Term profiles

The main idea behind our term classification process is to measure the degree of similarity between candidate terms and the known bioinformatics terms by comparing their lexical (constituents) and contextual (pragmatics) profiles.

Lexical profiles. Each candidate term is assigned a lexical profile, represented by all possible left-linear combinations of the word-level substrings present in a term (Nenadic and Ananiadou, 2006). For example, the lexical profile of the term *protein sequence alignment* is the following set: {*protein, sequence, alignment, protein sequence, sequence alignment, protein sequence alignment*}. These profiles are then compared (as sets) to the profiles of the seed and ontological terms. The hypothesis here is that – since scientific sublanguages are characterised by words and their collocations which appear more frequently in a given domain (Kittredge, 1982) – we can use lexical correlations to suggest potential candidates.

We have employed two different approaches: comparing a candidate term profile using an “average” bioinformatics seed/ontology term (LR_1, formula (1) below) and finding the best match (LR_2, formula (2)). In both cases we use a Dice-like coefficient to measure the lexical relevance. If LP(*t*) represents the lexical profile of a term t, and LP(s_i) and LP(o_i) lexical profiles of a seed and ontological term respectively, then lexical relevance of term *t* is calculated as follows:

$$\text{LR\_1(t)} = \frac{1}{2\text{n}} \sum_{i=1}^{n} \left(\frac{2(\text{LP(t)} \cap \text{LP}_\text{i}(s_i))}{|\text{LP(t)}| + |\text{LP}_\text{i}(s_i)|} \right) + \frac{1}{2m} \sum_{j=1}^{m} \left(\frac{2(\text{LP(t)} \cap \text{LP}_\text{j}(o_j))}{|\text{LP(t)}| + |\text{LP}_\text{j}(o_j)|} \right) \quad (1)$$

$$\text{LR\_2(t)} = \max_{\substack{\text{i=1 to n} \\ \text{j=1 to m}}} \left(2 \frac{\text{LP(t)} \cap \text{LP}_\text{i}(s_i)}{|\text{LP(t)}| + |\text{LP}_\text{i}(s_i)|}, 2 \frac{\text{LP(t)} \cap \text{LP}_\text{j}(o_j)}{|\text{LP(t)}| + |\text{LP}_\text{j}(o_j)|} \right) \quad (2)$$

Here, *n* and *m* represent the total number of seed terms and ontological terms respectively. In case of LR_1, we estimate lexical relevance on the basis of its relative similarity to the whole domain, whereas, in case of LR_2, we focus on maximal similarity to a seed or ontological term.

Contextual profiles. Target terms may have no lexical resemblance to the seed/ontological terms. For example, *fisheye* is a name of a tool that cannot be identified as a relevant bioinformatics term based only on its lexical properties. We therefore consider contexts (namely sentences) in which candidate terms occur in order to profile their behaviour using co-occurring nouns and verbs, as well as lexico-syntactic patterns in which candidate terms occur. A contextual profile of each term comprises its noun sub-profile, verb sub-profile and context pattern sub-profile. Similarly, contextual profiles of the seed terms are built using the literature from which they have been extracted.

Contextual elements are identified using a POS tagger and lemmatiser (the Genia tagger[9] was used), parser (Stanford parser[10]) and the TerMine service. As a result of pre-processing and filtering non-content bearing units (including modals and adverbs), each sentence is repre-

[8] http://www.mygrid.org.uk/wiki/Mygrid/BiologicalWebServices

[9] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[10] http://nlp.stanford.edu/software/lex-parser.shtml

sented as a stream of lexico-syntactic (noun phrases, verb phrases, prepositions) and terminological units, with their relative positional information with respect to the candidate term. In our experiments, we have used two types of unit representation (see Table 1 for an example). In the first type, noun phrases and terms are represented by their class only (as NP and Term respectively), whereas verb phrases and prepositions are represented by their lemmas. We have not generalised verbs and prepositions since they are expected to carry useful information for classification of candidate terms (Spasic and Ananiadou 2004). The second type of pattern contains lemmas for all units, including NPs and terms. A pattern profile is then represented by left (LP**)** and right (RP**)** patterns, which represent units appearing on the left and right side of the candidate term in a sentence respectively.

Verb profile	*produce*
Noun profile	*Genscan, program, list, transcript*
LPs	Term, *produce,* NP, *of* *Genscan program, produce, a list, of*
RPs	*of,* NP *of, predicted transcripts*

Table 1: An example of contextual profiles of the term *nucleotide FASTA*, originated from the following sentence: *The Genscan program can produce a list of nucleotide FASTAs of predicted transcripts.* The first line in LPs and RPs rows represents the first pattern type (lemmas for verbs and prepositions only), while the other represents the second type (lemmas for all constituents).

Contextual profiles are used to measure contextual relevance (CR) of each candidate term by comparing them to the contextual profiles of the seed terms. Similarly to lexical relevance, we have used two formulae comparing a candidate term profile to the average seed term, and to the most similar one:

$$\mathrm{CRN\_1(t)} = \frac{1}{n}\sum_{i=1}^{n}\left(2\frac{\mathrm{CPN(t)} \cap \mathrm{CPN_i}(s_i)}{|\mathrm{CPN(t)}| + |\mathrm{CPN_i}(s_i)|}\right) \quad (3)$$

$$\mathrm{CRN\_2(t)} = \max_{s_i \in ST}\left(2\frac{\mathrm{CPN(t)} \cap \mathrm{CPN}\,(s_i)}{|\mathrm{CPN(t)}| + |\mathrm{CPN}\,(s_i)|}\right) \quad (4)$$

Here, CPN(*x*) represents a contextual noun profile of (a candidate or seed) term *x*. Relevance measures using verbs (CRV) and patterns (CRP) are calculated similarly. In addition to these term-term comparisons, we also consider *aggregate* contextual seed profiles comprising features (i.e. nouns, verbs, LPs, RPs) appearing in context of any seed term. Using these values, the aggregate contextual (noun) relevance is calculated as

$$\mathrm{CRN\_3(t)} = 2\frac{\mathrm{CPN(t)} \cap \mathrm{CPN\_A}(ST)}{|\mathrm{CPN(t)}| + |\mathrm{CPN\_A}(ST)|} \quad (5)$$

where CPN_A(*ST*) is the aggregate noun profile of the seed terms. Similar approaches are followed for verb and pattern profiles.

2.4 Building the controlled vocabulary

As described before, our main aim is to provide a methodology to automatically build a terminological resource containing terms that are similar lexically and pragmatically to a given set of terms from the knowledge resources. Our approach is based on combining the four types of profile similarities to estimate the overall relevance **(**OR) of a candidate term:

$$\mathrm{OR(t)} = \theta \cdot LR(t) + \alpha \cdot CRN(t) + \beta \cdot CRV(t) + \gamma \cdot CRP(t) \quad (6)$$

where LR(*t*), CRN(*t*), CRV(*t*) and CRP(*t*) represent relevance of term *t* based on lexical, contextual nouns, contextual verbs and contextual pattern profiles respectively. The parameters α, β, γ and θ can be used to assign different weights to the profiles' contributions. By applying term weighting, we can obtain a list of candidate terms with high OR, and extract/classify terms with OR above a certain threshold as relevant and consider them for the CV building. The threshold value and term post-filtering can be varied according to the user's requirement for precision and recall.

3 Experiments and Results

To assess the suggested method, we have performed three experiments. First, we have evaluated the performance (precision/recall) of term classification on a subset of documents. Then, we have evaluated the top 300 terms extracted by the system with regard to precision, and finally estimated the recall as compared to the myGrid bioinformatics ontology.

The knowledge resources used in the experiments are as follows. We used 250 seed terms and 440 ontological terms, for which lexical and contextual (ST only) profiles have been generated. The corpus from which we collected candidate terms consisted of 2120 full text open-access articles from *BMC Bioinformatics*[11] (published before March 2008). Full text is essential for this

[11] http://www.biomedcentral.com/bmcbioinformatics/

task, firstly because we expect to find many of the candidate terms in the methods section, and, secondly, as it is more likely to find detailed contexts for term classification in full text documents rather than in abstracts only. After applying the C-value method on the corpus, we have collected almost 100,000 candidate terms (see later Table 4 for detailed statistics) and generated their lexical and contextual profiles.

We used 135 additional bioinformatics terms manually extracted from the service describing literature cited on the myGrid website for tuning the system parameters (cf. formula (6)). A genetic algorithm iterative procedure given in (Spasic et al, 2004) has been performed to learn the parameters to optimise the results on the tuning terms so that the maximal number of the tuning terms ends up in the top 10% of the suggested candidate terms. We randomly varied the values of parameters through 1000 iterations, providing that $\alpha+\beta+\gamma+\theta=1$. In each optimisation cycle we have considered all individual profiles (e.g. LR_1, CRN_2, etc.) or their combinations (e.g. LR_1 & CRN_3 & CRV_2 & CRP_3) so to find the best performing values of the parameters. While the max-based lexical similarity (LR_2) was better than the average-based LR_1, there were no significant differences between various contextual formulae. Still, the best overall performance on the tuning terms was when we combined CRN_1, CRV_3, CRP_2 and LR_2 with the following parameter values: α = 0.355, β= 0.158, γ = 0.02 and θ = 0.462 (used as the default parameters further on).

Experiment 1: term classification performance. In order to estimate the precision and recall of the term classification part, we have randomly selected a subset of five documents, in which 375 terms appear (automatically recognised by TerMine). These have been manually classified by a domain expert as relevant/irrelevant. We have then evaluated the system performance (using the usual metrics for precision, recall and F-measure) on this set. The best performing individual metrics (LR_2, CRN_1, CRV_3 and CRP_2) are summarised in Table 2. Table 3 summarises the performance of three combined profiles with the best performance. The best results were obtained when CRN_1, CRV_3, CRP_2 and LR_2 were combined, with precision in the 70% range and recall in the 90% range (F-measure in the 80% range). Figure 2 summarises the results for the best performing metrics.

	LR_2	**CRN_1**	**CRV_3**	**CRP_2**
Precision	69.1	63.4	71.2	**80.6**
Recall	83.3	77.0	62.7	19.8
F-measure	75.5	69.5	66.7	31.8

Table 2: The performance of the best individual metrics on the test set (375 terms).

	Comb1	**Comb2**	**Comb3**
Precision	68.2	67.9	67.1
Recall	**92.1**	84.1	**92.1**
F-Measure	**78.4**	75.2	77.2

Table 3: The performance of combined metrics on the test set (375 terms). [Comb1 = CRN_1, CRV_3, CRP_2 and LR_2, with the default parameters; Comb2 using only CRN-1 and LR2 with α = 0.298 and θ = 0.702; Comb3 using CRN_1 and CRV-3 with β= 0.258 and θ = 0.742].

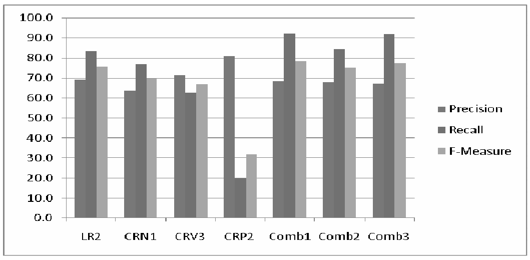

Figure 2: The performance of individual metrics

Experiment 2: the controlled vocabulary precision. Two domain experts evaluated the top 300 terms as suggested by the system. The results (see Fig. 3) have showed that the top 100 terms were highly relevant, with 93% of terms deemed suitable to make a direct entry into the bioinformatics CV. The precision for the top 200 terms fell to 83% and to 79% for the top 300 terms.

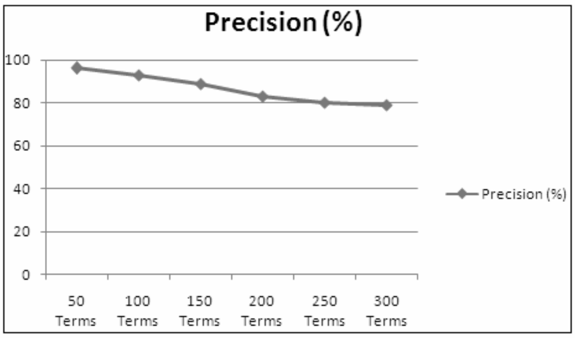

Figure 3: Precision of the top 300 CV terms

Experiment 3: "reconstructing" the myGrid bioinformatics ontology. In addition to estimating recall for the term classification task, we investigated to what extent the system could reconstruct the myGrid bioinformatics ontology. The experiments have shown that even 45 terms (10% of the myGrid ontology) appeared in the first 100 terms, totalling to 59 (13.4%) for the first 300 terms (see Figure 4). We have also found that the total of 20% of the suggested top 300 terms fully matched the corresponding myGrid concepts.

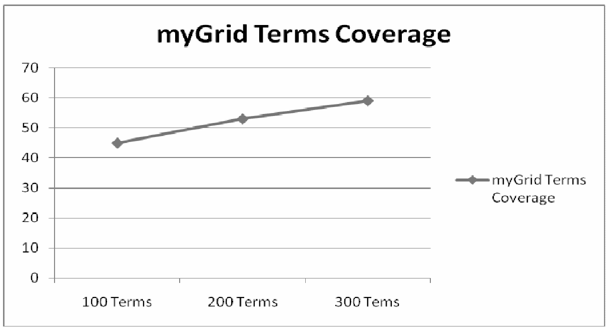

Figure 4: The number of myGrid terms recovered

4 Discussion

We have presented a generic methodology to automatically build and expand a controlled vocabulary in a domain of interest using literature mining. For this purpose, we employed a term classification approach that combines lexical and contextual properties of candidate terms and compares them to seed entities.

In the experiment 1 of the bioinformatics CV building task, the best individual performance (the best F-measure) was observed in the case of lexical relevance (see Figure 2 and Table 2). In addition to lexical properties of candidate terms (that typically give precise results but fail to identify some relevant terms), we also consider contextual profiles. Table 4 shows the number of terms being recognised using various metrics. Most of the top suggested terms made into the CV based on their lexical profiles, but there were still terms that were only contextually similar to the seed terms (e.g. terms such as *statistical approach* or *SVM classifier*; or *biological text* (e.g. as an input concept)).

Overall, in the case of lexico-syntactic patterns, the results show very good precision (see Table 2, last column). The reason is that the patterns originated from the seed terms were able to model the pragmatics of bioinformatics terms. Since the number of seed terms and their contexts were limited (250 terms and 1034 contexts), this has resulted in an accetably low recall. In the experiments we have varied the representation of patterns (considering the generic classes of neighbouring units and varying the length of the window of neighbouring words), and the best balance between precision and recall was obtained using three neighbouring units.

When lexical and contextual profiles were combined, a significant increase in recall was observed as compared to individual metrics (see Table 3 – recall of 92.1% as compared to 83.3% for lexical and 19.8% for contextual), with no significant drop in precision (if at all), resulting in the overall improved F-measure.

We have also varied the seed terms used by swapping 100 out of 250 terms with new terms collected using the same methodology from the service description literature. However, the results were similar, showing small variations of 1-2% in precision and recall.

Total number of candidate terms collected using ATR	98,986
Number of terms classified using lexical similarity	61,977
Number of terms classified using contextual nouns	84,412
Number of terms classified using contextual verbs	64,477
Number of terms classified using contextual patterns	17,638

Table 4: The number of terms suggested from the BMC corpus using different similarity metrics

The evaluation of the top 300 candidate terms revealed that there were three term types suggested. The first type relates to terms that refer to a generic concept related to bioinformatics services and can make a direct entry into the CV (direct true positives). More than half of all terms are in this category. The second category contains terms that would need slight modification before becoming part of the CV. For example, this type includes units that begin with a generic or non-specific modifier (e.g. *user friendly* in *user friendly Gpcr oligomerization knowledge base*), or wrongly identified terminological head (e.g. *compromise* in *tab delimited text file compromise*). We have applied simple rules to fix these issues, improving the number of (direct) entries by more than 11%. The third type contains names that refer to toolkits, workbench platforms, databases etc. (e.g. *protein visualization tool RASMOL*, *myGrid Taverna workbench*, etc.) They do not refer to generic concepts and thus are not direct entries to the CV, but are of interest for the service discovery process. Note that such terms were also included in the seed term set, so their contextual profiles were used as positive use cases. Overall, adding these terms improved the total precision to 79.3%.

The experiments with the myGrid ontology (experiment 3) were interesting in the sense that the suggested method was promising in both reconstructing the terms from an ontology (reasonable recall), but also in identifying new potential entries or synonyms that could be used (e.g. terms such as *life science identifier* (*LSI*), *systems biology mark up language* (*SMBL*), etc. have been suggested).

5 Related work

There have been several approaches to semi-automated building of controlled vocabularies and ontologies from literature (Grefenstette 1994). For example, Spasic et al. (2008) present a methodology for development of CVs for metabolomics. They employed an automatic term recognition method to identify candidate terms from a corpus and then filtered relevant terms on the basis of their semantic association to a set of manually chosen relevant concepts. In this case, the UMLS[12] was used as a (static) semantic model to identify properties to which target terms should conform.

Sabou et al. (2005) present an automatic method that learns domain ontologies for Web service descriptions from textual information attached to Web services. They annotated a corpus with linguistic information and then performed syntactic parsing and employed a set of syntactic patterns to identify and extract information from the corpus. The patterns are focused on domain concepts, their functionalities (verbs associated with concepts) and inter-relationships between concepts (via prepositions). This extracted information is then transformed into a structured ontology.

Automatic term classification is also related to our work, in particular for different biological entities (e.g. gene and protein mentions (Yeh et al, 2005)). The reported methodologies include keyword-driven approaches, where biomedical terms containing functional words such as *receptor*, *factor* or *radical* are used to assign term categories. These functional words may not always be discriminative, and determination of term class is not possible merely by comparing the functional words, which may lead to the ambiguity in term classification (Krauthammer and Nenadic, 2004). Statistical and machine learning approaches are also used (Collier et al, 2000; Lui and Friedman 2003). For example, an approach for disambiguation between proteins, genes, and mRNAs using different machine learning techniques (naïve Bayesian classification, decision trees and inductive learning) was reported by Hatzivassiloglou et al. (2001). They used different features for classification including words that appeared near a term, positional, morphological, distributional and shallow syntactic information about terms and reported an overall accuracy between 69.4% and 85% for a two-way classification task (gene/protein) and between 65.9% and 78.1% for a three-way classification task (gene/protein/ mRNA).

Apart from using morphological, lexical and syntactical properties of a term, key features from the context of a term occurrence can also be employed to determine the class of that term. For example, Al-Mubaid (2006) used mutual information and χ^2 as feature selection techniques to identify the best features from term contexts to build a term classifier. Similarly, Spasic et al. (2005) combined machine learning and domain knowledge (the UMLS thesaurus) to design a case-based reasoning system for term classification based on context alignment (using the edit distance similarity between syntactic and semantic constituents). In their previous work, Spasic and Ananiadou (2004) also used automatically learnt verbal preferences to support classification of biomedical terms.

6 Conclusions

Most bio text-mining approaches so far have focused on identification of biological and molecular entities from the literature to support knowledge acquisition and discovery in the life sciences (Jensen et al, 2006), with very few attempts to characterise the bioinformatics sublanguage and the terminology used to present technologies, experimental procedures and methodologies. In this paper we have focused on a controlled vocabulary that can be used to semantically annotate bioinformatics services and tools.

We have presented a term classification driven methodology to automatically build a CV for a domain represented with a set of seed terms and (optionally) a set of ontological descriptions. The methodology integrates lexical and contextual profiles of candidate terms, and compares them to the available resources. In the lexical ap-

[12] http://www.nlm.nih.gov/research/umls/

proach, we quantify the degree of sharing of constituents between candidate terms and the seed and ontology units. In the context-based profiling, we model textual behaviour of terms, using co-occurring nouns and verbs, or describing contexts using contextual patterns. While the ontological concepts are used only to capture the lexical dimension of the domain conceptual space, the seed terms are also used to describe pragmatics of the given domain through a set of “known” use cases.

The results of the methodology applied to the bioinformatics domain revealed that the approach is useful for a rapid creation of a CV. We have processed all of the BMC Bioinformatics articles, with the estimated best precision of around 70% and recall of 90%. The precision for the top 100 suggested terms was 93%.

The CV generation can be viewed as a first step towards facilitating the automation of the service description process by not only aiding in the provision of baseline terminologies, but also by providing a useful lexical resource that can then be utilised for other NLP tasks like information retrieval, named entity recognition and information extraction in the bioinformatics domain. Future work will include incremental learning of terms and identification of patterns that are relevant for service descriptions, as well as more detailed identification of roles that bioinformatics terms may have in a given context (e.g. service input, task/operation term, availability, etc.).

Acknowledgements

We are grateful to the myGrid team, in particular to Franck Tanoh and James Eales (the University of Manchester) for the evaluation of the results.

References

Al-Mubaid H. (2006). "Context-Based Technique for Biomedical Term Classification”. *Proc*. of the 2006 IEEE Congress on Evolutionary Computation, CEC-2006, pp.5726-5733

Cannata N, E Merelli and RB Altman (2005). "Time to Organize the Bioinformatics Resourceome". PLoS Comput Biol 1(7): e76.

Collier N, C. Nobata and J. Tsujii (2000). "Extracting the Names of Genes and Gene Products with a Hidden Markov Model" Proc. of *COLING 2000*, pp. 201-207.

Frantzi K, S. Ananiadou and H. Mima (2000). "Automatic Recognition of Multi-Word Terms: The C-value/ NC-value method." International Journal on Digital Libraries 3(2): 115-130.

Grefenstette G (1994). "Exploration in Automatic Thesaurus Discovery". Springer, Vol. 278, 1994.

Hatzivassiloglou V, P.A.. Duboue and A. Rzhetsky (2001). "Disambiguating Proteins, Genes, and RNA in Text: A Machine Language Approach." Bioinformatics 17(1): 97-106.

Jensen JL, J. Saric and P. Bork (2006). "Literature mining for the biologist: from information retrieval to biological discovery". Nature Reviews Genetics

Kageura K and B. Umino (1996). "Methods of automatic term recognition: a review". Terminology 1996; 3:259–289.

Kittredge R (1982). "Sublanguages". Comput Linguist 8(2): 79-84.

Krauthammer M. and G. Nenadic (2004). "Term identification in the biomedical literature". Journal of Biomedical Informatics 37(6): 512-526.

Liu H. and C. Friedman (2003). "Mining Terminological Knowledge in Large Biomedical Corpora". Proc. of 8th PSB, p. 415-426

Nenadic G. and S. Ananiadou (2006). "Mining Semantically Related Terms from Biomedical Literature" ACM Transactions on ALIP 5(1): 22-43.

Oinn T, P. Li., DB Kell, C. Goble, A. Goderis, M. Greenwood, D. Hull, R. Stevens, D. Turi and J. Zhao (2007). "Taverna/myGrid: aligning a workflow system with the life sciences community." In Dennis et al. (Eds), Workflows for e-Science: scientific workflows for Grids, Springer, 300-319.

Sabou M, C. Wroe, C. Goble and G. Mishne (2005). "Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics." Proc of 14th Int. Conf. on WWW, p. 190-198

Spasic I. and S. Ananiadou (2004). "Using automatically learnt verb selectional preferences for classification of biomedical terms". Journal of Biomedical Informatics (Named Entity Recognition in Biomedicine), Vol. 37, No. 6, pp. 483-497

Spasic I., G. Nenadic and S. Ananiadou (2004). “Learning to Classify Biomedical Terms through Literature Mining and Genetic Algorithms”. Proc. of *IDEAL 2004*, pp: 345-351.

Spasic I, S. Ananiadou and J. Tsujii (2005). "MaSTerClass: a case-based reasoning system for the classification of biomedical terms." Bioinformatics 21(11): 2748-2758.

Spasic I, D. Schober, SA Sansone, D Rebholz-Schuhmann, D Kell and N Paton (2008). "Facilitating the development of controlled vocabularies for metabolomics technologies with text mining." BMC Bioinformatics 9(Suppl 5):S5

Wolstencroft K, P. Alper, D. Hull, C. Wroe, P.W. Lord, R.D. Stevens and C. Goble (2007). "The myGrid Ontology: Bioinformatics Service Discovery". International Journal of Bioinformatics Research and Applications, 3(3):326 – 340, 2007.

Yan J (2002). "GeneSOM—self-organizing map package", Version 0.2-5, 2002. Available from http://cran.R-project.org.

Yeh A, A. Morgan, M. Colosimo and L. Hirschman (2005). "BioCreAtIvE Task 1A: gene mention finding evaluation". BMC Bioinformatics 2005; 6(Supp1):S2.