Mining Semantic Descriptions of Bioinformatics Web Resources from the Literature

Hammad Afzal, Robert Stevens, and Goran Nenadic

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

Abstract. A number of projects (myGrid, BioMOBY, etc.) have recently been initiated in order to organise emerging bioinformatics Web Services and provide their semantic descriptions. They typically rely on manual curation efforts. In this paper we focus on a semi-automated approach to mine semantic descriptions from the bioinformatics literature. The method combines terminological processing and dependency parsing of journal articles, and applies information extraction techniques to profile Web services using informative textual passages, related ontological annotations and service descriptors. Service descriptors are terminological phrases reflecting related concepts (e.g. tasks, approaches, data) and/or specific roles (e.g. input/output parameters, etc.) of the associated resource classes (e.g. algorithms, databases, etc.). They can be used to facilitate subsequent manual description of services, but also for providing a semantic synopsis of a service that can be used to locate related services. We present a case-study involving full text articles from the BMC Bioinformatics journal. We illustrate the potential of natural language processing not only for mining descriptions of known services, but also for discovering new services that have been described in the literature.

1 Introduction

The bioinformatics domain has recently witnessed a number of tools and data sources available on the Web that can be used to retrieve or carry out on-line data analysis. For example, as indicated by the BioCatalogue[1] project, there are around 200 Web servers that provide Web Service interfaces that are becoming both an essential and a critical part of bioinformatics research. These resources need to be organised and their functionalities semantically described to make them accessible by both bioinformaticians and search/discovery engines. A number of projects[2] (e.g. myGrid, BioMOBY, BioCatalogue, myExperiment etc.) have been initiated to comprehensively catalogue these resources and provide their semantic descriptions (see Table 1 for an example). Most of the cataloguing frameworks, however, rely on manual annotation that has resulted in a backlog of non-described services, reducing the chance of their discovery and use in the community [1]. In order to deal with the huge number of resources

[1] http://www.biocatalogue.org

[2] For details about these projects see: http://www.mygrid.org.uk/, http://www.biomoby.org/, and http://www.myexperiment.org/

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 535–549, 2009.

Table 1. A partial service description of service Emma as described for Feta [2]. The input and output of operations are specified in terms of parameters that have name, description and semantic type. Operation tasks and semantic types of parameters are linked to the myGrid ontology [3].

Service name:		Emma	
Description:		Performs a multiple alignment of nucleic acid or protein sequences using ClustalW program.	
Operation	**Name:**	Emma	
	Description:	Performs a multiple alignment of nucleic acid or protein sequences using ClustalW program.	
	Task	www.mygrid.org.uk/ontology/**multiple_local_aligning**	
	Input	**Parameter**	Name: **sequence_usa**
			Description: The Uniform Sequence Address, or USA, is a standard way of specifying a sequence to be read into a program in EMBOSS. …
			SemanticType: www.mygrid.org.uk/mygrid-moby-service#**simpleParameter**
	Output	**Parameter**	Name: **outseq**
			Description: Returns a multiple sequence alignment report
			SemanticType: http://www.mygrid.org.uk/ontology# **multiple_sequence_alignment_report**

(tools, databases, Web services etc.) that need semantic description, (semi)automated approaches are needed.

A number of resources (along with their typical use-cases) have been described and presented in various textual documents (scientific articles, blogs, documentation, user manuals, etc.). In this paper we focus on the extraction of functional descriptions of bioinformatics tools and Web services from scientific, full-text articles. In our approach, semantic descriptions include informative textual passages, related ontological annotations and service descriptors. Service descriptors are terminological phrases reflecting related concepts (e.g. tasks, roles, approaches) typically used with specific resource classes (e.g. algorithms, databases, etc.). These descriptions can be used not only to facilitate subsequent manual description of services, but also for providing a semantic synopsis of a service that can be used to locate related services. We present a case-study involving a subset of full text articles from the BMC Bioinformatics journal, and discuss the manually evaluated results.

2 Background and Related Work

Automatic semantic annotation of Web services is a relatively new area, involving description of functionality, input and output parameters, etc. A number of approaches have been based on invoking services automatically, or on using corresponding Web

Service Definition Language (WSDL) files and existing descriptions of similar services or workflows. For example, Carman and Knoblock presented an automatic approach to learn definitions and semantic descriptions of online information resources by invoking them and comparing the output they produce with that of known sources of information [4]. On the basis of this comparison, they used the metadata associated with known sources to add annotations to unknown sources. The method was evaluated on 25 services from five domains (geospatial, financial, weather, hotels and cars) – the precision ranged between 56% and 91%. Lerman et al. presented a meta-data based classification algorithm that used WSDL files to perform semantic labelling of inputs and outputs of Web services [5]. The semantic types were organised in the form of a domain ontology. The method was evaluated on 313 WSDL files from different domains, with an overall F-measure of approximately 80%. Previously, Hess and Kushmerick annotated Web services using machine learning to classify metadata used to describe those services [6]. Information about the services given in WSDL files was used in this work. Finally, Belhajjame et al. performed automatic annotation of parameters using workflow annotations, by inferring from their links to other (annotated) operation parameters within existing workflows [7].

These methods rely on existing annotations and classification methods in order to infer descriptions of new Web services. In general, the applied techniques demonstrated reasonable performance only in cases of highly focused domains with a large number of training examples (cf. [4, 6, 7]). Finally, neither relies on textual sources (journals articles, application notes etc).

Our work uses techniques from natural language processing, in particular terminological processing, phrase structure identification (e.g. noun and verb phrases) and predicate-argument structures (PAS) to extract service descriptions. For example, a PAS can comprise a verb along with its arguments, e.g. subject and object(s). The Stanford parser[3], for instance, identifies dependencies in the form *dep(abc, xyz)* where *dep* is the dependency type (e.g. nominal subject (nsub), direct object (dobj), etc.), and *abc* and *xyz* are arguments. For example, dependency statement *nsubj (applied, BLAST)* denotes *BLAST* as the nominal subject of *applied*. Wattarujeekrit et al. argued the significance of using PAS-based extraction compared to regular expressions applied on shallow parsed sentences, in particular for event extraction in the biomedical domain [8]. Similarly, Tateisi et al. demonstrated the need and benefits of recognising predicate-argument structures for improving the performance of information extraction systems in biology [9]. Our method also uses predicate-argument structures, but additionally incorporates the phrase-structure recognition and intensive terminological processing in the process of service profiling.

3 Methodology

Our method combines terminological processing and dependency parsing of documents, and applies information extraction techniques to profile Web tools and services. It is focused around the concepts of (1) *semantic classes* (SC) associated with

[3] More details on the Stanford parser: http://nlp.stanford.edu/software/lex-parser.shtml; a full list of typed dependencies: http://nlp.stanford.edu/software/dependencies_manual.pdf

bioinformatics resources of interest (e.g. algorithms, applications, etc.), and (2) *semantic descriptors* that represent semantic roles of related SC instances. Descriptors are used to profile a given resource and/or to link it to a domain ontology (e.g. frequent descriptors are *gene expression*, *phylogenetic tree*, *microarray experiment*, *hierarchical clustering*, *amino acid sequences*, *motif*, etc.).

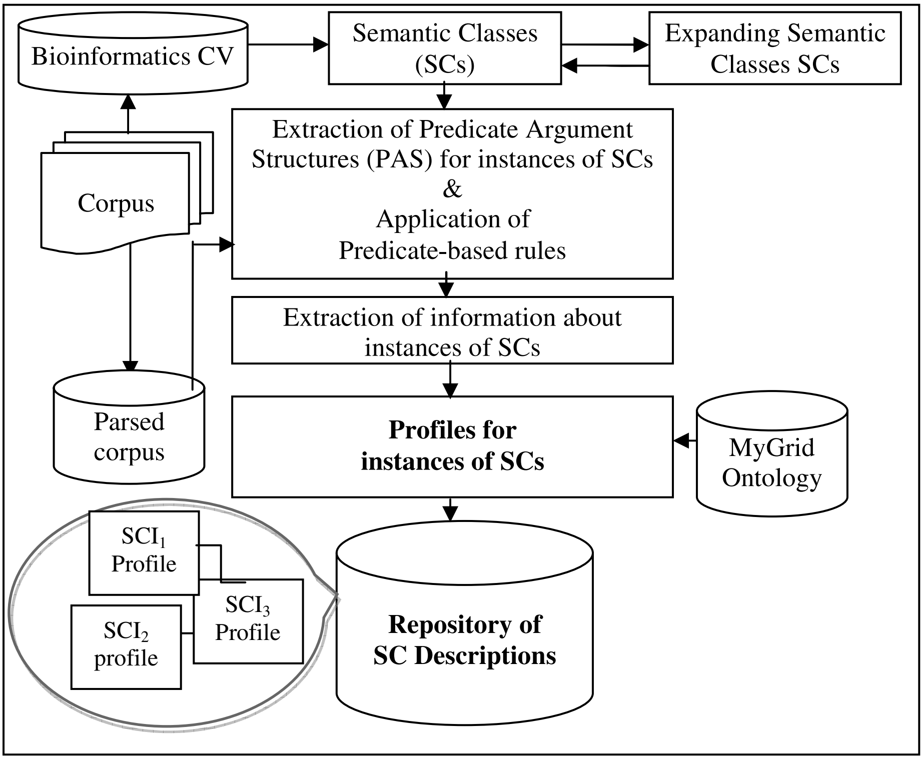

Bioinformatics CV
Semantic Classes (SCs)
Expanding Semantic Classes SCs
Corpus
Extraction of Predicate Argument Structures (PAS) for instances of SCs & Application of Predicate-based rules
Extraction of information about instances of SCs
Parsed corpus
Profiles for instances of SCs
MyGrid Ontology
SCI1 Profile
SCI3 Profile
SCI2 profile
Repository of SC Descriptions

Fig. 1. System Architecture

The method has 3 steps: in the first step, the instances of the semantic classes are identified in the literature in sentences that potentially contain their descriptions. In the second step, semantic descriptors are extracted from the sentences using string matching and a rule based approach, and are potentially mapped to a domain ontology that describes service functionalities. Finally, in the third step, the most relevant sentences are selected to complete the semantic profile of a service. All semantic profiles are collected into a repository that eventually can be used by human curators to provide annotations. The system architecture is presented in Figure 1.

3.1 Semantic Classes

Semantic classes represent the major concepts used in the description of bioinformatics services and tools. The instances of these classes are both extraction targets and can be the key components of semantic descriptions of other services. These concepts were engineered from the myGrid bioinformatics ontology, and include *algorithm, application, data*, *data resource* and *task* concepts (see example instances in Table 2). As opposed to other classes whose instances are expressed using noun phrases, tasks are typically represented by verbs or verbal phrases portraying functions related to other classes. We therefore followed two methods to identify SCs: for the first four classes we compiled dictionaries with instances, while for the task class we used information extraction rules (see Section 3.3). In the following text we refer to the non-task classes as SCs for simplicity, while tasks will be considered separately.

Table 2. Examples of semantic classes and their instances

Semantic class	Example instances
Algorithm	SigCalc *algorithm*, CHAOS local *alignment*, SNP *analysis*, KEGG Genome-based *approach*, GeneMark *method*, K-fold cross validation *procedure*
Application	PreBIND Searcher *program*, Apollo2Go *Web Service*, FLIP *application*, Apollo Genome Annotation curation *tool*, GenePix *software*, Pegasys *system*
Data	GeneBank *record*, Genome Microbial CoDing *sequences*, Drug Data *report*
Data resource	PIR Protein Information *Resource*, BIND *database*, TIGR *dataset*, BioMOBY Public Code *repository*

We used iterative terminological processing to collect and classify instances belonging to the SCs. In the first step, we extracted a number of instances by identifying typical terminological heads of a given SC (e.g. Smith-Waterman *algorithm* is an algorithm). The set of key terminological heads was obtained by exploring the first level of the myGrid ontology. We then examined a list of 100,000 bioinformatics terms previously collected in the process of building a controlled vocabulary[4] (CV) of bioinformatics [10]. We selected terms and associated them with the corresponding SC if their terminological head belonged to the following set:

Heads(application) = {*application*, *tool*, *service*, *software*, *system*, *program*}
Heads(algorithm) = {*algorithm*, *method*, *approach*, *procedure*, *analysis*, *alignment*}
Heads(data) = {*record*, *report*, *sequence*}
Heads(data resource) = {*resource*, *database*, *dataset*, *repository*}

In the second step, we collected instances that did not have these discriminative terminological heads, but appeared in specific contexts with other instances from a given SC. For example, we collected all instances that co-appeared in coordination and apposition expressions with other instances using patterns similar to those used by Hearst and Shütze [11].

In the final step, we identified instances that demonstrated behaviour similar to the instances identified in the previous step. Similar behaviour was modelled by analysing dependency structures that involved nsubj, dobj and nsubjpass (nominal subject in

[4] Available at: http://gnode1.mib.man.ac.uk/bioinf/CV

passive) dependencies. The motivation here was to firstly identify verbs with which known SC instances appeared in the capacity of nominal subject, direct object and passive nominal subject, and then – using these verbs – extract terms appearing in a similar context. For each SC, a list of frequent and discriminative verbs was automatically compiled (different frequency thresholds were set for different SCs). Newly identified instances were grouped with the corresponding SCs based on a majority vote.

3.2 Semantic Service Descriptors

Service descriptors are terminological phrases used in existing descriptions to refer to the related concepts and specific roles (e.g. input/output parameters, etc.) corresponding to the semantic classes. The reason behind collecting these terms was to identify terms that have been previously used in the description of Web services and tools. The hypothesis was that their presence in the "free" literature would indicate useful sentences for annotating SC instances.

We have used two sources to build a dictionary of service descriptors. The first resource was the list of terms collected from the bioinformatics ontology used in the MyGrid project [3]. This list contains 443 terms describing concepts in *informatics* (the key concepts of data, data structures, databases and metadata); *bioinformatics* (domain-specific data sources e.g. model organism sequencing databases, and domain-specific algorithms for searching and analyzing data e.g. the sequence alignment algorithm); *molecular biology* (higher level concepts used to describe bioinformatics data types, used as inputs and outputs in services e.g. protein sequence, nucleic acid sequence); and *tasks* (generic tasks a service operation can perform e.g. retrieving, displaying and aligning). The second resource contained the terms (recognised by the TerMine[5] service) and frequent noun phrases obtained from the existing descriptions of bioinformatics Web services available in Feta[6] and other WS providers' websites[7].

3.3 Extraction of Functional Service Descriptions

Our methodology for the extraction of functional descriptions (tasks) was predicate-centric, i.e. organised around verbs that appeared with the instances of the SCs. We applied a set of extraction patterns (see Appendix 1) on sentences that were parsed by the Stanford parser to obtain predicate-argument structures and phrase structures (separately). Here, our hypothesis was that by keeping one of the arguments to be an SC instance, the verb and other argument would provide a clue about the functionality of that instance.

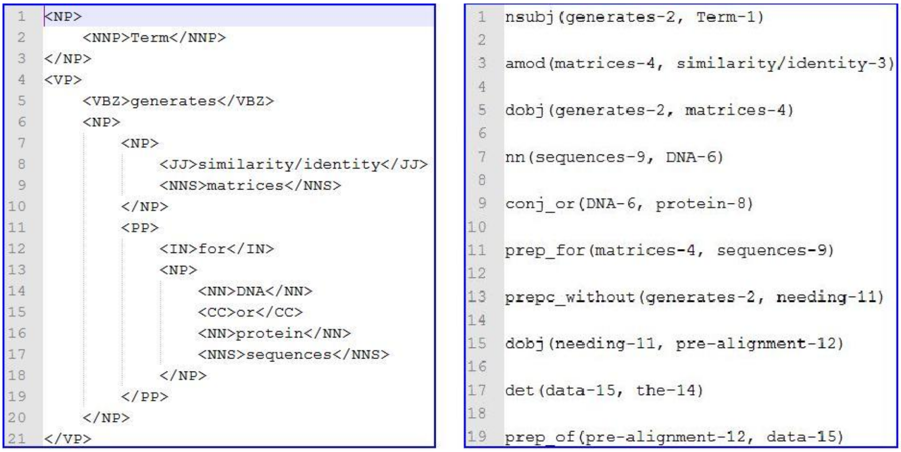
We employed two parsing methodologies to extract the arguments and potential functions: one by integrating the dependency parse and phrase-structure parse, and the other using only the phrase structure parse. In the first case, we derived one of the arguments (either Arg-1 or Arg-2) using *nsubj, dobj or nsubjpass* dependencies.

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

[6] http://www.mygrid.org.uk/feta/mygrid/descriptions/

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

These SC instances were the recognised terms that have been pre-marked in sentences. Finding an exact mention (i.e. scope) of the other argument, however, was not as straightforward. In addition, we were also interested in the sub-clause appearing with the PAS which might contain useful descriptive information for the SC instance. For this purpose, we integrated the phrase-structure parse with the associated verb phrase containing the predicate and sometimes, sub-clause information as well.



```
<NP>
    <NNP>Term</NNP>
</NP>
<VP>
    <VBZ>generates</VBZ>
    <NP>
        <NP>
            <JJ>similarity/identity</JJ>
            <NNS>matrices</NNS>
        </NP>
        <PP>
            <IN>for</IN>
            <NP>
                <NN>DNA</NN>
                <CC>or</CC>
                <NN>protein</NN>
                <NNS>sequences</NNS>
            </NP>
        </PP>
    </NP>
</VP>
```

```
nsubj(generates-2, Term-1)

amod(matrices-4, similarity/identity-3)

dobj(generates-2, matrices-4)

nn(sequences-9, DNA-6)

conj_or(DNA-6, protein-8)

prep_for(matrices-4, sequences-9)

prepc_without(generates-2, needing-11)

dobj(needing-11, pre-alignment-12)

det(data-15, the-14)

prep_of(pre-alignment-12, data-15)
```

Fig. 2. Integrating parsing results: part of the phrase-structure parsed tree (left) and dependency parsed tree (right) for the sentence "Matrix Global Alignment Tool MatGAT generates similarity/identity matrices for DNA or protein sequences"

Figure 2 provides an example and illustrates the approach. Here, the term *Matrix Global Alignment Tool MatGAT* was recognised as a tool instance, and was replaced by a Term in the original sentence that was then dependency parsed. The dependency parse identified that the term was a subject (nsubj) and *matrices* was a direct object (dobj) of the verb *generates*. In order to get the full dobj argument, we used the phrase structure parse to identify the full verb phrase (<VP>) involving the verb in question (*generate*), and then integrated the two. In this case, we derived that *Matrix Global Alignment Tool MatGAT* uses *DNA or protein sequence* to generate *similarity/identity matrices*.

We compiled a list of 108 verbs appearing with the SC instances along with their PASs as described above. The verbs providing similar semantic information or indicating the presence of similar functional "content" were manually grouped in clusters (e.g. the verbs related to inputs and/or outputs; see Table 3 for examples). For each of the groups, we devised a predicate-centric set of patterns for the extraction of functional service descriptions. For example, *App % accepts % Input* specifies a common pattern that provides information about input for a given application. The patterns created for the most frequent verbs are given in Appendix 1.

Table 3. Examples of verbs grouped according to the type of functional "content" they typically provide

Verbs	**Typical functional content**
applied, *access* [*to*], *achieve*, *align*, *allow*, *based*, *developed*, *implemented*, *present*, *provide*, *used*	general function description, task specification
accept, *applied*, *create*, *provide*, *query*, *retrieve*, *starts with*, *take* [*input*], *used*	inputs, outputs
outperform, *perform* [*better/worse*], *compare*	comparison, similar tasks
implemented [*in/using*]	technique, programming language
contained, *constructed*, *generated*	composition, subtasks

3.4 Semantic Profiling of Bioinformatics Resources

The ultimate aim was to provide a separate profile for each instance of semantic classes of interest, resulting in a repository of descriptions that could be used for annotation. Each profile contained two conceptual parts: related key-terms and textual descriptions. The related key-terms included: associated myGrid ontology concepts, service descriptors and other related terms (including other SC instances and input/output parameters), all identified by co-occurrence within the same sentence with a given SC instance. Textual descriptions contained the extracted structured information presenting functional content (obtained by applying the rules on predicate-argument structures) and associated sentences. Figure 3 gives a summary of semantic profile of an instance, while Appendix 2 provides a complete example (more examples are available in the supplementary materials[8]).

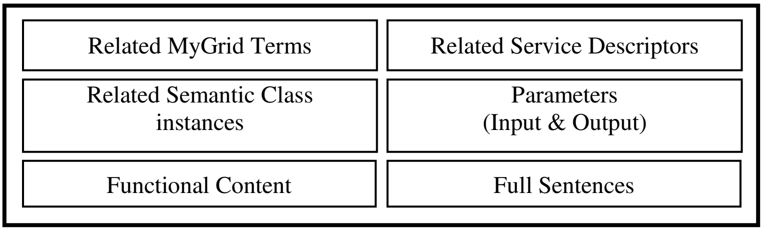

Related MyGrid Terms
Related Service Descriptors
Related Semantic Class instances
Parameters (Input & Output)
Functional Content
Full Sentences

Fig. 3. Semantic profile of a semantic class instance

4 Experiments and Results

We performed two experiments to evaluate the methodology presented above. We first examined to what extent we could reconstruct the existing bioinformatics service descriptions. In the second experiment, we examined how useful the extracted semantic service profiles were for the manual curation process. Both experiments have been done in the context of the MyGrid project [12].

[8] Available at: http://gnode1.mib.man.ac.uk/bioinf/descriptions

The semantic descriptors were collected from 471 descriptions from the Feta repository and 450 descriptions retrieved from various service providing websites. These descriptors were collected by applying automatic term recognition (TerMine was used) to these descriptions. Literature data have been compiled from 2120 full text open-access articles comprising the entire publication output of BMC Bioinformatics published before March 2008. It was obvious that this task requires full-text articles as resource descriptions are not likely to appear in abstracts (unless a publication introduces a resource).

Table 4 gives the number of service instance mentions that have been identified in the corpus using terminological head comparisons and coordination co-occurrences (Section 3.1). Table 5 presents the number of functional descriptions collected for each of the four semantic classes (Section 3.3). For each service instance, a complete profile was generated (Figure 3).

Table 4. The number of instances of the semantic class in the BMC Bioinformatics corpus

Semantic Class	# of instances identified using terminological head comparison	# of instances identified using coordination co-occurrences	Total # of instances
Algorithm	5658	64	5722
Application	1862	43	1905
Data	1424	18	1442
Data Resource	2307	34	2341

Table 5. The total number of descriptions compiled for each SC

	Algorithm	Application	Data	Data resource
General description, function	1222	316	468	226
Parameters (input/output)	122	50	90	38
Comparison (similar instances)	229	37	149	35
Implementation environment	10	5	1	3

The extracted description profiles were evaluated manually by a MyGrid bioinformatics service curator, who evaluated the following three components: the quality of the associated myGrid ontological terms, the quality of service descriptors and the quality of textual descriptions. The descriptions were assessed by their capability to be used for semantic description of a given bioinformatics service. Each component was scored as follows:

- 0: completely irrelevant (e.g. sentence *The HeatMapper tool has already proven to be very useful in several studies* is irrelevant for semantic description of HeatMapper).

- 1: partially useful description (e.g. sentence *To compare Kalign to other MSA programs, the following test sets were used* is partially useful as it specified that Kalign was a multiple-sequence alignment algorithm).
- 2: contains information useful for annotation (e.g. sentence *To add a new species to the COG system, the annotated protein sequences from the respective genome were compared to the proteins in the COG database by using the BLAST program and assigned to pre-existing COGs by using the COGNITOR program* explains the functionality of COGNITOR).

In the first experiment, we randomly selected five well-known bioinformatics services that were already manually curated and then evaluated their generated profiles against the existing "gold standard" descriptions. As an example, consider a manual description of the ClustalW service[9]:

> *ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments.*

The following description was extracted from the literature (for full results see the supplementary materials):

- **MyGrid Terms:** *phylogenetic tree*, *Clustalw format*;
- **service descriptors:** *ClustalW program, DNA sequence, phylogenetic tree, protein sequence, Phylogenetic Analysis, Classification tree, homology, …*
- **descriptive sentences** (selection)**:** (1) *Finally, a phylogenetic tree for the SARS-CoV isolates has been produced using the CLUSTALW program, showing high compatibility with former qualitative classification.* (2) *We also used the CLUSTALW program for multialignment as a control process, as well as for phylogenetic investigations.*

Overall (for the five existing services examined), the extracted sentences were deemed as fully useful (all textual service descriptions were scored as 2). While, semantic descriptors were considered as more than partially useful (the average score of 1.33), the extracted myGrid ontology terms were only partially useful (the average score was 1) – the main reason was that ontological terms did not appear frequently in the corpus, and thus were too sparse to be useful (see also discussion).

In the second experiment we considered five randomly chosen bioinformatics services that have not yet been manually annotated. For these services there were generally fewer documents containing mentions of the service. Still, the results were only slightly worse: the textual descriptions were deemed highly useful (the average score of 1.67), while semantic descriptors were uniformly given a score of 1. Interestingly, there were no MyGrid terms extracted for any of the services examined.

[9] This description is from the EBI's repository.

5 Discussion

The main aim of our work was to provide a technology to facilitate semantic description of bioinformatics services for both service curators and for automated annotation understandable in the context of the Semantic Web. The results of the application of our methodology suggested that text mining techniques could produce accurate and efficient descriptions of bioinformatics services. The quality of the extracted data was measured from the curator's perspective only (i.e. how useful extracted descriptions are for their task). As expected, sentences were most useful as these provided a broader context for the curator to annotate a service. Key service descriptors were only partially useful for curation, but – on the other hand – these are more likely to be appropriate for providing a semantic synopsis of a service that could be used to represent the service in the Semantic Web context and/or locate related services automatically (by a search agent). Although there are limitations for the accurate capturing of predicate-argument structures, translating an extracted PAS into a related Resource Description Framework (RDF) statement would make the description a direct component for the Semantic Web representation of services.

As expected, there were obvious limitations in retrieving related ontological terms from the literature. As previously shown for other ontologies, conceptual descriptions are not likely to appear in text in the form in which they appear in an ontology [13]. Therefore, additional string comparison methods (including similarity distance, n-grams, variation, etc.) would need to be employed to improve the coverage [14].

Despite the number of mentions identified, the recognition of tool and service instances by using only lexical and syntactical rules proved to be difficult. As we used terminological heads to identify instances, in some cases references to generic classes were also selected (e.g. *clustering algorithm*). Therefore, context-driven rule based patterns (including PAS) would be needed to recognise them more accurately and differentiate between mentions of generic concepts and specific instances.

Phrase structure parses enabled us to extract the information in a more descriptive form rather than just by key entities. This information could be used to describe specific aspects of services/tools and thus have a potential to bridge the gap between co-occurrence based service descriptors and full sentence extraction. Still, providing and integrating phrase structures with dependency parsing was challenging, since there were numerous verbs used to communicate similar information and thus this step was not as productive as expected. In addition, some of the semantic classes were similar to each other in their "textual" behaviour (e.g. both Algorithms and Applications perform a specific functionality on a particular input, generating a specific output), and therefore the extraction patterns would need to take into account both the semantic classes and verbs involved, as well as their specific roles for the given class.

6 Conclusion

In this paper we presented a literature mining approach to automatically extract information that semantically described bioinformatics Web services and tools. The work had two aims: reducing the amount of effort invested by domain experts in manual curation, and providing semantic synopses of services that could be used in the context of the Semantic Web.

The suggested methodology was based on the concept of major semantic classes associated to bioinformatics resources (e.g. applications, algorithms, data, data resources) that were compiled from the myGrid bioinformatics ontology. The instances of these classes were collected from the bioinformatics literature along with related sentences. Sentence filtering was followed by the extraction of semantic descriptors that referred to frequent semantic roles of tools and services in a particular context. The sentences were also dependency parsed and integrated with phrase structures separately recognised. The integrated dependency structures were used to identify specific functional content. A semantic profile of each service/tool was generated by combining descriptive sentences, semantic descriptors, links to other related tools, and semantic labels linked to the myGrid ontology. The presented case-study involving a subset of full text articles from BMC Bioinformatics illustrated the potential of natural language processing not only for mining descriptions of known services, but also for discovering and describing new services that have been mentioned in the literature.

To the best of our knowledge, this is the first attempt to mine semantic service description from the domain literature. Although the initial results are promising, there is room for improvements which would include a context-based recognition approach to service/tools mentions, improving specificity and coverage of PAS-based information extraction, and inferring semantic descriptions from textually related services. Finally, the utility of the proposed approach will be further tested within the BioCatalogue project, which is a registry that provides documentation, location and semantic annotations for Web Services for Life Sciences.

Acknowledgements

We are grateful to Franck Tanoh for the evaluation of the extracted service descriptions. This work was partially supported by the UK Biotechnology and Biological Science Research Council (BBSRC) via the projects “Mining Term Associations from Literature to Support Knowledge Discovery in Biology” and “BioCatalogue”. We are thankful to the National University of Sciences & Technology Pakistan’s Higher Education Program for additional support.

References

1. Cannata, N., Merelli, E., Altman, R.B.: Time to Organize the Bioinformatics Resourceome. PLoS Computational Biology 1, e76 (2005)
2. Lord, P., Alper, P., Wroe, C., Goble, C.: Feta: A Light-Weight Architecture for User Oriented Semantic Service Discovery. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 17–31. Springer, Heidelberg (2005)
3. Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P.W., Stevens, R.D., Goble, C.A.: The myGrid Ontology: Bioinformatics Service Discovery. International Journal of Bioinformatics Research and Applications 3, 326–340 (2007)

4. Carman, M.J., Knoblock, C.A.: Learning Semantic Descriptions of Web Information Sources. In: International Joint Conferences on Artificial Intelligence, Hyderabad, pp. 2695–2701 (2006)
5. Lerman, K., Plangrasopchok, A., Knoblock, C.A.: Automatically Labeling the Inputs and Outputs of Web Services. In: Proc. of AAAI 2006, Boston, MA, USA, pp. 149–181 (2006)
6. Hess, A., Kushmerick, N.: Learning to Attach Semantic Metadata to Web Services. In: Proc. 2nd International Semantic Web Conference, Sanibel Island, Florida, USA (2003)
7. Belhajjame, K., Embury, S.M., Paton, N.W., Stevens, R., Goble, C.A.: Automatic Annotation of Web Services based on Workflow Definitions. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 116–129. Springer, Heidelberg (2006)
8. Wattarujeekrit, T., Shah, P., Collier, N.: PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 5, 155 (2004)
9. Tateisi, Y., Ohta, T., Tsujii, J.: Annotation of Predicate-argument Structure on Molecular Biology Text. In: Workshop on the 1st International Joint Conference on Natural Language Processing (IJCNLP 2004) (2004)
10. Afzal, H., Stevens, R., Nenadic, G.: Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary. In: Proc. of the Third International Symposium on Semantic Mining in Biomedicine, Turku, Finland, pp. 5–12 (2008)
11. Hearst, M.A., Schutze, H.: Customizing a lexicon to better suit a computational task. In: Corpus processing for lexical acquisition, pp. 77–96. MIT Press, Cambridge (1996)
12. Oinn, T., Li, P., Kell, D.B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turi, D., Zhao, J.: Taverna/myGrid: aligning a workflow system with the life sciences community. In: Dennis, B., Ian, G., Taylor, J., Deelman, E., Shields, M. (eds.) Workflows for e-Science: scientific workflows for Grids, pp. 300–319. Springer, Guildford (2007)
13. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics 6, 239–251 (2005)
14. Rebholz-Schuhmann, D., Arregui, M., Gaudan, M., Kirsch, H., Jimeno, A.: Text processing through Web services: Calling Whatizit. Bioinformatics 24(2), 296–298 (2007)

Appendix 1: A Collection of Predicates, Patterns and Their Usage for Information Extraction

In these patterns, <SC_I> represents an instance of any SC, whereas <Algo>, <Data>, <App>, <Data Resource> depict instances of the corresponding SCs.

Predicate	**Pattern**
Accept	<App/Algo> % accepts % <Input>
Access	<App> provides % access to % <Data Resource>
achieve	<App/Algo> % achieves % <Performance-quality>
Align	<App/Algo> % aligns the % <Data>.
Allow	<SC_I> allows user to <Description>
Applicable	<App/Algo> % are/is applicable to % <Problem/Input>
Applied	< App >% applied % <Algo> % to <NP: Input(s)>/ <VP: Function>
Available	< App > is available at/on <Forum/Website>
Based	<Algo > is based on <Concept>
Called	Gives specific name of a generalized concept
Compare	<SC_I>compares<SC_I>
Constructed	<Data/Data Resource> constructed from/using <composition>
Constructed	<Data/Data Resource/Output> constructed by/using < App/Algo>
Contain	<Database > contains [information] <composition>
Create	<SC_I> used <SC_I> to create <Output>
Developed	< App /Algo> developed <Algo> that/which/to/for <Function>
Generated	<Data/Data Resource/Output> generated by < App /Algo>
Generated	< Data/Data Resource/Output> generated using <App/Algo>
Implemented	<App> implemented <Algo> to <Function>
Implemented	<Algo> implemented using/in/as <tech>/<language>/<part-of>
Outperform	< App/Algo> outperforms <App/Algo>
Outperform	<SC_I> outperforms <SC_I>
Perform	<SC_I> performs better than < SC_I>
Present	present <Algo> that/which<Function>
Querie (s/d)	<SC_I> queries/queries <Data Resource (Database)>
Retrieve	<SC_I> retrieves/retrieved <Output/Result>
Take	<App/Algo> takes input <Input>
Used	<SC_I> % is used to % <Function>

Appendix 2: Extracted Service Descriptions for the GeneClass Algorithm (Example)

Related service descriptors

Descriptor	Frequency of co-occurrence
GeneClass algorithm	5
Motif data	4
Reliable predictive model	2
genome-wide protein-DNA binding data	2
Differential gene expression	3
Transcriptional gene regulation	2

Descriptive sentences

1. *We also show how to incorporate genome-wide protein-DNA binding data from ChIP chip experiments into the GeneClass algorithm, and we use an improved noise model for gene expression data.* [PMCID:1810316]
2. *Target set: We extend the original GeneClass algorithm to use all target genes for which both motif and expression data is available.* [PMCID:1810316]
3. *Motif data versus CHIP chip data: In order to study different aspects of target gene regulation we use different sets of motifs and parents with the GeneClass algorithm.* [PMCID:1810316]
4. *The GeneClass algorithm for predicting differential gene expression starts with a candidate set of motifs; representing known or putative regulatory element sequence patterns and a candidate set of regulators or parentSS.* [PMCID:1810316]

Predicate-argument structures and associated functional content

Predicate	Subject	Object	Type
show	*We*	*how to incorporate genome-wide protein-DNA binding data from ChIP chip experiments into the GeneClass algorithm*	Functional Description
predict	*The GeneClass algorithm*	*differential gene expression starts with a candidate set of motifs x003bc*	Input/ Output
extend	*We*	*the original GeneClass algorithm to use all target genes for which both motif and expression data is available*	Functional Description