An exercise in kidney factomics: From article titles to RDF knowledge base

James M. Eales¹, George Demetriou¹ and Robert Stevens¹

¹School of Computer Science, University of Manchester, UK

ABSTRACT
Motivation: There are many existing resources that integrate data between databases; they do this either semantically by the use of RDF and triplestores (e.g. Bio2RDF), or with web links and ID mapping services (e.g. PICR, eUtilis). Results declared in the literature are, however, only rarely interlinked with existing databases and even more rarely interlinked with each other. We describe a method to take factual statements reported in the literature and turn them into semantic networks of RDF triples. We use a method based on finding titles of papers that contain positive, direct statements about the outcome of a biomedical investigation. We then use dependency parsing and an ontological perspective to create and combine graphs of knowledge about a domain. Our aim in this work is to collect knowledge from the literature for inclusion in the Kidney and Urinary Pathway Knowledge Base (KUPKB), which will be used in the e-LICO² project to illustrate the utility of data-mining methods for biomarker discovery and pathway modelling.

1 INTRODUCTION
A common approach for creating a knowledge base is to transform an existing database or data resource into triples and then model the relationships between these triples using either existing or newly developed ontological resources (Jupp et al. 2010, Croset et al. 2010). Once the ontological structure behind a knowledge base has been defined we can augment it with information from other sources, such as the literature or further databases. This augmentation takes a considerable amount of effort to identify statements in the literature and form them into a representation compatible with the knowledge base (Croset et al. 2010, Coulet et al. 2010a).

The Kidney and Urinary Pathway Knowledge Base (KUPKB) has been created from existing databases and ontologies, it also has its own ontology (KUPO) for modelling relationships between data (Jupp et al. 2010). Currently the KUPKB has been populated with experimental results manually extracted from the published literature. We extend this by providing triples for inclusion in the KUPKB that are automatically extracted from the literature and are known to describe an experimental result.

We want to identify a focused set of reliable statements on what is known about the KUP domain. Traditionally most text mining systems use article abstracts or full text; instead we use article titles. We are analysing titles because they are short and to-the-point. Titles can summarise the findings of a whole study into a single sentence. Titles are also the first thing a user sees when searching PubMed and they are therefore important for advertising an article to potential readers. If a study identifies a new piece of definable knowledge then the authors will usually want to present this clearly in the title. If a study finds a slightly less than clear result, then the language used to describe it is often softened and we can detect this using text mining methods. Our process for extracting these triples will involve the computationally expensive task of dependency parsing (Coulet et al. 2010a, Klein and Manning 2003), therefore it is important to limit the number and length of sentences in the text to be analysed. Titles can use complex language but are also quite short, this makes them a useful alternative to abstracts or full text documents.

Previous work in this area has also used dependency parsing (Coulet et al. 2010a, Coulet et al. 2010b) and has proven useful in the field of pharmacogenomics when looking for relationships between pre-defined sets of entities. Our approach will focus on identifying the facts presented from arbitrary biological articles, representing them as RDF triples, and then later matching these to entities in existing ontologies. Further work incorporating the use of semantic patterns for identifying entity relationships (Gaizauskas et al. 2003, Humphreys et al. 2000) has proven useful for capturing relationships describing protein structure, metabolic pathways and the function of enzymes. All of which have used a semantic framework to make the results of the analysis more widely usable, but also to make it easier to incorporate newly identified relationships with existing knowledge and then form queries based on the combined relationships; it is this flexibility that we seek in this work.

Our approach is to collect a set of titles, classify them into factual and non-factual groups and then extract sets of tri-
amples from the factual titles. We define a factual title here as “A positive, direct statement, about the outcome of a biomedical investigation”. An example of a factual title is:

“Bluetongue virus RNA binding protein NS2 is a modulator of viral replication and assembly.” (PubMed ID: 17241458)

We can see how the title does not contain “soft” or “hedged” language and instead clearly states a result of an investigation.

Such statements do not contain all the necessary contextual information to fully comprehend the implications of its finding, but this is not our aim; instead we hope to capture what is reported by the authors and then present this to other readers who can investigate further. An example of a non-factual title is:

“A role for NANOG in G1 to S transition in human embryonic stem cells through direct binding of CDK6 and CDC25A” (PubMed ID: 19139263)

This title contains many specifics (e.g. the NANOG protein, and the G1 and S cell cycle phases) and does allude to a role of NANOG in cell cycle transition, but the role is not explicitly defined, it merely suggests that a role exists. Such statements are important and could be used, but the lack of an explicit role would have to be recorded; this will be future work. Our work has revealed other kinds of titles, such as ones that report, “hedged” or possible results; describe tools or methods; and those that simply say what the article is about. In this work we concentrate on the positive, direct descriptions of an investigation to create a focused corpus as possible of statements on a given topic as a basis for triplification.

2 METHODS

The resources referred to in this paper are available as part of myExperiment pack 181.

Titles often contain multiple sentences and these can have distinct linguistic purposes. As we want to be able to distinguish between the factual and non-factual titles (but a single title can contain both factual and non-factual parts), we split all titles into their component sentences using the OpenNLP sentence detector and a set of heuristics that distinguish between the factual and non-factual purpose. As we want to be able to distinguish between the factual and non-factual parts of sentences, we identify subjects, objects and predicates by the application of heuristic rules. The dependencies are retrieved from the classification process and reused in the triplification process. The rules are applied in the following order:

1. Concatenate all compound noun (nn) and adjectival modifier (amod) dependencies on shared governor tokens.
2. Identify all nominal subject (nsubj) and nominal subject passive (nsubjpass) dependencies.
3. Identify all direct object (dobj) dependencies.
4. Attempt to join all subject and object dependencies by a common governor token. The shared governor token becomes the predicate of a new triple, with the dependent token of the subject and object dependency becoming the subject and object of the new triple respectively.
5. If the object of the new triple has a prepositional modifier (prep) then attempt to create a second new triple (Figure 1) using the dependent token of the prep as its object. The object of the first and the subject of the second triple are set as a new anonymous entity with a unique label.

Titles were collected through the eUtils interface to PubMed. A keyword search for ‘kidney’ or ‘renal’ in the title/abstract field returned 86,217 results (21/10/10), the title and PubMed ID for each of the citations was retrieved and stored for analysis. These titles were also split into a set of 91,626 sentences using the same method used for the training data.

For each sentence we derive a set of attributes to describe the title. These attributes fall into 5 groups: simple; word; phrasal; sentence and biological attributes. We use information on tokens, biological named entities, POS tags, chunks, the parse tree and the list of dependencies to profile each sentence. A full list of these attributes and the training data set can be found in our myExperiment pack. Of significant note is our use of the Whatizit (Rebholz-Schuhmann et al. 2008) named entity recognition service which provides database IDs for proteins, genes, diseases and chemicals; we use the number of matches for each entity type as attributes and we use the IDs to create URI references.

We build an SVM classifier model using the full set of profiles from the training data. We use the SMO implementation of an SVM classifier from Weka in RapidMiner.

2.1 Triplification

Our triplification process uses the dependency parse of the sentence, provided by the Stanford parser (de Marneffe et al. 2006, Klein and Manning 2003), to identify subjects, objects and predicates by the application of heuristic rules. The dependencies are retrieved from the classification process and reused in the triplification process. The rules are applied in the following order:

1. Concatenate all compound noun (nn) and adjectival modifier (amod) dependencies on shared governor tokens.
2. Identify all nominal subject (nsubj) and nominal subject passive (nsubjpass) dependencies.
3. Identify all direct object (dobj) dependencies.
4. Attempt to join all subject and object dependencies by a common governor token. The shared governor token becomes the predicate of a new triple, with the dependent token of the subject and object dependency becoming the subject and object of the new triple respectively.
5. If the object of the new triple has a prepositional modifier (prep) then attempt to create a second new triple (Figure 1) using the dependent token of the prep as its object. The object of the first and the subject of the second triple are set as a new anonymous entity with a unique label.

http://www.myexperiment.org/packs/181.html

http://rapid-i.com/content/view/202/206/lang.en/
(6) For each new triple look for conjunct (conj) dependencies with a token shared between the triple’s object and the dependency’s dependent token. Create new triples with shared subject and predicate tokens, but with the object set as the dependent token of the conjunct dependency (Figure 2).

(7) For each sentence look for abbreviation (abbrev) dependencies and create new triples with a “has_label” predicate.

(8) Create a separate ontological form of each extracted triple by nominalising the predicate. This results in two statements, the first linking the subject with the predicate (via a “participates_in” relationship) and the second linking the predicate with the object (by a “has_participant” relationship). The predicate becomes an instance of class “biological process”. This form of the triples allows a more ontological view, by nominalising the predicate verb as a biological process. This ontological form of the triples is not used in the graph visualisation or triple evaluation but can be used for ontology mapping.

Figure 1. Triplification output for example sentence including a prepositional modifier. “Mycophenolic acid inhibits the phosphorylation of NF-kappaB and JNKs and causes a decrease in IL-8 release in H2O2-treated human renal proximal tubular cells.”. Only the first preposition-derived triples are shown.

Figure 2. Triplification output for example sentence containing a conjunct dependency. “Nitric oxide diminishes apoptosis and p53 gene expression after renal ischemia and reperfusion injury.”. Only the conjunction-derived triples are shown.

3 RESULTS

3.1 Training data cross-validation

A 10-fold, stratified, cross-validation of the training data (Table 1) produced a weighted average F1 of 90.7% and the F1 for the factual (good) class of titles was 77.4%. 409 sentences were labeled as ‘good’ (21.1%) and 1,529 as ‘bad’ (78.9%).

The training data were annotated by RS and GD independently; their annotations were found to disagree on 38 (2%) sentences, giving an inter-annotator agreement (using Cohen’s kappa coefficient) of 0.936.

### Table 1. Classifier cross-validation output on training data

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure (F1)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>80.64</td>
<td>74.33</td>
<td>77.36</td>
<td>409</td>
</tr>
<tr>
<td>Bad</td>
<td>93.27</td>
<td>95.23</td>
<td>94.24</td>
<td>1529</td>
</tr>
<tr>
<td>Weighted average</td>
<td>90.60</td>
<td>90.82</td>
<td>90.68</td>
<td></td>
</tr>
</tbody>
</table>

3.2 KUP title classification

We classified each of the KUP title sentences using a model built using the full set of training data, this gave us 5,735 (6.3%) sentences classified as factual and the remaining 8,891 (93.7%) classified as non-factual. The proportion of ‘good’ titles varies considerably between the training (21.1%) and KUP title (6.3%) collections. In a preliminary manual analysis of the first 300 sentences classified as ‘good’, we found that 209 (70%) were true ‘good’ titles, this compares favourably with the ‘good’ classification accuracy on the training data of 74% (304 correct out of 409 sentences).

3.3 KUP title triplification

Using the list of dependencies for each sentence we apply the rules defined in section 2 to create a set of triples. This process created a set of 7,113 triples, containing 9,080 unique nodes, 6,989 edges of 1,255 unique edge types (triples are available in tab-delimited and RDF/XML format on myExperiment). These can be formed into a graph by connecting triples with shared subject and object entities. The largest connected component of this graph (see myExperiment for Cytoscape graph) contains 2,676 nodes and has 2,765 edges of 603 distinct types (see myExperiment for visualisation). The central region of this graph has several highly connected entities, the most highly connected being “rats”. Other highly connected entities are “kidney”, “renal function”, “angiotensin II” and “renal injury”. In a manual analysis of a sample of 150 triples extracted from the KUP titles, we found that 96 (64%) were correct. It should be emphasised that titles erroneously classified as “good” were found to commonly produce incorrect triples, thus compounding errors made before triplification. Furthermore the Stanford parser has not been trained on biomedical text, this can lead to parser errors and therefore dependency and triplification errors.

4 CONCLUSIONS

We have described a twin approach to putting facts from the literature into RDF triples. Our main goal was to create a corpus of fact-orientated statements about a particular do-
main. We did this by training a classifier to recognise titles that form positive, direct statements about the outcome of an investigation. We then turned these into co-ordinated sets of triples using a dependency parser, in which we expand key verb relationships into new triples containing anonymous entities to which other entities can be linked.

Our results so far are satisfactory in that we do create a focused corpus of titles of the right kind. It may be possible to optimise our features and our generation of the initial set of titles to improve performance. For example, we could include the names of disease, gene or protein entities found in the text. We have deliberately favoured precision over recall in an attempt to avoid too much "noise" in our resulting triples. This is obviously at the expense of recall, but this was a price worth paying to avoid a larger "tidying up" task.

We are also deliberately doing "factomics", where we retrieve and encode fact-like statements. Scientific papers are rich in context that is needed to fully interpret such facts (Mons 2009). This approach does not attempt any kind of full extraction of the scientific knowledge necessary for the interpretation of the facts. Instead, we have taken the approach that we are exposing the "headlines" of what has been said, and provided links back to the original paper for when a scientist finds a "fact of interest".

On inspection of a sample of triples we have found that 64% were correct. At each stage of our process, however, there will be unwanted titles and poor triplification and noise will accumulate. It seems that improvements in the title classification process should pay the greatest dividends, by providing a tighter and more focused set of genuinely factual titles to the triplification process.

To interlink our triples with the KUPKB, we intend to rewrite our text-based subject, predicate and object values using URIs derived from several sources. Using our existing named entity recognition results from Whatizit (see Methods) we will replace matching subject and object values with URIs for Uniprot (in the case of proteins) and ChEBI (for chemical entities). We will also use the NCBO BioPortal Annotator\(^6\) to replace further subject and object labels with URIs from the Mouse anatomy ontology\(^7\). We will also replace any matching predicate values found in the Molecular Interactions ontology\(^8\). We will use these predicate mappings to normalise each verb by expanding a single RDF statement into two. Finally each set of triples will be given an RDF context of the corresponding PubMed ID. All of these mapped ontologies are currently part of the KUPKB thus easing the integration of literature-derived knowledge into the knowledge base.

Literature of various sorts forms a vital repository of a domain’s knowledge. This knowledge needs to be exposed in integrated, computationally accessible forms. As well as integrating within the literature, we need to integrate with knowledge from resources such as databases. RDF forms an attractive means for doing this, especially when combined with the common vocabularies that are being developed by the community. Text mining offers a tempting means to expose this literature-based knowledge, yet can suffer from the need to create corpora of focused collections of desirable kinds of statements. We have presented one technique for creating a focused corpus of one kind of statement and turning this into triples for a domain knowledge base.

**ACKNOWLEDGEMENTS**

This work was funded by the e-LICO project - EU/FP7/ICT-2007.4.4.

**REFERENCES**


\(^6\) http://biportal.bioontology.org/annotator

\(^7\) http://purl.bioontology.org/ontology/MA

\(^8\) http://purl.bioontology.org/ontology/MI