Combining NLP and Semantics for Mining Software Technologies from Research Publications

Conference or Workshop Item

How to cite:


For guidance on citations see FAQs.

© 2016 The Authors

Version: Accepted Manuscript

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1145/2872518.2889358

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
Combining NLP and semantics for mining software technologies from research publications

Hélène de Ribaupierre\textsuperscript{1,2}  
\textsuperscript{1}Department of Computer Science,  
University of Oxford,  
Oxford OX1 3QD,  
United Kingdom  
helene.de.ribaupierre@cs.ox.ac.uk  

Francesco Osborne\textsuperscript{2}  
\textsuperscript{2}Knowledge Media Institute,  
The Open University  
Milton Keynes MK7 6AA, United Kingdom  
francesco.osborne@open.ac.uk  

Enrico Motta\textsuperscript{2}  
\textsuperscript{2}Knowledge Media Institute,  
The Open University  
Milton Keynes MK7 6AA, United Kingdom  
e.motta@open.ac.uk  

ABSTRACT  
The community of natural language processing (NLP) has developed a variety of methods for extracting and disambiguating information from research publications. However, they usually focus only on classic research entities such as authors, affiliations, venues, references and keywords. We propose a novel approach, which combines NLP and semantic technologies for generating from the free text of research publications an OWL ontology describing a number of software technologies used or introduced by researchers, such as applications, systems, frameworks, programming languages, and formats. The method was tested on a sample of 300 publications in the Semantic Web field, yielding promising results.

Author Keywords  
NLP, Semantic Web, Ontology, Digital Library, Information extraction, Data mining, Artificial Intelligence

ACM Classification Keywords  
I.2 Artificial Intelligence, H.3.7 Digital Library, I.2.7 Natural Language Processing

INTRODUCTION  
The Web contains a large mass of research publications, which is destined to grow even further due also to the success of the open access movement. The knowledge derived from these publications can be used for a variety of tasks such as producing research analytics, identifying experts, supporting researchers’ work, assessing the effectiveness of research policies and even for fostering the long-term ambition of creating systems able to reason on research problems. However, these publications are not in machine-readable formats and thus are not easy to process.

The problem of deriving sound knowledge from research publications was historically addressed from two mainly perspectives. On the one side, the community of semantic publishing has proposed machine-readable publication formats and created repositories of scholarly data adopting web standards such as RDF. On the other side, the community of natural language processing (NLP) has developed a variety of methods for extracting and disambiguating information from research papers and their metadata. In both cases the output is usually a machine-readable descriptions of entities such as authors, affiliations, venues, references and keywords. However, these classic research entities allow only for a coarse-grained analysis of the research environment. We still lack effective methods to extract and describe semantically a number of more structured and fine-grained entities, such as applications, framework, formats, scientific paradigms, algorithms, experiments and datasets and so on. For this reason, some recent efforts tried to address these complex entities, by extracting rhetorical entities [1] (e.g., claims, arguments), discourse elements [2] (e.g., methodologies, definitions, hypothesis), and chemicals [3].

This paper contributes to this challenge by introducing a novel approach, which combines natural language processing (NLP) and semantic technologies, to learn an OWL ontology defining the software technologies described in research publications, such as applications, systems, frameworks, programming languages and formats. This solution was developed for enriching further the knowledge base of Rexplore [4], a system which uses semantic technologies for supporting users in exploring research. One of the main assets of Rexplore is Klink-2 [5], an algorithm for generating large-scale and granular ontologies of research topics. We intend to combine the Klink-2 topic ontology with the ontology of software technologies to provide a more comprehensive representation of the research landscape. The resulting knowledge base will be used for a variety of tasks: searching for the applications used in a certain field and assess their popularity, analyzing the dynamics of the creation of new technologies, and supporting the reproducibility of scientific frameworks.

TECHNOLOGY EXTRACTION  
Our approach performs noun phrases detection on the title and the abstract of each research publication and outputs an OWL ontology describing the resulting technologies. For analyzing the free text we adopted GATE [6], a well-known open source NLP platform. We also exploited a number of GATE plugins: Ontology OWLIM2, a module for importing ontologies, ANNIE, a component that forms a pipeline composed of a tokenizer, a gazetteer, a sentence splitter and a part-of-speech tagger, and JAPE (Java Annotation Patterns Engine), a grammar language for operating over annotations based on regular expressions.

The approach takes in consideration all the sentences that contain a number of clue terms related to software technologies and the verbs usually adopted to introduce or describe them.
To evaluate the performance of the method we tested different inputs (only abstracts or both abstracts and titles) and filters (with, without, with the Klink-2 ontology, with both Klink-2 ontology and WordNet): 1) only abstracts (A), 2) abstracts with Klink-2 ontology (AK), 3) abstract with Klink-2 ontology and WordNet (AKW), 4) titles and abstracts (TA), 5) titles and abstracts with Klink-2 ontology (TAK), 6) titles and abstracts with Klink-2 ontology and WordNet (TAKW). The performance of the last three versions is comparable in terms of F1 score, yielding different precision/recall tradeoffs (Table 2). The adoption of the two semantic knowledge bases as filters raises considerably the precision, but lowers the recall. In fact a number of technologies have common names which can be filtered out by WordNet and some of the most popular ones can actually be considered research topics (e.g., OWL). For this reason, we intend to improve further the ability of the approach to recognize technologies by considering also a number of additional entities derived from publication metadata, such as venues and authors.

Table 2: Results of the evaluation on 300 abstracts

<table>
<thead>
<tr>
<th>True Positive</th>
<th>A</th>
<th>AK</th>
<th>AKW</th>
<th>TA</th>
<th>TAK</th>
<th>TAKW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.68</td>
<td>0.75</td>
<td>0.92</td>
<td>0.73</td>
<td>0.81</td>
<td>0.96</td>
</tr>
<tr>
<td>Recall</td>
<td>0.52</td>
<td>0.50</td>
<td>0.42</td>
<td>0.75</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>F1</td>
<td>0.59</td>
<td>0.60</td>
<td>0.58</td>
<td>0.74</td>
<td>0.74</td>
<td>0.72</td>
</tr>
</tbody>
</table>

CONCLUSION

The evaluation suggests that a combination of NLP and semantic technologies is effective in extracting software technologies from research publications and can be customized for yielding different compromises between precision and recall. We plan to work on improving recall and on the automatic generation of semantic relationships between technologies and other research entities.

REFERENCES