Combining NLP and semantics for mining software technologies from research publications

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

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

Enrico Motta[2]
[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, 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 asses 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.

Sentence 1	We	describe	several	advanced	functionalities	of	Magpie	a	tool
Part-of-speech tags	PRP	VPB	JJ	VBD	NNS	IN	NNP	DT	NN
Sentence 2	We	developed					FSAD	a	prototype
Part-of-speech tags	PRP	VBD					NNP	DT	NN
JAPE rule	(Pronoun)	(ListVerbOntology)	Token(0,4)				(CandidateNoun)	DT	(CatToolsOntology)

Table 1: Example of JAPE rule for detecting tools derived by two annotated sentences.

To this end, we crafted an ontology defining the categories of software technologies (e.g., “application”, “implementation”, “system”, “prototype”) and a second one including the different verbs used for describing these technologies (e.g., “describe”, “develop”, “implement”). For example, the following sentence introduces Magpie, a semantic web browser: “We describe several advanced functionalities of Magpie, a tool that assists users with interpreting the web resources”. The position of the noun “Magpie” in the context of the sentence, followed by the clue term “tool” and object of “describe”, suggests that it may be the name of an application. Of course, the syntactic structure of a sentence for describing a technology can vary a lot. The technology name can be a proper noun, a common noun or a compound noun, and is not necessarily the subject or the object of the sentence. For this reason, we need a wide range of clue terms and rules able to adapt to the variety of contexts in which technologies can be referred. We reuse a methodology that we introduced in [2] to construct JAPE rules from annotated examples representative of the variety of ways in which technologies can be referred. This method clusters sentences that have similarities in the sequence of deterministic terms (e.g., the categories of technologies and verbs described in the ontologies), then replaces these terms with either a JAPE macro or an ontology concept and non-deterministic terms with a sequence of optional token (Table 1). We produced 18 JAPE rules to identify similar syntactic constructions and extract related technologies. After this initial learning phase, the approach performs the following steps: 1) it splits titles and abstracts into sequences of tokens and assigns them with part-of-speech tags (e.g., noun, verb and adverb) using ANNIE; 2) it selects the sentences including the clue terms using a sequence of JAPE rules; 3) it applies the eighteen previously defined JAPE rules to generate a list of candidate tools; 4) it runs a number of filters on the list and outputs an OWL ontology which associates the detected technologies with the sentences in which they were described and the publications from which they were extracted.

To improve the performance of the method we tested different kinds of filters. In particular, we used WordNet and Wiktionary for excluding some categories of common names and the Klink-2 ontology of Computer Science for filtering out research topics that may be confused with technologies (e.g., “NLP”, “Semantic Web”). We also applied additional domain heuristics; for example, we did not filter animal names because a good number of applications in Computer Science are named after animals. Since many technologies appear in titles, we also tried to run the approach in two phases. We first processed the titles and generated a gazetteer of technology names, and then we analyzed the abstracts, using the gazetteer to find further sentences associated with these technologies, in addition to the standard JAPE rules.

EVALUATION

To evaluate the performance of our method, we tested it on a gold standard of 300 manually annotated abstracts (downloaded from Microsoft Academic Search) comprising 702 sentences and 259 software technologies in the field of Semantic Web. The ontologies adopted in the prototype and the evaluation data are available at http://cui.unige.ch/~deribauh/WWW2016results/.

We compared six versions of the approach using different inputs (only abstracts or both abstracts and titles) and filters (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.

	A	AK	AKW	TA	TAK	TAKW
Precision	0.68	0.75	0.92	0.73	0.81	**0.96**
Recall	0.52	0.50	0.42	**0.75**	0.68	0.57
F1	0.59	0.60	0.58	**0.74**	**0.74**	0.72

Table 2: Results of the evaluation on 300 abstracts

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

1. B. Sateli and R.Witte. 2015. What's in this paper?: Combining Rhetorical Entities with Linked Open Data for Semantic Literature Querying. WWW 2015 Companion.
2. H. de Ribaupierre and G. Falquet. 2014. User-centric design and evaluation of a semantic annotation model for scientific documents. IKNOW2014. ACM, USA.
3. P. Corbett and A. Copestake. 2008. Cascaded classifiers for confidence-based chemical named entity recognition. BMC bioinformatics, 9 (Suppl 11).
4. F. Osborne, E. Motta, and P. Mulholland. 2013. Exploring Scholarly Data with Rexplore. 12th International Semantic Web Conference, Bethlehem, USA.
5. F. Osborne, E. Motta. 2015. Klink-2: integrating multiple web sources to generate semantic topic networks. 14th International Semantic Web Conference, Sydney, Australia.
6. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. ACL 2002.