Natural Language Processing for Intelligent Access to Scientific Information

Horacio Saggion and **Francesco Ronzano**
DTIC
Universitat Pompeu Fabra
Carrer Tanger 122, Barcelona (08018), Barcelona, Spain
{name.surname}@upf.edu

Abstract

During the last decade the amount of scientific information available on-line increased at an unprecedented rate. As a consequence, nowadays researchers are overwhelmed by an enormous and continuously growing number of articles to consider when they perform research activities like the exploration of advances in specific topics, peer reviewing, writing and evaluation of proposals. Natural Language Processing Technology represents a key enabling factor in providing scientists with intelligent patterns to access to scientific information. Extracting information from scientific papers, for example, can contribute to the development of rich scientific knowledge bases which can be leveraged to support intelligent knowledge access and question answering. Summarization techniques can reduce the size of long papers to their essential content or automatically generate state-of-the-art-reviews. Paraphrase or textual entailment techniques can contribute to the identification of relations across different scientific textual sources. This tutorial provides an overview of the most relevant tasks related to the processing of scientific documents, including but not limited to the in-depth analysis of the structure of the scientific articles, their semantic interpretation, content extraction and summarization.

1 Introduction

During the last decade the amount of scientific information available on-line increased at an unprecedented rate. Recent estimates reported that a new paper is published every 20 seconds. PubMed includes more than 26M papers with a growth rate of about 1,370 new articles per day. Elsevier Scopus and Thomson Reuthers ISI Web of Knowledge respectively index more than 57 and 90 million papers. The Cornell University Library arXiv initiative provides access to over 1M e-prints from various scientific domains. At the same time, more and more papers can be freely read on-line since they are published as Open Access content: the full text of 27% of PubMed publications and more than 17% of the articles indexed by Scopus and ISI Web of Knowledge is available on-line for free and this percentages are considerably growing. The Directory of Open Access Journals, one of the most authoritative indexes of high quality, Open Access, peer-reviewed publications, lists more than 9,200 journals and 2.3M papers. Sometimes between 2017 and 2021, more than half of the global papers are expected to be published as Open Access content. Moreover, several top conferences are making their articles freely available through dedicated archives even before the conference takes place. Social networks are by no means outside of this picture: research networks like ResearchGate, Academia.edu or Mendeley are rapidly expanding, facilitating scientific information sharing.

In this scenario of scientific information overload, researchers, as well as any other interested actor, are overwhelmed by an enormous and continuously growing number of articles to consider. Understanding recent advances in specific research fields, new methods and techniques, peer reviewing, writing and evaluation of research proposals and, in general, any activity that requires a careful and comprehensive assessment of scientific literature has turned into an extremely complex and time-consuming task for scientists world-wide. In this context, the Natural Language Processing community plays a central role in investigating and improving new approaches to the analysis of scientific information, thus uncovering incredible opportunities for contributions and experimentations. The extraction and integration of

information from scientific papers constitute a key factor for the development of rich scientific knowledge bases which can be leveraged to support structured and semantically-enabled searches, intelligent question answering and personalized content recommendation. Summarization techniques can help to identify the essential contents of publications thus generating automatic state of the art reviews while paraphrase or textual entailment can contribute to identify relations across different scientific textual sources.

The objective of this tutorial is to provide a comprehensive overview of the most relevant problems we have to face when we mine scientific literature by means of Natural Language Processing Technologies, thus identifying challenges, solutions, and opportunities for our community. In particular, we consider approaches and tools useful to analyze and characterize a wide range of structural and semantic peculiarities of scientific articles, including document formats, layout-dependent information, discursive structure and networks of citations. We discuss relevant scenarios where the availability of structured, semantically-annotated publications improves the way we benefit from scientific literature, including article summarization, scientific content search, selection and aggregation, and publication impact assessment. Related tools, applications, datasets and publication venues are also reviewed.

2 Tutorial Outline

The topics of the tutorial are described below. The tutorial will be combined with demonstrations of existing technologies.

1. **Introduction**

   Overwhelmed by scientific publications (patents, research articles, tutorials, presentations, etc.)

   Challenges & opportunities of scientific information overload

2. **Analyzing the Structure of Scientific Publications**

   Available formats & contents

   Retrieving textual contents from PDF publications

   Document structure analysis

   Patent analysis

3. **Mining the Semantics of Scientific Publications**

   Text organization

   Rhetorical structure analysis

   Citation networks

   Interpretation of citation purpose and polarity

4. **Extracting Information from Scientific Literature**

   Scientific entities and their identification (names, formulas, numbers, drugs, genes, etc.)

   Relation extraction problems (interactions, causal relations)

5. **Summarizing Scientific Information**

   Classic summarization approaches to scientific document

   Classification-based approaches

   Citation-based approaches

   Summarizing patents

6. **Language Resources for Scientific Text Analysis and Representation**

   Available scientific corpora for experimentation

   Lexical Resources in specialized domains

   Ontologies for scientific information modelling

7. **Social Media and Science: new Opportunities**

   Socially connected scientific entities

   Social Media metrics to assess research impact

8. **Applications, Challenges and Projects**

   Scientific literature on-line portals

   Discussion venues and challenges

   Relevant projects

3 Tutorial Web Site

More details on this tutorial can be accessed on-line at: `http://taln.upf.edu/pages/coling2016tutorial/`.

Acknowledgments

We acknowledge (partial) support by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and the European Project Dr. Inventor (FP7-ICT-2013.8.1 - Grant no: 611383).