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 (Munroe, 2013). 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 (Björk et al., 2014): 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 (Lewis, 2012). 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 (Bar-Ilan et al., 2012; Thelwall et al., 2013): research networks like ResearchGate, Academia.edu or Mendeley are rapidly expanding, facilitating scientific information sharing (Haustein et al., 2014).

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.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts,
pages 9–13, Osaka, Japan, December 11-17 2016.

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 (Lipinski et al., 2013; Guo et al., 2011; Ronzano and Saggion, 2016b) 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 (He et al., 2010; Huang et al., 2012). 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 (Teufel and Moens, 2002; Abu-Jbara and Radev, 2011; Ronzano and Saggion, 2016a; Saggion and Lapalme, 2002; Saggion, 2008).

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 (Constantin et al., 2013; Lopez, 2009), layout-dependent information (Ramakrishnan et al., 2012; Luong et al., 2012; Councill et al., 2008), discursive structure (Liakata et al., 2010; Teufel et al., 2009; Fisas et al., 2015) and networks of citations (Teufel et al., 2006; Athar, 2011; Abu-Jbara et al., 2013). 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 Outline

The half-day tutorial will review the following nine of top-level topics. For each topic, some of the core themes to discuss is specified.

1. **Scientific Information Overload: Challenges and Opportunities**
   - Overwhelmed by scientific publications (research articles, patents, 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
9. **Mining scientific articles with the Dr. Inventor Framework**
    - The Dr. Inventor project
    - Overview and demo of the Dr. Inventor Scientific Text Mining Framework

3 Tutorial Web Site

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

4 Organizers

Horacio Saggion holds a PhD in Computer Science from Universite de Montreal, Canada. He obtained his BSc in Computer Science from Universidad de Buenos Aires in Argentina, and his MSc in Computer Science from UNICAMP in Brazil. Horacio is an Associate Professor at the Department of Information and Communication Technologies, Universitat Pompeu Fabra (UPF), Barcelona. He is a member of the Natural Language Processing group where he works on automatic text summarization, text simplification, information extraction, sentiment analysis and related topics. His research is empirical combining symbolic, pattern-based approaches and statistical and machine learning techniques. Before joining Universitat Pompeu Fabra, he worked at the University of Sheffield for a number of UK and European research projects (SOCIS, MUMIS, MUSING, GATE, CUBREPORTER) developing competitive human language technology. He was also an invited researcher at John Hopkins University for a project on multilingual text summarization. He is currently principal investigator for UPF in several EU and national projects.

Horacio has published over 100 works in leading scientific journals, conferences, and books in the field of human language technology. He organized four international workshops in the areas of text summarization and information extraction and was co-chair of STIL 2009. He is co-editor of a book on multilingual, multisource information extraction and summarization published by Springer in 2013. Horacio is member of the ACL, IEEE, ACM, and SADIO. He is a regular programme committee member for international conferences such as ACL, EACL, COLING, EMNLP, IJCNLP, IJCAI and is an active reviewer for international journals in computer science, information processing, and human language technology. Horacio has given courses, tutorials, and invited talks at a number of international events including LREC, ESSLLI, IJCNLP, NLDB, and RuSSIR.

Francesco Ronzano holds a PhD in Information Engineering from the University of Pisa, Italy. Francesco is currently a Researcher of the Natural Language Processing Group (TALN) at the Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, where he deals with machine learning approaches for information extraction and text summarization, with special focus on scientific publishing and social media analysis. Francesco has several years of research experience mainly in the context of National (Italian and Spanish) and European Research Projects related to the exploitation of machine learning approaches and Web technologies to foster Language Technologies. His research interests include on-line data semantics, machine learning, knowledge representation and Semantic Web applications. Francesco has coordinated the development of the Dr. Inventor Text Mining Framework, an software tool to mine a wide range of facets of scientific publications.

Francesco has contributed to more than 40 publications among book chapters, journal articles, conference papers. He acted as reviewer to international conferences including AAAI, EMNLP, LREC, RANLP, etc.

Acknowledgements

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).

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc. of 49th Annual Meeting of the ACL: Human Language Techologies*, pages 500–509. ACL, June.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *HLT-NAACL*, pages 596–606.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*, pages 81–87. Association for Computational Linguistics.

Judit Bar-Ilan, Stefanie Haustein, Isabella Peters, Jason Priem, Hadas Shema, and Jens Terliesner. 2012. Beyond citations: Scholars' visibility on the social web. *arXiv preprint arXiv:1205.5611*.

Bo-Christer Björk, Mikael Laakso, Patrik Welling, and Patrik Paetau. 2014. Anatomy of green open access. *Journal of the Association for Information Science and Technology*, 65(2):237–250.

Alexandru Constantin, Steve Pettifer, and Andrei Voronkov. 2013. Pdfx: fully-automated pdf-to-xml conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180. ACM.

Isaac G Councill, C Lee Giles, and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In *LREC*, volume 8, pages 661–667.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2015. On the discoursive structure of computer graphics research papers. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 42.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283. Association for Computational Linguistics.

Stefanie Haustein, Vincent Larivière, Mike Thelwall, Didier Amyot, and Isabella Peters. 2014. Tweets vs. mendeley readers: How do these two social media metrics differ? *IT-Information Technology*, 56(5):207–215.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM.

Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1910–1914. ACM.

David W Lewis. 2012. The inevitability of open access. *College & Research Libraries*, 73(5):493–506.

Maria Liakata, Simone Teufel, Advaith Siddharthan, Colin R Batchelor, et al. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. 2013. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 385–386. ACM.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International Conference on Theory and Practice of Digital Libraries*, pages 473–474. Springer.

Minh-Thang Luong, Thuy Dung Nguyen, and Min-Yen Kan. 2012. Logical structure recovery in scholarly articles with rich document features. *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, 270.

Randall Munroe. 2013. The rise of open access. *Science*, 342(6154):58–59.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. 2012. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1):1.

Francesco Ronzano and Horacio Saggion. 2016a. An empirical assessment of citation information in scientific summarization. In *International Conference on Applications of Natural Language to Information Systems*, pages 318–325. Springer.

Francesco Ronzano and Horacio Saggion. 2016b. Knowledge extraction and modeling from scientific publications. *Proceedings of the Semantics, Analytics, Visualisation: Enhancing Scholarly Data Workshop*.

Horacio Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with sumum. *Comput. Linguist.*, 28(4):497–526, December.

Horacio Saggion. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, 49(2):103–125.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1493–1502. Association for Computational Linguistics.

Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. 2013. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5):e64841.