COLING 2014

Computerm 2014
4th International Workshop on Computational Terminology

Proceedings of the Workshop

August 23, 2014
Dublin, Ireland

ISBN 978-1-873769-34-8
Proceedings of the 4th International Workshop on Computational Terminology (Computerm)
Patrick Drouin, Natalia Grabar, Thierry Hamon and Kyo Kageura (eds.)

Introduction

Computational Terminology covers an increasingly important aspect in Natural Language Processing areas such as text mining, information retrieval, information extraction, summarisation, textual entailment, document management systems, question-answering systems, ontology building, etc. Terminological information is paramount for knowledge mining from texts for scientific discovery and competitive intelligence. Scientific needs in fast growing domains (such as biomedicine, chemistry and ecology) and the overwhelming amount of textual data published daily demand that terminology is acquired and managed systematically and automatically; while in well established domains (such as law, economy, banking and music) the demand is on fine-grained analyses of documents for knowledge description and acquisition. Moreover, capturing new concepts leads to the acquisition and management of new knowledge.

The aim of this fourth CompuTerm workshop is to bring together Natural Language Processing researchers to discuss recent advances in computational terminology and its impact in many NLP applications. The topics addressed in this workshop are wide ranging:

- term extraction, recognition and filtering, which is the core of the terminological activity that lays basis for other terminological topics and tasks;
- event recognition and extraction, that extends the notion of the terminological entity from terms meaning static units up to terms meaning procedural and dynamic processes;
- acquisition of semantic relations among terms, which is also an important research topic as the acquisition of semantic relationships between terms finds applications such as the population and update of existing knowledge bases, definition of domain specific templates in information extraction and disambiguation of terms;
- term variation management, that helps to deal with the dynamic nature of terms, their acquisition from heterogeneous sources, their integration, standardisation and representation for a large range of applications and resources, is also increasingly important, as one has to address this research problem when working with various controlled vocabularies, thesauri, ontologies and textual data. Term variation is also related to their paraphrases and reformulations, due to historical, regional, local or personal issues. Besides, the discovery of synonym terms or term clusters is equally beneficial to many NLP applications;
- definition acquisition, that covers important research and aims to provide precise and nonambiguous description of terminological entities. Such definitions may contain elements necessary for the formal description of terms and concepts within ontologies;
- consideration of the user expertise, that is becoming a new issue in the terminological activity, takes into account the fact that specialized domains contain notions and terms often nonunderstandable to non-experts or to laymen (such as patients within the medical area, or bank clients within banking and economy areas). This aspect, although related to specialized areas, provides direct link between specialized languages and general language;
- systematic terminology management and updating domain specific dictionaries and thesauri, that are important aspects for maintaining the existing terminological resources. These aspects become crucial because the amount of the existing terminological resources is constantly increasing and because their perennial and efficient use depends on their maintenance and updating, while their re-acquisition is costly and often non-reproducible;

- monolingual and multilingual resources, that open the possibility for developing cross-lingual and multi-lingual applications, requires specific corpora, methods and tools which design and evaluation are challenging issues;
- robustness and portability of methods, which allows to apply methods developed in one given context to other contexts (corpora, domains, languages, etc.) and to share the research expertise among them;
- social netwoks and modern media processing, that attracts an increasing number of researchers and that provides challenging material to be processed;
- utilization of terminologies in various NLP applications, as they are a necessary component of any NLP system dealing with domain-specific literature, is another novel and challenging research direction.

In the call for paper, we encouraged authors to submit their research work related to various aspects of computational terminology, ranging from term extraction in various languages (using verb co-occurrence, information theoretic approaches, machine learning, etc.), translation pairs extracting from bilingual corpora based on terminology, up to semantic oriented approaches and theoretical aspects of terminology. Besides, experiments on the evaluation of terminological methods and tools are also encouraged since they provide interesting and useful proof about the utility of terminological resources:

- direct evaluation may concern the efficiency of the terminological methods and tools to capture the terminological entities and relations, as well as various kinds of related information;
- indirect evaluation may concern the use of terminological resources in various NLP applications and the impact these resources have on the performance of the automatic systems. In this case, research and competition tracks (such as TREC, BioCreative, CLEF, CLEF-eHealth, I2B2, *SEM, and other shared tasks), provide particularly fruitful evaluation contexts and proved very successful in identifying key problems in terminology such as term variation and ambiguity.

The Computerm 2014 workshop received 14 submissions from 10 countries and 3 continents addressing issues on 12 languages. Further to a double-blind peer-review process, 6 papers were accepted for oral presentations and 7 as posters. The acceptance rate for oral presentations is 40% and the overall acceptance rate is 86.66%. We believe this workshop will be a nice place for fruitful research discussions, and the emergence of new research topics and collaborations. The objective of the combined oral and poster presentations is to strengthen this point.

Acknowledgments

First of all, we would like to thank the members of the program committee for the quality of their reviews. We are particularly grateful to Noemie Elhadad for accepting to give an invited talk. We are grateful to the COLING organizers, in particular Jennifer Foster, John Judge and Joachim Wagner for their help in the workshop organisation. And last but not the least, we would like to thank all the authors for the quality of their submission and their hard work. We also thank the participants to the Computerm 2014 workshop. Without all of them, this 4th workshop on Computational Terminology would not take place.

Workshop Organization

Organizers:

Patrick Drouin, Observatoire de linguistique Sens-Texte, Université de Montréal, Montréal, Canada
Natalia Grabar, CNRS UMR 8163 STL, Université Lille 1&3, Villeneuve d'Ascq, France
Thierry Hamon, LIMSI-CNRS, Orsay, France & Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
Kyo Kageura, Library and Information Science Laboratory, University of Tokyo, Tokyo, Japan

Program Committee:

Sophia Ananiadou, University of Manchester, National Centre for Text Mining, UK
Olivier Bodenreider, NLM, USA
Beatrice Daille, IRIN, France
Éric Gaussier, LIG, Université Joseph Fourier, France
Gregory Grefenstette, Clairvoyance Corp, France
Marie-Claude L'Homme, University of Montréal, Canada
Philippe Langlais, RALI, Canada
John McNaught, UMIST & National Centre for Text Mining, UK
Rogelio Nazar, Pontificia Universidad Católica de Valparaíso, Chile
Goran Nenadic, University of Manchester, UK
Jorge Vivaldi Palatresi, University Pompeu Fabra, Spain
Selja Seppälä, University at Buffalo, USA
Karin Verspoor, University of Melbourne, Australia
Pierre Zweigenbaum, LIMSI, France

Invited Speaker:

Noemie Elhadad, Department of Biomedical Informatics, Columbia University, USA

Table of Contents

Workshop Program

Saturday August 23, 2014

(8:45) Opening Remarks

(9:00) Session 1

9:00 *Generalising and Normalising Distributional Contexts to Reduce Data Sparsity: Application to Medical Corpora*
Amandine Périnet and Thierry Hamon

9:30 *Assigning Terms to Domains by Document Classification*
Robert Gaizauskas, Emma Barker, Monica Lestari Paramita and Ahmet Aker

10:00 *Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation*
Mihael Arcan, Claudio Giuliano, Marco Turchi and Paul Buitelaar

10:30 by Coffee Break

(11:00) Invited Speaker: Noemie Elhadad

11:00 *Terminology Questions in Texts Authored by Patients*
Noémie Elhadad

12:30 by Lunch Break

(14:00) Session 2

14:00 *NPMI Driven Recognition of Nested Terms*
Malgorzata Marciniak and Agnieszka Mykowiecka

14:30 *Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation*
Rejwanul Haque, Sergio Penkale and Andy Way

15:00 *The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics*
Behrang Q. Zadeh and Siegfried Handschuh

15:30 by Coffee Break

Saturday August 23, 2014 (continued)

(16:00) Poster Session

Building the Interface between Experts and Linguists in the Detection and characterisation of Neology in the Field of Neurosciences
Jesús Torres-del-Rey and Nava Maroto

A comparative User Evaluation of Terminology Management Tools for Interpreters
Hernani Costa, Gloria Corpas Pastor and Isabel Durán Muñoz

Automatic Annotation of Parameters from Nanodevice Development Research Papers
Thaer M. Dieb, Masaharu Yoshioka, Shinjiroh Hara and Marcus C. Newton

Evaluating Term Extraction Methods for Interpreters
Ran Xu and Serge Sharoff

Unsupervised Method for the Acquisition of General Language Paraphrases for Medical Compounds
Natalia Grabar and Thierry Hamon

Identifying Portuguese Multiword Expressions using Different Classification Algorithms - A Comparative Analysis
Alexsandro Fonseca, Fatiha Sadat and Alexandre Blondin Massé

Towards Automatic Distinction between Specialized and Non-Specialized Occurrences of Verbs in Medical Corpora
Ornella Wandji Tchami and Natalia Grabar

(17:00) Closing Session