INTERNATIONAL WORKSHOP

ON DEFINITION EXTRACTION

held in conjunction with the International Conference

RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by
Gerardo Sierra, Mara Pozzi and Juan-Manuel Torres

Borovets, Bulgaria

18 September 2009

International Workshop

ON DEFINITION EXTRACTION

PROCEEDINGS

Borovets, Bulgaria
18 September 2009

ISBN 978-954-452-013-7

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Foreword

In the last few years the automatic extraction of definitions from textual data has become a common research topic in several domains of Natural Language Processing. These include:

- Definition extraction as a methodological resource for fields as different as computational semantics, information extraction, text mining, ontology development, WEB semantics and e-learning.
- The conception of definition extraction as a self-challenging task, in particular in computational lexicography and terminography, fields oriented towards the design and implementation of electronic tools such as lexical knowledge bases, machine-readable dictionaries, terminological databases, thesauri, machine translation systems or question-answering systems.

However, in contrast to the general use of definition extraction in multiple domains, there is no specific forum for sharing information about methodologies, tools, evaluation techniques or applications related to this field. Therefore, the goal of this workshop is to provide an opportunity to discuss theoretical and applied issues regarding definition extraction, such as:

- Contributions concerning the state of the art in definition extraction.
- Concrete applications of definition extraction in scientific or technical fields.
- The newest techniques to recognise and extract definitions candidates from running text using symbolic or statistical methods.
- Demonstration of computational tools for extracting definitions from large corpora.

The ten accepted papers report recent research initiatives on the topic of definition extraction and its applications.

The first paper, A formal scope on the relation between definitions and verbal predications by César Aguilar and Gerardo Sierra, outline a formal description of grammatical relations found in definitional contexts in Spanish and describe syntactic patterns relating definitions and predications and the usefulness of these patterns for the identification of definitions in technical corpora.

In the paper Description and evaluation of a definition extraction system for Spanish language, Rodrigo Alarcón, Gerardo Sierra and Carme Bach present a description and evaluation of a pattern-based approach for definition extraction in Spanish specialised texts based on the search for definitional verbal patterns related to analytical, extensional, functional and synonymical definitions.

Enriching a lexicographical tool with domain definitions: Problems and solutions by María Barrios, Guadalupe Aguado de Cea y José Ángel Ramos describes the problems faced by definition extraction methods due to poor definition construction and proposes some solutions.

In the paper Extraction of author's definitions using indexed reference identification, Marc Bertin, Iana Atanassova and Jean-Pierre Descles explore the establishment of relations between definitions and authors by using indexed references based on a linguistic ontology for the extraction of definitions from multilingual corpora of scientific texts.

The paper Evolutionary algorithms for definition extraction, by Claudia Borg, Mike Rosner and Gordon Pace, explores the use of machine learning methods to extract definitions. It reports the positive results obtained by the use of genetic programming and genetic algorithms to learn the relative importance of typical linguistic forms of definitions.

Language independent system for definition extraction: First results using learning algorithms, by Rosa Del Gaudio and António Branco, presents several language-independent approaches to deal with unbalanced data sets applied to two corpora in different languages for definition extraction using machine learning algorithms.

Gerard de Melo and Gerhard Weikum's paper Extracting Sense-Disambiguated Example Sentences From Parallel Corpora investigates to what extent sense-specific example sentences can be extracted from parallel corpora using lexical knowledge bases for multiple languages as a sense index to disambiguate word senses.

In her paper A proposal for a framework to evaluate feature relevance for terminographic definitions, Selja Seppälä proposes a theoretical and methodological terminology framework to evaluate relevant features obtained from definition extraction procedures for terminographical purposes.

In the paper Linguistic realization of conceptual features in terminographic dictionary definitions, Esperanza Valero Doménech and Amparo Alcina Caudet report the result of manual analysis of specialised dictionary definitions to identify relevant conceptual features and their linguistic realisation to extract and generate definitions.

Finally, Eline Westerhout's paper entitled Definition extraction using linguistic and structural features presents a promising approach to definition extraction in Dutch using a combination of linguistic (n-grams, type of article, type of noun) and structural information (layout, position).

We hope that this workshop will provide a forum for interaction among members of different research communities, a means for participants to increase their knowledge and understanding of the potential of definition extraction and a means for promoting definition extraction as a consolidated domain of NLP.

September 2009

Gerardo Sierra
María Pozzi
Juan-Manuel Torres

Chair

Gerardo Sierra, Universidad Nacional Autónoma de México, Mexico City, Mexico

Programme Committee

Teresa Cabré, Universitat Pompeu Fabra, Barcelona, Spain
Patrick Drouin, Université de Montréal, Montréal, Canada
Thierry Fontenelle, Microsoft Corporation, USA
Adam Kilgarriff, University of Sussex, Sussex, United Kingdom
John McNaught, National Center for Text Mining, Manchester, United Kingdom
Véronique Malaisé, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
Alfonso Medina, Universidad Nacional Autónoma de México, Mexico City, Mexico
Paola Monachesi, Universiteit Utrecht, Utrecht, The Netherlands
María Pozzi, El Colegio de México, Mexico City, Mexico
Paolo Rosso, Universitat Politcnica de València, Valencia, Spain
Juan-Manuel Torres, Université d'Avignon, Avignon, France

Organising Committee

César Aguilar, Universidad Autnoma de Quertaro, Queretaro, Mexico
Rodrigo Alarcón, Universidad Nacional Autónoma de México, Mexico City, Mexico
Carme Bach, Universitat Pompeu Fabra, Barcelona, Spain
Héctor Jiménez, Universidad Autónoma Metropolitana, Mexico City, Mexico
Horacio Saggion, University of Sheffield, Sheffield, United Kingdom

Table of Contents

Workshop Program

Friday, September 18, 2009

9:20–9:30 Welcome and opening Remarks

9:30–10:00 *Language independent system for definition extraction: first results using learning algorithms*
Rosa Del Gaudio and António Branco

10:00–10:30 *A Proposal for a framework to evaluate feature relevance for terminographic definitions*
Selja Seppälä

11:00–11:30 *Linguistic realization of conceptual features in terminographic dictionary definitions*
Esperanza Valero and Amparo Alcina

11:30–12:00 *A formal scope on the relations between definitions and verbal predications*
César Aguilar and Gerardo Sierra

12:00–12:30 *Extraction of author's definitions using indexed reference identification*
Marc Bertin, Iana Atanassova and Jean-Pierre Descles

14:00–14:30 *Evolutionary algorithms for definition extraction*
Claudia Borg, Mike Rosner and Gordon Pace

14:30–15:00 *Enriching a lexicographic tool with domain definitions: problems and solutions*
María A. Barrios, Guadalupe Aguado de Cea and José Ángel Ramos

15:00–15:30 *Description and evaluation of a pattern based approach for definition extraction*
Rodrigo Alarcón, Gerardo Sierra and Carme Bach

16:00–16:30 *Definition extraction using linguistic and structural features*
Eline Westerhout

16:30–17:00 *Extracting sense-disambiguated example sentences from parallel corpora*
Gerard de Melo and Gerhard Weikum