Communications in Computer and Information Science 811

More information about this series at http://www.springer.com/series/7899

Samir Mbarki · Mohammed Mourchid
Max Silberztein (Eds.)

Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications

11th International Conference, NooJ 2017
Kenitra and Rabat, Morocco, May 18–20, 2017
Revised Selected Papers

Springer

Editors

Samir Mbarki
Université Ibn Tofail de Kénitra
Kenitra
Morocco

Mohammed Mourchid
Université Ibn Tofail de Kénitra
Kenitra
Morocco

Max Silberztein
Université de Franche-Comté
Besançon
France

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-73419-4 ISBN 978-3-319-73420-0 (eBook)
https://doi.org/10.1007/978-3-319-73420-0

Library of Congress Control Number: 2017962897

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Editors' Preface

NooJ is a linguistic development environment that provides tools for linguists to construct linguistic resources that formalize a large gamut of linguistic phenomena: typography, orthography, lexicons for simple words, multiword units and discontinuous expressions, inflectional, derivational and agglutinative morphology, local, structural dependency and transformational syntax, and semantics. For each elementary linguistic phenomenon to be described, NooJ proposes a set of computational formalisms, the power of which ranges from very efficient finite-state automata to very powerful Turing machines as well as a rich toolbox that allows linguists to construct, maintain, test, debug, accumulate, and reuse linguistic resources. This makes NooJ's approach different from most other computational linguistic tools that typically offer a unique formalism to their users and are not compatible with each other.

NooJ provides parsers that can apply any set of linguistic resources to any corpus of texts, to extract examples or counter-examples, annotate matching sequences, perform statistical analyses, and so on. Because NooJ's linguistic resources are neutral, they can also be used by NooJ's generators to produce texts. By combining NooJ's parsers and generators, one can construct sophisticated NLP (natural language processing) applications such as MT (machine translation) systems, paraphrases generators, etc.

Since its first release in 2002, NooJ has been enhanced with new features every year. Linguists, researchers in social sciences, and more generally all professionals who analyze texts have contributed to its development and participated in the annual NooJ conference. In 2013, a new version of NooJ was released, based on the JAVA technology and available to all as an open source GPL project. Moreover, several private companies are now using NooJ to construct business applications in several domains, from business intelligence to opinion analysis. To date, there are NooJ modules available for over 50 languages; more than 3,000 copies of NooJ are being downloaded each year.

The present volume contains 20 articles selected from the papers and posters presented at the International NooJ 2017 conference at the Ibn Tofail University and the ENSIAS, in Morocco. These articles are organized in four parts: "Vocabulary and Morphology" containing five articles; "Syntactic Analysis" containing six articles, "Natural Language Processing Applications" containing seven articles, and "NooJ's Future" containing two articles.

The articles in the first part involve the construction of electronic dictionaries and the description of morphological phenomena, as well as a bilingual comparison of verb tenses information that can be used by MT software:

– Masako Watabe's article "A NooJ Dictionary for the Rromani Language: Toward a NooJ-Relevant Sorting of Morphosyntactic Tags" aims at defining a set of morphosyntactic tags that can be used to describe the properties of substantive in Rromani's four dialects.

- Maximiliano Duran's article "Morphological Grammars to Generate and Annotate Verb Derivation in Quechua" presents the formalization of bi- and tri-suffixed agglutination of verbs in Quechua.
- Cristina Mota, Lucília Chacoto, and Anabela Barreiro's article "Integrating the Lexicon-Grammar of Predicate Nouns with Support Verb *fazer* into Port4NooJ" describes the formalization of 3,000 predicate nouns in Portuguese, and its application in an automatic paraphrase generator.
- Rafik Kassmi, Mohammed Mourchid, Abdelaziz Mouloudi, and Samir Mbarki's article "Processing Agglutination with a Morpho-Syntactic Graph in NooJ" shows how agglutinative morphological grammars (rather than inflectional ones) could be used to formalize Arabic tenses.
- Ilham Blanchete, Mohammed Mourchid, Samir Mbarki, and Abdelaziz Mouloudi's article "Formalizing Arabic Inflectional and Derivational Verbs Based on Root and Pattern Approach Using NooJ Platform" describes a system of dictionary to formalize the Arabic vocabulary, based on roots rather than on lemmas.

The articles in the second part describe the construction of sophisticated syntactic grammars and the use of such grammars to help students learn Italian or Spanish as a second language:

- Nadia Ghezaiel Hammouda and Kais Haddar's article "Arabic NooJ Parser: Nominal Sentence Case" presents a formalization of Arabic nominal sentences and its implementation using NooJ grammars.
- Said Bourahma, Mohammed Mourchid, Samir Mbarki, and Abdelaziz Mouloudi's article "The Parsing of Simple Arabic Verbal Sentences Using NooJ Platform" presents a parser for simple Arabic verbal sentences that uses a dependency grammar to produce parse trees.
- Krešimir Šojat, Kristina Kocijan, and Božo Bekavac's article "Identification of Croatian Light Verb Constructions with NooJ" presents a set of linguistic resources used to formalize light verb constructions in Croatian.
- Maddalena della Volpe, Annibale Elia, and Francesca Esposito's article "Semantic Predicates in the Business Language" presents a set of syntactic grammars that recognize simple sentences in the language used for business, and produce the corresponding semantic predicates.
- Ignazio Mauro Mirto and Emanuele Cipolla's article "Invalid Syntax: NooJ Assisted Automatic Detection of Errors in Auxiliaries and Past Participles in Italian" presents a formalization of compound tenses that can be used to help Italian learners select auxiliary verbs and apply past participle agreements correctly.
- Andrea Rodrigo, Silvia Reyes, and Rodolfo Bonino's article "Some Aspects Concerning the Automatic Treatment of Adjectives and Adverbs in Spanish: A Pedagogical Application of the NooJ Platform" presents a formalization of Spanish Adjective phrases and its pedagogical applications to help Spanish learners.

The articles in the third part involve the construction of software applications capable of parsing and extracting meaningful information:

- Azeddin Rhazi and Ali Boulaalam's article "Corpus-Based Extraction and Translation of Arabic Multi-Words Expressions (MWEs)" presents a hybrid system capable of extracting Arabic multi-word expressions automatically from bilingual corpora.

- Hajer Cheikhrouhou's article "The Automatic Translation of French Verbal Tenses to Arabic Using the Platform NooJ" shows the differences between the Arabic and the French verbs tense systems, and proposes a bilingual set of linguistic resources to translate automatically conjugated French verbs of communication and movement to Arabic.
- Tong Yang's article "Automatic Extraction of the Phraseology Through NooJ" presents a system capable of automatically recognizing and extracting multiword units and semi-frozen expressions from a corpus of French texts in the culinary domain.
- Yuraś Hiecevič, Alena Kryvaltsevich, Nastassia Kazloŭskaja, Anastasija Drahun, Jaŭhienija Zianoŭka, and Aliaksandr Ščarbakoŭ's article "Sentiment Analysis Algorithms for the Belarusian NooJ Module in the Touristic Sphere" presents a software application and its linguistic resources capable of performing automatic sentiment analyses of touristic texts.
- Imen Ennasri, Sondes Dardour, Héla Fehri, and Kais Haddar's article "Question–Response System Using the NooJ Linguistic Platform" presents a question-answering application in the medical domain capable of parsing users' questions in Arabic, which retrieves the potential answers in two corpora of texts: one in Arabic and one in English.
- Mario Monteleone, Raffaele Guarasci, and Alessandro Maisto's article "NooJ Morphological Grammars for Stenotype Writing" presents a system that automatically detects and correct typos in stenotype writing.
- Carmela Scoppetta, Anastasia Alfieri, Flavio Merenda, Sonia Lay, Annalisa Colasanto, and Raffaele Manna's article "From Language to Social Perception of Immigration" presents a system that automatically analyzes a corpus of journalistic texts and a corpus of comments on these texts, on the topic of immigration in Italy.

The articles in the last part involve the development of two companion applications for NooJ: a Web-based graphical interface as well as an industrial-strong linguistic engine:

- Zineb Gotti, Samir Mbarki, Sara Gotti, and Naziha Laaz's article "Nooj Graphical User Interfaces Modernization" presents a theoretical approach to software modernization, and applies it to modernize NooJ's graphical user interface.
- Max Silberztein's article "A New Linguistic Engine for NooJ: Parsing Context-Sensitive Grammars with Finite-State Machines" presents a set of algorithms that can be used to apply linguistic resources developed with NooJ in a very efficient way.

This volume should be of interest to all users of the NooJ software because it presents the latest development of its linguistic resources as well as its future enhancements.

Linguists as well as computational linguists who work on Arabic, Belarusian, Croatian, French, Italian, Portuguese, Quechua, Rromani, or Spanish will find advanced, up-to-the-minute linguistic studies for these languages in this volume.

We think that the reader will appreciate the importance of this volume, both for the intrinsic value of each linguistic formalization and the underlying methodology as well as for the potential for developing NLP applications along with linguistic-based corpus processors in the social sciences.

December 2017

Samir Mbarki
Mohammed Mourchid
Max Silberztein

Organization

Program Committee

Max Silberztein (Program Chair)	Université de Franche-Comté, France
Xavier Blanco	Autonomous University of Barcelona, Spain
Mohammed El Hannach	Sidi Mohammed Ben Abdellah University, Morocco
Mohammed Essaidi	ENSIAS, Morocco
Héla Fehri	University of Gabes, Tunisia
Yuras Hetsevich	United Institute of Informatics Problems, Belarus
Kristina Kocijan	University of Zagreb, Croatia
Svetla Koeva	University of Sofia, Bulgaria
Peter Machonis	Florida International University, USA
Samir Mbarki	Ibn Tofail University, Morocco
Slim Mesfar	University of Manouba, Tunisia
Mohammed Mourchid	Ibn Tofail University, Morocco
Mario Monteleone	University of Salerno, Italy
Johanna Monti	University of Sassari, Italy
Fadoua Ataa Allah	Institut Royal de la Culture AMazighe, Morocco
Jan Radimský	University of South Bohemia, Czech Republic
Azeddine Rhazi	Cadi Ayyad University, Morocco
François Trouilleux	Université Blaise-Pascal, France

Contents

Vocabulary and Morphology

Syntactic Analysis

Natural Language Processing Applications

NooJ's Future

List of Contributors

Anastasia Alfieri Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Anabela Barreiro L2F/INESC-ID, Lisbon, Portugal

Božo Bekavac Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

Ilham Blanchete MIC Research Team, Laboratory MISC, IbnTofail University, Kenitra, Morocco

Rodolfo Bonino IES N° 28 "Olga Cossettini", Rosario, Argentina

Ali Boulaalam FLSF, Med Ben Abdellah University, Fes, Morocco

Said Bourahma MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Lucília Chacoto Universidade do Algarve, Faro, Portugal; CLUL, Lisbon, Portugal

Hajer Cheikhrouhou University of Sfax, LLTA, Sfax, Tunisia

Emanuele Cipolla Consiglio Nazionale delle Ricerche, Palermo, Italy

Annalisa Colasanto Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Sondes Dardour MIRACL Laboratory, University of Sfax, Sfax, Tunisia

Anastasija Drahun The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus

Maximiliano Duran Université de Franche-Comté, Besançon, France

Annibale Elia Department of Political, Social and Communication Sciences, University of Salerno, Fisciano, Italy

Imen Ennasri MIRACL Laboratory, University of Sfax, Sfax, Tunisia

Francesca Esposito Department of Political, Social and Communication Sciences, University of Salerno, Fisciano, Italy

Héla Fehri MIRACL Laboratory, University of Sfax, Sfax, Tunisia

Sara Gotti Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Zineb Gotti Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Raffaele Guarasci Dipartimento di Scienze Politiche, Sociali e della Comunicazione, Università degli Studi di Salerno, Salerno, Italy

Kais Haddar MIRACL Laboratory, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia

Nadia Ghezaiel Hammouda Miracl Laboratory, Higher Institute of Computer and Communication Technologies of Hammam Sousse, Sousse, Tunisia

Yuras Hetsevich The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus

Rafik Kassmi MISC, Ibn Tofail University, Kénitra, Morocco

Nastassia Kazloŭskaja The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus

Kristina Kocijan Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

Alena Kryvaltsevich The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus

Naziha Laaz Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Sonia Lay Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Alessandro Maisto Dipartimento di Scienze Politiche, Sociali e della Comunicazione, Università degli Studi di Salerno, Salerno, Italy

Raffaele Manna Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Samir Mbarki MIC Research Team, MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Flavio Merenda Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Ignazio Mauro Mirto Università di Palermo, Palermo, Italy

Mario Monteleone Dipartimento di Scienze Politiche, Sociali e della Comunicazione, Università degli Studi di Salerno, Salerno, Italy

Cristina Mota L2F/INESC-ID, Lisbon, Portugal

Abdelaziz Mouloudi MIC Research Team, MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Mohammed Mourchid MIC Research Team, MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Silvia Reyes Facultad de Humanidades y Artes, Universidad Nacional de Rosario, Rosario, Argentina

Azeddin Rhazi FLSH, Qadi Ayyad University, Marrakech, Morocco

Andrea Rodrigo Facultad de Humanidades y Artes, Universidad Nacional de Rosario, Rosario, Argentina

Aliaksandr Ščarbakoŭ The Belarusian State University of Informatics and Radio-electronics, Minsk, Belarus

Carmela Scoppetta Dipartimento di Scienze Sociali Politiche e della Comunicazione, Università di Salerno, Fisciano, Italy

Krešimir Šojat Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

Max Silberztein Université de Franche-Comté, Besançon, France

Maddalena della Volpe Department of Business, Management and Innovation System, University of Salerno, Fisciano, Italy

Masako Watabe Paris-Sorbonne University, Paris, France

Tong Yang DILTEC (Didactiques des langues, des textes et des cultures), Université Sorbonne Nouvelle (Paris 3), Paris, France

Jaŭhienija Zianoŭka The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus