Communications in Computer and Information Science

1153

Commenced Publication in 2007

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

Héla Fehri · Slim Mesfar · Max Silberztein (Eds.)

Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications

13th International Conference, NooJ 2019
Hammamet, Tunisia, June 7–9, 2019
Revised Selected Papers

Springer

Editors
Héla Fehri
University of Sfax
Sfax, Tunisia

Slim Mesfar
Manouba University
Manouba, Tunisia

Max Silberztein
University of Franche-Comté
Besançon, France

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-030-38832-4 ISBN 978-3-030-38833-1 (eBook)
https://doi.org/10.1007/978-3-030-38833-1

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

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, phrase-structure and dependency grammars, as well as transformational and semantic grammars. For each linguistic phenomenon to be described, NooJ proposes a set of computational formalisms, the power of which ranges from very efficient finite-state automata (that process regular grammars) to very powerful turing machines (that process unrestricted grammars). NooJ also contains a rich toolbox that allows linguists to construct, maintain, test, debug, accumulate, and share 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 analyzes, 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, abstracts and paraphrases generators, etc.

Since its first release in 2002, several private companies have used NooJ's linguistic engine 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 140,000 copies of NooJ have been downloaded.

NooJ has also been enhanced with new features to respond to the needs of researchers who need to analyze texts in various domains of Human and Social Sciences (history, literature and political studies, psychology, sociology, etc.), and more generally of all the professionals who analyze texts. In 2013, a new version for NooJ was released, based on the JAVA technology and made available to all as an open source GPL project, distributed by the European Metashare platform.

This volume contains 18 articles selected from the papers and posters presented at the International NooJ 2019 conference in Hammamet, Tunisia. The following articles are organized in three parts: "Development of Linguistic Resources" contains six articles; "NLP Applications" contains five articles; and "NooJ for the Digital Humanities" contains seven articles.

The six articles in the first part involve the construction of electronic dictionaries and grammars to formalize various linguistic phenomena:

We would like to thank Ms. Olfa Ben Amor from the University of Monastir for her editing work on this volume.

- In their article “Recognition of Arabic Phonological Changes by Local Grammars in NooJ,” Rafik Kassmi, Mohammed Mourchid, Abdelaziz Mouloudi, and Samir Mbarki present a set of local grammars used to recognize Arabic phonological changes.
- In “Lexicon-Grammar Tables Development for Arabic Psychological Verbs,” Asmaa Amzali, Asmaa Kourtin, Mohammed Mourchid, Abdelaziz Mouloudi, and Samir Mbarki show how they organized and formalized Arabic psychological verbs using lexicon-grammar tables.
- In “The Identification of English Non-Finite Structures Using NooJ Platform,” Ben Amor Olfa and Faiza Derbel present a set of local grammars used to identify non-finite clauses in an English corpus of business-related texts with a impressive recall rate (96%).
- In “Automatic Recognition and Translation of Polysemous Verbs Using the Platform NooJ,” Hajer Cheikhrouhou presents the linguistic information associated with communication and movement verbs extracted from Dubois & Dubois-Charlier’s LVF (*Les Verbes Français*) dictionary and implements a set of NooJ grammars to identify and disambiguate them. Once disambiguated, they can be safely translated into Arabic.
- In “Negation of Croatian Nouns,” Natalija Žanpera, Kristina Kocijan, and Krešimir Šojat present a set of morphological grammars for Croatian that can recognize the various prefixes used to express negation and compute the words polarity accordingly.
- Finally, in “The Automatic Generation of NooJ Dictionaries from Lexicon-Grammar Tables,” Asmaa Kourtin, Asmaa Amzali, Mohammed Mourchid, Abdelaziz Mouloudi, and Samir Mbarki present a set of computer tools aimed at helping linguists to implement the wealth of information contained in lexicon-grammar tables in the form of NooJ electronic dictionaries.

The five articles in the second part describe the implementation of spectactular NLP software applications:

- In “The Data Scientist on LinkedIn: Job Advertisement Corpus Processing with NooJ,” Maddalena della Volpe and Francesca Esposito present an application capable of parsing a large number of job advertisements collected by LinkedIn to extract skills required by companies and organizations.
- In “Recognition and Analysis of Opinion Questions in Standard Arabic,” Essia Bessaies, Slim Mesfar, and Henda Ben Ghezala show how a question/answering system for Arabic can be structured in four processes (Question Analysis, Text Segmentation, Passage Retrieval, and Answer Extraction) and how NooJ can be used to perform tasks in these four processes.
- In “Disambiguation for Arabic Question-Answering System,” Sondes Dardour, Héla Fehri, and Kais Haddar focus on the problem of solving ambiguities in Medical Question-Answering systems for Arabic. Both Arabic written texts in general, and medical questions in particular are highly ambiguous; the authors present a set of local grammars used to solve these two types of ambiguity.
- In “A NooJ Tunisian Dialect Translator,” Roua Torjmen, Nadia Ghezaiel Hammouda, and Kais Haddar show how they built an automatic Tunisian dialect to

modern standard Arabic translator, using bilingual dictionaries, morphological grammars, and local translation grammars.
- In “Automatic Text Generation: How to Write the Plot of a Novel with NooJ,” Mario Monteleone presents a system capable of generating novel plot templates automatically using dictionaries that use narrative-related tags (such as “cloth”, “game”, “currency”) and local grammars.

The seven articles in the last part involve applications of NooJ in the Digital Humanities, three pedagogical applications of NooJ and three analysis of various discourses, university communication, Amazon reviews, and detection of hate crime and terroris threats:

- In “Arabic Learning Application to Enhance the Educational Process in Moroccan Mid-High Stage using NooJ Platform,” Ilham Blanchete, Mohammed Mourchid, Samir Mbarki, and Abdelaziz Mouloudi present a pedagogical application that helps students analyze nouns, in terms of linguistic information (e.g. lemma, root, semantic domain, etc.).
- In “Causal Discourse Connectors in the Teaching of Spanish as a Foreign Language (SLF) for Portuguese Learners using NooJ,” Andrea Rodrigo, Silvia Reyes, Cristina Mota, and Anabela Barreiro present a pedagogical application that helps teach Spanish causal discourse connectors to Portuguese students.
- In “Construction of Educational Games with NooJ,” Héla Fehri and Ines Ben Messaoud present two educational games: ProMoNooJ (a multilingual game in which players must classify terms) and AlphaNooJ (a word-building game). While playing, the users actually learn terms and linguistics concepts.
- In “Mining Entrepreneurial Commitment in University Communication: Evidence from Italy,” Maddalena della Volpe and Francesca Esposito process websites of 91 Italian Universities to analyze how they communicate their strategic intentions.
- In “Dealing with Producing and Consuming Expressions in Italian Sentiment Analysis,” Nicola Cirillo has parsed a corpus of product reviews from Amazon in order to extract sentiment terms and compute their polarities.
- In “Detecting Hate Speech Online: A Case of Croatian,” Kristina Kocijan, Lucija Košković, and Petra Bajac present a system capable of finding and categorizing hate speech in Croatian online texts from Facebook news pages. The crucial issue to solve was to adapt the standard Croatian linguistic resources to parse texts that are not necessarily used in informal communication.
- In “Rule Based Method for Terrorism, Violence and Threat Classification: Application to Arabic Tweets,” Elahsoumi Wissam, Boujelben Ines, and Keskes Iskander have parsed tweets in Arabic in order to find inappropriate messages promoting terrorism, violent messages, and threats automatically.

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 a large variety of applications.

Linguists as well as Computational Linguists who work on Arabic, Croatian, French, Italian, Portuguese, Argentinian Spanish, or Tunisian dialects will find advanced, up-to-date linguistic studies for these languages.

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 of developing NLP applications along with linguistic-based corpus processors in the Social Sciences.

June 2019

Héla Fehri
Slim Mesfar
Max Silberztein

Contents

Development of Linguistic Resources

Natural Language Processing Applications

NooJ for the Digital Humanities