Communications in Computer and Information Science 607

Commenced Publication in 2007
Founding and Former Series Editors:
Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

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

Tatsiana Okrut · Yuras Hetsevich
Max Silberztein · Hanna Stanislavenka (Eds.)

Automatic Processing of Natural-Language Electronic Texts with NooJ

9th International Conference, NooJ 2015
Minsk, Belarus, June 11–13, 2015
Revised Selected Papers

Springer

Editors
Tatsiana Okrut
United Institute of Informatics Problems
Minsk
Belarus

Max Silberztein
Université de Franche-Comté
Paris
France

Yuras Hetsevich
United Institute of Informatics Problems
Minsk
Belarus

Hanna Stanislavenka
United Institute of Informatics Problems
Minsk
Belarus

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-42470-5 ISBN 978-3-319-42471-2 (eBook)
DOI 10.1007/978-3-319-42471-2

Library of Congress Control Number: 2016944166

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

NooJ is a linguistic development environment that provides tools for linguists to construct and accumulate linguistic resources that formalize a large gamut of linguistic phenomena: typography, orthography, lexicons for simple words, multiword units and discontinuous expressions, inflectional and derivational morphology, local, structural, and transformational syntax, and semantics. In 2011, the European Metanet initiative endorsed NooJ, and it is now distributed as an open source software on the European Metashare platform. More than 3,000 copies of NooJ are downloaded each year.

Currently, there are NooJ modules available for over 50 languages. Linguists, researchers in social sciences, and more generally all professionals who analyze large corpora of texts have contributed to its development, year after year, and participated in the annual international NooJ conference.

NooJ 2015 received 51 submissions. The present volume contains 20 articles selected from the 35 papers that were presented at the International NooJ 2015 Conference, which was held during June 11–13 at the United Institute of Informatics Problems of the National Academy of Sciences in Minsk, Belarus. These articles are organized in three parts: "Corpora, Vocabulary and Morphology" contains four articles; "Syntax and Semantics" contains seven articles; "Applications" contains nine articles.

The articles in the first part involve the construction of large texts and dictionaries:

- Ivan Reentovich, Yuras Hetsevich, Valery Varanovich, Evgenia Kachan, and Hanna Kazlouskaya's article "First One Million Corpora for Belarusian NooJ Module" describes the construction of a large corpus and its organization in sections (fiction, medical, scientific texts, etc.) that has been parsed with NooJ in order to perform several experiments.
- Dhekra Najar and Slim Mesfar's article "A Large Terminological Dictionary of Arabic Compound Words" presents a new dictionary for Arabic that formalizes the morphology of multiword expressions in Arabic and classifies them into 20 semantic domains.
- Maximiliano Duran's article "The Annotation of Compound Suffixation Structure of Quechua Verbs" presents a morphological set of rules that expand a dictionary of 1,500 verbs in Quechua by adding sequences of suffixes to each verb. The resulting dictionary has been applied to a corpus of texts to evaluate the precision and recall of the morphological rules.
- Serena Pelosi's article "Morphological Relations for the Automatic Expansion of Italian Sentiment Lexicons" proposes a method to help construct a lexicon of sentiment terms rapidly, using a basic dictionary of 5,000 adjectives of sentiments, suffixation rules, as well as syntactic clues in order to detect new terms automatically in texts.

The articles in the second part involve the construction of syntactic and semantic grammars:

- Max Silberztein's article "Transformational Analysis of Transitive Sentences" discusses the feasibility of implementing a large-coverage set of transformational rules in order to take into account all the possible syntactic operations that might be applied to an elementary sentence.
- Xavier Blanco Escoda's article "A Hierarchy of Semantic Labels for Spanish Dictionaries" presents a comprehensive system of semantic hierarchy used to describe dictionary entries in Spanish. Currently, this system contains 700 semantic labels that are actual words rather than abstractions.
- Yuras Hetsevich, Tatsiana Okrut, and Boris Lobanov's article "Grammars for Sentence into Phrase Segmentation: Punctuation Level" describes a prosodic system that can annotate phrases according to four types of intonation: finality, non-finality, interrogation, and exclamation.
- Mario Monteleone's article "Local Grammars and Formal Semantics: Past Participles vs. Adjectives in Italian" presents a set of local grammars that can solve ambiguities between adjectives and past participles in Italian.
- Kristina Kocijan and Sara Librenjak's article "Recognizing Verb-Based Croatian Idiomatic MWUs" presents a set of grammars that can detect idioms in Croatian and evaluates it by applying it to a Web corpus sample.
- Paula Carvalho, Cristina Mota, and Anabela Barreiro's article "Paraphrasing Human Intransitive Adjective Constructions in Port4NooJ" presents the new linguistic resources of the Port4NooJ system, which aims at generating paraphrases in Portuguese automatically. These new resources include 15 lexicon-grammar tables that provide extremely rich information.
- Nadia Ghezaiel and Kais Haddar's article "Study and Resolution of Arabic Lexical Ambiguity Through the Transduction on Text Automaton" presents a set of local grammars that are applied to Arabic texts in cascade in order to solve various types of ambiguities.

The articles in the third part describe various NLP software applications built with NooJ:

- Vadim Zahariev, Stanislau Lysy, Alena Hiuntar, and Yury Hetsevich's article "Grapheme-to-Phoneme and Phoneme-to-Grapheme Conversion in Belarusian with NooJ for TTS and STT Systems" proposes a grammar-based system to automatically transcribe oral texts (transcribed in a phonetic representation) into an orthographical transcription, using syntactic grammars whose linguistic units are phonemes rather than words.
- Maria Pia di Buono's article "Semi-Automatic Indexing and Parsing Information on the Web with NooJ" presents a search engine capable of parsing users' queries in natural language, using a representation model both for the users' queries and for the documents. The system is tested on the Italian Wikipedia database.

- Lesia Kaigorodova, Yuras Hetsevich, Kiryl Nikalaenka, U.A. Sychou, R.A. Prakapovich, and S. Gerasuto's article "Language Modelling for Robots–Human Interaction" presents a system that manages robot–human interactions, based on a simplified language whose syntax and semantics are formalized using NooJ grammars.
- Alessandro Maisto and Raffaele Guarasci's article "Morpheme-Based Recognition and Translation of Medical Terms" proposes a system capable of analyzing complex medical terms in English (e.g., "otolaryngology") and in Italian, using morpho-semantic rules.
- Yuras Hetsevich and Julia Borodina's article "Using NooJ for Processing of Satellite Data" presents a system that can parse satellite telemetry data and translate it into Belarusian.
- Mohamed Aly Fall Seideh, Héla Fehri, Kais Haddar, and Abdelmajid Ben Hamadou's article "Named Entity Recognition from Arabic–French Herbalism Parallel Corpora" presents a system that uses parallel texts in Arabic and French and uses bilingual dictionaries as well as syntactic grammars to find the corresponding translation of terms in Botanical medicine.
- Alena Veka and Yauheniya Yakubovich's article "Automatic Translation from Belarusian into Spanish Based on Using NooJ's Linguistic Resources" describes a Belarusian–Spanish dictionary that can be used to automatically translate words, idiomatic expressions, and phrases, from Belarusian to Spanish.
- Farida Yamouni's article "A French–Tamazight MT System for Computer Science" describes the construction of an MT system that can process multiword terms in the vocabulary of computer science, using a French–Tamazight electronic dictionary as well as translation grammars.
- Valérie Collec Clerc's article "Mixed Prolog and NooJ Approach in Japanese Benefactive Constructions" presents a system that recognizes Japaneses sentences that contain benefactive auxiliaries, using a set of NooJ linguistic resources (dictionaries and grammars for named entities), associated with a set of Prolog pragmatic rules that take the relationship between the speaker and the listener into account in order to produce the resulting correct sentences.

This volume should be of interest to all users of the NooJ platform because it presents the latest development of the software, its latest linguistic resources, as well as new software applications.

In particular, linguists and computational linguists who work on Arabic, Belarusian, Croatian, English, French, Italian, Japanese, Portuguese, Spanish, Quechua, or Tamazight will find in this volume state-of-the-art 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 for developing NLP applications.

June 2016 The Editors

Organization

The International NooJ 2015 Conference was organized by the United Institute of Informatics Problems of the National Academy of Sciences of Belarus in cooperation with Université de Franche-Comté and the NooJ International Association.

General Conference Chair

Max Silberztein	Université de Franche-Comté, France

Organizing Committee

Yuras Hetsevich	United Institute of Informatics Problems, Belarus
Max Silberztein	Université de Franche-Comté, France
Barys Lobanov	United Institute of Informatics Problems, Belarus
Tatsiana Okrut	United Institute of Informatics Problems, Belarus
Julia Baradzina	United Institute of Informatics Problems, Belarus
Dzmitry Dzenisiuk	United Institute of Informatics Problems, Belarus
Alena Hiuntar	United Institute of Informatics Problems, Belarus
Stanislau Lysy	United Institute of Informatics Problems, Belarus
Lesia Kaigorodova	United Institute of Informatics Problems, Belarus

Scientific Committee

Xavier Blanco	Autonomous University of Barcelona, Spain
Krzysztof Bogacki	University of Warsaw, Poland
Héla Fehri	University of Gabès, Tunisia
Yuras Hetsevich	United Institute of Informatics Problems, Belarus
Svetla Koeva	University of Sofia, Bulgaria
Peter Machonis	Florida International University, USA
Slim Mesfar	University of Manouba, Tunisia
Johanna Monti	University of Sassari, Italy
Max Silberztein	Université de Franche-Comté, France
Marko Tadic	University of Zagreb, Croatia
François Trouilleux	Université Blaise-Pascal, France
Simonetta Vietri	University of Salerno, Italy
Igor Sovpel	Belarusian State University, Belarus
Alexander Zubov	Minsk State Linguistic University, Belarus
Vladimir Golenkov	Belarusian State University of Informatics and Radioelectronics, Belarus
Alena Hiuntar	United Institute of Informatics Problems, Belarus
Stanislau Lysy	United Institute of Informatics Problems, Belarus
Lesia Kaigorodova	United Institute of Informatics Problems, Belarus


Оипи


ELLIADD


Association Internationale Sans But Lucratif
www.nooj-association.org

Contents

Application