UNITED INSTITUTE OF INFORMATICS PROBLEMS
OF THE NATIONAL ACADEMY OF SCIENCES OF BELARUS

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts “NooJ’2015”

NOOJ 2015

Abstracts

June 11–13, 2015, Minsk, Belarus

Minsk
UIIP NASB
2015

УДК 004.91

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015" : Abstracts (11–13 June, 2015, Minsk, Belarus). – Minsk : UIIP NASB, 2015. – 80 p.
ISBN 978-985-6744-89-4.

This volume contains the abstracts of the International conference "NooJ 2015". The research presented covers different aspects of natural language processing using NooJ, including formalizing such levels of linguistic phenomena as syllabification, phonemic and prosodic transcription, multiword units and discontinuous expressions, local and structural syntax; transformational syntax and paraphrase generation, semantic analysis and machine translation, etc.

Abstracts are published in the form presented by authors.

У дадзеным зборніку прадстаўлены тэзісы дакладаў Міжнароднай канферэнцыі "NooJ 2015". Разглядаюцца розныя аспекты апрацоўкі натуральнай мовы з выкарыстаннем лінгвістычнага асяроддзя распрацоўкі NooJ, улічваючы фармалізаванне такіх напрамкаў лінгвістычнага аналізу як склададзяленне, фанетычная і прасадычная транскрыпцыі, устойлівыя выразы і дыскрэтныя слоўныя канструкцыі, лакальны і структурны сінтаксісы, трансфармацыйны сінтаксіс і перафразаванне, семантычны аналіз і машынны пераклад і г. д.

Тэзісы друкуюцца ў выглядзе, пададзеным аўтарамі.

Scientific Editors:
DSc in Engineering B.M. Lobanov,
PhD in Engineering Yu.S. Hetsevich

ISBN 978-985-6744-89-4

CONTEXT-SENSITIVE HOMOGRAPH DISAMBIGUATION WITH NOOJ IN BELARUSIAN AND RUSSIAN ELECTRONIC TEXTS

T. Okrut [1], B. Lobanov [1], Y. Yakubovich [2]
[1] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[2] Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: tatberrie@gmail.com

When we read, we may encounter words having different phonological representations associated with singular orthographic representation. In speech synthesis, disambiguation of such words, or homograph disambiguation, serves an obstacle to overcome at the stage of text preprocessing. There are several major types of homographs we deal with in Belarusian and Russian: different lexemes of the same part of speech, different forms of the same lexeme and different lexemes of different parts of speech. Moreover, these homographic groups may be divided into the subgroups on the base of grammatical similarity of homographic word pairs [1]. Elements in such pairs differ at least in one grammatical feature, which influences the stress position in a word. For example, in the following homographic pairs of the group "different forms of the same lexeme" elements differ in number:

ГО́ДА (singular, "year") – ГОДА́ (plural, "years"),
О́ЗЕРА (singular, "lake") – ОЗЁРА (plural, "lakes").

Such similarity allows developing of one context-sensitive disambiguation algorithm for a number of homographic pairs at once. The authors have already developed a Russian syntactic NooJ grammar for disambiguation of 58 homographs referring to the homographic subgroup "Singular nouns of masculine or neuter gender in genitive case – Plural nouns in accusative or nominative case". Therefore, the goal of this research is to improve the grammar mentioned above and to develop a similar Belarusian disambiguation grammar using a context-sensitive approach.

References

1. Вырашэнне амаграфіі з дапамогай NOOJ для больш чым 50 амографаў рускай мовы / Т.І. Окрут [і інш.] // Контрастивные исследования и прикладная лингвистика: материалы Междунар. науч. конф., Минск, 29–30 окт. 2014 г. : в 2 ч. / М-во образования Респ. Беларусь, Минский гос. лингв. ун-т; редкол. : А.В. Зубов [и др.]. – Минск, 2014. – Ч. II. – С. 83–87.

CONTENTS