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

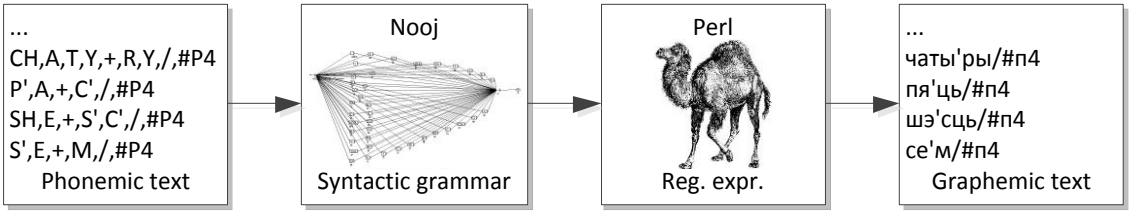
GRAMMARS FOR MAKING WRITTEN ORTHOGRAPHIC WORDS FROM TRANSCRIBED SPOKEN LANGUAGE

A. Hiuntar, V. Zahariev
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: lena205593@gmail.com, zahariev@bsuir.by

There are two main problems for natural language processing on the transition step between morphological and phonetic levels of speech. The first one is a transformation from signed words to phonetically transcription for further processing with code signals corresponding to phonetic units [1]. This task is very common for example in text-to-speech synthesis systems [2]. The second one is an inverse problem: building of written orthographic words from transcribed spoken language. This task is very important item within the framework of automatic speech recognition (ASR) systems. In large vocabulary ASR systems, it is hard if not impossible to train separate statistical models for all words. In such systems, words are described as sequences of phonemes in a pronunciation lexicon, and statistical modeling is applied to phonemic units. In such case the problem of correct phoneme to grapheme transformation has a great significance.

In our work we suggest solution of this problem based on Nooj framework [3]. A key feature of our finding is the use of syntactic level grammar for processing of phonetic units that are in fact atomic linguistic units of morphological, not syntactic level of speech. This fact allows us to build a more flexible model for phoneme-to-grapheme conversion, given the great opportunity of syntactic grammars, within the constraints imposed by the form of the incoming flow of phonetic units from the ASR speech analysis module. A higher level of abstraction, given by syntactic grammars, makes it possible to handle blank positions and pauses in the phonetic text, which is actually difficult to achieve, through the using of morphological grammars. Since its use implies a mandatory meaningful units, which can't be, for example, a space or any absence of the sign.

The general sequence of phonemic processing text as follows (fig.). Source phonemic (or allophonic) text derived from the resolver module of the speech signal is converted into a text object of Nooj system. This text is treated with a prepared set of syntactic level grammars. With these grammars performed a linguistic analysis of the text, and the results are exported from Nooj in XML format. Next, using a special Perl script is processing the given file, using a set of regular expression searching elements that Nooj recognized as grapheme units. On the output we get a graphemic text.



General processing scheme

Construction of grammars based on rules-based phoneme-to-grapheme conversion, both for basic variants of these types of transformations and complex options take into account the right and left contexts are considered in this report. Study and analysis of performance data grammars in terms of the number of errors, computational complexity and speed of the proposed algorithms convert the corresponding language resources are made. Practical results of this work we are going to use in natural language dialog module within mobile robot.

References

1. Dutoit, T. Applied Signal Processing: A Matlab-based Proof of Concept / T. Dutoit, F. Marques. – Springer Science, Buisness Media, LLC, 2009. – 456 p.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 03.01.2015.

3. Transcription Generator [Electronic resource]. – 2013. – Mode of access : http://corpus.by/transcriptionGenerator. – Date of acces : 03.01.2015.

CONTENTS