Filozofická fakulta
Faculty of Philosophy

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Université de Franche-Comté, Besançon

The NOOJ Association

NOOJ 2016

International Conference

Book of Abstracts

České Budějovice (Czech Republic)

9-11 June 2016

Scientific Committee

Xavier Blanco (Autonomous University of Barcelona, Spain)

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)

Mario Monteleone (University of Salerno, Italy)

Johanna Monti (University of Sassari, Italy)

Karel Pala (Masaryk University, Czech republic)

Vladimír Petkevič (Charles University, Prague)

Jan Radimský (University of South Bohemia, Czech republic)

Max Silberztein (Université de Franche-Comté, France)

Marko Tadić (University of Zagreb, Croatia)

François Trouilleux (Université Blaise-Pascal, France)

Organization Committee

Jan Radimský (University of South Bohemia, Czech Republic)

Max Silberztein (Université de Franche-Comté, France)

Zuzana Nevěřilová (Masaryk University, Czech Republic)

Petr Kos (University of South Bohemia, Czech Republic)

Mario Monteleone (Università degli studi di Salerno, Italy)

NOOJ 2016 International Conference – Book of Abstracts

Edited by Jan Radimský

Published by the University of South Bohemia in České Budějovice (2016)

Table of Contents

Knowledge-based system for the solution of the phoneme-to-grapheme problem of belarusian speech recognition using NOOJ

Lesia Kaigorodova
United Institute of Informatics Problems of the NAS of Belarus, Minsk, BELARUS
lesia.piatrouskaya@gmail.com

Anna Karpenko
United Institute of Informatics Problems of the NAS of Belarus, Minsk, BELARUS
rfe.karpenko@gmail.com

Yury Hetsevich
United Institute of Informatics Problems of the NAS of Belarus, Minsk, BELARUS
yury.hetsevich@gmail.com

Speech recognition for some groups of languages, i.e. Slavic, is still of poor performance compared to some other widely used languages, i.e. English, Spanish, French, etc. Some new methods for language modeling, phone-to-word task and phoneme recognition should be proposed for speech recognition of Belarusian, Ukrainian and even for Russian language. For Russian language this problem is being actively solved by some academia and speech technology companies, but for Belarusian language (for which this work is addressed) it is still a challenging task.

The classic design of the decoder that is the core of any speech recognition system is represented by a chain of finite state transducers that solve complex phoneme-to-grapheme problem (P2G) as well as language modeling task.

The aim of the proposed discussion is to describe how the transcribed data of speech corpus are collected and organized using NooJ [1] in order to solve phoneme-to-grapheme problem for speech recognition task for Belarusian language.

The overall problem of creation of transcribed data is split into two simpler problems, i.e. generation of lexicon for the speech corpus and generation of triphone database for the same corpus based on the obtained lexicon and corresponding text corpus.

Lexicon is the dictionary that is based on unique words of the text corpus that corresponds to the speech corpus and unique phonemic transcriptions of these words. Phonemic transcription of units for the lexicon is generated using the text corpus and Text-to-Speech Synthesizer [2]. NooJ Syntactic and Morphological grammars and some Python script are used to process text corpus and phonemic transcriptions in order to obtain lexicon.

Generation of triphone database is based on the transcription of words from obtained lexicon and 2-word and 3-word sequences from the text corpus. Here NooJ Syntactic grammars are also used to process the data in order to generate the database. This database is the source for creating weighted finite state transducer that is supposed to solve phoneme-to-grapheme problem.

Given that all the components, i.e. speech and text corpus of Belarusian language, lexicon and triphone database, are collected and organized in a specific format, it can be considered that the basis for the NooJ knowledge-based system for the P2G problem of Belarusian speech recognition task is created.

References

Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net – Date of access: 01.02.2016.

Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – 2013. – Mode of access: http://corpus.by/tts3. – Date of access: 01.02.2016.