Automatic Text-checking for Slovak Language

Ondrej ČIČKÁN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xcickano@fiit.stuba.sk

We encounter text-checking almost in every word processing program, web browser or other applications. Late detection of spelling mistakes in the curriculum vitae, book or diploma work can be unpleasant for the author. The function of text-checking tool is to automatically detect these errors and propose corrections. It may also be useful in other programs, which require that the input text is written correctly.

Our goal is to offer a tool that automatically checks a text in the Slovak language and detects the largest possible percentage of error. We decided to use a statistical method, where we use language and error model. These models will help us to choose the correct word from list of multiple suggested corrections for misspelled word. This method also allows us to correct the real-word errors.

Our solution is based on existing tools for the text-checking *Korektor* developed at Charles University in Prague by Michal Richter [1]. This tool provides modularity for adding support for other languages. Original authors set free access to various script and tutorials that are necessary for incorporating new language to this tool.

We have to create language models and error model for Slovak language. Process of creating language model needs a lot of text, preferably with no errors. We collected newspaper articles published online, which have been already proofread. *Korektor* also supports use of multiple different language models. Therefore, we decided to use combination of language models based on word forms, lemmas and morphological tags (defined part of speech of the word) which should contribute to better results [2].

In our solution, we clear the collected text in Slovak language and then assign lemma and morphological tag to each word using annotated dictionary. After we have prepared texts, we use a toolkit for building statistical language models *SriLM*.

Very important is also error model, which reflect probabilities of making certain mistakes in text. In contrast to the language model, process of creating error model needs the text corpus, which contains as much errors as possible. Various blogs published on the Internet meet this condition. In these texts, we identify spelling errors.

* Supervisor: Marián Šimko, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 63–64.

When all models are created, we set it up to work with *Korektor* following instructions published on official website.

Part of our work is to develop freely available web service. We will create REST API that can be accessed directly via web browser or via any programming tool that support standard HTTP request methods and JSON for output handling. This API will provide web services for auto correction, spelling suggestions and diacritics completion for any given input text.

An important aspect of our work is to state how good our models incorporated in Korektor are in correcting text. We evaluate the accuracy, precision and recall of our solution on different sets of test data. There was not found any text in Slovak language with annotated errors, which could be used for easy evaluation. Therefore, we chose two approaches to create our own testing data:

- Test data created by rewriting text, which was read aloud
- Test data created by a script making random mistakes in words

We compare result of *Korektor* to results achieved by existing tools for automatic text-checking for Slovak language like Hunspell and build-in spellchecker in Microsoft Word.

We also examine result of our solution only on real-words errors, because today there is no freely available spellchecker that correct these types of error effectively. We also evaluate the influence of different types of language models on results, especially on real word errors.

Our contribution is creation of language and error models for Slovak language and incorporate it to existing text-checking tool *Korektor*. This tool using statistical method for word correction has potential to achieve better result than any public known spellchecker for Slovak language. We also create public accessed API for error correction that can be used in future projects.

References

[1] Richter, M., Straňák, P., Rosen, A.: *Korektor – A System for Contextual Spell-checking and Diacritics Completion* In Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), pages 1-12, Mumbai, India, 2012

[2] Staš, J., Hládek, D., Juhár, J., Incorporating Grammatical Features in the Modeling of the Slovak Language for Continuous Speech Recognition. In: Modern Speech Recognition Approaches with Case Studies, S. Ramakrishnan (Ed.), InTech Open Access, Rijeka, Croatia, 2012, pp. 257—276, ISBN 978-953-51-0831-3