Error Correction for Information Retrieval of Czech Documents

Jiří Martínek[1], Pavel Král[1,2]
[1]*Dept. of Computer Science & Engineering*
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2]*NTIS - New Technologies for the Information Society*
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
{*jimar, pkral*}*@kiv.zcu.cz*

Keywords: Czech, Error Correction, Fulltext, Language Model, OCR

Abstract: This paper proposes a novel system for information retrieval over a set of scanned documents in the Czech language. The documents are in the form of raster images and thus they are first converted into the text form by optical character recognition (OCR). Then OCR errors are corrected and the corrected texts are indexed and stored into a fulltext database. The database provides a possibility of searching over these documents. This paper describes all components of the above mentioned system with a particular focus on the proposed OCR correction method. We experimentally show that the proposed approach is efficient, because it corrects a significant number of errors. We also create a small Czech corpus to evaluate OCR error correction methods which represent another contribution of this paper.

1 INTRODUCTION

At present, a number of printed documents are scanned into electronic form. These scans are created for instance in order to save historical documents, to reduce the manipulation labours during document manipulation/management, to fulfill some government laws which order digitization, etc. Unfortunately, the documents are usually saved in the form of raster images and it is thus impossible to search appropriate information. Therefore, optical character recognition (OCR) techniques emerged and the documents are converted into plaintext form.

Unfortunately, despite the claims of many commercial OCR players, the resulting results are far from perfect and therefore error correction methods are beneficial. There are many projects dealing with OCR correction techniques. These projects are usually focused on English or other specific language such as Arabic or Chinese. However, to the best of our knowledge, only few work deals with Czech. Moreover, there is no system able to search information in a set of pdf documents in a form of raster images.

Therefore, the main goal of this paper consists in proposing of a novel system to handle this issue and searching information over scanned documents in the Czech language. The scanned documents are first converted to plaintext and then OCR errors are corrected using a method proposed below. These texts are indexed and saved into a fulltext database. This paper describes all components of the above mentioned system with a particular focus on the OCR correction approach.

We also created a small Czech corpus for evaluation of error correction methods. This corpus is freely available for research purposes at `http://ocr-corpus.kiv.zcu.cz` and represents another contribution of this paper.

The rest of the paper is organized as follows. The following section describes some interesting OCR correction approaches. Section 3 deals with the architecture of the proposed system, while Section 4 details the proposed approach for error correction. Section 5 first describes our document collection and then presents the results of experiments realized on this data. The last section concludes the paper and proposes some future research directions.

2 SHORT REVIEW OF OCR ERROR CORRECTION

There are several ways to improve accuracy of the OCR systems. The simplest approaches are rule-based and use a set of manually defined rules. Another group of methods uses for error correction manually defined lexicons followed by a distance measure to choice a closest word for replacing. The third group of methods use usually statistical methods with machine learning. The above mentioned approaches are briefly described for instance in a survey (Kukich, 1992). We describe next more in detail some interesting methods.

Zhidong et al. propose in (Zhidong et al., 1999) a language-independent OCR system which recognizes text from most of the world's languages. Their approach uses hidden Markov models (HMM) to model each character. The authors empoy unsupervised adaptation techniques to provide the language independence. The paper also describes the relationship between speech recognition and OCR.

Perez-Cortes et al. describe in (Perez-Cortes et al., 2000) an interesting method to post-process the OCR results in order to improve the accuracy. The authors propose a solution based on finite-state Markov model and modified Viterbi algorithm.

Another approach (Pal et al., 2000) focuses on the Indian language and *non–word* errors. The authors use for OCR error correction morphological parsing. A set of rules for the morphological analysis is presented. Unfortunately, is is not clear, whether this approach is applicable for any language OCR.

The authors of (Afli et al., 2016) use language models and statistical machine translation (SMT). This work is focused on historical texts. The purpose of the SMT is to translate words in source language into another words in a target language. The main idea is to translate OCR outputs into corrected texts using both language models.

Kissos and Dershowitz propose (Kissos and Dershowitz, 2016) a method involving a lexical spellchecker, a confusion matrix and a regression model. The confusion matrix and regression model are used for choosing good correction candidates.

3 SYSTEM ARCHITECTURE

The proposed system has a modular architecture as depicted in Figure 1 and is composed of three main modules.

The first module is used for OCR conversion of the document in the raster image form. Tesseract open source OCR Engine[1] is used as a core of our OCR analysis. The input of this module are raster images and the output is a so called *confidence matrix* which contains the possible recognized characters with confidence values.

The second module is dedicated to the correction of the OCR errors. Its input is the confidence matrix provided by the previous module and the output is the corrected text. This module combines probabilities of character language model with the values from the confidence matrix. A rule-based approach with Levenstein distance is also implemented in this module. The methods integrated in this module are described more in details in the following section.

The last module is used for document storage, indexing and retrieval. The open source search engine Apache Solr[2] is used for this task. The input is the corrected text obtained by the previous module. This module provides the possibilities of searching over the pdf data.

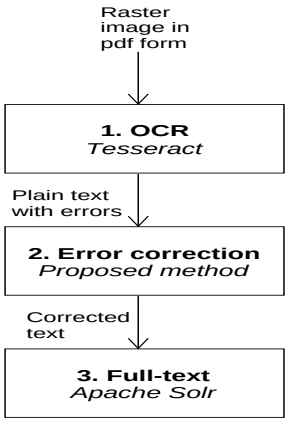

Figure 1: Modular architecture of the proposed system

4 PROPOSED METHOD

The proposed error correction method is at the character level. It uses first a rule-based approach for correction of the regular errors. Then, we use a statistical algorithm which combines the output of the Tesseract with language models. The last step consists in using dictionary-based Levenstein method as a post-processing of the previous step.

[1] https://github.com/tesseract-ocr

[2] http://lucene.apache.org/solr/

4.1 Rule-based Approach

This approach employs a set of manually defined rules to replace some characters by the other ones. For example the *in-word* character "0" (zero) is replaced by the character "O" or the character "1" (one) is replaced by the character "l", etc. Then, the result is checked against the manually defined dictionary. This approach can reduce a set of incorrect words and speed up the whole correction process.

4.2 Statistical Algorithm

This approach combines the scores of statistical n-gram language models with the confidence matrix values obtained by the Tesseract system using linear combination:

$$\delta = wT + (1-w)\pi \tag{1}$$

where w is the weight of the confidence matrix ($w \in [0;1]$), T is the character confidence obtained by the Tesseract system and π is the character probability provided by the language model. We use 3-gram language models with smoothing trained on Czech Wikipedia corpus (csWiki) (Suchomel, 2012).

We must identity the best (highest) probability values for all characters in the analyzed word. We use Viterbi algorithm (Forney, 1973) for this task. This algorithm creates several character possibilities during the *forward* step. The most probable character sequence is determined during the *backward* step by chosing the maximal value for each node. This is depicted in Figure 2.

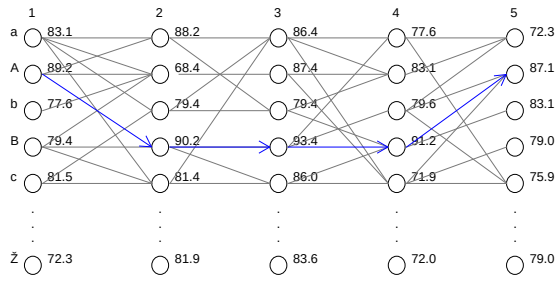

Figure 2: Viterbi algorithm

This figure shows several possibilities in the character space. Viterbi algorithm then choices the best path (blue colour with maximal values) in this lattice.

4.3 Dictionary-based Levenstein Method

If a word is not contained in the dictionary, we can use Levenstein distance (Levenshtein, 1966) for a further correction. This metric computes a distance between the corrected and the dictionary words as the minimum number of insertions, deletions or substitutions required to change the target word into the source one. Then, we replace the corrected word by the closest one (i.e. with minimum distance) from our dictionary. Note, that this last step is optional.

5 EXPERIMENTS

5.1 Czech Document Dataset

Unfortunately, there is no Czech freely available dataset to evaluate error correction methods.

Therefore, we collected the documents from the Czech Wikipedia. Every document was printed and scanned. The scanning was done with the different resolution, we chose 150, 300 and 600 DPI, respectively. For each scan we saved the correct text from the Wikipedia, which will be used as gold data for evaluation of our methods.

The final corpus is composed of the scans of 20 Czech documents in the pdf format. The documents have maximum one page of the text and differ in word number. The longest document has 523 words, while the shortest one has 119 words only. The average word number is 299.

This corpus is freely available for research purposes at `http://ocr-corpus.kiv.zcu.cz` and represents another contribution of this paper.

5.2 Evaluation Metrics

The main metric used for evaluation of the experiments is the standard *Word error rate (WER)*. It is defined as follows:

$$WER = \frac{S+D+I}{N} \tag{2}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of all words in the gold data.

Another metric is the *Word accuracy* defined as $1-WER$.

We further use the *Character error rate (CER)* which is computed similarly as WER, however the words are replaced by the characters.

The last metric used in this paper is the *Accuracy*. It is calculated as the number of correctly recognized words divided by the number of all words in the reference data.

5.3 Impact of the Document Resolution

According to the Tesseract documentation the resolution of the documents should be at least 300 DPI[3].

In the first experiment, we would like thus to validate this claim and determine the optimal resolution of the scanned documents in the Czech language. Another important property strictly related to the document resolution is execution time. Therefore, we also measure and report this value. The results of this experiments are shown in Table 1.

Resolution [DPI]	150	300	600
WER [%]	24.6	19.6	19.5
Execution time [ms]	10 575	8569	8405

Table 1: Tesseract OCR results depending on the document resolution

This table shows that Tesseract achieves the worst results for the document resolution 150 DPI. On the other hand the documents in resolutions 300 and 600 DPI achieve comparable WER. However, document with 600 DPI is in average 4 times bigger than document with 300 DPI which represents an important issue for data storage.

This table further shows that the time for processing of the images in resolution 300 and 600 DPI are comparable, however for conversion of images in resolution 150 DPI is needed significantly more of time.

This experiment proved that the resolution 300 DPI is sufficient for Tesseract OCR system and therefore we chose this value for the following experiments.

5.4 Evaluation of Error Correction

In the second experiment, we would like to evaluate the performance of the proposed OCR error correction module.

First we would like to find the optimal weight w (see Equation 1) of the combination of the Tesseract and the language models. We explore the values of $w \in [0;1]$ where the extreme value 0 means that only the language models are used, while the 1 value signifies that only the Tesseract output is used. The following two figures show the results of this experiment. Figure 3 shows the results with the Levenshtein distance correction, while Figure 4 depicts the results without this correction.

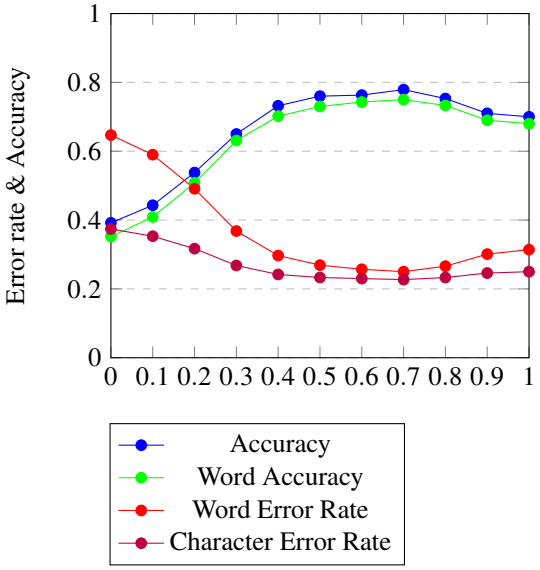

Figure 3: OCR results depending on the w value. Levenstein distance is used.

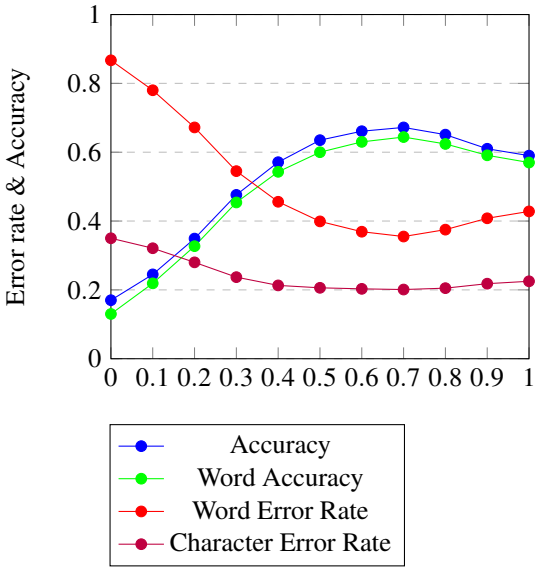

Figure 4: OCR results depending on the w value without Levenstein distance.

This experiment shows that the curves in both figures have very similar behaviour and that the best results are in both cases achieved by using the $w = 0.7$.

These figures also shows that the impact of the Tesseract system is more important, than of the language models. However, it is also evident that the usage of language models plays a positive role for error correction. The Word Accuracy is improved in both cases by 7% in absolute value (from 57.0% to 64.4% in the case without Levenstein distance and

[3]https://github.com/tesseract-ocr/tesseract/wiki/FAQ

from 68.0% to 75.0% with Levenstein method).

This experiment also shows, that Levenstein distance plays a positive role for error reduction. The best obtained Word Accuracy is about 75.0% with Levenstein distance.

6 CONCLUSIONS & FUTURE WORK

In this paper, we describe a novel system for information retrieval over a set of scanned documents in Czech language with a particular focus on the OCR error correction. We have experimentally shown that the proposed approach is efficient, because it corrects a significant number of errors.

Another contribution of this paper represents a new small Czech corpus which we created for evaluation of our OCR error correction method. This corpus is freely available for research purposes.

Our current document dataset is very small. Therefore, our first perspective consists in the extension of this corpus by other raster documents. The documents can be classified into several classes as for instance invoices, contracts, agreements, etc. Another perspective thus consists in creation of the class dependent language models. We assume that these language models should correct better OCR errors, because they will be adapted to the document types.

ACKNOWLEDGEMENTS

This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

REFERENCES

Afli, H., Qiu, Z., Way, A., and Sheridan, P. (2016). Using smt for ocr error correction of historical texts. In *LREC*.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Kissos, I. and Dershowitz, N. (2016). Ocr error correction using character correction and feature-based word classification. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, pages 198–203. IEEE.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Pal, U., Kundu, P. K., and Chaudhuri, B. B. (2000). Ocr error correction of an inflectional indian language using morphological parsing. *J. Inf. Sci. Eng.*, 16(6):903–922.

Perez-Cortes, J. C., Amengual, J.-C., Arlandis, J., and Llobet, R. (2000). Stochastic error-correcting parsing for ocr post-processing. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 405–408. IEEE.

Suchomel, V. (2012). Recent czech web corpora. In *6th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU*, pages 77–83.

Zhidong, L., Issam, B., Kornai, A., John, M., Prem, N., and Richard, S. (1999). A robust, language-independent ocr system.