The Same is Not The Same - Postcorrection of Alphabet Confusion Errors in Mixed-Alphabet OCR Recognition*

Christoph Ringlstetter[1], Klaus U. Schulz[1], Stoyan Mihov[2] and Katerina Louka[3]
[1]CIS, [3]IPSK, University of Munich, [2]IPP – Bulgarian Academy of Sciences, Sofia
Contact: kristof@cis.uni-muenchen.de

Abstract

Character sets for Eastern European languages typically contain symbols that are optically almost or fully identical to Latin letters. When scanning documents with mixed Cyrillic-Latin or Greek-Latin alphabets, even high-quality OCR-software is often not able to correctly separate between Cyrillic (Greek) and Latin symbols. This effect leads to an error rate that is far beyond the usual error rates observed when recognizing single-alphabet documents. In this paper we first survey similarities between Latin and Cyrillic (Greek) letters and words for distinct languages and fonts. After briefly introducing a new and public corpus collected by our groups for evaluating OCR-technology over mixed-alphabet documents, we describe how to adapt general algorithms and tools for postcorrection of OCR results to the new context of mixed-alphabet recognition. Experimental results on Bulgarian documents from the corpus and from other sources demonstrate that a drastic reduction of error rates can be achieved.

Keywords: Optical character recognition, postcorrection methods, mixed alphabets, free corpora for evaluation, reduction of error rates.

1 Introduction

Assume after scanning a Bulgarian document you find the sequence *Иван ора нивата* (pronounced: *Ivan ora nivata*, English: *Ivan plowed the field*) in the OCR output. To be on the safe side, you visually check correctness and find that *Иван ора нивата* perfectly coincides with the corresponding sequence *Иван ора нивата* found in the aligned original document. Since you need an English version of the document you send the OCRed text to a computer aided translation system. Surprisingly, you find that *opa* (*plowed*) is translated into *grandfather*. Applying other text processing routines you again obtain completely unexpected results for *opa*. With some background knowledge, there is a simple explanation: your OCR software, which was prepared to read Bulgarian texts with German passages, simply confused the Bulgarian word *opa*, written in Cyrillic letters, with the German word *opa* (*grandfather*), written in Latin letters.

Our scenario points to a general and serious problem that arises when scanning mixed-alphabet documents with Cyrillic-Latin or Greek-Latin characters. Depending on the background font, Cyrillic (Greek) and Latin characters and even words may look completely identical. Current OCR-software may be configured for recognition of mixed-alphabet input. However, the close similarity of symbols and words from different alphabets often leads to a surprising number of recognition errors where character sets are confused. Error rates are far beyond standard rates for single-alphabet input (cf. e.g., [7, 2]). Interestingly, OCR output often contains tokens with symbols from two alphabets. In general, these invisible errors become only apparent after applying electronic text processing and document analysis routines to the OCRed text. Depending on the kind of application, the OCRed text may then turn out to be inappropriate (cf. e.g., [1, 6, 9, 8]).

After the opening of most countries of the former East to the Western world, a strong movement can be observed to unify legislative principles and other

*Funded by German Research Foundation DFG and by VolkswagenStiftung.

areas of public life, and to simplify and improve economic exchange. In this contact, a large and growing relevance of mixed-alphabet documents can be observed in many Eastern European countries. The results presented below show that OCR recognition on documents with mixed Cyrillic-Latin alphabets is still far from optimal. In this special context, general postcorrection strategies developed for English [4, 1, 10] are of limited use. Refined strategies for mixed-alphabet input, which might be directly integrated into future OCR systems, represent a more promising basic step for successfully recognizing, analyzing and processing these documents.

Contributions and structure of this paper. We isolate alphabet confusion (ac-) errors as a new class of OCR errors and examine its sources - similarities between characters of distinct alphabets - for a variety of Eastern-European languages (Section 2). We briefly describe a large OCR test corpus with real-life mixed-alphabet documents from distinct areas and genres, mainly written in Bulgarian language. This corpus is freely available for the academic community (Section 3). On the documents of the corpus (as well as on many other mixed-alphabet documents), standard commercial OCR software produces many ac-errors. We show how to adapt our system for postcorrection of OCR results to mixed-alphabet input (Section 4) using simple techniques. For mixed-alphabet documents from the corpus and from other sources we analyse accuracy and ac-error rates resulting (1) from plain OCR recognition and (2) after postcorrection. A significant improvement of accuracy and reduction of error rates is achieved (Section 5).

2 Sources for ac-errors

Formally, by an *alphabet confusion error* (ac-error), we mean an OCR recognition error where a symbol of a given alphabet is erroneously classified as a symbol of another alphabet. Assuming a reasonable quality of printed documents, OCR software only confuses letters from distinct alphabets if these letters have a similar form. Similarities that give rise to ac-errors in particular exist between Latin characters on the one hand side and Cyrillic or Greek characters on the other hand. Today, Cyrillic letters are used in Russia, Ukraina, Byelorussia (White Russia), Bulgaria, Serbia, Macedonia and in many other countries of Eastern and Central Asia. The alphabets of all these languages are very similar and differ only by a small number of special characters. Hence problems and results reported below for Bulgarian mixed-alphabet documents probably can be generalized to other languages with Cyrillic alphabet. Greek letters are used in Greece and on Cyprus.

Similarities between symbols of distinct alphabets depend on the background font. For Latin and Cyrillic (Bulgarian alphabet) letters, ac-errors in particular arise from documents printed in Universum or in Times New Roman Cursive. Both fonts are popular, e.g., in Bulgaria. Ca. 20% of the documents in the corpus described below are written in these fonts. Furthermore, a large number of ac-errors can also be found in typewriter documents and in documents printed in Arial.

Figure 1 lists equivalences between letters of the two alphabets for the fonts Universum and Times New Roman Cursive that are used in the refined postcorrection strategy described below. Note that in many cases, letters of the two alphabets are optically completely identical. The tables do not capture all similarities between the respective alphabets, and for other fonts, variants of the tables might be more useful. Figure 2 lists equivalences between Latin and Greek symbols. Here, e.g., Times New Roman Cursive and Verdana Cursive cause many ac-errors, as we found in experiments with Greek texts (s.b).

3 Resources

The *Sofia-Munich* corpus, [5], which is freely available for academic groups and use, was recently built up by our teams in the framework of a two-years project.[1] The project was centered around postcorrection of OCR results, with a special focus on problems caused by a mixed Cyrillic-Latin input alphabet. Hence the major part of the corpus consists of Bulgarian documents. The following brief description concentrates on this part and ignores corpus units built from 128 German documents (including ground truth data for 312 pages). A more detailed description of the complete corpus can be found in a forthcoming paper [5].

The corpus is structured along the standards of the Brown Corpus ([3]) and includes multi-page excerpts

[1]Funded by VolkswagenStiftung

Latin	*A*	*B*	*C*	*E*	*H*	*K*	*M*	*O*	*P*	*T*	*X*	*Y*	*a*	*c*	*e*	*g*	*k*	*m*	*n*	*o*	*p*	*u*	*x*	*y*
Cyrillic	*А*	*В*	*С*	*Е*	*Н*	*К*	*М*	*О*	*Р*	*Т*	*Х*	*У*	*а*	*с*	*е*	-	-	*т*	*п*	*о*	*р*	*и*	*х*	*у*
Latin	A	B	C	E	H	K	M	O	P	T	X	Y	a	c	e	g	k	m	n	o	p	u	x	y
Cyrillic	A	B	C	E	H	K	M	O	P	T	X	У	a	c	e	g	k	m	n	o	p	u	x	y

Figure 1: Latin-Cyrillic transition table for the fonts Times New Roman Cursive (upper table) and Universum (lower table)

Latin	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*a*	*y*	*n*	*i*	*o*	*p*	*u*	*w*
Greek	*Α*	*Β*	*Ε*	*Ζ*	*Η*	*Ι*	*Κ*	*Μ*	*Ν*	*Ο*	*Ρ*	*Τ*	*Υ*	*α*	*γ*	*η*	*ι*	*ο*	*ρ*	*υ*	*ω*
Latin	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*a*	*y*	*n*	*i*	*o*	*p*	*u*	*w*
Greek	*Α*	*Β*	*Ε*	*Ζ*	*Η*	*Ι*	*Κ*	*Μ*	*Ν*	*Ο*	*Ρ*	*Τ*	*Υ*	*α*	*γ*	*η*	*ι*	*ο*	*ρ*	*υ*	*ω*

Figure 2: Latin-Greek transition table for the fonts Times New Roman Cursive (upper table) and Verdana Cursive (lower table)

from 630 documents that cover distinct topics and almost all genres of written language. Documents were collected in printed paper form from enterprises and organisations and thus come with all kinds of real-life problems such as images, strokes, signatures, stamps over text, etc. From each document we randomly extracted an excerpt of ca. 5 pages for the corpus. Meta properties of each excerpt are characterized in a table using 46 attributes (e.g., date, font, font size, formatting, languages etc.). Image files in png format (for scanning we used 256 scales of grey at 600 dpi) represent one (for singular sheets) or two pages (for books etc.) of a given excerpt. We have 546 image files containing informative prose (newspapers, magazines, textbooks, learning material, religion) 678 files of imaginative prose (general fiction, mystery, adventure, love, humor,...) 680 files from private organisations and government, and 402 files from enterprises (services, trade, industry). For each image file, the corpus contains the parallel file obtained via OCR recognition with one of the leading commercial OCR systems. For 223 excerpts with an OCR error rate between 1% and 30% (word level) we prepared ground truth data for one file, manually correcting OCR results.

Bulgarian EC corpus. Since the *Sofia-Munich corpus* contains a large variety of distinct document types, training of postcorrection methods is difficult. Postcorrection experiments where also made with a homogeneous corpus of official EC documents (Bulgarian version) where training has better effects.

Greek-Latin corpus. For experiments with mixed Greek-Latin alphabet we took a corpus of 20 articles of Greek online newspapers including English names and small English text passages. We randomly selected one page from each document and printed it using Times Cursive and Verdana Cursive. Printed versions were scanned and analysed using two standard commercial OCR software packages.

4 Postcorrection method

When analyzing OCR recognition results for the mixed-alphabet documents in the aforementioned corpora we found that ac-errors represent a serious problem (results are given in Section 5). We then adapted our existing system for postcorrection of OCR results to mixed-alphabet input. In this section we briefly describe our approach to lexical postcorrection and the new variant.

For each token w_i^{ocr} recognized by the given OCR engine and each dictionary D in a repository $\mathcal{D}$ of relevant correction dictionaries we preselect a list of n *correction candidates* in D. For preselection, the standard Levenshtein distance d_0 is used. Entries v of D where $d_0(w_i^{ocr}, v)$ is small are preferred. The background dictionary system contains large-scale dictionaries for Bulgarian, German, English, Greek, as well as specialized dictionaries for proper names, geographic names, abbreviations and acronyms.

After the above preselection, for each correction

Corpus/OCR	tokens	error rate OCR → pc	ac-error rate OCR → pc
SM-OCR1-Tr	8110	11.22→ 6.57%	5.42→ 1.25%
SM-OCR1-Te	7923	10.59→ 6.25%	5.44→ 1.26%
SM-OCR2-Tr	5099	37.24→ 15.87%	9.96→ 1.75%
SM-OCR2-Te	5115	43.63→ 16.81%	10.28→ 2.01%
EC-OCR1-Tr	6571	15.05→ 5.68%	10.94→ 1.81%
EC-OCR1-Te	6230	16.44→ 7.03%	13.00→ 3.74%
EC-OCR2-Tr	6571	48.52→ 9.0%	27.71→ 4.14%
EC-OCR2-Te	6230	48.81→ 11.35%	27.50→ 3.23%

Table 1: Number of tokens (composed of standard letters only), error rate (word level) for plain OCR recognition and postcorrection, ac-error rate for plain OCR recognition and postcorrection for Bulgarian Sofia-Munich (SM) corpus (Universum and Times New Roman Cursive), Bulgarian EC corpus (Universum), recognition with software OCR1 and OCR2, training (Tr) and test (Te) data.

candidate v of w_i^{ocr} we compute a normalized similarity value $s(v, w_i^{ocr})$ based on (1) a variant of the Levenshtein distance with flexible edit weights obtained from symbol confusion statistics and (2) a normalized collocation frequency value $f(v, w_{i-1}^{ocr}, w_{i+1}^{ocr})$ based on trigram frequencies in a large subcorpus of the World Wide Web. Note that in this way, sentence context is taken into account.

Balance and Score. The *score* of a correction candidate v for a token w^{ocr} is $score(v) := \alpha \cdot s(v, w^{ocr}) + (1-\alpha) f(v)$. The *balance parameter* α is a value in $[0, 1]$ that determines the relative weight of similarity versus frequency. Since frequency and distance values are of distinct nature and normalized in distinct ways, this gives only a basic intuition.

Threshold. If the error rate of the OCR is low, it does not make sense to automatically replace each OCR-token that is not found in the dictionary by the best correction candidate. Instead, we override the OCR result only in the presence of additional confidence. The *threshold parameter* τ defines the minimal score which has to be achieved in order a correction to take place.

Parameter Optimization. If ground truth data for training are available, optimized values for the parameters α and τ are computed using a simple hill climbing procedure. For an initial pair of standard values (α_0, τ_0), the system automatically computes the correction accuracy that is obtained. Fixing the value for τ_0 we use the system to compute a value α_1 that leads to optimal correction accuracy for the given threshold τ_0. Fixing then α_1, we compute a value τ_1 that leads to optimal correction accuracy for the given value of the balance parameter, α_1. In this way we continue until a local maximum is found.

Generation of training data. When postcorrecting long documents or large corpora of similar documents we produce ground truth data for training and parameter optimization via interactive correction of partial recognition results.

Adaption to mixed alphabet input. The background dictionaries used for correction are either purely Cyrillic or purely Latin. Before preselecting correction candidates from distinct dictionaries for a token w_i^{ocr} recognized by the given OCR engine, we produce two variants of w_i^{ocr}. The first (second) variant is obtained replacing each Latin (Cyrillic) character σ of w_i^{ocr} by an optically "equivalent" Cyrillic (Latin) letter (if it exists). Equivalence is encoded in a table, cf. Figures 1 and 2. If there does not exist an equivalent letter in the other alphabet, we leave σ unmodified. For the search in the relevant Cyrillic (Latin) background dictionaries, the variant where we try to translate letters into Cyrillic (Latin) is used. Since search is approximate, input words with alien symbols make sense.

5 Experiments and evaluation results

Texts of the aforementioned corpora where scanned and processed with two standard high-quality commercial OCR software systems, respectively called OCR1 and OCR2. Afterwards we applied our postcorrection strategy. Table 1 presents results for the Bulgarian part of the *Sofia-Munich* corpus and for the EC corpus. Results on the *Sofia-Munich* corpus characterize error rates in a random selection of documents in the fonts Universum and Times New Roman Cursive. The EC-corpus was printed in Universum. Error rates on the word level are given for plain OCR recognition and after postcorrection, both for arbitrary errors and for ac-errors.

It becomes obvious that ac-errors represent a problem for both OCR systems, the problem being much more serious for OCR2. Leaving ac-errors aside, OCR2 has principal difficulties with mixed-alphabet texts in the given fonts. Recognition results are lower than those known from English texts, and even recognition of pure Cyrillic texts is far from optimal. Our postcorrection strategy significantly reduces both ac-error rate and the general error rate. A drastic reduction is achieved for the corpus of EC documents, which is more homogeneous w.r.t. contents and layout. We also analyzed documents of the *Sofia-Munich* corpus in other fonts than those mentioned above. Here ac-error rates are lower, but remain unacceptable.

Table 2 gives an error statistics for OCR recognition results for the Greek corpus (8410 tokens). Here we did not apply our postcorrection machinery. Again OCR2 has serious problems with mixed-alphabet input.

Corpus/OCR	err.r.	ac-errors	ac-err. r.
GR-OCR1-Ti	2.37%	62	0.74%
GR-OCR1-Vd	2.12%	49	0.58%
GR-OCR2-Ti	12.00%	550	6.54%
GR-OCR2-Vd	10.87%	457	5.43%

Table 2: OCR recognition results for Greek-Latin input. Error rate (word level) for plain OCR recognition, number of ac-errors, and ac-error rate for Greek Newspaper corpus in cursive Times(Ti) and cursive Verdana(Ve) font.

References

[1] A. Dengel, R. Hoch, F. Hönes, T. Jäger, M. Malburg, and A. Weigel. Techniques for improving OCR results. In H. Bunke and P. S. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 227–258. World Scientific, 1997.

[2] ISRI. OCR accuracy produced by the current DOE document conversion system. Technical Report 2002-06, Information Science Research Institute University of Nevada Las Vegas, 2002.

[3] H. Kucera and W. N. Francis, editors. *Computational Aspects of Present-Day American English*. Brown University Press, 1967.

[4] K. Kukich. Techniques for automatically correcting words in texts. *ACM Computing Surveys*, pages 377–439, 1992.

[5] S. Mihov et al. A corpus for comparative evaluation of OCR software and postcorrection techniques. In *Proc. of ICDAR'05*, 2005.

[6] S. Mori, H. Nishida, and H. Yamada. *Optical Character Recognition*. John Wiley & Sons, New York, 1999.

[7] S. V. Rice, G. Nagy, and T. A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, 1999.

[8] K. Taghva, J. Borsack, and A. Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing and Management*, 32(3):317–327, 1996.

[9] K. Taghva, J. Borsack, A. Condit, and S. Erva. The effects of noisy data on text retrieval. *Journal of the American Society for Information Science*, 45:50–58, 1994.

[10] K. Taghva and E. Stofsky. OCRSpell: an interactive spelling correction system for OCR errors in text. *International Journal of Document Analysis and Recognition*, 3:125–137, 2001.