A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques*

Stoyan Mihov[1], Klaus U. Schulz[2], Christoph Ringlstetter[2], Veselka Dojchinova[1], Vanja Nakova[1], Kristina Kalpakchieva[1],Ognjan Gerasimov[1], Annette Gotscharek[2] and Claudia Gercke[2]

[1]IPP – Bulgarian Academy of Sciences, Sofia, [2]CIS, University of Munich

Contact: stoyan@lml.bas.bg

Abstract

We describe a new corpus collected for comparative evaluation of OCR-software and postcorrection techniques. The corpus is freely available for academic groups and use. The major part of the corpus (2306 files) consists of Bulgarian documents. Many of these documents come with Cyrillic and Latin symbols. A smaller corpus with German documents has been added. All original documents represent real-life paper documents collected from enterprises and organizations. Most genres of written language and various document types are covered. The corpus contains the corresponding image files, rich meta-data, textual files obtained via OCR recognition, ground truth data for hundreds of example pages, and alignment software for experiments.

Keywords: Optical character recognition, postcorrection of OCR results, public corpora, comparative evaluation, ground truth data, Cyrillic documents, mixed-alphabet documents, meta-data.

1 Introduction

Evaluating OCR software on public corpora [RJN96, PCH93] helps to compare distinct OCR systems, to obtain a better picture of the accuracy that can be expected in distinct application contexts, it points to existing deficiencies and shortcomings. Similarly the evaluation of postcorrection techniques [Kuk92, RNN99, ISR02, DHH+97] on freely available OCR data sets helps to compare the strengths and weaknesses of distinct postcorrection strategies. Successful postcorrection techniques can be integrated into future generation OCR systems.

Still, a comparative evaluation of OCR software and techniques for postcorrection of OCR results is difficult since only a small number of corpora are freely available that are appropriate for this task [PH95, RN96, KV00]. Most of these corpora are composed of English documents. On the other hand, problems for OCR recognition and interesting post-correction tasks in particular arise from special languages and alphabets [BW97]. Here test corpora are available only in a few cases [PH95, DH97]. The need to prepare OCR test corpora for non-English languages is emphasized, e.g., in [GHHP97]. In the ideal case, these corpora should be representative w.r.t. contents and formats, which means that a broad range of contents, genres and documents types should be covered. The collection of new corpora along this line is time consuming and expensive.

In this paper we describe the *Sofia-Munich* corpus of scanned paper documents, which was designed and built up for the above-mentioned tasks in our groups in the framework of a two-years project centered around postcorrection of OCR results. Motivated by the need to support document analysis techniques in Eastern European countries, the project had a special focus on problems for OCR systems caused by mixed Cyrillic-Latin alphabet input. The major part of the corpus (2306 files) consists of Bulgarian documents. Many of these documents come with Cyrillic and Latin symbols. A smaller corpus with German documents has been added. The complete corpus is

freely available for academic groups and use.

The corpus comes with a file that collects meta-data for each document. As a special feature that drastically simplifies all kinds of evaluation tasks, reconstructions of the original texts (ground truth data) have been prepared for hundreds of pages. Since the corpus contains image files and OCR recognition results, a comparative evaluation of both OCR systems and postcorrection methods is directly supported by aligning ground truth data with recognition results. Suitable alignment software can be obtained from the authors.

In the remainder of the paper we describe the structure of the corpus and add technical details about file formats and other kind of useful meta-information. We also summarize and illustrate the typical problems and errors that were observed when applying modern commercial OCR software to convert the documents to symbolic electronic form. Interestingly, in the Bulgarian part, special errors caused by Cyrillic letters represent a serious problem. This again illustrates the need to have suitable OCR test data sets for distinct languages and alphabets.

2 Size and composition of the corpus

The *Sofia-Munich* corpus is structured along the standards of the Brown Corpus ([KF67]). The *Bulgarian subcorpus* includes 2306 image files representing excerpts (ca. 5 pages, or approximately 2000 words) from 630 real-life documents that cover almost all genres of written language. We have

1. 546 files with informative prose (newspapers, magazines, textbooks, learning material, religious texts)

2. 678 files with imaginative prose (general fiction, mystery, adventure, western, love, humor,...)

3. 680 files with material from private organisations and government, and

4. 402 files with business texts (faxes, invoices, etc.) from enterprises (services, trade, industry).

The 630 background documents were collected in paper form from a large number of distinct enterprises and organisations. Many documents come with distinct real-life problems such as logos, strokes, signatures, stamps over text, text within images, etc.

210 documents date back to 1980-1989, 179 (241) documents date back to 1990-1999 (2000-2004).

The *German subcorpus* contains excerpts (312 image files) from 128 documents collected from enterprises and organisations.

The complete collection includes faxes, type writer documents, laser and matrix printer texts, and copies. For each document, a written agreement/declaration (signed by a representative of the organisation or company that contributed the document) exists stating that the document may be distributed for evaluation.

3 Technical features

Each image files is stored in Portable Network Graphic (png) format and either represents one page from a document representing a collection of separate sheets, or two consecutive pages from a magazine, newspaper, journal, or book. Document excerpts have been scanned with 256 scales of grey at 600 dpi. The used scanner was HP Scanjet 5470C driven by HP Precisionscan Pro. Symbolic textual files are stored in Ascii. Our alignment software is written in Java.

4 Meta information on paper documents

Detailed information on each image file/document is stored in a table with 20 attributes that cover all kinds of useful meta data. The attributes are: the unique identifier of the image file, the name of the background document, the author of the document (if available), the year where the document was created (if available), the number of pages of the complete document, the number of scanned pages, the source (company, organisation or person that provided the document), the number of the agreement/declaration (cf. Sec. 2) for the document, a Boolean value "translated" indicating if the document is translated, the name, size and color of the main font used in the document, a Boolean value indicating if different font

Error rate	1980-1989	1990-1999	2000-2004
0-1%	28	106	169
1-30%	101	55	67
30-100%	81	18	5

Table 1: Error rates and period of creation for Bulgarian documents.

sizes are used, a Boolean value indicating if special formatting such as bold, italics, etc. is used, a Boolean value "tables/pictures" indicating if there are tables and pictures in the document, the number of text columns, a list of other languages than the standard language (Bulgarian or German) used in the text, the document type (book, magazine, newspaper, fax, sheet, flyer), the paper color, defects (paper flexion, handwritten notes, stamps, etc.).

Meta-data for Bulgarian documents include additional attributes. Some of these attributes encode information on deficiencies of the OCR recognition for the file.

5 OCR-recognition, ground truth data and error analysis

Each png file was processed using a standard commercial OCR software. The resulting textual file containing the OCR recognition result is included in the corpus. For 312 image files from German documents, a reconstruction of the original text (ground truth data) has been prepared using interactive postcorrection.

In the Bulgarian part, before preparing ground truth data, the error rate (percentage of words with recognition errors) was estimated, manually comparing OCR output and original files. Table 1 shows the error rates that were observed for documents from distinct periods[1]. For each excerpt with an estimated error rate between 1% and 30%, ground truth data for one image file have been prepared. The number of such excerpts is 223.

---

[1] For some documents, periods had to be estimated.

5.1 Error analysis for Bulgarian documents with error rate 1-30%

The recognition errors for 223 Bulgarian document excerpts with error rate 1-30% have been analyzed in detail. In what follows we first list characteristic sources of errors. Afterwards we classify the observed error patterns. Illustrating examples from the corpus are added.

Error sources

Cyrillic letters. Recognition errors for Cyrillic letters represent the most important class of errors. For details, see below.

Inadequate positioning of books. When scanning pages from the middle of a thick book, image regions from the folding area are distorted. Tokens from these areas are often garbled in the OCR result. Sometimes the separation between the two pages is not perfect.

Paper quality, hand writing, stamps. Poor paper quality may result in additional dots, stars and other symbols in the text. Hand writing is not correctly recognized. If hand writings or stamps overlap with regular text, errors result.

Text inside pictures, text as picture. Sometimes textual regions with large letters were classified as pictures. Text inside pictures was often not correctly recognized.

Tables. As a general rule, textual contents of tables were not recognized very well.

Column segmentation. Wrong column segmentation often arises from newspaper and magazine pages were pictures induce a difficult page segmentation. In some cases, two columns were interpreted as one.

Low contrast and blurring. For texts with low contrast and blurring, recognition results in general are poor.

Low print quality. Some type writer texts produce blurred letters and irregular spaces. This may lead to recognition results where letters are wrongly classified and merged. Also words may be merged (s.b.).

In what follows it should be noted that the input documents are dominated by Cyrillic letters. Hence, not surprisingly, many error patterns are of a specific nature, demonstrating that recognition of Cyrillic letters in specific fonts is still far from optimal.

ния летен театър. Другият показател за почитта и любовта към най-стария балетен форум е стълпотворението от знаменитости във Варна - те дойдоха тук като педагози или просто като гости на конкурса. Ще изброя само някои от тях - Оливър Мац и Щефи Шерцер - лауреати на Варна и премиер-солисти на Дойче опер в Берлин, Ото Бубеншчек - лауреат на Лозана и премиер-солист на балета на Ноймаер в Хамбург, Николай Цискаридзе - лауреат на Москва и премиер-солист на Болшой, Любов Кунакова - лауреат на Варна и звезда на Мариинския театър в Петербург, легендарната прима на Болшой - Наталия Безсмерт-нова. "Звезден" бе и съставът на журито, в което влизаха емблематични личности на световния балет като Юри Григорович, Мишел Денар, Дмитрий Симкин. Всички изброени имена са на действени личности, които са в центъра на внима-нието на публиката и медиите чрез работата си в различни области на балет-ното изкуство.

ния летен театър. Другият показател за почитта и л!обоВта ^м най-стария балетен форум е стълпотВорението от знаменитости ВъВ Варна - те дойдоха тук като педагоzи или просто като гости на конкурса. Ще изброя само ™кои от тях - ОлиВър Мац и Щефи Шерцер - лауреати на Варна и премиер-солисти на Дойче опер В Берлин, Ото Вубени4ек - лауреат на Лозана и премиер-солист на балета на Ноймаер В Хамбург, HukoAau LJuckapugse - лауреат на Моск?а и премиер-солист на Болшой, ЛЬобоВ Kynako?a - лауреат на Варна и зВезда на Мариин-с^я театър В Петербург, легендарната прима на Болшой - Наталия Безсмерт-ноВа. "ЗВезден" бе и съставът на Журито, В което Влизаха емблематични личности на световния балет като Юри Григорович, Мишел Денар, Дмитрий CuMkun. Bcu4ku изброени имена са на действени личности, koumo са в центъра на вниманието на nyOAukama и медиите чрез работа си в различни области на балетното U3kycm6o.

Figure 1: On the snapshot typical situations of Cyrillic to Latin symbol substitutions in the "Universum" font are seen (see the framed words).

Error patterns

Cyrillic to Latin symbol substitution. OCR software was configured for mixed Bulgarian-English input. Nevertheless, in almost 50% of the sample documents, Cyrillic letters were substituted by Latin letters. In one article, almost 200 of these errors were found. This confusion of alphabets sometimes affects single symbols, in other cases complete words. Most problematic are the fonts "Universum" (largely used in Bulgarian newspapers and magazines) and "Times new Roman".

Cyrillic to Cyrillic/unknown symbol substitution. Errors of this kind are again frequently found in documents printed in "Universum".

Cyrillic to digits, capitals, symbols. Cyrillic letters were also recognized as digits, punctuation marks, numbers, and lower-case Cyrillic letters were recognized as uppercase (Cyrillic or Latin) letters.

Symbol merges and splits. Merged symbols are typical for typewriter text because of distinct spaces between consecutive letters. We found up to 10 errors of this form per document. Split letters mainly arise from "Universum", where one Cyrillic letter is recognized as two Latin letters. Sometimes, numerical symbols and others are found in the recognition result.

Word merges and splits. These errors are rather exceptional and again typical for type writer texts.

	Ti	Ar	Un	TW	Mix
S-ER %	4.42	1.70	5.03	5.81	4.27
W-ER %	4.90	3.25	13.73	7.92	3.91
C/L-ER %	21.29	4.25	11.37	4.23	2.64
Excerpts	21	12	30	71	11

Table 2: Symbol error rate (S-ER), word error rate (W-ER), Cyrillic-Latin symbol confusion rate (C/L-ER), and number of excerpts for distinct fonts (Ti = Times, Ar = Arial, Un = Universum) and document types (TW = type writer, Mix = mixed font documents).

Hand writing between the words may lead to merges. Split words occur in texts printed with a matrix printer, where sometimes all letters of a word are separated.

False friends. We observed 89 false friends (erroneous words accidently representing entries of a dictionary).

Table 2 summarizes the analysis of 145 representative excerpts of a particular font/document type.

5.2 Error analysis for German documents

Among 312 image files of German documents, for 139 we observed a word error rate up to 1%. For 129 (44) files, the error rate is between 1%-30% (beyond 30%).

6 Conclusion

The preliminary evaluation results described above show that special errors caused by Cyrillic letters – in particular the confusion of Cyrillic letters with Latin letters – seriously deteriorate OCR accuracy. This problem illustrates that results on OCR accuracy for English texts cannot be simply generalized to Eastern European texts written in Cyrillic alphabet, and emphasizes again the need to have suitable OCR test data sets where recognition of Cyrillic letters and mixed alphabets can be studied. The *Sofia-Munich* corpus represents an initial step. With the free distribution of the corpus for academic institutions and use we hope to motivate other groups to contribute to

future extensions. After finding the aforementioned errors in the corpus, we also looked at suitable post-correction techniques. With an adaption and refinement of our existing postcorrection system to mixed-alphabet input, a significant reduction of error rates could be achieved. These results are described in a forthcoming paper [RSML05].

References

[BW97]   Horst Bunke and Patrick S.P. Wang. *Handbook of Character Recognition and Document Image Analysis*. World Scientific, Singapure, 1997.

[DH97]   R. B. Davidson and R. L. Hopley. Arabic and Persian OCR training and test data sets. In *Proc. Symp. Document Image Understanding Technology (SDIUT97)*, pages 303–307, 1997.

[DHH+97] Andreas Dengel, Rainer Hoch, Frank Hönes, Thorsten Jäger, Michael Malburg, and Achim Weigel. Techniques for improving OCR results. In Horst Bunke and Patrick S.P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 227–258. World Scientific, 1997.

[GHHP97] Isabelle Guyon, Robert M. Haralick, Jonathan J. Hull, and Ihsin T. Phillips. Data sets for OCR and document image understanding research. In Horst Bunke and Patrick S.P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 779–799. World Scientific, 1997.

[ISR02]   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.

[KF67]   Henry Kucera and W. Nelson Francis, editors. *Computational Aspects of Present-Day American English*. Brown University Press, 1967.

[Kuk92]   Karen Kukich. Techniques for automatically correcting words in texts. *ACM Computing Surveys*, pages 377–439, 1992.

[KV00]   Paul B. Kantor and Ellen M. Voorhees. The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2(2/3):165–176, 2000.

[PCH93]   Ihsin T. Phillips, Su Chen, and Robert M. Haralick. CD-ROM document database standard. In *Proc. of the 2nd International Conference on Document Analysis and Recognition (ICDAR 93)*, pages 478–483, 1993.

[PH95]   Ihsin T. Phillips and Robert M. Haralick. University of Washington English/Japanese document image database II: A database of document images for OCR research. CD-ROM, 1995.

[RJN96]   Stephen V. Rice, Frank R. Jenkins, and Thomas A. Nartker. The fifth annual test of OCR accuracy. Technical Report TR-96-01, Information Science Research Institute University of Nevada Las Vegas, 1996.

[RN96]   Stephen V. Rice and Thomas A. Nartker. The ISRI analytic tools for OCR evaluation. Technical Report TR-96-02, Information Science Research Institute University of Nevada Las Vegas, 1996.

[RNN99]   Stephen V. Rice, George Nagy, and Thomas A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, 1999.

[RSML05] Christoph Ringlstetter, Klaus U. Schulz, Stoyan Mihov, and Katerina Louka. The same is not the same - postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. In *Proc. of the Eigth International Conference on Document Analysis and Recognition(ICDAR 05)*, 2005.