

Oprava chýb v textoch vytvorených pomocou OCR

Petra Galuščáková
galuscakova@gmail.com

Abstrakt

OCR (Optical Character Recognition) je technika, ktorá sa používa pri prevode textu z písanej podoby do elektronickej formy. Je založená na rozpoznávaní daných geometrických tvarov, napr. čiar, kruhov a polkruhov. Pri takomto spracovaní textu vznikajú niektoré typické chyby. Dôležitá je preto posteditácia textu, ktorá sa však často vynecháva. Cieľom práce je vytvoriť program, ktorý by aspoň čiastočne nahradil posteditáciu a automaticky opravil čo najviac chýb. Program by mal prípadne pomôcť korektorovi pri manuálnej oprave textu. Program je jazykovo nezávislý a využíva trigramové modely. Ďalej je v programe zakomponovaná oprava technickejších chýb, ako sú napríklad úvodzovky.

1. Úvod

1. 1. Ciele práce

Pri automatickom preklade je dôležitým bodom získať dáta v zdrojovom a v cieľovom jazyku. Dobrým zdrojom takýchto dát sú zdigitalizované knihy. Často je možné nájsť tú istú knihu zdigitalizovanú v rôznych jazykoch, čo je veľmi cenný zdroj dát pre automatický preklad. Tieto knihy bývajú skoro vždy vytvorené pomocou zdigitalizovania ich papierovej podoby, čiže bývajú naskenované a následne spracované pomocou OCR softwaru. Tento postup ďalej vyžaduje, aby knihy ešte následne prešli korektúrou. To je však časovo náročné, preto sa korektúra v praxi často vynecháva. V textoch potom ostávajú chyby, ktoré môžu spôsobovať problémy pri použití v automatickom preklade. Základným cieľom práce je vytvoriť program, ktorý by pomohol vyčistiť texty vytvorené pomocou OCR softwaru tak, aby mohli byť lepšie použiteľné v automatickom preklade. Druhým cieľom práce je pomoc pri manuálnej korekcii textu.

1. 2. Požiadavky na program

Z cieľov práce vyplývajú aj základné požiadavky na program. Program by mal byť command-line a mal by fungovať dostatočne rýchlo, aby bolo možné veľké množstvo textov čo najrýchlejšie automaticky opraviť. Pre využitie v automatickom preklade je dôležité opraviť najmä chybné slová. Pri manuálnej korekcii sú však najčastejšie opravované tzv. technické chyby. Technickými chybami sú napríklad čísla stránok, predely medzi stránkami, spojovníky a nekonzistencia úvodzoviek. Okrem automatickej opravy chýb by mal program aj zvýrazniť podozrivé miesta, ktoré nie je možné automaticky opraviť. Vhodným programovacím jazykom na vytvorenie programu sa zdá byť Perl.

1. 3. Typy chýb

OCR software dosahuje dnes už veľmi vysokú presnosť. Kvalita výstupu však veľmi záleží na kvalite vstupných obrázkov. Keďže sú OCR programy založené na rozpoznávaní geometrických útvarov ako čiara alebo kruh, vznikajú tak pri digitalizácii niektoré typické chyby. Niektoré z týchto chýb sú v Tabuľke 1.

Pôvodný znak	Rozpoznané znaky
d	cl
h	li
m	iri
y	v

e	c
l	!
t	f
z	/

Tabuľka 1: Typické chyby, ktoré vznikajú pri spracovaní textu pomocou OCR softwaru

Ďalšími častými chybami sú chýbajúce dĺžne a mäkčene. Pri oprave týchto chýb sa používajú trigramy.

Prípadné znečistenia textu sa niekedy prejavia ako znaky navyše. Často sa stretávame aj s chýbajúcou alebo nadbytočnou interpunkciou. Z textu je potrebné odstrániť aj spojovníky na konci riadkov a oddeľovače, ktoré software pridáva medzi po sebe nasledujúce strany. Veľmi často bývajú nesprávne rozpoznané úvodzovky. Otváracie úvodzovky sú často rozpoznané ako dve čiarky, ukončujúce úvodzovky bývajú rozpoznané ako apostrofy. Veľmi častá je nekonzistencia úvodzoviek, keď sú niekde úvodzovky rozpoznané ako slovenské ukončovacie (“), inde ako anglické úvodzovky ("). Tri bodky je v niektorých prípadoch lepšie nahradiť špeciálnym symbolom. [1] Tieto typy chýb môžeme nazývať technické a je lepšie ich spracovávať inak ako pomocou trigramov.

2. Implementácia

2.1. Korektúra pomocou trigramov

Prevažná väčšina programov, ktoré slúžia na opravu chýb z OCR softwaru, je založená na slovníkovom prístupe. To znamená, že ak sa slovo nenachádza v slovníku, potom je chybné, inak je správne. V našom programe je využitý prístup založený na kombinácii phonotaxe a štatistiky. [2] Pri prístupe, ktorý využíva phonotax, sa vyhľadávajú v texte trigramy, ktoré sa v danom jazyku nevyskytujú. Takéto trigramy sa v ďalšom texte nazývajú **zakázané**. Zakázané trigramy je možné získať z textov, ktoré nevznikli pomocou OCR a neobsahujú chyby. Takýto text budeme ďalej nazývať **korpus**.

Štatistický prístup vyhľadáva trigramy, ktoré sa vyskytujú častejšie v textoch získaných pomocou OCR, ako v ostatných textoch. Tieto trigramy by teda mali reprezentovať chyby OCR softwaru. V ďalšom texte sa nazývajú **chybové** trigramy. Text z OCR a text spracovaný inak však musia byť identické. V opačnom prípade získané chybové trigramy nezávisia od typických chýb, ktoré vytvára OCR software, ale od vlastností týchto textov. Preto je tieto trigramy potrebné vytvoriť porovnaním jedného súboru, ktorý je získaný pomocou OCR softwaru a následne opravený. Takéto súbory je však väčší problém získať.

2.2. Riešenie

Ak je daný zoznam chybových a zakázaných trigramov, vyhľadajú sa tieto trigramy v súbore, ktorý sa má kontrolovať. Na vyhľadávanie je použitá perlovská knižnica Aho-Corasick [3], ktorá implementuje daný vyhľadávací algoritmus. Táto knižnica značne zrýchliла prácu celého programu. Daný trigram sa označí ako podozrivý ak je zakázaný a zároveň chybný. Tento prístup sa zdá byť vhodným kompromisom medzi počtom falošných alarmov a správne nájdených zlých trigramov, vyžaduje však dostatočný počet tréningových dát. Podozrivé trigramy sa buď vyznačia, alebo sa skúsia automaticky opraviť.

Súčasťou programu je súbor s častými chybami, ktoré vznikajú v OCR software. Pre každý chybový reťazec je uvedený odpovedajúci korektný reťazec. Pri oprave sa prechádza tento zoznam a skúšajú sa nahradiť jednotlivé znaky v označenom trigrame. Následne sa skontroluje počet výskytov takto upraveného trigramu v opravenom texte. Zo všetkých možných opráv sa vyberie na základe počtu výskytov ten najvhodnejší.

2. 3. Trénovacia fáza

Na to, aby program správne fungoval, je potrebné mať tri súbory

- jednojazyčný korektný korpus
- výstup OCR softwaru
- korektúru výstupu OCR softwaru

Na vytvorenie súboru zakázaných trigramov je potrebná znalosť abecedy daného jazyka. Trojice znakov z tejto abecedy, ktoré sa v korpuse nenachádzajú alebo majú malý počet výskytov, sú označené ako zakázané.

Zo súboru vytvoreného pomocou OCR softwaru a opraveného súboru je vytvorený zoznam trigramov, ktoré sa v oboch súboroch nachádzajú. V zozname sú aj s počty trigramov v každom z týchto súborov. Z týchto počtov je možné pri ďalšom spracovaní vytvoriť relatívne frekvencie daného trigramu v pôvodnom a opravenom súbore. Ak sa teda trigram nachádza viackrát v pôvodnom súbore ako v opravenom, tak je pravdepodobne trigram chybový.

Súčasťou programu je skript, ktorý zo zadaných súborov vytvorí potrebné súbory so zakázanými a chybovými trigramami.

2. 4. Ďalšia korektúra

Ďalším kontrolným mechanizmom, ktorý je v programe zabudovaný, je kontrola neštandardných znakov uprostred slov. Hľadajú sa veľké písmená, čísla a nealfanumerické znaky uprostred, prípadne na konci slova.

Veľká väčšina chýb, ktoré sa v súboroch vyskytujú, je ale technického charakteru. V textoch sa vyskytujú napríklad nadbytočné medzery, nekorektné úvodzovky, čísla strán, oddeľovače stránok alebo spojovníky na konci riadkov. Tieto chyby je možné často opraviť pomocou regulárnych výrazov. To je však už špecifické pre daný dokument alebo aspoň typ OCR.

Regulárne výrazy, ktoré sú v programe zakomponované, by mali byť univerzálne pre všetky jazyky, ktoré používajú latinu. Regulárne výrazy, ktoré môžu byť špecifické pre niektoré jazyky (napríklad úvodzovky), je však možné pomocou prepínačov vypnúť.

Program neumožňuje iba automatickú opravu, ale aj zvýraznenie všetkých nájdených chýb. Chyby sa najskôr vyhľadajú aj pomocou trigramových modelov aj pomocou regulárnych výrazov. Ak užívateľ zvolí možnosť automatickej opravy, potom sa program pokúsi chybu opraviť. V opačnom prípade ju iba zvýrazní. Pre niektoré chyby však možnosť opravy nemusí existovať. V takomto prípade si užívateľ môže zvoliť, či chce takéto chyby zvýrazniť.

2. 5. Trénovacie dáta

Program by mal fungovať pre slovenčinu, češtinu a angličtinu. Pre každý z týchto jazykov je potrebné získať korektný korpus, výstup OCR softwaru a korektúru výstupu. Najväčší problém je získať dostatočne veľkú korektúru výstupu OCR softwaru.

Pre slovenčinu boli všetky tri súbory získané zo Zlatého fondu Slovenskej literatúry [4]. Zlatý fond sa zaoberá digitalizáciou diel slovenskej literatúry. Dobrovoľní digitalizátori spracovali doteraz takmer 900 diel. Každý zoskenovaný text kontrolujú po sebe navzájom traja digitalizátori, preto sú opravené súbory veľmi kvalitné. Korektný korpus pre slovenčinu obsahuje 1169441 slov. Opravený text obsahuje takmer 100000 slov.

Pre češtinu a angličtinu bol korektný korpus získaný z korpusu Czeng [5], vybrané boli texty pre češtinu a angličtinu z Project Syndicate. Text pre češtinu má 1513837 slov, text pre angličtinu 1698701 slov. Opravené texty majú naproti slovenčine iba 7704 slov pre češtinu a 5272 slov pre angličtinu.

3. Výsledky

3. 1. Testovacie súbory

Úspešnosť OCR softwaru sa u rôznych programov líši. Dôležitá je pritom kvalita vstupných dát. OCR softwary majú už korekciu výsledných textov v rôznej miere zabudovanú. Často používajú slovníkový prístup.

Pre slovenčinu bol program testovaný na šiestich súboroch, ktoré boli vytvorené pomocou dvoch rôznych softwarov, ABBY FineReader a ReadIris, pričom kvalita vstupných textov bola rôzna. Pre češtinu a angličtinu boli prevedené po dva testy. Testovacie súbory podstatne líšia počtom chýb v texte.

3. 2. Výsledky testov

Výsledky, ktoré boli dosiahnuté pri automatickej oprave chýb na testovacích textoch pre slovenčinu, češtinu a angličtinu, sú uvedené v Tabuľke 2.

Súbor	Jazyk	Počet slov	Počet chýb	Správne opravených	Nesprávne opravených	Precision	Recall
Test 1	SK	317	37	20	2	90,91%	54,05%
Test 2	SK	299	36	22	4	84,62%	61,11%
Test 3	SK	491	2	2	0	100,00%	100,00%
Test 4	SK	637	4	4	1	80,00%	100,00%
Test 5	SK	319	19	11	0	100,00%	57,89%
Test 6	SK	343	26	18	3	85,71%	69,23%
Test 7	CZ	321	18	4	0	100,00%	22,22%
Test 8	CZ	544	36	9	5	64,29%	25,00%
Test 9	EN	460	3	1	0	50,00%	33,33%
Test 10	EN	297	14	3	3	50,00%	21,43%

Tabuľka 2: Výsledky automatickej opravy chýb

Prevažná väčšina chýb, ktoré sa vyskytovali v súboroch, bola technického charakteru. Výsledky veľmi záležia od veľkosti trénovacích dát. Pre češtinu bol súbor s korektúrou výstupu OCR softwaru výrazne menší ako pre slovenčinu. Hodnoty precision a recall teda klesli. Pre angličtinu bol tento trénovací súbor ešte menší ako pre češtinu a hodnoty precision a recall sa ešte viac znížili. Všetky opravy v angličtine boli technického charakteru. Pre rôzne jazyky však zvolený prístup môže fungovať odlišne, preto výsledky pre rôzne jazyky záležia aj na charakteru daného jazyka.

4. Záver

Hlavným cieľom práce bolo vytvoriť program, ktorý by automaticky opravil chyby v dokumentoch vytvorených pomocou OCR softwaru. Bol použitý postup založený na trigramových modeloch. Tento postup neumožňuje opraviť všetky chyby, ale môže zlepšiť kvalitu textu. Pri dostatočnom počte trénovacích dát je zlepšenie kvality textu znateľné. Tým môže program pomôcť pri získavaní kvalitnejších dát pre automatický preklad. Program tiež umožňuje automatickú opravu niektorých technických chýb, ktoré sa často v týchto dokumentoch objavujú a je potrebné ich pracne opravovať. Program tak môže uľahčiť prácu aj korektorom, ktorí takéto dokumenty opravujú ručne.

5. Literatúra

- [1] *Z papíru do čtečky 5: Technická korektura*. Dostupné na internete: <<http://www.pepak.net/e-books/z-papiru-do-ctecky-5-technicka-korektura/>>
- [2] Stina Nylander: *Statistics and Graphotactical Rules in Finding OCR-errors*. Dostupné na internete: <<http://stp.ling.uu.se/exarb/arch/2000-001.pdf>>
- [3] *Aho-Corasick*. <<http://search.cpan.org/~vbar/Algorithm-AhoCorasick-0.02/lib/Algorithm/AhoCorasick.pm>>
- [4] *Zlatý fond*. <<http://zlatyfond.sme.sk/>>
- [5] *Czeng* <<http://ufal.mff.cuni.cz/czeng/>>