BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2022
http://excel.fit.vutbr.cz

Uspořádání zpřeházených řádků s pomocí jazykového modelu

Michael Holubec*

Abstrakt

Práce se zabývá stanovením posloupnost čtení (Reading order) textových regionů u digitalizovaných dokumentů. Identifikace posloupnosti čtení je jednou z důležitých součástí při rekonstrukci a extrakci obsahu digitalizovaných dokumentů. Kromě zavedených metod využívajících k sestavení posloupností čtení prostorových informací, práce zkoumá také možnost využití textového obsahu dokumentu a jeho analýzy pomocí jazykového modelu.

Na datasetu třinácti novinových článků porovnává úspěšnost identifikace správné posloupnosti čtení pomocí prostorové analýzy, jazykové analýzy a kombinované analýzy. Prostorová analýza dle provedených experimentů dosahuje 85 % úspěšnosti. Samotný jazykový model poskytuje velmi omezené výsledky (úspěšnost 16 %), jeho užití v kombinaci s prostorovou analýzou však zvyšuje úspěšnost z původních 85 % na 89 %.

Výstupem práce jsou mechanismy identifikující posloupnost čtení, které mohou sloužit pro dodatečné zpracování digitalizovaných dokumentů. Rovněž poskytují robustní základ pro případná další rozšíření a vylepšení přesnosti identifikace posloupnosti čtení.

Klíčová slova: Reading order — Posloupnost čtení — Prostorová analýza — Jazyková analýza — Jazykový model — LSTM — OCR

Přiložené materiály: GitHub

*xholub31@stud.fit.vutbr.cz, *Fakulta informačních technologií Vysokého učení technického v Brně*

1. Úvod

Systémy pro automatický přepis textu z obrázků (Optical Character Recognition, OCR) poskytují zpravidla přepis odstavců po řádcích [1]. Pro čtení přepisu je žádoucí odstavce seřadit tak, jak za sebou logicky následují; systémy pro detekci odstavců v obraze k tomu však zpravidla nejsou trénovány. Proto je explicitní tvorba posloupnosti čtení samostatným krokem při tvorbě plnohodnotného přepisu stránky [2].

Identifikace posloupnosti čtení je jednou ze součástí komplexního procesu nazývaného jako porozumění dokumentu (Document understanding), během kterého jsou klasifikovány prvky stránky (titulek, odstavec, autor) a vyhledávány vztahy mezi nimi, třeba jako posloupnost čtení (Reading order) těchto prvků [3]. Z hlediska posloupnosti čtení mohou být dokumenty různě komplexní: od souvislého textu v knize, přes novinové stránky obsahující řadu článků a nesouvisející inzerce sázené do více sloupců, až po velmi strukturovaná data ve fakturách nebo matrikách. Rozpoznání posloupnosti čtení u dokumentů s jedním sloupcem (standardně knihy) je poměrně přímočaré. Naopak rozpoznání posloupnosti u novinových článků může být značně problematické a z důvodu nezávislosti člán-

ků a dalších elementů také nejednoznačné, respektive může existovat více správných variant.

Jednou z metod pro rozpoznání posloupnosti čtení je prostorová analýza. Ta využívá prostorových údajů jak textových regionů, tak také různých pomocných prvků, například optických oddělovačů odstavců. Mezi prostorové údaje patří souřadnice jednotlivých regionů, tvar, velikost, poměr hustoty písmen proti velikosti plochy regionu, vzdálenost mezi regiony, sousednost, překrývání regionů a mnoho dalších. Tyto vlastnosti jsou použity pro definování vzájemných vztahů mezi regiony a určení uspořádání [4].

Protože se prostorová analýza spoléhá pouze na prostorové údaje, stává se, že v některých případech jsou v posloupnosti bezprostředními sousedy prvky, které spolu po jazykové a obsahové stránce nijak nesouvisí. Proto se kromě využití prostorových údajů nabízí využít samotného obsahu textových regionů.

Mechanismem, který je schopný analyzovat textový obsah, je jazykový model. Jazykový model je schopný určit pravděpodobnost výskytu sekvence [5]. Tato práce proto popisuje princip a implementaci dalších dvou metod, kterými jsou jazyková analýza a kombinovaná analýza. Jazyková analýza odhaduje posloupnost čtení čistě na základě pravděpodobností návaznosti dvou textových regionů. Kombinovaná analýza sdružuje prostorovou analýzu a jazykovou analýzu, kdy je výstup prostorové analýzy ovlivněn výpočty pravděpodobností jazykového modelu.

Článek podrobněji popisuje všechny tři zmíněné metody a představuje výsledky metod na připraveném datasetu 13ti novinových článků.

2. Metriky pro měření úspěšnosti detekce posloupnosti čtení

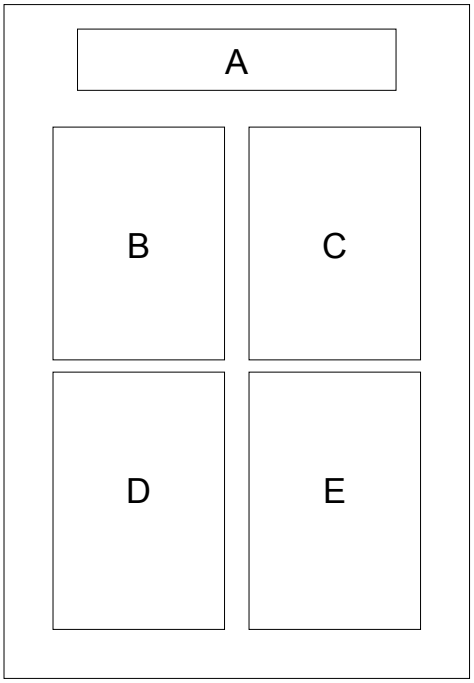
Posloupnost čtení lze definovat jako binární relací, pro kterou platí:

$$\forall a,b \in X, aRb \vee bRa, \tag{1}$$

kde X je množina všech textových regionů stránky. Taková relace definuje pořadí jednotlivých regionů v posloupnosti čtení. Dvojice (a,b) vyjadřuje vztah, kdy po regionu a následuje v posloupnosti region b, tedy že b je *bezprostředním následníkem* a. Následník, který není bezprostřední, je *prostým následníkem*.

2.1 Recall

Pro měření úspěšnosti identifikované posloupnosti čtení je užito dvou metrik; Recall a Prima. Recall (viz. 2) je základní metrika, která udává podíl počtu správně identifikovaných dvojic TP vůči celkovému počtu dvojic v referenční posloupnosti čtení P. Jednoduchost



Obrázek 1. Příklad rozložení stránky. U takového rozložení lze identifikovat dvě posloupnosti čtení: $\{(a,b),(b,c),(c,d),(d,e)\}$ nebo $\{(a,b),(b,d),(d,c),(c,e)\}$. Bez podrobnějších informací není možné určit, která posloupnost je správná. Pro tuto stránku platí, že b je bezprostředním následníkem a a že e je prostým následníkem a.

výpočtu této metriky je také její nevýhodou, protože bere v úvahu pouze bezprostřední následníky a neřeší případy, kdy b je prostým následníkem a.

$$\text{Recall} = \frac{TP}{P} \tag{2}$$

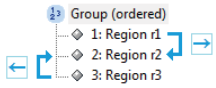
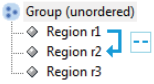
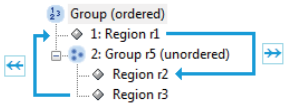
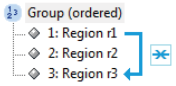
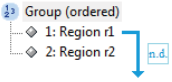
2.2 Prima

Prima [6] je komplexní metrika, která rozlišuje 7 různých vztahů, viz 2. Zavádí pojem skupiny (Group). Skupina může být uspořádaná (Ordered, záleží na pořadí regionů) nebo neuspořádaná (Unordered, nezáleží na pořadí regionů). Jednotlivé skupiny je možné různě zanořovat a vytvářet stromovou strukturu. Tato metoda hlídá a penalizuje případné nesprávné zařazení regionu do příslušné skupiny, je schopna zohlednit vztahy jak bezprostředních, tak prostých následníků. Metrika definuje chybovou matici M o velikosti 7x7, která porovnává identifikovaný vztah x vůči vztahu ground truth y a určuje velikost chyby p této dvojice.

$$p = M_{xy} \tag{3}$$

Celková chyba e je sumou hodnot dílčích chyb jednotlivých identifikovaných dvojic i.

$$e = \sum_i p_i \tag{4}$$

→	Direct successor	
←	Direct predecessor	
--	Fully unordered relation (e.g. both in same unordered group)	
→→	Somewhere before (but unordered group involved)	
←←	Somewhere after (but unordered group involved)	
-x-	Neither direct nor unordered relation	
n.d.	Relation not defined (one or both regions not in reading order tree)	

Obrázek 2. Tabulka vztahů, které jsou definovány a kontrolovány metrikou Prima. Převzato z [6].

Ta je dále normalizována maximální hodnotou p_{max} dílčí chyby a počtem regionů v referenční posloupnosti čtení n_{GT} a vyjádřena procentuální hodnotou s [6].

$$e_{50} = \frac{p_{max} * n_{GT}}{2} \tag{5}$$

$$s = \frac{1}{e * \frac{1}{e_{50}} + 1} \tag{6}$$

3. Prostorová analýza posloupnosti čtení

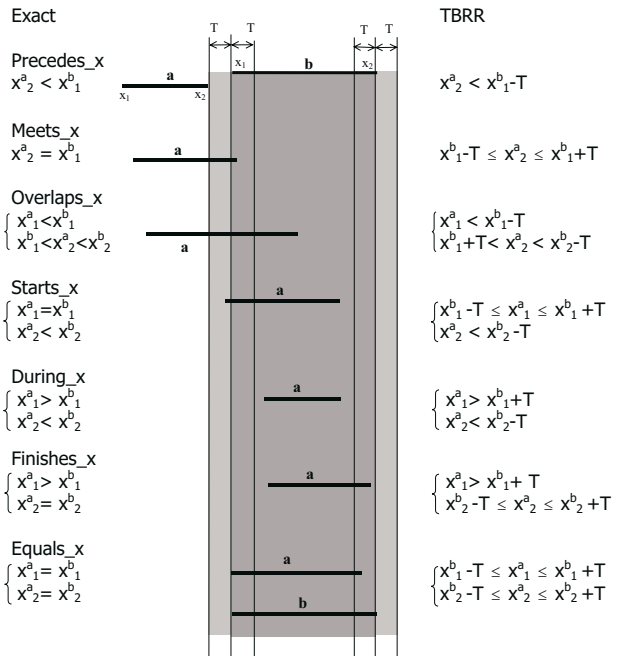
Implementace prostorové analýzy byla inspirována článkem *Document Understanding for a Broad Class of Documents* [4]. Je založena na prostorových vztazích mezi jednotlivými regiony. Definuje celkem 13 vztahů, z nichž 7 je standardních, viz Obrázek 3. Zbývajících 6 vztahů je inverzních vůči standardním (vztah *equals* je inverzní sám k sobě). Pro každý region je určen vztah vůči všem ostatním regionům na stránce. Výsledkem vyhodnocení je struktura, která nese dvě matice vztahů – jedna matice pro vztahy regionů na ose x, druhá matice pro vztahy na ose y. Příklad takové matice se vztahy je uveden v Tabulce 1, ve které jsou specifikovány vztahy na x-ové ose regionů z Obrázku 1.

Matice vztahů je vstupem algoritmu, který definuje uspořádání jednotlivých regionů. Výstupem algoritmu je ostře uspořádaná množina dvojic $\{(a,b),(b,c),(a,c),...\}$, kde platí, že region a předchází regionu b, b předchází c, atd.

Výstup lze interpretovat jako acyklický orientovaný graf, kde a, b a c jsou vrcholy a (a,b) je orientovaná hrana a tedy je možné použít topologické řazení,

Tabulka 1. Matice prostorových vztahů na ose x pro jednotlivé regiony z Obrázku 1. Vztahy s prefixem i_ vyjadřují inverzní vztah, například i_precedes vyjadřuje vztah následníka, zatímco precedes je vztah předchůdce. Dle obrázku B předchází C => precedes, ale C následuje po B => i_precedes.

X	A	B	C	D	E
A	equals	i_overlaps	overlaps	i_overlaps	overlaps
B	overlaps	equals	precedes	equals	precedes
C	i_overlaps	i_precedes	equals	i_precedes	equals
D	overlaps	equals	precedes	equals	precedes
E	i_overlaps	i_precedes	equals	i_precedes	equals



Obrázek 3. Ilustrace možných vztahů dvou regionů pro jednu osu. Region a je definován na intervalu $[x^a_1, x^a_2]$, region b pak na intervalu $[x^b_1, x^b_2]$. Aby spolu dva regiony byly v určitém vztahu, musí jejich vzájemné polohy splnit určité podmínky. Například regiony jsou spolu ve vztahu *precedes* (a předchází b) pouze tehdy, pokud platí, že $x^a_2 < x^b_1$ (interval regionu a končí dříve, než začíná interval regionu b). Pro eliminaci menších nepřesností při identifikaci hranic regionů je zavedená pomocná hodnota T, pomocí které jsou intervaly regionů při hledání vzájemného vztahu uměle rozšířeny. Převzato z [4].

jehož výstupem je seřazený seznam vrcholů interpretovatelný jako posloupnost čtení textových regionů na stránce.

Obecnou vlastností topologického řazení je, že výstup řazení nemusí být unikátní, respektive možných výstupů topologického řazení může být více. Z toho důvodu bylo přikročeno k úpravě analýzy, která spočívá ve sloučení takových textových regionů, u kterých je

identifikováno, že patří do jednoho celistvého sloupce. Sloupce jsou rozpoznány na základě minimální vzdálenosti mezi regiony, vztahů na ose x (vztah equals) i ose y (bezprostřední sousedi) a podobě. Sloučením do sloupců došlo k redukci počtu regionů, pro které je potřeba vyhodnotit vztahy, a také k redukci prvků ostře uspořádané množiny. Tím je omezen výčet možných výstupů topologického řazení.

Při identifikaci vztahů jednotlivých regionů se může objevit problém nepřesnosti, kdy vztah logicky navazujících regionů nemusí být dokonale rozpoznán. Může k tomu dojít například kvůli nepřesnostem při detekci hranic regionů. Pro odstranění tohoto problému je použita hodnoty prahová hodnota T, viz Obrázek 3. Prahová hodnota T vnáší do vyhodnocení vzájemných vztahů jistou míru benevolence a to tak, že hranice regionů při analýze dle potřeby rozšíří nebo zúží. Je vypočtena z nejmenšího objektu na stránce, konkrétně hodnota nepřesahuje polovinu délky nejmenší strany takového objektu. Pro každou stránku je nutné počítat vlastní prahovou hodnotu, která je pro tuto stránku konstantní (je platná pro všechny regiony).

4. Statistické jazykové modelování a jazyková analýza

Statistické jazykové modelování je obor zabývající se zkoumáním a tvorbou modelů, které jsou schopné určit pravděpodobnost výskytu sekvence o dané skladbě slov. Pravděpodobnost sekvence slov $w_1 \dots w_n$ je pravděpodobnost společného výskytu slov $P(w_1 \dots w_n)$. Takovou pravděpodobnost je možné rozdělit na jednotlivé složky:

$$P(w_1 \dots w_i) = P(w_1) \cdot P(w_2|w_1) \cdot \dots \\ \dots \cdot P(w_i|w_1 \dots w_{i-1}), \quad (7)$$

kde $P(w_1)$ je pravděpodobnost výskytu slova w_1, $P(w_2|w_1)$ podmíněná pravděpodobnost výskytu slova w_2 pokud tomuto slovu předcházelo slovo w_1 a tak dále [5].

Mezi jazykové modely řadíme například n-gramové modely, které aproximují pravděpodobnost omezenou délkou kontextu nebo moderní neuronové sítě, které jsou schopny aproximovat pravděpodobnost v delším časovém horizontu. V rámci této práce byla použita neuronová síť LSTM (*Long-Short Term Memory*) [7]. Pro tokenizaci byl pomocí nástroje SentencePiece [8] vytvořen subword model, pomocí kterého probíhal převod textu na tokeny, které neuronová síť zpracovávala.

V rámci této práce byla implementovaná jazyková analýza, která využívá jazykový model a obsah textových regionů pro identifikaci posloupnosti čtení. Metoda pracuje v cyklu, ve kterém porovnává podmíněnou pravděpodobnost všech textových regionů N. Vzájemným porovnáním vznikne matice o velikosti $N \times N$, která obsahuje výsledky podmíněných pravděpodobností všech textových regionů. Pro určení posloupnosti čtení jsou vybrány takové dva regiony X a Y, jejichž podmíněná pravděpodobnost $P(Y|X)$ je, v porovnání s ostatními, nejvyšší. Tyto dva regiony poté tvoří v posloupnosti čtení bezprostřední následníky (X,Y) a do dalšího kola vyhodnocení vstupují jako nový řetězec Z, který je výsledkem konkatenace těchto dvou regionů $Z = X \cdot Y$. Původní regiony X a Y jsou z vyhodnocení vyjmuty. Cyklus je ukončen jakmile je každý region zařazen v posloupnosti čtení a tím identifikována posloupnost pro všechny regiony.

5. Kombinovaná analýza posloupnosti čtení

Kombinovaná analýza, implementovaná v této práci, kombinuje výstup prostorové analýzy s výstupem jazykového modelu. Je rozdělena na dvě fáze. V první fázi je provedena kompletní prostorová analýza tak, jak je popsáno v kapitole 3. V druhé fázi je výstup prostorové analýzy podroben doplňující jazykové analýze.

Oproti vzájemnému porovnání všech regionů, jak je popsáno v sekci 4, je jazykový model použit k ohodnocení podmíněné pravděpodobnosti několika málo regionů. Pro každý *zdrojový* region, například konec sloupce, jsou identifikovány *kandidátní* regiony, kterými je možné na zdrojový region navázat. Mezi kandidátní regiony patří regiony v bezprostředním okolí, které by mohly logicky navazovat. Každý kandidátní region Y_i je společně se zdrojovým regionem X předložen jazykovému modelu. Ten vypočte podmíněnou pravděpodobnost $P(Y_i|X)$ pro všechny kandidátní regiony Y. Pokud je výstup jazykové analýzy odlišný od výstupu prostorové analýzy a výsledná pravděpodobnost překročí určitý práh (80 % pro dva kandidátní regiony, 65 % pro tři a více), pak dvojice (X,Y_i) tvoří nový prvek posloupnosti.

Výstupem kombinované analýzy je posloupnost identifikovaná prostorovou analýzou, která je dodatečně upravena jazykovým modelem.

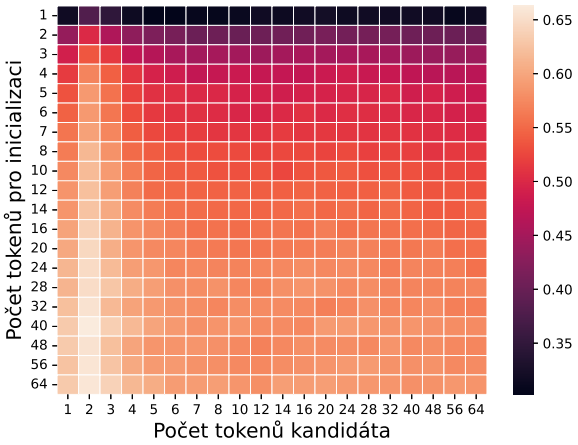
6. Vytvoření a vyhodnocení úspěšnosti jazykového modelu

Jazykový model, použitý při jazykové a kombinované analýze, byl postaven na LSTM architektuře a trénovaný na korpusu české Wikipedie. Korpus byl rozdělen na

trénovací (725MB), validační (3MB) a testovací (3MB) sadu. Pro převod korpusu byl natrénován subword model s velikostí 20 000 tokenů. Trénování sítě probíhalo po batchích, jeden batch sestával z 20 sekvencí, každá sekvence obsahovala 35 tokenů. Samotné trénování probíhalo na GPU, proběhlo celkem 28 epoch a na testovací sadě bylo dosaženo perplexity 32,38.

Protože různé délky sekvencí mají různý vliv na úspěšnost odhadu návaznosti, respektive výpočet podmíněné pravděpodobnosti, byl proveden experiment pro zjištění vlivu délek a zjištění ideální konfigurace pro odhady pravděpodobností. Konfigurací je míněno ideální nastavení délky *zdrojové* sekvence a délky *kandidátní* sekvence pro maximalizaci úspěšné identifikace posloupnosti čtení. V rámci experimentu byly zkoumány kombinace obsahující 1–64 tokenů. Pro každou takovou kombinaci bylo provedeno 10 000 testů, při každém testu bylo rozhodnuto o návaznosti jednoho z 15ti možných kandidátů. Jeden kandidát byl skutečným následníkem, ostatní byly náhodně vybrané. Měřen byl počet správných rozhodnutí jazykového modelu o následníkovy zdrojové sekvence. Experiment byl proveden na testovací sadě a výsledek celého experimentu je znázorněn na Obrázku 4.

Z experimentu vyplynulo, že ideální konfigurací je co nejdelší zdrojová sekvence s omezenou délkou kandidátní sekvence, nejlépe o velikosti dvou tokenů. Tato konfigurace byla použita jak v případě jazykové analýzy, tak i v případě kombinované analýzy a je zahrnuta i ve všech experimentech popsaných v sekci 7.



Obrázek 4. Grafické znázornění průměrné úspěšnosti jazykového modelu v odhadu návaznosti dvou sekvencí. Délka *zdrojové* sekvence je znázorněna na ose y, délka *kandidátních* sekvencí na ose x. Škála udává procentuální úspěšnost správně identifikované návaznosti z celkem 15 kandidátů. Z obrázku je zřejmé, že největší úspěšnosti jazykový model dosahuje, pokud je délka *zdrojové* sekvence co největší a zároveň délka *kandidátní* sekvence je omezena na dva tokeny.

Tabulka 2. Výsledky jednotlivých metod, kterých bylo dosaženo na připraveném datasetu novinových článků.

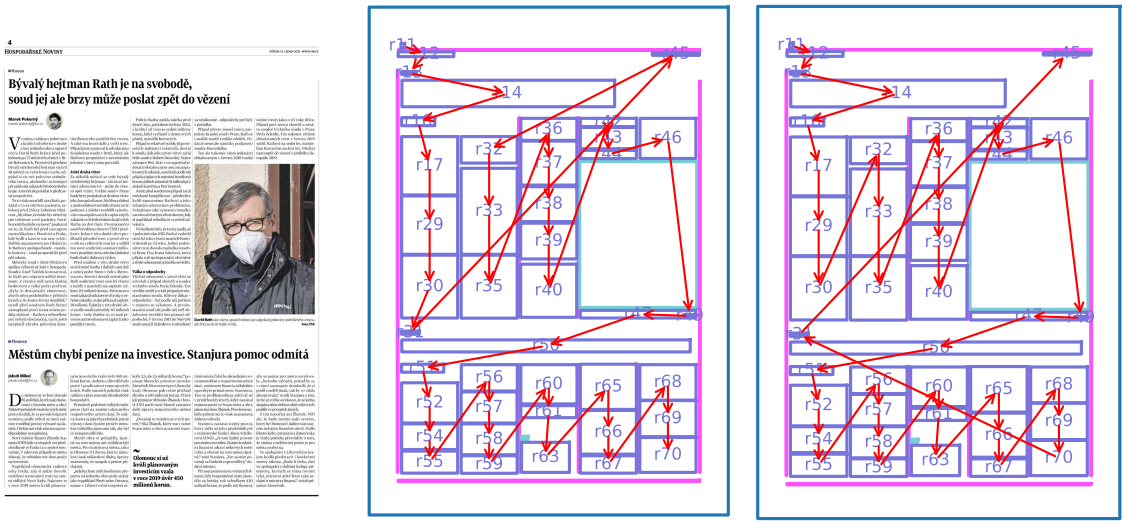
	Prima	Recall
Kombinovaná analýza	**92,34** %	**88,87** %
Prostorová analýza	89,83 %	84,89 %
Tesseract	78,59 %	70,98 %
Jazyková analýza	51,40 %	16,11 %
Top-to-bottom	50,73 %	11,69 %

7. Výsledky vyhodnocení na vytvořeném datasetu

Algoritmus prostorové, jazykové a kombinované analýzy byl testován na sadě 13ti novinových stránkách Hospodářských novin vydaných 12. 01. 2022. Jednotlivé novinové články jsem předzpracoval nástrojem Aletheia [9]. Tímto nástrojem jsem také provedl OCR analýzu, jejímž výstupem byly kromě textových řetězců a identifikovaných obrázku, také prostorové údaje jednotlivých regionů (polygony představující ohraničení textových regionů, pozice v prostoru a další). Dále jsem anotoval posloupnosti čtení jednotlivých článků, které při vyhodnocení implementovaných algoritmů sloužily jako referenční posloupnost čtení.

Na připravené datové sadě (PageXML novinových článků) byla změřena úspěšnost identifikace posloupnosti čtení pro všechny tři algoritmy. Každým algoritmem bylo vyhodnoceno všech 13 stránek datasetu, výsledky byly následně zprůměrovány. Na dané datové sadě dosáhl nejlepších výsledků algoritmus kombinované analýzy. Prostorová analýza dosáhla, v porovnání s kombinovanou analýzou, mírně horšího výsledku. Úspěšnost samotné jazykové analýzy je velmi nízká. Pro porovnání byla do výsledku zahrnuta také konvenční metoda Top-to-bottom (metoda zleva doprava, shora dolů) a posloupnost čtení identifikována nástrojem Tesseract (OCR engine, implicitní pořadí textových regionů lze interpretovat jako pořadí čtení). Celkové výsledky jsou uvedeny v Tabulce 2.

Kombinovaná analýza dosahuje nejlepších výsledků nejspíše z toho důvodu, že pro definici posloupnosti čtení využívá jak prostorových, tak jazykových informací. Proti jazykové analýze má kombinovaná analýza tu výhodu, že pokud provádí analýzu návaznosti pomocí jazykového modelu, pak jej provádí jen pro kandidáty z nejbližšího okolí. V případě jazykové analýzy jsou testu posloupnosti podrobeny i regiony, které spolu logicky nijak nesouvisí.



(a) Originál **(b)** Prostorová analýza **(c)** Kombinovaná analýza

Obrázek 5. Obrázek znázorňuje posloupnosti čtení identifikované prostorovou a kombinovanou analýzou. Pro ilustraci je přiložena originální stránka. Lze poukázat na region *r30* (vlevo uprostřed), který v případě prostorové analýzy 5b má jako bezprostředního následníka v kontextu článku nesouvisející region. Aplikováním jazykového modelu v kombinované analýze 5c došlo k přehodnocení a opravě bezprostředního následníka.

8. Závěr

Práce zkoumala možnosti identifikace posloupnosti čtení digitalizovaných dokumentů s komplexnějším rozložením, primárně novinovými stránkami. Navrhla metodu jazykové analýzy a kombinované analýzy, která mírně vylepšuje chování prostorové analýzy.

Jazyková analýza definuje posloupnost čtení na základě porovnání podmíněných pravděpodobností všech textových regionů na stránce. Kombinovaná analýza pracuje ve dvou fázích. V první fázi provede prostorovou analýzu a na základě vzájemných vztahů regionů definuje posloupnost čtení. V druhé fázi je vytvořeným jazykovým modelem výstup prostorové analýzy korigován na základě jazykových vlastností regionů. Výstupem je posloupnost čtení kombinující přístup prostorové a jazykové analýzy.

Na připravené datové sadě bylo provedeno měření úspěšnosti jednotlivých metod. Nejnižší průměrné úspěšnosti dosáhla jazyková analýza, která v metrice Prima dosáhla 51,4 % úspěšnosti a Recall 16,11 %. Mnohem lepší průměrné úspěšnosti dosáhla prostorová analýza, která v Prima metrice dosáhla 89,83 % a v Recall 84,86 %. Nejlepších průměrných výsledků na datasetu novinových článků dosáhla kombinovaná metrika, která využívá obou předchozích metod. V Prima metrice dosahuje hodnoty 92,34 % a v Recall 88,87 %.

Práce představila tři metody pro identifikaci posloupnosti čtení, z nichž dvě (prostorovou a kombinovanou analýzu) je možné použít v reálném prostředí při extrakci obsahu digitalizovaných dokumentů.

Poděkování

Rád bych poděkoval vedoucímu práce Ing. Karlu Benešovi za odborné vedení práce, cenné rady, trpělivost a čas, který mi věnoval v rámci konzultací.

Literatura

[1] Sargur N. Srihari, Ajay Shekhawat, and Stephen W. Lam. *Optical Character Recognition (OCR)*, page 1326–1333. John Wiley and Sons Ltd., GBR, 2003.

[2] Anoop Namboodiri and Anil Jain. *Document Structure and Layout Analysis*, pages 29–48. 03 2007.

[3] Leon Todoran, Marco Aiello, Christof Monz, and Marcel Worring. Logical structure detection for heterogeneous document classes. 11 2000.

[4] Marco Aiello, Christof Monz, Leon Todoran, and Marcel Worring. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition*, 5(1), 08 2002.

[5] Joshua Goodman. A bit of progress in language modeling. *CoRR*, cs.CL/0108005, 2001.

[6] Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. The significance of reading order in document recognition and its evaluation. 08 2013.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[8] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018.

[9] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia - an advanced document layout and text ground-truthing system for production environments. In *2011 International Conference on Document Analysis and Recognition*, pages 48–52, 2011.