OCR Post-Correction with Heuristics

Michał Marcińczuk

michal.marcinczuk@samurailabs.ai

Samurai Labs / Wrocław University of Science and Technology

October 2021

0 Outline

1 Outline

1 Task

OCR output

Znałem go dobrze—mówił J n ksa—\nprzyjaźniliśmy się,
należeli do jednego gro - \nna;

Expected post-correction

Znałem go dobrze — mówił Jaksa — przyjaźniliśmy się,
należeli do jednego grona;

2 Outline

2 Data analysis — *difflib*

```
i: Znałem go dobrze-mówił Jnksa-\nprzyjaźniliśmy się, należeli do jednego gro-\textbackslashnna;
e: Znałem go dobrze - mówił Jaksa - przyjaźniliśmy się, należeli do jednego grona;

equal     i[0:16]  --> e[0:16]    Znałem go dobrze --> Znałem go dobrze
insert    i[16:16] --> e[16:17]                 '' --> ' '
equal     i[16:17] --> e[17:18]                  - --> -
insert    i[17:17] --> e[18:19]                 '' --> ' '
equal     i[17:24] --> e[19:26]           mówił J --> mówił J
replace   i[24:25] --> e[26:27]                 n --> a
equal     i[25:28] --> e[27:30]               ksa --> ksa
insert    i[28:28] --> e[30:31]                 '' --> ' '
equal     i[28:29] --> e[31:32]                  - --> -
replace   i[29:30] --> e[32:33]                 \textbackslashn --> ' '
equal     i[30:73] --> e[33:76] przyjaźniliśmy się, należeli do jednego gro
                                --> przyjaźniliśmy się, należeli do jednego gro
delete    i[73:75] --> e[76:76]              -\textbackslashn --> ''
equal     i[75:78] --> e[76:79]              na; --> na;
```

2 Categories of errors [1/3]

Hyphenation

Some words at the end of the line are broken between lines — the parts were separated with hyphen (-). In some cases, OCR recognized hyphens as other punctuation marks — equal sign (=) or guillemets («, »).

Punctuation

- opening round bracket was recognized as upper letter C,
- exclamation mark (!) was recognized as pipe (|),
- double quotations mark („) was recognized as two commas (,,),
- full stop (.) was recognized as comma (,) and vice-versa.

2 Categories of errors [2/3]

Character swaps

- more than 40 pairs of characters — typical OCR errors, for example:
  - a →ą, u
  - n →u, m
- applied only for words not present in the dictionaries and if there was no ambiguity,
- SGJP dictionary and a list of words from the training dataset (frequency > 10).

Word and phrase swaps

- swaps for characters were extended to words and bigrams to handle cases were single character replacement was ambiguous,

2 Categories of errors [3/3]

Trashes

- OCR systems tend to generate semi-random sequences of characters when they encounter non-text elements on the image — stains, marks, and so on.
- samples:
  > NYR UR UR UR UR UR UR UR UB UR UB UR UN WA YN UN UN UN,
  > «+ «+ «0556655 0% rsa .
  > = esz órie Fade JE3A.

3 Outline

3 Test-A and Test-B

	test-A	**test-B**
OCR output	18.726	-
hyphenation	9.750	-
punctuation	9.122	-
word swaps	8.945	-
char swaps	8.368	-
trashes	8.237	-
phrase swaps	8.062	-
Final	8.062	8.217

Table: WER on the test-A and test-B datasets

3 Leaderboard

	test-B
Mateusz Piotrowski	3.744
Krzysztof Wróbel	4.302
Paweł Dyda	5.328
Zbigniew Bronk	7.208
Michał Marcińczuk	**8.217**

Table: WER on the test-B datasets

3 Conclusions

- reduced the WER by a half, but the results were still much lower than the top ones — 8.217 vs 3.744 WER,
- fast processing time
  - 210 documents per second on a single CPU,
  - 40 seconds to process the whole dataset (8435 documents).
- there is still place for improvement but it might be more time-consuming :-)

Thank you for your attention