PDF Extraction Based on Lexical Analysis for Thai Texts

Santipong Thaiprayoon

Outlines

- Introduction and Problem
- A survey of PDF Extracting Tools
- The Overall Process of Thai PDF-PP
- Experiments and Discussion
- Conclusion and Future Work

Introduction and Problems

- PDF has become the universal exchange format for digital document.
- PDF is the ability to create and share documents across platforms with different operating systems and hardware environments.
- PDF has some drawbacks on search and extraction abilities.
- Extracting original text from PDF documents can be a very challenge problem.

Introduction and Problem (cont'd)

- Most of the open-source PDF extracting tools can handle low-level parsing and manipulation of objects in PDF documents. However, these tools do not fully support recovery of original text in reading order and complex structures.
- The errors are mainly caused by the misplacement of some characters in the resulting texts.
- For Thai language, the problem is more intensified due to the complex lexeme structure, i.e., character composition, of Thai words
- We propose an approach called PDF-PP (Thai PDF Post Processor), which performs text cleansing based on the lexical analysis.

A Survey of PDF Extracting Tools

- we compare between two open-source tools, PDFBox and Xpdf, which are widely used among developers.
- To compare the performance between PDFBox and Xpdf, we apply three most popular PDF generating tools: Microsoft Word, Adobe Acrobat Distiller and Open Office.

A Survey of PDF Extracting Tools (cont'd)

Example text: มากินกุ้งปิ้งในถ้ำ (Come and eat grilled shrimp in a cave)

PDF Generator	PDF Extractor	Result
Acrobat Distiller	PDFBox	ม า ก ิ น ก ุ ง ้ ป ้ ิ ง ใ น ถ ้ ำ
	Xpdf	ม ␣ า ก ิ น ก ุ ้ ง ป ิ ้ ง ใ น ถ ้ ำ
MS Word	PDFBox	ม า ก ิ น ก ุ ้ ง ป ิ ง ้ ใ น ถ ␣ า ้
	Xpdf	ม ␣ า ก ิ น ก ุ ้ ง ป ิ ้ ง ใ น ถ ้ า
Open Office	PDFBox	ม า า ก ิ น ก ้ ุ ง ป ้ ิ ง ใ น ถ ำ ้ า
	Xpdf	ม า า ก ิ น ก ุ ้ ง ป ิ ้ ง ใ น ถ ำ ้ า

Fig. 1. Result comparison from different PDF extracting tools

Type I error	**The vowel, Sara Am (ำ) is incorrectly converted into Sara Aa (า). For example, "สำคัญ" is incorrectly converted as "สาคัญ".**
Type II error	There is an inserted space before the vowel, Sara Aa (า.(For example, "กระเป๋า" is incorrectly converted as "กระเป๋ า".
Type III error	There is a randomly inserted space before any consonant. For example, "ป้องกัน" is incorrectly converted as "ป้อ งกัน".
Type IV error	The vowel, Sara Am (ำ) is incorrectly separated into two characters, Nikhahit (ำ) and Sara Aa (า).
Type V error	Sara Ae (แ) is incorrectly separated into two Sara E (เ).
Type VI error	There is an inserted Sara Aa (า) right after Sara Am (ำ).
Type VII error	The tonal mark is misplaced with the vowel. For example, the tonal mark, Mai Tho (้) is misplaced with Sara Uu (ู).
Type VIII error	The above and below vowels are shifted into the wrong position. For example, the sentence "ถ่านหินกาาซธรรมชาติก้าลังขาดัแคลนจัวนเจัยนจัะหมดัโลก".

Fig. 2. Error type from different PDF generating tools

The Overall Process of Thai PDF-PP

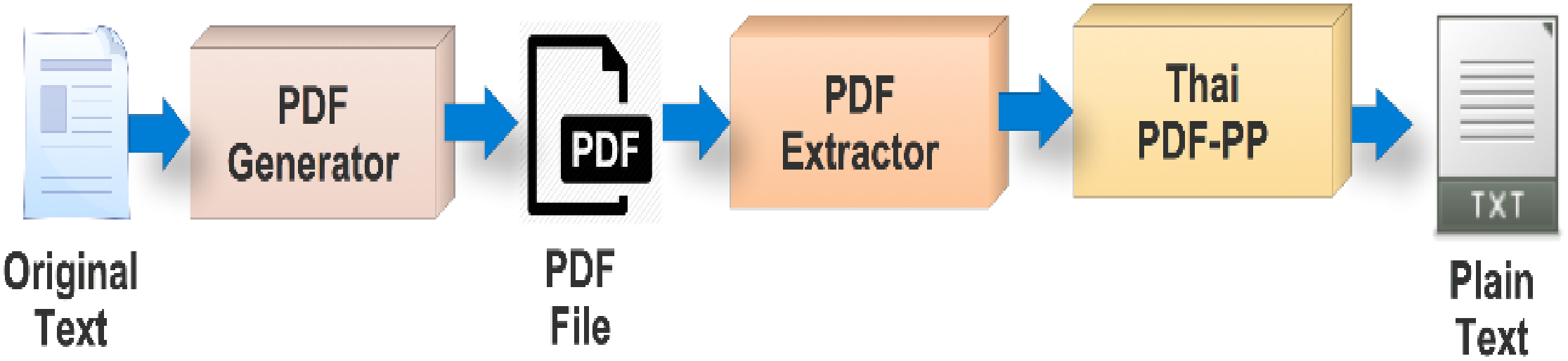

Fig. 3. PDF extraction process with the Thai PDF-PP

Type I error solution	**Automatically detect the words with missing Nikhahit (◌ं) from Sara Am (◌ं า), then try to insert Nikhahit (◌ं) and parse text. If the process yields a valid word, then the Nikhahit (◌ं) is correctly inserted.**
Type II error solution	Automatically detect Sara Aa (า) and remove the space in front of it. For example, the word “กระเป๋ า” is converted into “กระเป๋า”.
Type III error solution	Automatically detect space characters, then try to remove and parse text. If the process yields a valid word, then the space is correctly removed.
Type IV error solution	Automatically detect Nikhahit (◌ं) followed by Sara Aa (า), then merge them into Sara Am (◌ำ).
Type V error solution	Automatically detect two Sara E (เ), then merge them into Sara Ae (แ).
Type VI error solution	Automatically detect Sara Aa (า) right after Sara Am (◌ำ), then delete Sara Aa (า).
Type VII error solution	Automatically detect the order of the tonal mark and the vowel. If it is an invalid order then switch the order.
Type VIII error solution	No solution for Type VIII error yet.

Fig. 4. Solution type of error

Experiments and Discussion

- we perform experiments on eleven documents using the BEST corpus.
- The corpus contains approximately 500,000 words.
- We applied the Xpdf for extracting a PDF file to plain text.
- To measure the correctness of extracted content, we find common word sequence between the original text document in the BEST corpus and the extracted content.
- We then compare the performance of baseline (i.e., extracted content from Xpdf) and our proposed approach using accuracy measurement.

Experiments and Discussion (cont'd)

PDF Generator	Accuracy (%)	
	Baseline	Thai PDF-PP
Acrobat Distiller	96.93	99.78
MS Word	96.99	99.64
Open Office	60.79	65.54

Table1. Experimental results on extracted texts

Conclusion

- we propose an approach called Thai PDF-PP (Thai PDF Post Processor) which performs text cleansing based on the lexical analysis.
- The proposed approach helps increase the accuracy for all PDF generators by approximately 3 -5%.
- For future work, we plan to improve the accuracy of our approach by applying n-gram model to help reduce word ambiguity.

Thank You!

For more information please contact email.