Preparing books for alignment

André Santos
andrefs@cpan.org

http://slidesha.re/qXvZ9j

November 2011

Part I

Talking about ebooks and alignment

1. Introduction
2. Ebook sources
3. Ebook formats
4. Alignment challenges
5. Per-fide tools

1. Introduction

2. Ebook sources

3. Ebook formats

4. Alignment challenges

5. Per-fide tools

Introduction

- Alignment process results are highly dependent on the quality of the documents being aligned
- Several problems can affect the alignment: document format, page structure, text encodings, document sections, . . .
- Specific pre-processing can help!

This presentation

- Will be focused on preparing books for alignment
- Most of the methods described can be applied to other types of documents

1 Introduction

2 Ebook sources

3 Ebook formats

4 Alignment challenges

5 Per-fide tools

Ebook sources

- Project Gutenberg
- Websites and online communities
- Publishers
- Other sources

Project Gutenberg

- www.gutenberg.org
- Project Gutenberg offers over 36,000 ebooks free to download, available in several different formats.
- Books are free because their copyright has expired.
- Books are digitized and proofread by volunteers (Distributed Proofreaders).

Websites and online communities

- Websites which make available the national literary heritage
- Groups of people devoted to making ebooks available online for free
- Frequently motivated by other causes
  - e.g. visually impaired people can use speech synthesizers to "read" plain text books

Publishers

- Publishers selling books in electronic formats is becoming more frequent
- Not that many free books
- Amazon Kindle Store, Apple iBookStore, Barnes & Noble, O'Reilly, . . .

Other sources

- Even copyrighted books are available online “for free”
- Torrents, file sharing sites, book sharing groups, . . .
- Books shared by friends
- Frequently illegal

1 Introduction

2 Ebook sources

3 Ebook formats

4 Alignment challenges

5 Per-fide tools

Ebook formats

- Mobi (Kindle)
- Epub (Apple)
- Nook (B&N)
- Fiction Book (.fb2)
- HTML
- PDF
- ...

`http://en.wikipedia.org/wiki/Comparison_of_e-book_formats`

1. Introduction
2. Ebook sources
3. Ebook formats
4. Alignment challenges
5. Per-fide tools

Alignment challenges

- Document format
- Noisy text
- Ghost sections
- Results evaluation

Formats

- Before alignment, documents must be converted to plain text
- Different formats require different conversion tools
- Different documents may require different conversion options
- Plain text format lacks the notion of document structure, page structure, text format, etc
- Different notations are used to represent these elements

Noisy text

- pagination – page numbers, headers, footers, . . .
- previous text formatting – sub/superscript, bold, italics, . . .
- sections
- paragraphs
- translineations and transpaginations
- footnotes
- text encoding
- . . .

Ghost sections

- Sections which can only be found in one of documents being aligned
- Often enough to derail the alignment process
- Happens frequently with Prefaces, Prologues, Translator Notes, . . .

Results evaluation

- How to evaluate the results of an alignment?
- How to compare two different alignments of the same documents?

Per-fide tools

`bookcleaner`: cleans plain text books

`pairbooks`: finds duplicated books and candidate pairs

`syncbooks`: aligns books at section level

`tmx_inspect`: generates a small summary of a TMX file

`tmx_compare`: compares two TMX files

Per-fide tools

`bookcleaner`: cleans plain text books

`syncbooks`: aligns books at section level

`tmx_inspect`: generates a small summary of a TMX file

`tmx_compare`: compares two TMX files

Part II

Hands-on

Enough talking, let's do stuff!

Instructions

- Open browser in `http://shell.andrefs.info`
- login: **andrefs**
- password: **Perfide**
- `cd workshop/dir...`

Book noise problems – Example

(. . .)

```
gaiement. Sur le devant s<92>'ouvrait la porte
d<92>'entrée, donnant accès dans la salle commune.
Une légère véranda, qui en proté-

           <96>- 86 <96>-
^L geait la partie antérieure contre l<92>'action
des rayons solaires, reposait sur de sveltes bambous.
Le tout était peint d<92>'une fraîche
```

(. . .)

La Jangada, Jules Verne

Download

This was ~~awesome~~ OK, where can I get these tools?

- These tools are available in CPAN - `http://search.cpan.org/~andrefs`
- You'll need a Linux machine and a recent distribution of Perl
- Also, several Perl modules dependencies. Install with:

```
cpanm Text::Perfide::BookCleaner
```

Preparing books for alignment

André Santos
andrefs@cpan.org

http://slidesha.re/qXvZ9j

November 2011