Search is not only the Web
IR Applications

Walid Magdy
School of Informatics
University of Edinburgh

26 Oct 2022

Objectives

- Main objective of IR
- Two search tasks
  - Printed documents search
  - Patent search
- Possible ideas for Group Project 😊
**Information Retrieval Objective**

- IR is finding material of an unstructured nature that **satisfies an information need** from within large collections.

  - **Information need**
    - Expected search scenario?
    - Modeling the task?
  
  - **Data nature**
    - Approach?
    - Scalable? Fast?
  
  - **User Satisfaction**
    - More relevant documents?
    - Effective evaluation?

---

**Printed Documents Retrieval**
Printed Documents Retrieval

- **Documents:**
  text on printed papers (books)

- **Information need:**
  Information within these books

- **Challenge:**
  It is an image of text

- **Common Approach:**
  OCR $\rightarrow$ Recognized text $\leftarrow$ Search

- **Challenges in Common Approach:**
  OCR $\rightarrow$ Text with mistakes (WER$_A = 40\%$)
  OCR $\rightarrow$ Not available for all languages

---

Problem

- **Text with errors (sometime many errors)**

---

![Image of text with errors and a diagram showing IR effectiveness with different qualities of OCR]
n-gram Char Representation of OCR

- Original: example sentence
- OCR output: example sentence
- 3-gram char representation:
  $ex\ exa\ xar\ arn\ rmp\ npl\ ple\ le$ $se\ sen\ enl\ lcn\ cnc\ nce\ ce$
- Query: example sentence → $ex\ exa\ xam\ amp\ mpl\ ple\ le$ $se\ sen\ ent\ nte\ ten\ enc\ nce\ ce$
- Matching:
  $ex\ exa\ xar\ arn\ rnp\ npl\ ple\ le$ $se\ sen\ enl\ lcn\ cnc\ nce\ ce$
  $ex\ exa\ xam\ amp\ mpl\ ple\ le$ $se\ sen\ ent\ nte\ ten\ enc\ nce\ ce$

OCR Correction using Error Model

- OCR text
- Generate Candidates
- Best Fitting Word Selection
- Corrected text
- Select part and correct
- Manually corrected version

Use for search
Query Garbling using Error Model

Query → Generate possible errors → Query, Query, Query, ....

OCR Correction using Edit Distance

OCR text → Generate Candidates → Best Fitting Word Selection → Corrected text

Use for search
Multi-OCR Text Fusion

OCR text 1

Word Alignment

OCR text 2

Best Fitting Word Selection

Fused text

Language Model

Use for search

OCR Search

- Recognition errors in OCR text degrades retrieval
- Different methods of text processing can overcome the negative effect on retrieval and improves search
- n-gram character representation improves retrieval, but not that much
- Some training and resources are needed which can be manual correction, trained language model, or both
- Previous methods fail when errors are large (WER>50%)
Solution – back to Information Need

- Information need: the printed papers
- Question: Why convert image to text?
- Related work: Word Spotting

Modeling the Problem

Text Domain | Image Domain
---|---
Query | information
Draw | OCR
### OCRless Search

![Diagram of OCRless Search process]

1. **Indexing**
   - Input: Document/
   - Output: Index of IDs

2. **Create IDs**
   - Input: Document/
   - Output: Clusters

3. **Clustering**
   - Input: Clusters
   - Output: Segments to elements

4. **Segment to Elements**
   - Input: Segments to elements
   - Output: Document

5. **Draw Query**
   - Input: Search Index of IDs
   - Output: List of ranked documents

6. **Replace with Candidate IDs and Formulate Query**
   - Input: List of ranked documents
   - Output: Synonyms

7. **Synonyms**
   - Input: Synonyms
   - Output: Replacement candidates

---

### Solution – OCRless Search

![Diagram of Solution process]

1. **Draw Query**
   - Input: مسلم
   - Output: List of ranked documents

2. **Replace with Candidate IDs and Formulate Query**
   - Input: List of ranked documents
   - Output: Synonyms

3. **Synonyms**
   - Input: Synonyms
   - Output: Replacement candidates

---

**Institute for Language, Cognition and Computation**

**ILCC**

**The University of Edinburgh**
Solution – OCRless Search

- Effective and fast
- Robust to OCR errors (*vIdea*)
- No training resources required
- Language independent

- Microsoft TechFest Demo
  The same engine for searching printed documents in: Arabic, English, Chinese, Hebrew, and Hieroglyphic

Printed Documents Retrieval

- Text-based solutions: correction
- Image-based: clustering

- Current State-of-the-art?

- Information need → Approach
Patent Search

- Given a patent application, check if the invention described is novel

Query → Search → Results list

Patent application

Several languages

Many results to check 100-600 docs/search
Patent Search – User Satisfaction

- NTCIR, CLEF, TREC
- Recall-oriented → Try not to miss a relevant document
  - Recall is the objective
- Precision is also important
- Huge # documents checked (100-600 documents)

- Evaluation: average precision (AP)!!
  - Focuses on finding relevant docs early in ranked list
  - Less focus on recall

Example

For a topic with 4 relevant docs and 1st 100 docs to be examined:

System1: relevant ranks = \{1\}
System2: relevant ranks = \{50, 51, 53, 54\}
System3: relevant ranks = \{1, 2, 3, 4\}

\[
\begin{align*}
\text{AP}_{\text{system1}} &= 0.25 \\
\text{AP}_{\text{system2}} &= 0.0481 \\
\text{AP}_{\text{system3}} &= 1
\end{align*}
\]

\[
\begin{align*}
\text{R}_{\text{system1}} &= 0.25 \\
\text{R}_{\text{system2}} &= 1 \\
\text{R}_{\text{system3}} &= 1
\end{align*}
\]

- We need a metric that reflects recall and ranking quality in one measure
**PRES: Patent Retrieval Evaluation Score**

\[
PRES = 1 - \frac{\sum r_i - \frac{n(n+1)}{2}}{N_{\text{max}}} \quad n: \text{number of relevant docs}  \\
r_i: \text{rank of the } i^{\text{th}} \text{ relevant document}  \\
N_{\text{max}}: \text{max number of checked docs}
\]

- Derived from \( R_{\text{norm}} \) (Rocchio, 1964)
- Gives higher score for systems achieving higher recall and better average relative ranking
- Dependent on user’s potential/effort \((N_{\text{max}})\)
- Robust to incomplete relevance judgements

---

**PRES: as a cumulative gain**

Value added to score when finding relevant document

<table>
<thead>
<tr>
<th>( \frac{1}{n} )</th>
<th>( N_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>10</td>
</tr>
<tr>
<td>0.33</td>
<td>8</td>
</tr>
<tr>
<td>0.25</td>
<td>6</td>
</tr>
<tr>
<td>0.20</td>
<td>5</td>
</tr>
<tr>
<td>0.16</td>
<td>4</td>
</tr>
<tr>
<td>0.14</td>
<td>3</td>
</tr>
<tr>
<td>0.12</td>
<td>2</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
</tr>
</tbody>
</table>

- **MAP**
- **PRES**
- **Recall**
**Patent Search – CLIR**

- Query: Full patent application
- Common approach: MT (the best)
- Challenge: training recourses + speed!
- Ideal: Query + Document translation

---

**Patent Search – CLIR – Objective?**

- Manual translation
  - It is a great idea to apply stemming in information retrieval
- MT output
  - he are an great ideas to applied stem by information retrieving
- MT evaluation: MT sucks
- IR evaluation: MT rocks 😊

- MT4IR: An efficient MT that neglects morphological and syntactic features of output
Ordinary MT vs. MT4IR

Query (lang x)

Process

Translate

Query (lang x, no stop words, and stemmed)

Index (lang y)

Search

Results (lang y)

MT Model (lang x→y)

Train MT

Parallel Corpus

Patent Search – MT4IR

Retrieval effectiveness for a Patent CLIR En-Fr task

Google Translate

MT4IR

Ordinary MT

<table>
<thead>
<tr>
<th>Training size</th>
<th>PRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2k</td>
<td>0.30</td>
</tr>
<tr>
<td>8K</td>
<td>0.33</td>
</tr>
<tr>
<td>80K</td>
<td>0.36</td>
</tr>
<tr>
<td>800K</td>
<td>0.40</td>
</tr>
<tr>
<td>8M</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Patent Search – MT4IR

E.g. play, plays, played, playing

Patent Search – MT4IR

Translation speed for a Patent CLIR En-Fr task
Summary

- The objective is IR is “User Satisfaction”
- Understand the user needs well
- Design the IR task carefully
- You do not have to stick to the path in the literature
- Are you sure performance is measured correctly?

Readings