Search is not only the Web
IR Applications

Walid Magdy
School of Informatics
University of Edinburgh

27 Oct 2021

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 → Recognized text ← Search

- **Challenges in Common Approach:**
  OCR → Text with mistakes (WER$_{\text{Ar}} \approx 40\%$)
  OCR → Not available for all languages

---

**Problem**

- **Text with errors (sometime many errors)**
n-gram Char Representation of OCR

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

OCR Correction using Error Model

- OCR text → Generate Candidates → Best Fitting Word Selection → Corrected text
- Select part and correct
- Manually corrected version
- Character Error Model
- Language Model
- Use for search
Query Garbling using Error Model

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

Character Error Model

Use for search

OCR Correction using Edit Distance

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

Edit Distance

Language Model

Dictionary

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
**OCRless Search**

- **Segment to elements**
- **Create IDs document**
- **Cluster ID**
- **Cluster**

213 31 32 2 213 31 33 41
1190 23 802 ...

**Index of IDs**

---

**Solution – OCRless Search**

الإيمان

- **Draw query**
- **Replace with candidate IDs and formulate query**

List of ranked documents

Search Index of IDs

syn(1284, 21, 673, 1208)
syn(430, 4, 6412, 3094)
syn(231, 9011, 32, 721)
syn(40, 110, 2213, 2214)
Solution – OCRless Search

- Effective and fast
- Robust to OCR errors (v1de0)
- 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 Collection

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}

\[ AP\text{\_system1} = 0.25 \]
\[ AP\text{\_system2} = 0.0481 \]
\[ AP\text{\_system3} = 1 \]

\[ R\text{\_system1} = 0.25 \]
\[ R\text{\_system2} = 1 \]
\[ R\text{\_system3} = 1 \]

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

\[ PRES = 1 - \frac{\sum r_i - n(n+1)}{2 N_{\text{max}}} \]

- \( n \): number of relevant docs
- \( r_i \): rank of the \( i \)th relevant document
- \( N_{\text{max}} \): 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
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**
- **Index** (lang y) → **Search** → **Results** (lang y)

**Train MT**

**MT Model** (lang x → y)

**Parallel Corpus**

---

**Patent Search – MT4IR**

Retrieval effectiveness for a Patent CLIR En-Fr task

![Graph showing retrieval effectiveness for different training sizes.](image)

- **Google Translate**
- **MT4IR**
- **Ordinary MT**

---

**Institute for Language, Cognition and Computation**

**THE UNIVERSITY of EDINBURGH**
Patent Search – MT4IR

E.g. play, plays, played, playing

Patent Search – MT4IR

Translation speed for a Patent CLIR En-Fr task

20 times faster
9 times faster
5 times faster
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