Search is not only the Web
IR Applications

Walid Magdy
School of Informatics
University of Edinburgh

23 Oct 2019

Objectives

• Main objective of IR
• Different tasks in IR
  • Printed documents search
  • Patent search
  • Social 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$_{AR}$ ≈ 40%)
  - OCR → Not available for all languages

---

Problem

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

---

**IR effectiveness with different qualities of OCR**
**n-gram Char Representation of OCR**

- **Original:**
  example sentence

- **OCR output:**
  examplne sencnace

- **3-gram char representation:**
  \$ex exa xar arn rnp npl ple le$ $se sen enl lnc lcn cnc nce ce$

- **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$ $se sen enl lnc lcn cnc nce ce$
  \$ex exa xam amp mpl ple le$ $se sen ent nte ten enc nce ce$

---

**OCR Correction using Error Model**

1. **OCR text**
2. **Generate Candidates**
3. **Best Fitting Word Selection**
4. **Corrected text**

- Select part and correct
- Manually corrected version
- Train Error Model
- e → c
t → l
m → rn
d → cl

*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
Multi-OCR Text Fusion

OCR text 1

OCR text 2

Word Alignment

Best Fitting Word Selection

Language Model

Fused text

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**

- **Query**
- **Draw**
- **Information**

**Image Domain**

- **OCR**
OCRless Search

Qa'awati and sad faygina la ilaah. Qa'awati and sad faygina la ilaah.

Create IDs
document

Indices of IDs

Segment to elements

Clustering

Indexing

213 31 32 2 213 31 3341
1190 23 802 ...

Clustering

Cluster ID

Cluster

Institute for Language, Cognition and Computation
ILCC

Solution – OCRless Search

الإيمان

Draw query

Replace with candidate IDs and formulate query

List of ranked documents

Search Index
doors 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 \(\rightarrow\) Approach
Patent Search

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

Query → Search → Results list

A System and Method for ...

Several languages

Patent application

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}

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

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

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

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

- \( n \): 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

- Official score in CLEF-IP since 2010
- Adapted in many Recall-oriented IR tasks

- User Satisfaction \( \rightarrow \) Objective function
**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

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

Patent Search – MT4IR

Retrieval effectiveness for a Patent CLIR En-Fr task

Institute for Language, Cognition and Computation
ILCC

THE UNIVERSITY
of EDINBURGH
Patent Search – MT4IR

![Graph showing OOV rate vs training size for MT4IR and Ordinary MT](image)

E.g. play, plays, played, playing

Patent Search – MT4IR

Translation speed for a Patent CLIR En-Fr task

![Bar chart showing decoding time vs training size for MT4IR and Ordinary MT](image)

- 5 times faster
- 9 times faster
- 20 times faster
Social Search

Microblog (Twitter) Search

- TREC Microblog track → Ad-hoc, filtering

- User’s information need?
- Search scenario? Task?
- Boolean? News updates?
Social Media & News

- News websites are biased

- People use social media to
  - Report news
  - Comment on news
  - Discuss different views on events

- Discussions on social media, reflects public interest

- Can social media answer the question: “What is happening in <region>?”

TweetMogaz

A News Portal of Tweets
TweetMogaz Components

Tweets Streaming

Information Filtering

Presentation

Index (Solr)

Tweets Analysis

Story Detection

Classification

Adaptive Information Filtering

Online Stream

Predefined Queries

Boolean Filter

Random Sampling

Term Selection

(positive)

(positive)

Train Classifier

Filtered Stream

Classifier

• Information need → Retrieval 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?

- Beating the baseline is always desirable, just be sure you are moving in the right direction

Readings

- Magdy W., K. Darwish, and M. El-Saban. Efficient Language-Independent Retrieval of Printed Documents without OCR. *SPIRE 2009*