Rethinking Scholarly Communication with AI:

Keywords Are Not the Answer To Your Search

2019-06-21
Linda Andersson

OAI 11 – The CERN-UNIGE Workshop on Innovations in Scholarly Communication
Behind the “Search” Button

The good exterminator & Co.

Socio-organization and cultural context

Work Task Context

Seeking Context - e.g. reading, formulating,

IR Context

Documents

Request

Index

Query

Match

Result

Search

Info about removing mice without killing them

Get rid of mice in a humane and correct way

Evaluation Criteria

Recall, precision, efficiency, quality of information process

Usability, quality of information/process

Quality of info & work process/result

Socio-cognitive relevance; quality of work task result

Modification and slide created by:
Dr. Mihai Lupu | Studio Director | Research Studio Data Science

Model first presented in:
Evolution of Indexing Technologies

The Golden Age of Artificial Intelligence

- Pure counting
- Term frequency
- Position in sentence
- SMART
- IDF
- Cosine similarity
- and many more...

Expert Systems, Knowledge bases (e.g. Cyc)

- Latent Semantic Analysis
- WWW appears
- Semantic Web appears
- Probabilistic models for IR
- Language Models

Inference on billions of tuples
on trillions

1950
1960
1970
1980
1990
2000
2010
2020

Deep Learning
Speech Vision
NLP IR

WWW appears

Slides courtesy to Dr. Mihai Lupu | Studio Director | Research Studio Data Science
Evolution of the Searching Functionalities

Information needs expressed by metadata keywords (KWIC)
“Information Needs” Potpourri of Various Needs

Therefore Identify & Categorize

Known item search

Explorative search
JoinSimilarity = \sum_{i,j=1,n}^{N} \frac{\cos \left( \vec{w}_i, \vec{w}_j \right)}{N}

---

Use Case: Patent Text Mining

---

Use Case: Patent Text Mining
• Meta data
  – Bibliographic data
  – Classification Schema

• Text Discourse
  – Section Title, description and Claim
  – Multilingual
  – Technical terms
  – heterogeneous distribution
  – Patentese

• Information Need
  – Prior art since mid 1900-century
  – Information extraction
    • Paragraphs, and semantic relations

The Information Need is Complex

Searching for:
- Problem and Solution.
- Scope of the invention
- Specific technical details

Engine function

Steering mechanism

No 37435 Benz Patent – Moterwagen 1886
(Patent) Paragraph Search

Document Index
~3M Doc

Paragraph Index
~61M Paragraphs

Domain NLP Pipeline

Automatic Query Generation (QG)

Learning Termhoodness

Merged Results

50 Topics, each topic QG 220
CLEF-IP 2013 Collection
Learning Termhoodness & Query Expansion

Ontology population

- Domain Rules
- NLP pipeline (general NLP)
- Corpus statistic
- Taxonomy statistics
- Filters to identify Relation and Technical terms
- Hyponymy

Terminology identification

- IPC-distributional-values
- Using Machine Learning models

FACULTY OF INFORMATICS
**Distributional Semantics** identifies similarities between different words

- “Underwear” similar to “underpants”, “undergarment”, “panties”, “underclothes”
- “Strength” similar to “strengths”, “strength”, “toughness”, “stronger”, “strength”

(Rekabsaz et al 2016 and Rekabsaz et al 2017)

But technical semantic relations are a mixture of single words and phrases

\[
\text{JoinedSimilarity} = \sum_{i,j=1,n}^{N} \frac{\cos \left( \vec{w}_i, \vec{w}_j \right)}{N}
\]

- \( w_i, w_j \) represent each word vector pair cosine similarity of a MWT
- \( N \) is the number of words for a MWT

- **synthetic fibers** synonym to **polyester fibers**
- **thrips** hypernym to **bulb fly larvae**
Example of Automatic Query Formulation

Automatic query expansion terms from ontologies

- brake pedal
  - vehicle operating pedal,
  - conventional hydraulic brake system
  - pedal devices
  - position brake actuating member
  - brake actuating member
  - hydraulically-assisted rack & pinion steering gear
  - brake operating member
  - conventional braking system
  - pedal pedals
- accelerator pedal
- case pedal device
- pedal device

Automatic Query Formulation

```
<QUERY>
(conjured OR clutch OR connectability OR nmofs OR fclp OR dnr OR slippage OR anda OR rotational OR acceleration OR backlash OR subordinate OR estimating OR ure OR brake OR torque OR stopped OR vehicle OR wheel OR command OR outputting OR estimate OR shock OR nsr OR driving OR pedal OR wheels OR shaft OR prohibiting OR determining OR sensor OR tws OR drive OR occurrence OR estimated OR prescribed OR stopping OR elapsed OR motor OR speed OR gdv OR instruction OR input OR output OR controller OR rotating OR accelerator OR electric OR force OR flag)
AND
("vehicle driving force control apparatus" OR "drive wheel" OR "rotational speed" OR "4wd controller" OR "clutch connection command" OR "rear wheel" OR "four-wheel drive state" OR "torque transfer path" OR "output rotational speed" OR "input rotational speed" OR "clutch control section" OR "detected parameter" OR "generation load torque" OR "torque fluctuation" OR "brake operation" OR "determination occurrence" OR "four-wheel drive vehicle" OR "input shaft" OR "4wd controller proceed" OR "wheel speed sensor" OR "output shaft" OR "response delay" OR "clutch input shaft" OR "backlash elimination" OR "drive mode switch" OR "brake pedal" OR "accelerator pedal" OR "targeted range" OR "transition time" OR "wheel speed" OR "rotational speed difference" OR "clutch connection" OR "motor torque" OR "generator load torque" OR "vehicle driving force control" OR "high rate" OR "electric motor" OR "throttle opening" OR "external disturbance" OR "vehicle driving force" OR "connected state" OR "previous equation" OR "prescribed range" OR "electric power" OR "prescribed rotational speed difference" OR "12-volt battery" OR "connection command" OR "disconnected state" OR "electric clutch" OR "four-wheel drive")
</QUERY>
```
Domain Knowledge Makes AI *Smart*

**Domain NLP Pipeline**
- Query Generation
- Domain Knowledge
- Normalization Component
- Query Expansion
- IPC-distributional-values

**Learning Termhoodness**

**Query Expansion**

**Noise Filtering**

**Query**

**Document Index**

**Passage Index**

**Post Rank MERGED**

*Distributional (DS) models word2vec*

(*Rekabsaz et al 2016 and Rekabsaz et al 2017*)

**Jointed Similarity**

\[
\text{Jointed Similarity} = \sum_{i,j=1, i \neq j}^{N} \frac{\cos \left( \frac{\vec{w_i}}{N}, \frac{\vec{w_j}}{N} \right)}{N}
\]
Use Case: Medical Information Need

Multi-lingual retrieval

Infrastructure for medical information retrieval

Asking clinical questions using PICO entities
**Clinical Questions**: I would like to know if Vitamin E together with aspirin is an efficient preventive treatment regimen in patients with transient ischemic attacks

**Text Example**: Vitamin E plus aspirin compared with aspirin alone in patients with transient ischemic attacks.

<table>
<thead>
<tr>
<th><strong>Population</strong></th>
<th><strong>Intervention</strong></th>
<th><strong>Comparison</strong></th>
<th><strong>Outcome</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>patients with transient ischemic attacks</td>
<td>Vitamin E plus aspirin</td>
<td>aspirin</td>
<td>vitamin E and aspirin significantly enhances the efficacy of the preventive treatment regimen</td>
</tr>
</tbody>
</table>
Complex text discourse
- Linguistic characteristics
- Very complex noun phrases with post modifier clauses

7 Months to Beta Prototype
- Developing team: 1 computational linguist, 1 software engineer
- Including 4 months for annotation
  - 4 domain experts (40h each)
  - 2 Linguistic students (100h each)

Design Challenges

- The Linguists identified the complex noun phrases, while the domain expert were better in identifying the population without person entity.
A successful text mining solution does not only focus on developing the technology or the best AI algorithm, it is much more complexed.

Keywords Are Not the Answer to Your Search

- User acceptance
- Identification
- Categorisation

Complexity

Language

Task

Domain
Rethinking Scholarly Communication with AI

Artificial Researcher in Science

Artificial Researcher in Open Access

Reading text
keeping accuracy high. For details on the implementation and parameter settings refer to Hoy et al. (2013) and Johnson (2007).

Output example
The model parameters are $\theta_i$, $\beta_j$, $\gamma_i$, and $\epsilon_{ij}$, specifying the probability of transferences for annotation $i$, the probability that the annotation $i$ is not sponsored, and the error rate of the transference. We used the following parameters for the model: $\theta_i = 0.5$, $\beta_j = 0.1$, $\gamma_i = 0.3$, and $\epsilon_{ij} = 0.2$. The learned value of $\theta_i$ will prove useful later when we try to identify suitable annotations (see Section 6).

Requests
Find relevant paragraphs
Retrieve summary
Extract specific citations

Output example

FACULTY OF INFORMATICS
Contact Information

PhD Linda Andersson
andersson@ifs.tuwien.ac.at
+43 (0) 699 17 821 600
http://www.ifs.tuwien.ac.at/~andersson/

Artificial Researcher

Your Scientific Knowledge Provider

linkedin.com/company/artificialresearcher
www.artificialresearcher.com

Artificial Researcher IT GmbH | Vienna - Austria
Linda Andersson - Résumé

Academic merits

- Information design 1998-2001
  - General Linguistics 2001-2003
  - Computational Linguistics 2006-2009
- Language Engineering 2004
- Library & Information Science 2004-2006
- Computer Science 2009-2019

Research fields

- Natural Language Understanding
- Natural Language Processing
- Information Extraction
- Information Retrieval

Awards

- 2018 Commercial Viability award by the Austrian Angel Investors Association.
- 2017 PhD high potential R&D award, i2c Award, Vienna University of Technology
- 1999 Finalist in a Venture Cup held at Mälardalens University College.
- Recognized Reviewer Award at 41st European Conference on Information Retrieval, ECIR2019 Cologne
Experiments in the Presentation


- Linda Andersson, Navid Rekabsaz, Allan Hanbury (2017) Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information The first WiNLP Workshop co-located with with the Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver

Other papers on patent text mining


Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik, Jon Brassey (2018) Medical Entity Corpus with PICO Elements and Sentiment Analysis Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan

Markus Zlabinger, Linda Andersson, Jon Brassey, Allan Hanbury (2018) Extracting the Population, Intervention, Comparison and Sentiment from Randomized Controlled TrialsBuilding Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth - Proceedings of {MIE}, Medical Informatics Europe, Göteborg, Sweden, April

