Search Matters 2017

Towards semantic search at the European Patent Office
About me

Alexander Klenner-Bajaja

- Bioinformatics at Goethe University Frankfurt
- PhD ETH Zurich & Goethe University – in Cheminformatics
- PostDoc at Fraunhofer Society SCAI in Cologne
  - Chemical entity recognition in patents
- Data Scientist, Search & Knowledge, DG2, EPO
  - automated search
  - search benchmarking
  - new search technologies
Anecdotal Evidence and the need for a Benchmarking Environment

- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence
Anecdotal Evidence and Benchmarking

- How can two different search strategies be compared?

- Historically decisions are taken based on expert feedback

- Valuable information that can be complemented with measurements

Distribution shown as Box-plot

Median: 5 citations in search reports
Search Benchmarking

- We implemented a holistic prototyping and benchmark system
- We have access to all distributed system in the EPO IT landscape and can connect them in "visual" workflows
Automated Queries exploiting the shown tools

Application → Automated Query Formulation → Search → PriorArt → relevant documents

Get Gold Standard → Query Generation → AnseraQueryMarkerQuery → Evaluate Search Results

Evaluation and Feedback
Benchmarking Environment

- From anecdotal to statistical evidence
- "Towards Automated Prior Art Search" (TAPAS)

Data Set Generation

Applications

Method 1

Search Corpus

Method 2

Retrospective Search

Patent Corpus

Citations

Split Date

Ø Recall@100: 0.25

trec_eval

Ø Recall@100: 0.17
Benchmarking Environment

- Retrieval success measured by different metrics
- Computed by the trec_eval tool

Top 50 Documents Returned

Cited Documents not Returned

- Average over all \( n \) simulated applications

Recall @ 50
\[
\frac{5}{10} = 0.5
\]

Hit Rate @ 50
\[\text{recall} > 0? = 1\]
Benchmarking Environment

- Search reports for about 40 million simple patent families
- The relevant documents are mentioned in the search report as either
  - X or A citations
- Only few citations per document (2 X-category, 3 A-category)
Benchmarking Environment

- Search reports for about 40 million simple patent families
- The relevant documents are mentioned in the search report as either
  - X or A citations
- Only few citations per document (2 X-category, 3 A-category)
- But more "soft" information available: Returned during search (R), put aside for detailed inspection (Ins), printed (Pr), stored for citation (Cs)
- Anecdotal Evidence and the need for a Benchmarking Environment

- **Keywords and query generation in (automatic) search scenarios (Project ERa)**

- Introducing machine learned semantic search technologies and (automatic) query expansion

- Introducing terminology based semantics within the APL project

- **(Automated) Search result confidence**
Term-Based Search

- Can we hope to find all citations with keywords?

  ➞ overlap in vocabulary (non-stopwords)

Application    Citation

![Venn diagram showing overlap between application and citation terms.](image)
Term-Based Search

- However: also overlap with random documents
- Random: Y-Scrambling
Term-Based Search fully automated

- How good can we get?
- More Like This on Title / Abstract / Description / Claims
- Percentage of citations found in the top $k$
Term-Based Search fully automated

- Can we close the gap?

... using semantic search?
- Anecdotal Evidence and the need for a Benchmarking Environment

- Keywords and query generation in (automatic) search scenarios (Project ERa)

- Introducing machine learned semantic search technologies and (automatic) query expansion

- Introducing terminology based semantics within the APL project

- (Automated) Search result confidence
Introducing "Semantics"

Virus
Automated Query Expansion

- Analyze word embeddings

  - Words with similar meaning have similar context

- Train neural network $\Rightarrow$ words represented by vectors

  apple = $[0.253, 1.793, \ldots, 5.555, 3.142]$

- Words with similar meaning have similar vectors

  - allows distance calculations between words
  - semantic similarity
Automated Query Expansion

Patent Corpus

CPC

- A01B
- A02B
- A02C
- H05K
- All

w2v model

- M_1
- M_2
- M_3
- M_{N-1}
- M_N

Query Word

top X relevant words per class
A virus is a small infectious agent that replicates only inside the living cells of other organisms.

A computer virus is a type of malicious software program ("malware") that, when executed, replicates by reproducing itself (copying its own source code) or infecting other computer programs by modifying them.

hiv, viral, human immunodeficiency, viral genome, genome, replication, pathogen, adenovirus, replicating

infected, virus infection, malicious, anti-virus, malware, virus-scanning, virus worm, macro virus, virus scanner
Semantics through Machine Learning

Example Queries: descriptions from Wikipedia

A **virus** is a small infectious agent that replicates only inside the living cells of other organisms.

A **computer virus** is a type of malicious software program ("malware") that, when executed, replicates by reproducing itself (copying its own source code) or infecting other computer programs by modifying them.

![CPC Breakdown](image1)

- Preparations for medical, dental or toilet purposes

![CPC Breakdown](image2)

- Electrical digital data processing
Automated Query Expansion

Inter model similarity in CPC A Human Necessities based on vocabularies

A47L  A45D  A47K
Automated Query Expansion

- Word2Vec Models applied to (100) extracted TF-IDF Keywords

"Touch sensing system and display apparatus"

EP2869168A120150506

<table>
<thead>
<tr>
<th>TF-IDF Keywords (10 of 100)</th>
<th>TF-IDF Scores</th>
<th>Word2Vec Enrichments (G06F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>possessory</td>
<td>49.01</td>
<td>possessory</td>
</tr>
<tr>
<td>sensing</td>
<td>33.04</td>
<td>sensing OR sensed OR sensor OR touch</td>
</tr>
<tr>
<td>touch</td>
<td>31.36</td>
<td>touch OR touched OR touch screen OR touches OR touch panel OR touching</td>
</tr>
<tr>
<td>bridge</td>
<td>27.88</td>
<td>bridge OR bridges OR pci bus OR bus OR interfaces 22a-22b OR pci-pci bridge</td>
</tr>
<tr>
<td>nodes</td>
<td>27.59</td>
<td>nodes OR node OR nodes n2</td>
</tr>
<tr>
<td>shared</td>
<td>26.86</td>
<td>shared OR share OR sharing OR non-shared</td>
</tr>
<tr>
<td>electrodes</td>
<td>25.52</td>
<td>electrodes OR electrode OR electrodes y2</td>
</tr>
<tr>
<td>controller</td>
<td>23.18</td>
<td>controller OR control OR controllers OR bus OR unit OR processor</td>
</tr>
<tr>
<td>instructing</td>
<td>22.11</td>
<td>instructing OR instruct OR instructs</td>
</tr>
<tr>
<td>masters</td>
<td>21.23</td>
<td>masters</td>
</tr>
<tr>
<td>memory</td>
<td>20.39</td>
<td>memory OR ram OR non-volatile OR memories OR processor OR flash</td>
</tr>
<tr>
<td>sensor</td>
<td>20.38</td>
<td>sensor OR sensors OR sensing OR humidity sensor OR sensed OR sensor senses</td>
</tr>
<tr>
<td>algorithm</td>
<td>20.02</td>
<td>algorithm OR algorithms</td>
</tr>
<tr>
<td>senses</td>
<td>20.02</td>
<td>senses OR sensed</td>
</tr>
<tr>
<td>data</td>
<td>19.91</td>
<td>data OR stored OR stores OR store OR storing</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Automated Query Expansion

- Query Enrichment brings increased performance

"Touch sensing system and display apparatus"

EP2869168A120150506

Keyword Extraction → TF-IDF Query
Recall@100: 0 (0 of 2)

W2V Enrichment → W2V Query
Recall@100: 1 (2 of 2)
Automated Query Expansion

- Documents found with increased Term Overlap with Query

<table>
<thead>
<tr>
<th>TF-IDF Query</th>
<th>W2V Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall@100: 0 (0 of 2)</td>
<td>Recall@100: 1 (2 of 2)</td>
</tr>
<tr>
<td>39100958</td>
<td>39</td>
</tr>
<tr>
<td>48143418</td>
<td>43</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>
Automated Query Expansion

- **Word2Vec Expansion Terms in Overlap**

<table>
<thead>
<tr>
<th>New Overlap</th>
<th>Original Keyword</th>
<th>Word2Vec Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>outputs</td>
<td>signal</td>
<td>OR signals OR circuit OR output OR clock OR outputs</td>
</tr>
<tr>
<td>apparatus</td>
<td>controlling</td>
<td>OR control OR controlled OR apparatus</td>
</tr>
<tr>
<td>screen</td>
<td>display</td>
<td>OR screen OR displays OR displaying OR displayed OR lcd</td>
</tr>
<tr>
<td>increasing</td>
<td>increases</td>
<td>OR increase OR decreases OR increased OR increasing OR decrease</td>
</tr>
<tr>
<td>serially</td>
<td>parallel</td>
<td>OR increasing OR decrease OR increase OR decreases OR decreased OR serially</td>
</tr>
</tbody>
</table>

- **Re-Suggestion of known Query Terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Original Keyword</th>
<th>Word2Vec Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>transmission</td>
<td>transmitted</td>
<td>OR transmits OR received OR transmit OR transmission OR transmitting</td>
</tr>
<tr>
<td>reception</td>
<td>transmitted</td>
<td>OR transmission OR transmitted OR transmits</td>
</tr>
<tr>
<td>transmitting</td>
<td>transmission</td>
<td>OR receiving OR transmitted OR transmission OR sending OR transmit</td>
</tr>
</tbody>
</table>

- Should this result in higher weight?
- Models heavily suggest inflections
Automated Query Expansion

- Query Expansion (Word2Vec, 647 models at CPC subclass level)
  - Top 70 TFIDF keywords, CLEF-IP

### Word2Vec

<table>
<thead>
<tr>
<th>Term</th>
<th>Overlap</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>polysaccharides</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>polymer</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>low molecular</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>hydrophilic</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>hyaluronic acid</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

### Top 30 Documents

<table>
<thead>
<tr>
<th>Number of Expansion Terms</th>
<th>Overlap</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Term</td>
<td>84 %</td>
<td>5 %</td>
</tr>
<tr>
<td>2 Terms</td>
<td>69 %</td>
<td>11 %</td>
</tr>
<tr>
<td>3 Terms</td>
<td>66 %</td>
<td>12 %</td>
</tr>
<tr>
<td>4 Terms</td>
<td>96 %</td>
<td>2 %</td>
</tr>
<tr>
<td>5 Terms</td>
<td>78 %</td>
<td>6 %</td>
</tr>
</tbody>
</table>
- Anecdotal Evidence and the need for a Benchmarking Environment

- Keywords and query generation in (automatic) search scenarios (Project ERa)

- Introducing machine learned semantic search technologies and (automatic) query expansion

- Introducing terminology based semantics within the APL project

- (Automated) Search result confidence
Introducing semantics through annotations

We are in the process of annotating the full prior art collection with normalised
- Chemical Entities
- Physical Units
- Citations
- Controlled Terminologies (e.g. MeSH)
to enable a semantic search of those entities.
Introducing annotations and semantics

- Our annotation platform's key elements
  - Based on the open source framework UIMA
  - Plug and play of new components (analysis engines)
  - Quality Evaluation Workbench
  - Knowledge Base for controlled vocabulary
  - Scaling architecture
  - Data and annotations stored in noSQL databases

Application Citation

- Aspirin (D001241)
- acetylsalicylic acid (D001241)

Matching terms through controlled vocabulary

MeSH Unique ID: D001241
Introducing annotations and semantics

Project in Execution - Platform installed and tested at this moment

1. Define (sub)set of documents to work on

```
Sources: Connector Management
Functional Testing 1

Connectors

<table>
<thead>
<tr>
<th>Connector</th>
<th>Type</th>
<th>Active</th>
<th>Schedules</th>
<th>Statistics</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>kime_o_all</td>
<td>kimeConnectorType</td>
<td>0:24</td>
<td>110057 / 110057 / 110057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>some_families</td>
<td>kimeConnectorType</td>
<td>0:24</td>
<td>40 / 40 / 40 / 40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

2. Define an annotation Pipeline

```
<table>
<thead>
<tr>
<th>Pipeline Name</th>
<th>State</th>
<th>Instances</th>
<th>Pre configured</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentences_nodist</td>
<td>STARTED</td>
<td>0</td>
<td></td>
<td>44 / 897.96</td>
</tr>
<tr>
<td>sentences_pos_nodist</td>
<td>STARTED</td>
<td>0</td>
<td></td>
<td>44 / 1,157.89</td>
</tr>
</tbody>
</table>
```
Word2Vec & APL

Aspirin = acetylsalicylic acid

Matching terms through controlled vocabulary

Virus = Virus

Separating identical terms with different meaning

Aspirin = acetylsalicylic acid (would have been False Negative)

Virus = Virus (would have been False Positive)
▪ Anecdotal Evidence and the need for a Benchmarking Environment

▪ Keywords and query generation in (automatic) search scenarios (Project ERa)

▪ Introducing machine learned semantic search technologies and (automatic) query expansion

▪ Introducing terminology based semantics within the APL project

▪ (Automated) Search result confidence
Identify Successful (Automated) Searches

Machine-Learning: look at the documents

Word2Vec Heatmap

Sequence Alignment (Bioinformatics)

Citation

Random
Can we identify successful searches? (Machine learned)

- Prediction Model for **Citability** based on Distribution Statistics of heat maps
  
  [mean, variance, skewness, kurtosis, min, q1, q2, q3, max]

- 10-fold Cross Validation (Y-Scrambling)

![Diagram showing error rates for different rounds with Training Set and Validation Set](image)

**Multi-Layer Perceptron**

- Ensemble of decision trees
- Bootstrap aggregating
- Random selection of features

**Random Forest**

- ~8% Predicted Relevance
- ~5% Error Rate
Identify Successful (Automated) Searches

- **Meta Data:** same author as in application occurs in result set
- **Interesting Item Set Mining & Subgroup Discovery**

![MLT Search Graph](image)

- Searches with same author in result set (35%)
  - Recall@100: 0.2357
- All searches (100%)
  - Recall@100: 0.1608
- Searches without same author in result set (65%)
  - Recall@100: 0.1209
- Only documents from same author
  - Recall@100: 0.0884
Complementariness of Results

Query: "Touch sensing system and display apparatus"

Method 1

Method 2
Complementariness of Results

Query

"Touch sensing system and display apparatus"

Method 1

Method 2

Method 3

... Method i
Complementariness of Results

Query
"Touch sensing system and display apparatus"

Method 1

Method 2

Method 3

\ldots

Method i
Conclusion and outline

- We are exploiting state-of-the-art semantic- and other search technologies

- We have created a prototyping and benchmarking environment to evaluate the performance and quality of a new search algorithm learning from prior searches

- Bringing all these new technologies together in a productive search environment is the biggest challenge ahead

- Semantic (text) search is one small mosaic piece in the complex search environment at the EPO
Acknowledgments

- The Search & Knowledge Directorate
  - Domenico Golzio, Director
  - Stefan Klocke, Head of Department
  - Volker Hähnke, Ranking and Confidence
  - Matthias Wirth, Neo4J, CPC classification
  - Pavel Goncharik, Ansera, ranking, lucene

- Examiners
  - Robert Herman, ERa
  - Ype Kingma, ERa
  - Markus Arndt, ERa
  - Christos Stergio, APL