The Essence of Patent Text Mining

Linda Andersson
SEMANTiCS Vienna 2018
Thursday 13th of September
My 20 years journey of exploring Information Literacy

Library and Information Science 2004-2006

Computational Linguistics 2006-2009

UPPSALA UNIVERSITET

Language Engineering 2004

Computer Science 2009-2018

Information design 1998-2001

General Linguistics 2001-2003

TU WIEN

MÅLABALENES HÖGSKOLA ESKILSTUNA VÄSTERÅS

Artificial Researcher In Science

SEMANTiCS Vienna 2018

FACULTY OF INFORMATICS
What is the Essence of Patent Text Mining

- Patent Retrieval
  - For document
  - For paragraph (Passage)

- Text summarization

- Automatic term extraction

- Ontology population
  - Hypernym & Hyponym
  - Acronyms

- Domain Claim Graphs (DCG)
  - domain adaptation of Natural language processing (NLP) tools

- Factoid search
At the core of patent text mining

- **Task Complexity**
  - The information need differs during the patent life cycle
  - Finding relevant documents, paragraphs and relations
    - Solution: observe how patent searcher construct queries and to what type of information need prior art, invalidity search

- **Language Complexity**
  - Word formation of new words are particular important for the patent text genre.
    - Solution: incorporate linguistic information

- **Domain Complexity**
  - Mixture of technical terms and legal terms, lengthy documents
    - Solution: incorporate domains specific linguistic information
The Patent Information Need

Not only identify relevant document but relevant paragraph that invalidate a new patent claim application.

When conducting prior-art search it is essential to find different aspects of a patent?
  - Each aspect can be divided into term pairs consisting of a general term and a specific term.

Consequently, if we have three aspects A, B and C each of these three aspects’ pairs need to be combined in the search process.

The search strategy in patent search consist of many complex queries targeting the main topic, as well as sub topics of patents.

(Adams, PatOlympic 2011)
No 37435 Benz Patent – Moterwagen (1886)

You search for the entire invention but also on specific details.

Engine function

Steering mechanism
The linguistic complexity increases

- Verbose query (very long)

- Ambiguity among word and phrases
  - Bus (bus slot card), (double-vehicle bus)

- Creative paraphrases
  - Patent: electronic still camera, electronic imaging apparatus
  - Mainstream: digital camera

- The multi-word unit (MWU) constitute a majority of all technical terms in technical dictionaries
  - The technical multi-word unit consist mainly of *noun phrases* composed of common word adjectives, nouns and occasionally prepositions (e.g. ‘of’)
But general NLP tools have limitations [1]

- Source (news text) versus Target data (patent text)

Verb participles were discovered to be erroneous in patent text

Task 1: Question and Answering

*What substance have a melting point of about 61° C?*

A *Tilidine Mesylate*, according to any one of claims 6 to 9, having a melting point of about 61°C as determined by DSC.

Task 2: Automatic Terminology Extraction

Every local **bus slot card** willing to master the bus will have to mimic 030, so it appears the 040-to-030 cycles translation adapter will always be in between the CPU and the local bus, no matter be it 040 or 060.

**Infrared radiation drying** is a method used to process food.
Ground truth:
**Subject:** Tilidine Mesylate **Predicate:** having **Object:** melting point of about 61°C

Google
**Subject:** Tilidine **Predicate:** *Mesylate**  **Object:** according ..... 

Stanford (corenlp.run)
**Subject:** Tilidine Mesylate **Predicate:** having
**Object:** a melting point of about 61°C as determined by DSC.

Open IE:
Ground truth: Infrared radiation drying

**Stanford**
Infrared radiation

**Google**
Infrared radiation drying
When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval?

Andersson et al. CIKM 2016

[2]
Query Generation: 220 versions for each topic

(freezing OR start OR liquid OR dough OR glucose OR
bake-off OR coating OR foodstuff OR pre-glaze OR syrum)
AND
("complex sugar"~5 OR "glucose syrup"~5 OR "dough
product"~5 OR "dough mixture"~5 OR "form liquid"~5
OR "pre-glaze composition"~5 OR "coating step"~5 OR
"coating part outer surface dough mixture"~9)
Finding Termhoodness among phrases

- State-of-the-art: C-Value

- The C-value reflects a phrase technical significance:
  - To what degree a noun phrase should be consider a technical concept.

- Computation consists of two parts,
  - Linguistic filter -> Natural language Processing (NLP)
  - Statistical-based evidence for terminological unit by computing nested NPs
## What is a technical term and what is not?

Too many false positive

<table>
<thead>
<tr>
<th>Candidate Term</th>
<th>Word2Vec</th>
<th>C-Value</th>
<th>Pointwise Mutual Information</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote communication</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Communication link</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Resin particle</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>washed washing</td>
<td>No/Yes (0.642)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Bar code</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Wet strength</td>
<td>Not</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Different features to Learning Termhoodness
(Random sample of 637 noun phrases 222 negative, 451 positive)

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature combination</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NLP</strong></td>
<td></td>
</tr>
<tr>
<td>syntax</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>Syntax frequency</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>n-gram</strong></td>
<td></td>
</tr>
<tr>
<td>Phrase length</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>NLP &amp; Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>C-value</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>Co-occurrence</strong></td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>Probability</strong></td>
<td></td>
</tr>
<tr>
<td>PMI</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>C-value and IPC</strong></td>
<td></td>
</tr>
<tr>
<td>IPC:CValue</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td><strong>IPC-distributional-values</strong></td>
<td></td>
</tr>
<tr>
<td>IPC:sum</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>IPC:count</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>IPC:mean</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>IPC:median</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>IPC:variance</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>IPC:stddev</td>
<td>x x x x x x x x x</td>
</tr>
<tr>
<td>Correctly Classified</td>
<td>77 77 77 78 76 71 77 78 77 71 77 77 76 77 75 75 69 70 76 67 71 73 68 68 65 68 66 71</td>
</tr>
</tbody>
</table>
Our assumption

*Phrases having a homogenous distribution of International Patent Classification categories will reflect the termhoodness compared to phrases with heterogeneous distribution*

Performance: F1 score of 0.845 in accuracy when detection Technical terms among 4400 candidates
<table>
<thead>
<tr>
<th>Run</th>
<th>Query lengths</th>
<th>IR model</th>
<th>PRES</th>
<th>Recall</th>
<th>MAP</th>
<th>MAP(P)</th>
<th>Prec(P)</th>
<th>Post Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 3 best methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLP, Expanded, Word, Technical terms (IPC), skip-gram (NLP1)</td>
<td>100</td>
<td>LMJM</td>
<td>0.544</td>
<td>0.631</td>
<td>0.285</td>
<td>0.112</td>
<td>0.218</td>
<td>Merged</td>
</tr>
<tr>
<td>Statistical, Expanded, unigram, bigram</td>
<td>100</td>
<td>LMJM</td>
<td>0.492</td>
<td>0.574</td>
<td>0.300</td>
<td>0.114</td>
<td>0.208</td>
<td>Merged</td>
</tr>
<tr>
<td>Statistical, only claim, unigram</td>
<td>100</td>
<td>LMJM</td>
<td>0.444</td>
<td>0.560</td>
<td>0.187</td>
<td>0.146</td>
<td>0.282</td>
<td>Merged</td>
</tr>
<tr>
<td>Baseline - unigram</td>
<td>100</td>
<td>LMJM</td>
<td>0.536</td>
<td>0.622</td>
<td>0.226</td>
<td>0.132</td>
<td>0.229</td>
<td>Merged</td>
</tr>
<tr>
<td>Best official runs clef-IP 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document, word, hyphened MWUs, Upper bound IDF</td>
<td>N/A</td>
<td>BM25</td>
<td>0.433</td>
<td>0.540</td>
<td>0.191</td>
<td>0.132</td>
<td>0.213</td>
<td>N/A</td>
</tr>
<tr>
<td>Document, word, hyphened MWUs, No upper bound IDF</td>
<td>N/A</td>
<td>BM25</td>
<td>0.432</td>
<td>0.540</td>
<td>0.190</td>
<td>0.132</td>
<td>0.214</td>
<td>N/A</td>
</tr>
</tbody>
</table>
What can learn in terms of Recall?

- Topics have different preferences in terms of QF methods which is related to their technical fields.

- With a larger topic set we can apply ML methods to learn which QF to use for the different technology domains in order increase Recall.
But manual patent search query is more complexed

- The vocabulary in a patent document is a mixture of
  - hypernym (broad) and hyponym (specific) terms
  - the hyponymy relation is a mixture of single words and phrases
    - **thrips** is a hypernym to **bulb fly larvae**
    - Different technical fields
The process can be applied to any text collections with bibliographic data.
**Query Expansion terms suggestion**

**JoinedSimilarity**

- **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
  - pair pedals

**Word2vec – threshold 0.7** [3]

- **brake:**
  - brake, brakes, braking, braked, pedal, antilock, clutch

- **pedal**
  - pedal, pedals, brake, braking

- **bus**
  - buses, busses, memory

---

<table>
<thead>
<tr>
<th>PRES</th>
<th>Recall</th>
<th>MAP</th>
<th>MAP (P)</th>
<th>Precision (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,563</td>
<td>0,653</td>
<td>0,271</td>
<td>0,106</td>
<td>0,207</td>
</tr>
</tbody>
</table>


*Only refers to the semiconductor concept.*
- Does “network lan” and “communication link” have (hyponymy) relation? Yes
- Does ”mechanical stress” and “communication link” have a (hyponymy) relation? No

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

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

But patent searching also include non-patent literature

Now we transferring domain know-how into the general scientific publications text mining
We are in the zone of domain specific text mining. 

Our previous expertise: Patent documents

Our patent text mining research is now state-of-the-art

Given: A patent claim
Required: Relevant passages in patent citations.

Solution considered the language, domain, and retrieval task complexity. All are necessary for a successful domain-specific text mining application.

Steps to attain efficient retrieval in the patent domain are:
1) Extracting lexico-semantic relations from patent data
2) Automatic query formulation for a user selected information need
3) Passage retrieval

Thank you for Your Attention
andersson@ifs.tuwien.ac.at