FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

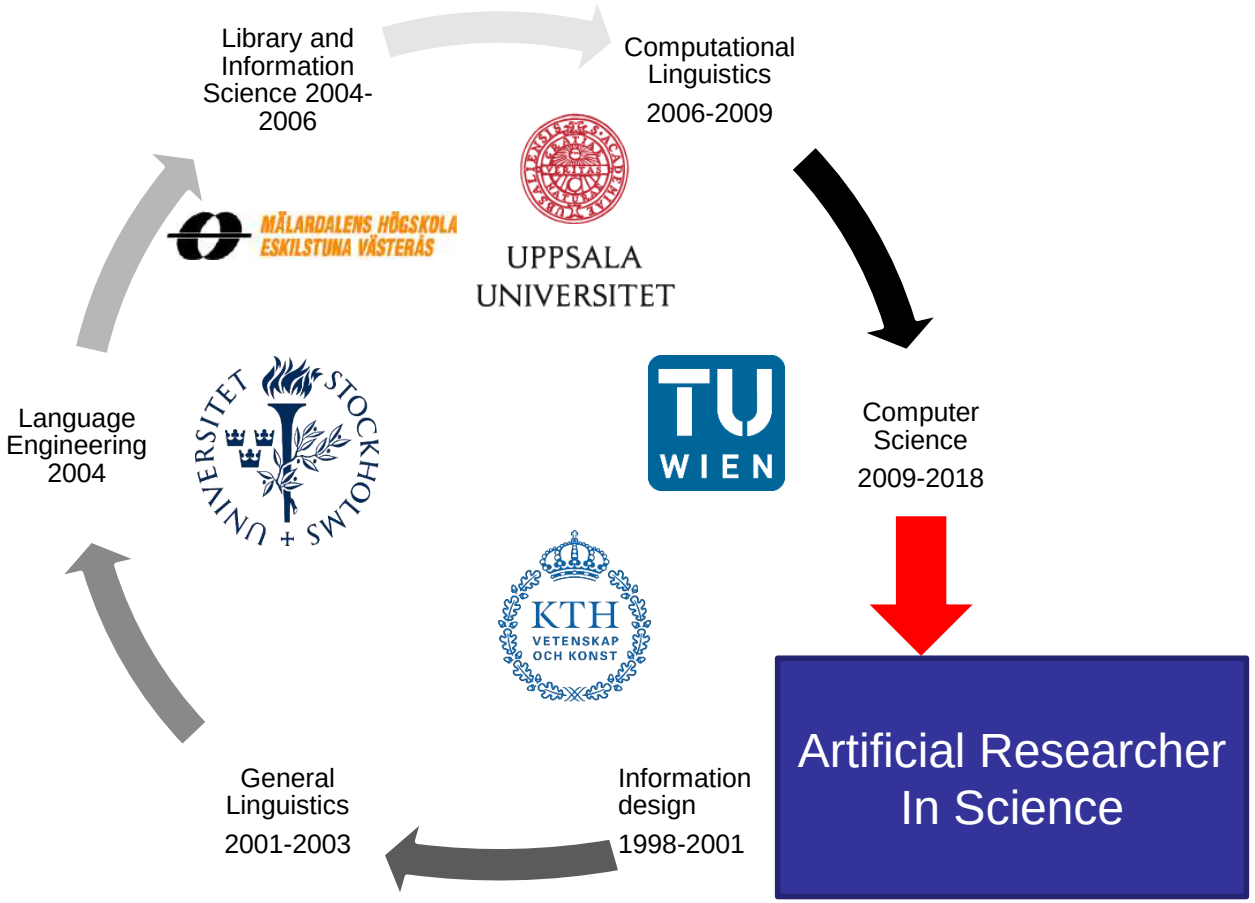
The Essence of Patent Text Mining

Linda Andersson

SEMANTiCS Vienna 2018

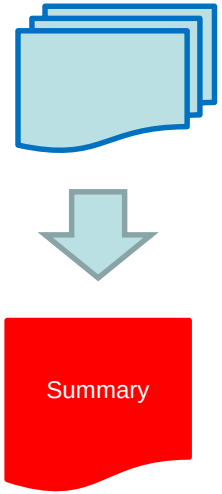
Thursday 13th of September

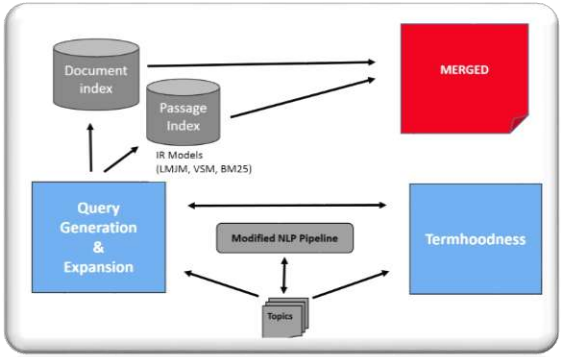
My 20 years journey of exploring Information Literacy

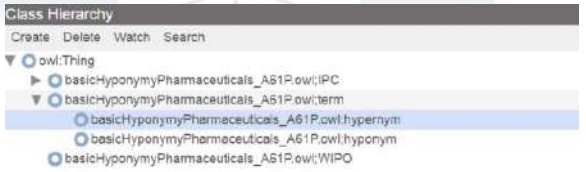


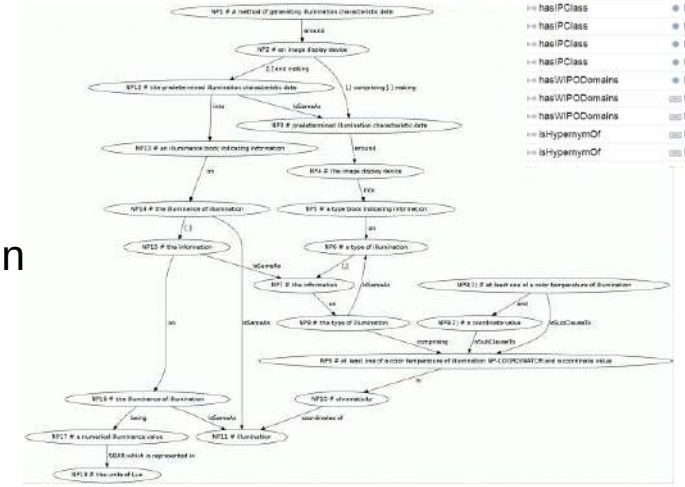
What is the Essence of Patent Text Mining

- 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 processin (NLP) tools
- Factoid search









- **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*

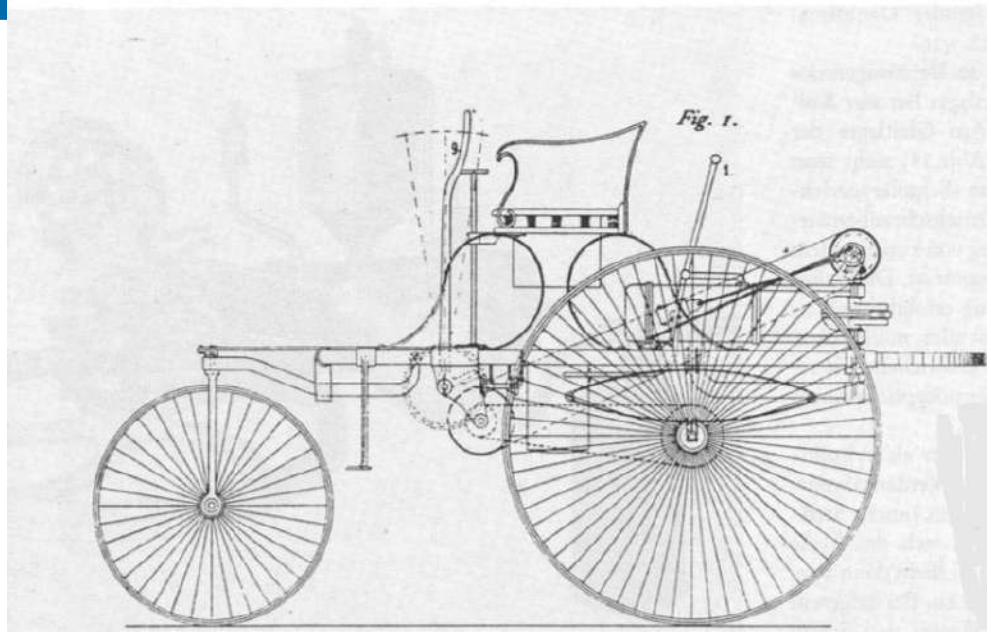
- *When conducting prior-art search it is essential to find different aspects of a patent?*
  - *Each aspect can be divided into term pairs consisting of a general term and a specific term.*

- *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.*

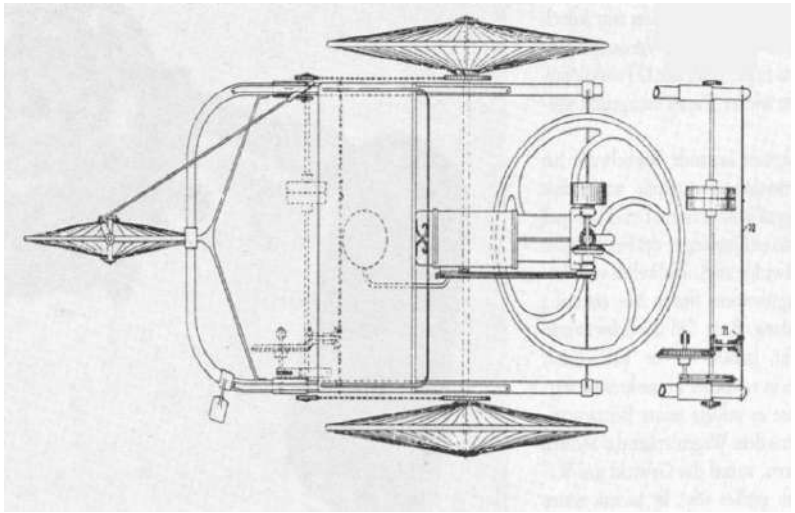
(Adams, PatOlympic 2011)

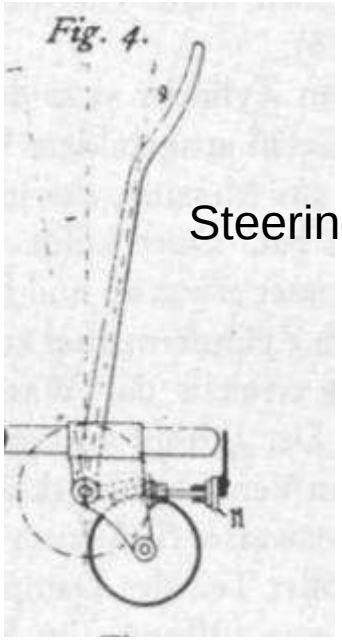
No 37435 Benz Patent – Moterwagen (1886)



You search for the entire invention but also on specific details

Engine function





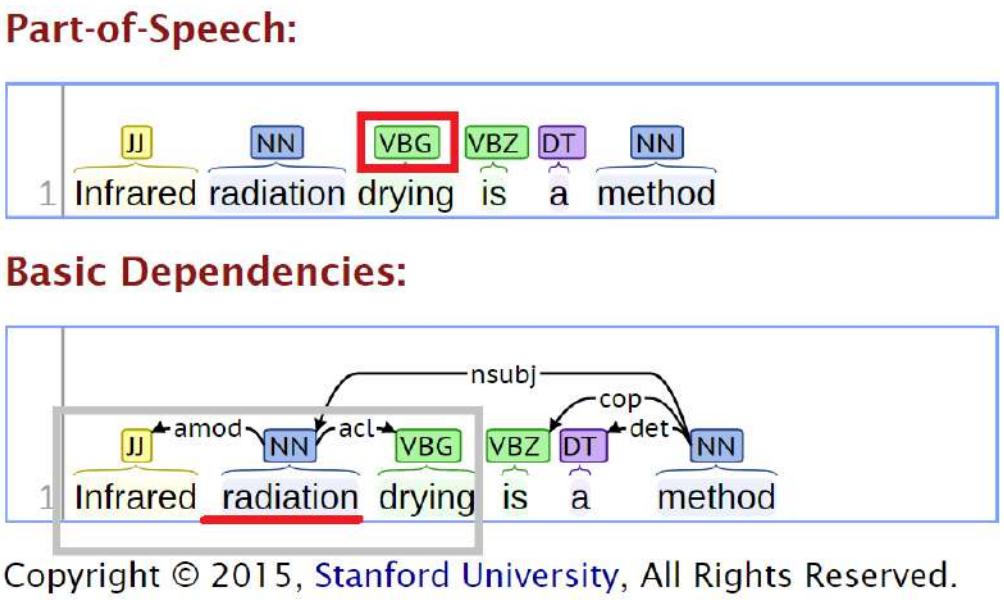
Steering mechanism

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’)

Natural Language Processing tools (NLP) identify Noun Phrases

- **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*

[1] *Domain Adaptation of General Natural Language Processing Tools for a Patent Claim Visualization System* Linda Andersson, et al , Multidisciplinary Information Retrieval, Lecture Notes in Computer Science, 2013

Task 1: Question and Answering

*What substance have a metling 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.

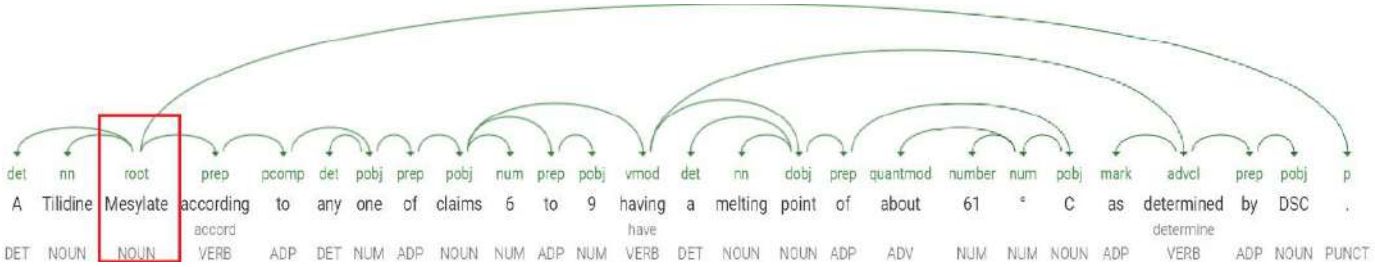
Application 1: Quantity linking

Ground truth:

Subject: *Tilidine Mesylate* **Predicate:** *having* **Object:** *melting point of about 61°C*

Google

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

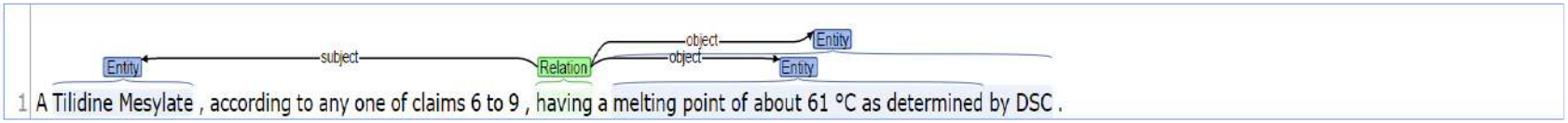


Stanford (corenlp.run)

Subject:*Tilidine Mesylate* Predicate:*having*

Object: *a melting point of about 61°C as determined by DSC.*

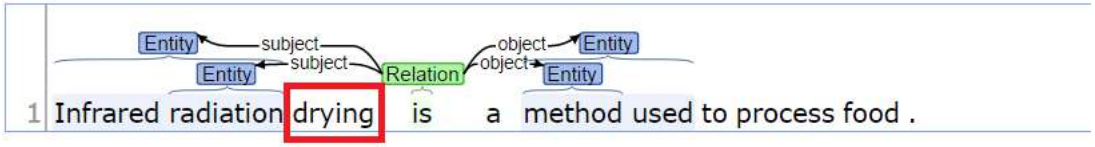
Open IE:



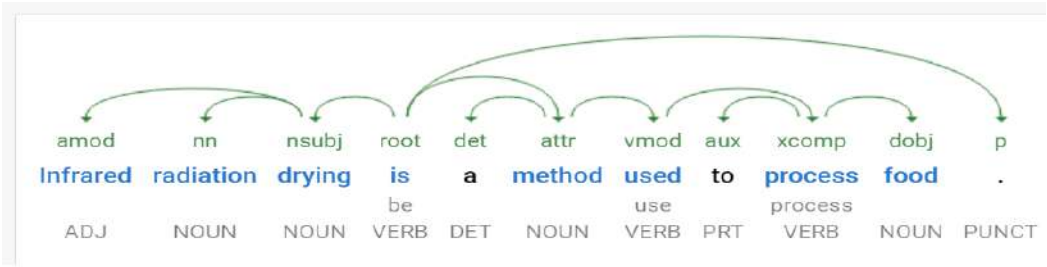
Ground truth: Infrared radiation drying

Stanford
Infrared radiation

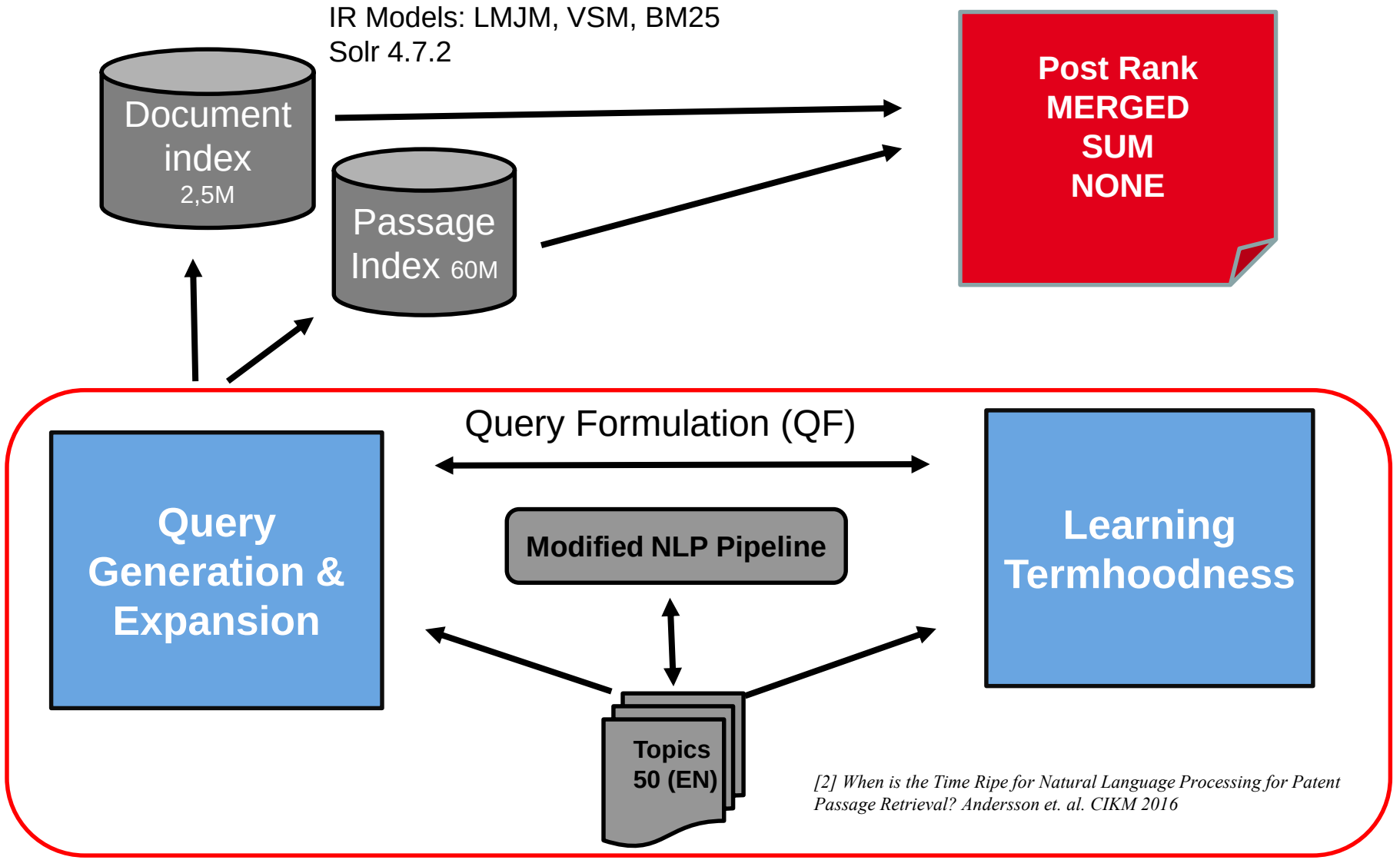
Open IE:



Google
Infrared radiation drying

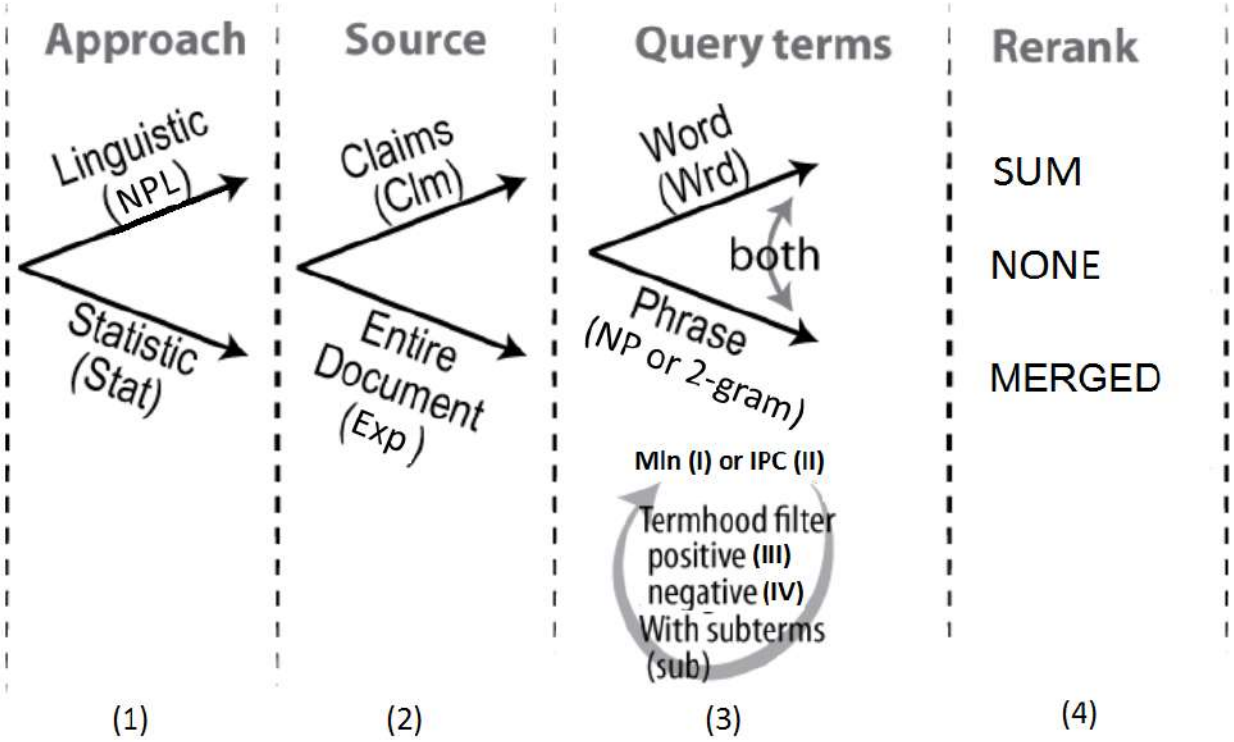


CLEF-IP 2013: Patent Passage Retrieval [2]



[2] When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval? Andersson et. al. CIKM 2016

Query Generation: 220 versions for each topic




```
<topic ucid: EP-1287743-A2 query: PSG-47>
  (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

Candidate Term	Word2Vec	C-Value	Pointwise Mutual information	Human
Remote communication	Yes	No	No	No
Communication link	No	Yes	Yes	Yes
Resin particle	No	Yes	No	Yes
washed washing	No/Yes (0.642)	Yes	No	No
Bar code	No	Yes	No	Yes
Wet strength	Not	Yes	No	Yes

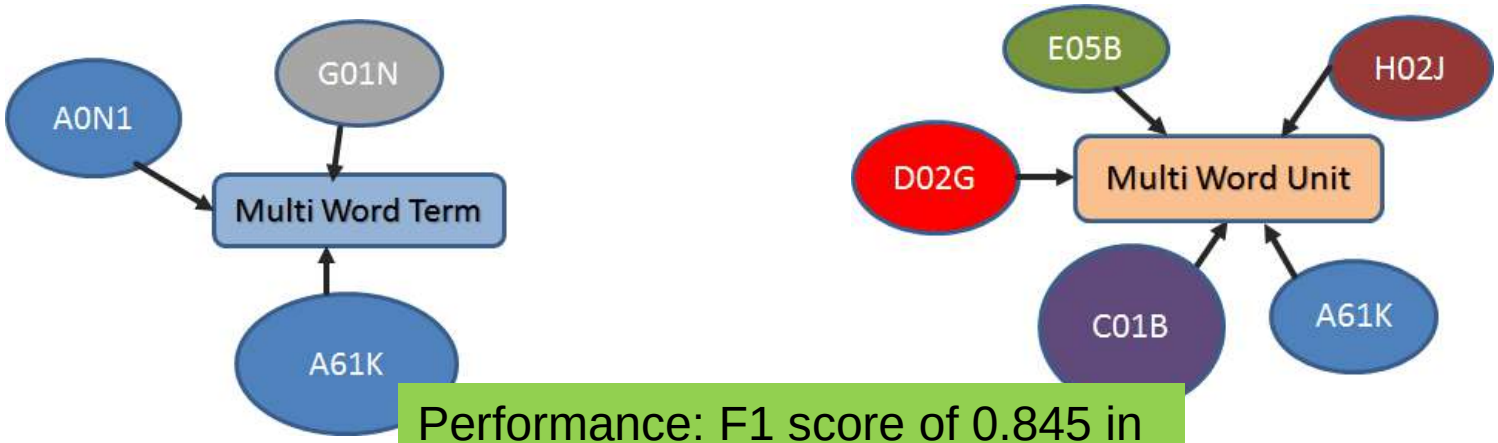
Different features to Learning Termhoodness

(Random sample of 637 noun phrases 222 negative, 451 positive)

Features		Feature combination																												
NLP	syntax	x	x	x	x	x	x	x	x	x			x	x	x	x	x	x		x	x		x	x	x				
	Syntax frequency	x	x	x	x	x		x	x	x			x	x	x	x				x	x								
n-gram	Phrase length	x	x	x	x	x		x	x	x			x	x	x	x				x			x						
NLP & Statistics	C-value	x	x	x										x	x	x	x					x						x	
Co-occurrence	DF	x	x		x	x		x	x				x	x	x		x	x	x								x		
Probability	PMI	x	x		x	x		x	x	x			x	x			x	x	x					x		x			
C-value and IPC	IPC:CValue	x	x	x					x		x											x							x
IPC-distributional-values	IPC:sum	x	x	x	x	x	x	x	x	x	x	x																	
	IPC:count	x	x	x	x	x	x	x	x	x	x	x																	
	IPC:mean	x	x	x	x	x	x	x	x	x	x	x																	
	IPC:median	x	x	x	x	x	x	x	x	x	x	x																	
	IPC:variance	x	x	x	x	x	x	x	x	x	x	x																	
	IPC:stddev	x				x	x				x	x																	
Correctly Classified		77	77	77	78	76	71	77	77	78	77	71	77	76	77	75	75	69	70	76	67	71	73	68	68	65	68	66	71

- Our assumption

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

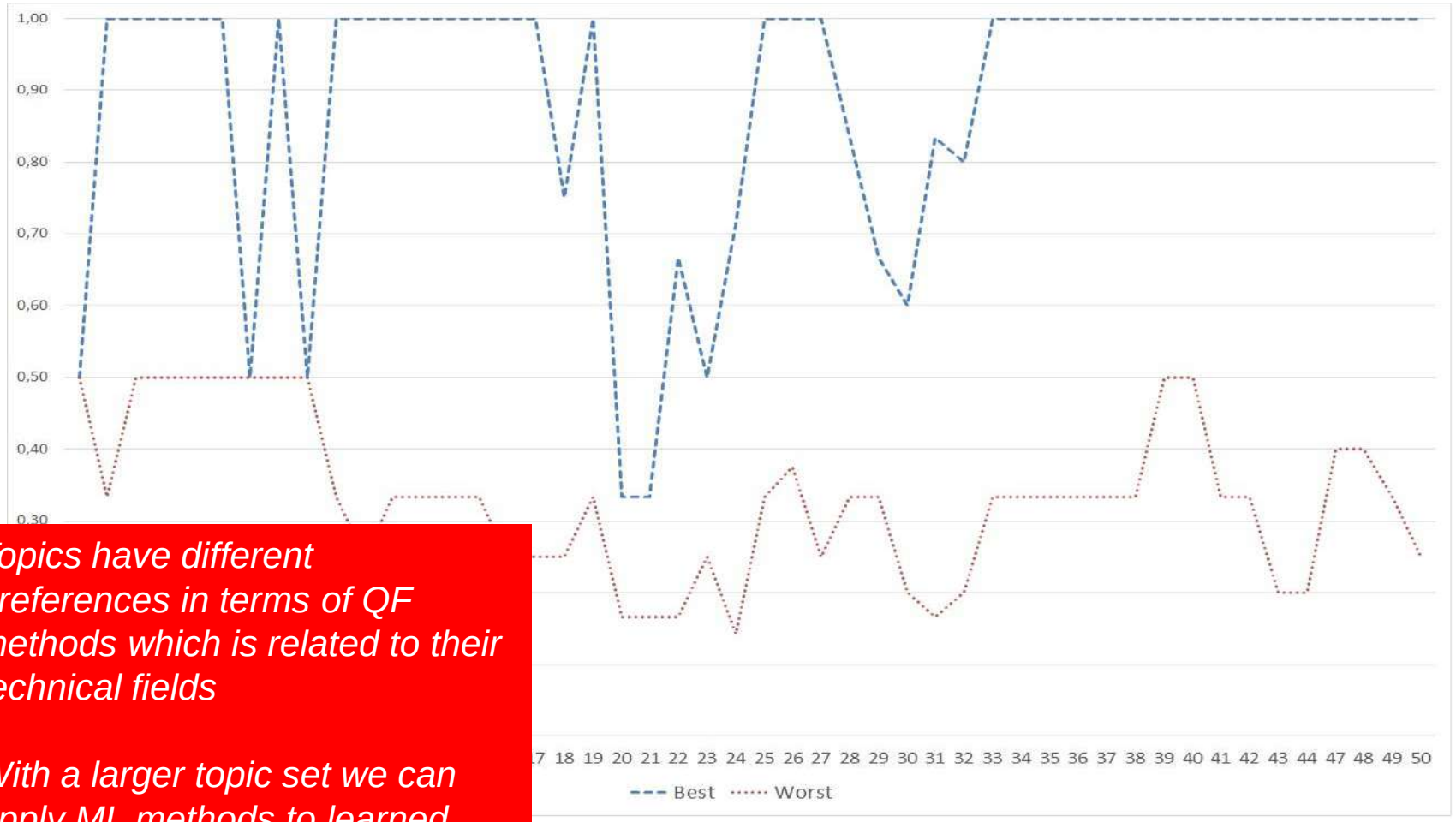


Performance: F1 score of 0.845 in accuracy when detection Technical terms among 4400 candidates

Patent Passage Retrieval CLEF-IP 2013

	Run	Query lengths	IR model	PRES	Recall	MAP	MAP(P)	Prec(P)	Post Ranking
Top 3 best methods	NLP, Expanded, Word, Technical terms (IPC), skip-gram (NLP1)	100	LMJM	***0.544***	***0.631***	0.285	0.112	0.218	Merged
	Statistical, Expanded, unigram, bigram	100	LMJM	0.492	0.574	**0.300**	0.114	0.208	Merged
	Statistical, only claim, unigram	100	LMJM	0.444	0.560	0.187	**0.146**	**0.282**	Merged
Baseline - unigram		100	LMJM	0.536	0.622	0.226	0.132	0.229	Merged
Best official runs clef-IP 2013	Document, word, hyphened MWUs, Upper bound IDF	N/A	BM25	0.433	0.540	0.191	0.132	0.213	N/A
	Document, word, hyphened MWUs, No upper bound IDF	N/A	BM25	0.432	0.540	0.190	0.132	0.214	N/A

What can learn in terms of Recall?

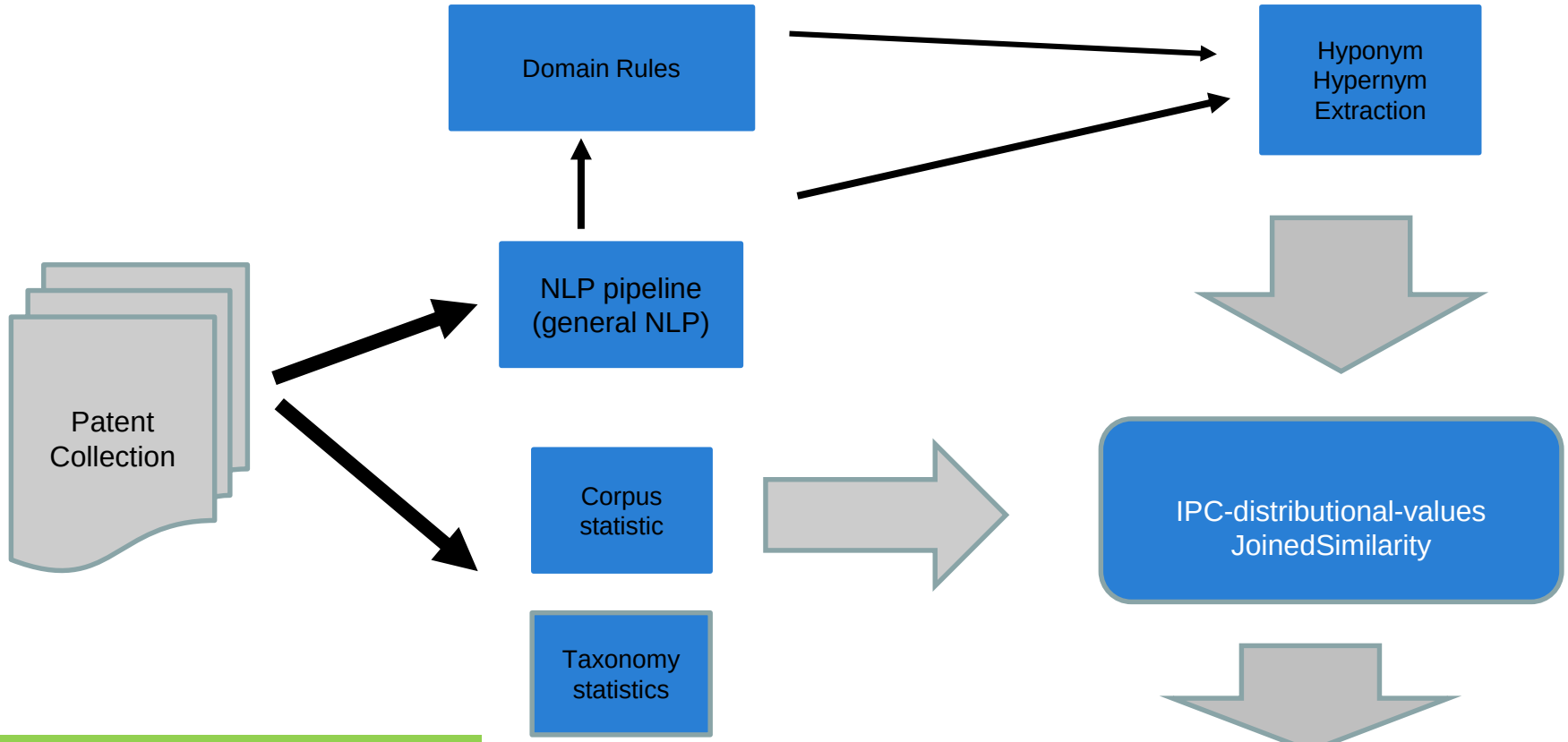


- *Topics have different preferences in terms of QF methods which is related to their technical fields*
- *With a larger topic set we can apply ML methods to learned which QF to use for the different technology domains in order increase Recall*

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

Automatic Knowledgebase Ontology



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

PRES	Recall	MAP	MAP (P)	Precison (P)
0,563	**0,653**	0,271	0,106	0,207

Word2vec – threshold 0.7 [3]

- **brake:**
  - brake, brakes, braking, braked, pedal, antilock, clutch
- **pedal**
  - pedal, pedals, brake, braking
- *bus**
  - *buses, busses, memory*

*Only refers to the semiconductor concept.

[3] Generalizing Translation Models in the Probabilistic Relevance Framework, Rekabsaz et al., CIKM 2016

JoinedSimilarity [4]

- Does “network lan” and “communication link” have (hyponymy) relation? Yes
- Does ”mechanical stress” and “communication link” have a (hyponymy) relation? No

$$JoinedSimilarity = \sum_{\substack{i,j=1,n \\ i \neq j \\ i<j}}^{N} \frac{\cos\left(\overrightarrow{w_i}, \overrightarrow{w_j}\right)}{N}$$

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

[4] Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information. Andersson et al WiNLP 2017

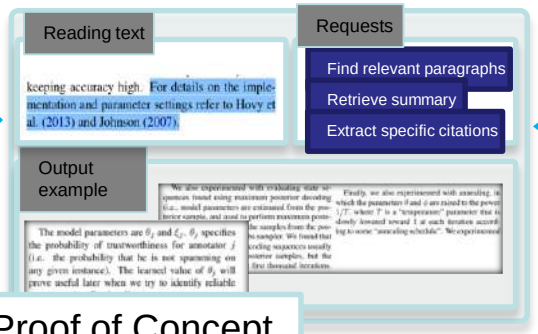
FAKULTÄT
FÜR !NFORMATIK

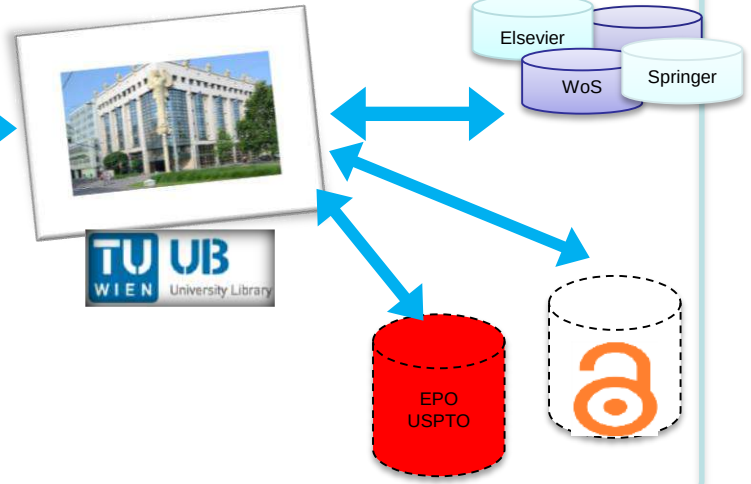
Faculty of Informatics

But patent searching also include non-patent literature

Now we transferring domain know-how into the general scientific publications text mining

Artificial Researcher In Science





Repository Processing – Proof of Concept

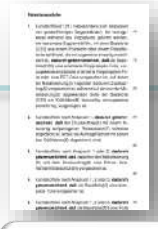
We are in the zone of domain specific text mining.

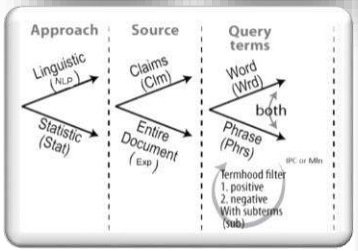
Our previous expertise: Patent documents
Our patent text mining research is now state-of-the-art

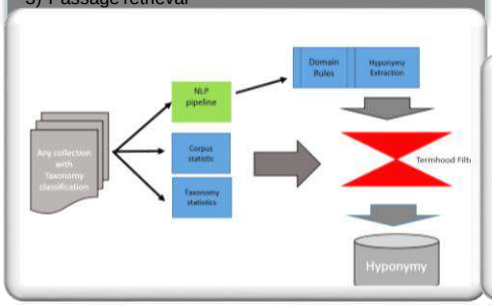
Given: A patent claim
Required: Relevant passages in patent citations.

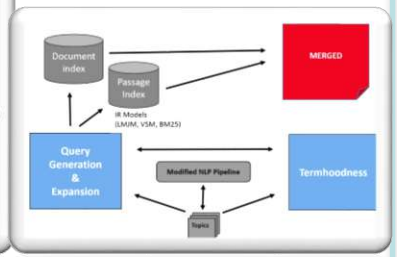
Solution considered the language, domain, and retrieval task complexity. All are necessary for a successful domain-specific text mining application.

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

FACULTY OF !INFORMATICS