Artificial Researcher

From A-Z, towards a patent text mining application

Domain Ontology Population

September 2019

The 1st PatentSemTech

Linda Andersson
andersson.ar.it@gmail.com
andersson@ifs.tuwien.ac.at
+43 (0) 699 17 821 600
http://www.ifs.tuwien.ac.at/~andersson/
https://www.artificialresearcher.com/

Linda Andersson - Résumé

Artificial Researcher

Academic merits

- Information design
  1998-2001

- General Linguistics
  2001-2003
- Computational Linguistics 2006-2009

- Language Engineering
  2004

UPPSALA
UNIVERSITET

- 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.

Artificial Researcher

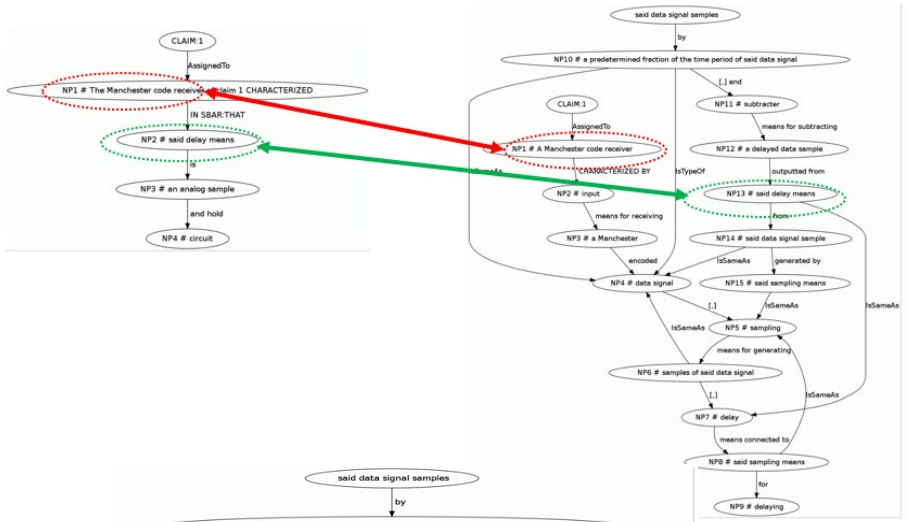
Outline

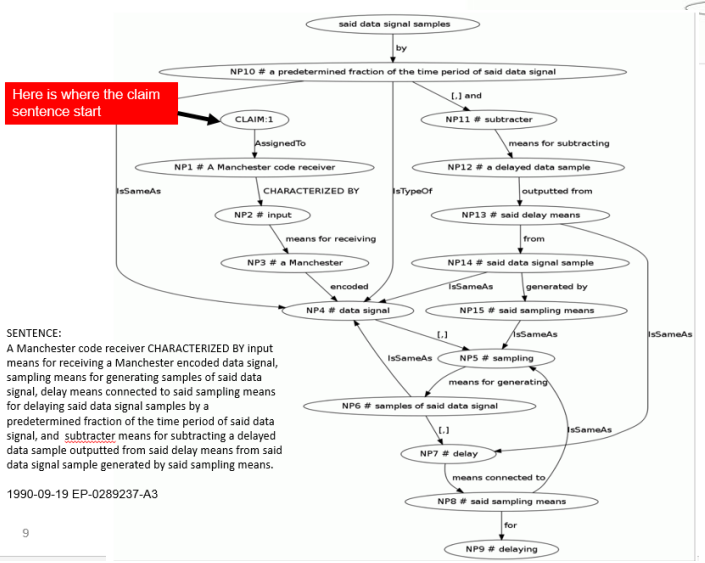
- Introduction and Motivations
- The limitation of general Natural Language Processing (NLP)
- Automatic Terminology extraction
- Domain ontology population

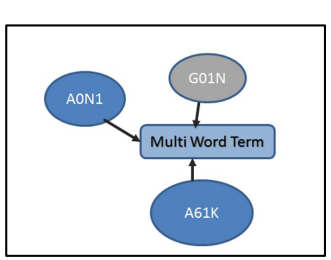
The Essence of Patent Text Mining

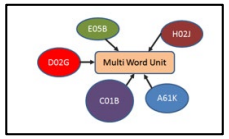
(Andersson 2019*)

Artificial Researcher

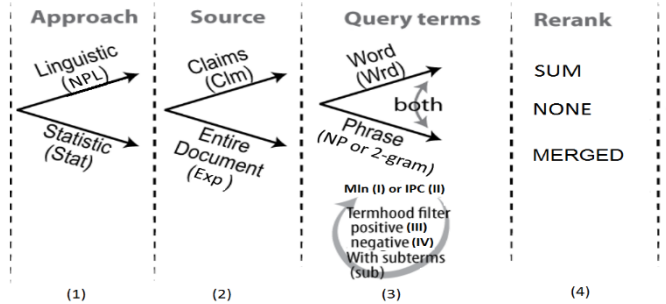


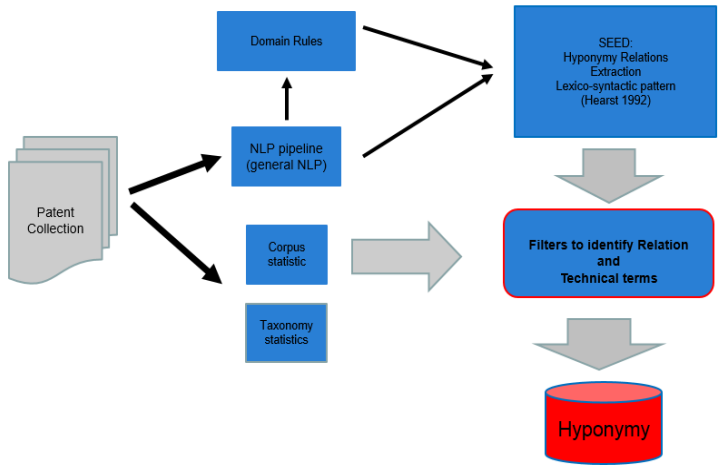


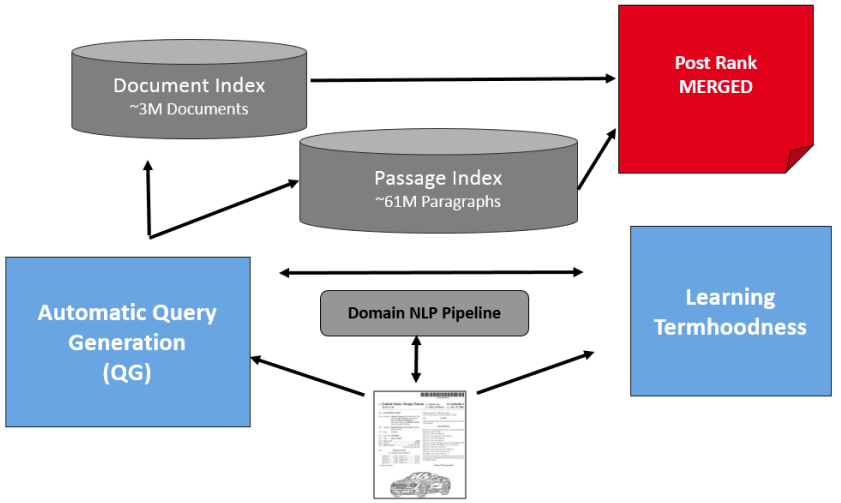




$$JoinedSimilarity = \sum_{\substack{i,j=1,n \\ i \neq j \\ i<j}}^{N} \frac{\cos\left(\overrightarrow{w_i}, \overrightarrow{w_j}\right)}{N}$$







TU WIEN FAKULTÄT FÜR !NFORMATIK Faculty of Informatics

SELF OPTIMIZER

In short: The patent text genre

Artificial Researcher

- Meta data
  - Bibliographic data (citing prior art, assignee, inventor, date)
  - Different Classification Schema
    - e.g. International Patent Classication (IPC).
    - reflects a semantic interpretation of the technical domains
    - taxonomy structure
- Linguistic Characteristics
  - No text normalization
  - Text section: Title, Abstract, Description, Claim
  - A mixture of technical term and legal terms
  - Patent genre consist of several sub languages

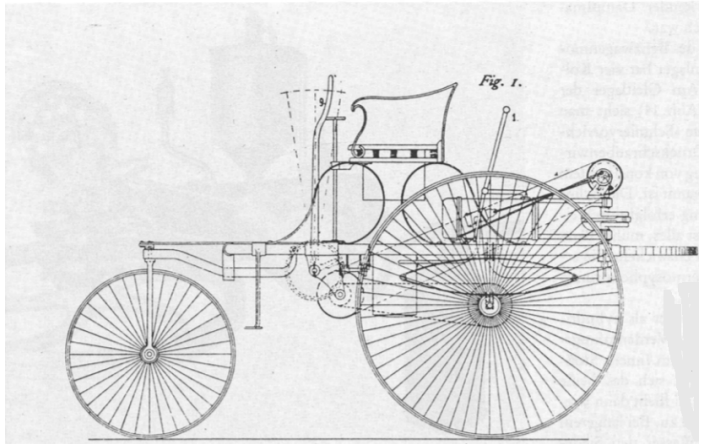
Motivation:
From the user requirement

- Use case: Provide a tool, which suggest related terms in the query formulation process

> *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 complexes queries targeting the main topic as well as sub topics of patents.*
>
> (S. Adams, Personal correspondence, PatOlympics 2011, Vienna).

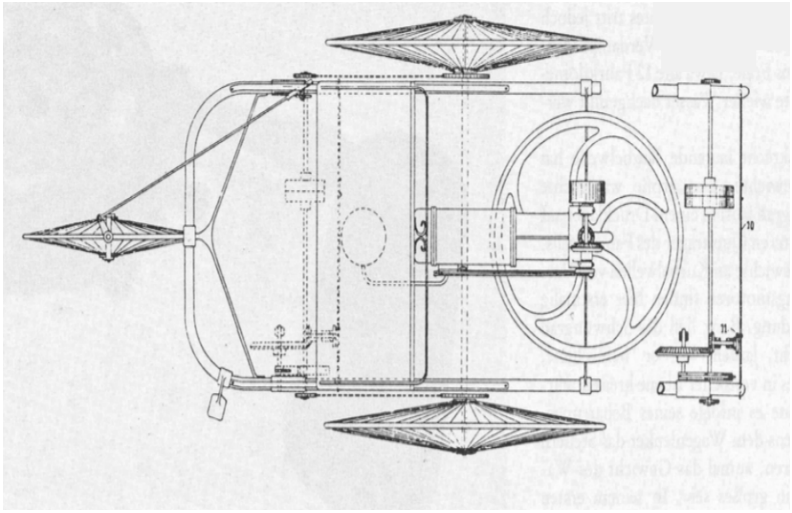
Motivation:

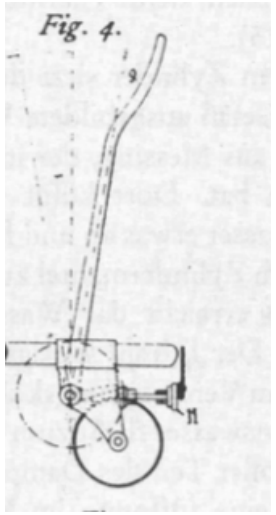
No 37435 Benz Patent – Moterwagen (1886)



You search for the entire invention but also on specific details

Engine function





Steering mechanism

Example of Automatic Query Formulation

Artificial Researcher

Automatic query expansion terms from ontologies

position brake actuating member
brake actuating member
hydraulically-assisted rack pinion steering gear
brake operating member
conventional braking system
pair pedals

accelerator pedal
case pedal device
pedal device

(12) United States Design Patent (10) Patent No.: US D500,000 S
Dyson et al. (45) Date of Patent: ** Dec. 21, 2004

(54) AUTOMOBILE BODY

(75) Inventors: Andrew P Dyson, West Bloomfield, MI (US); Joseph S Dehner, Bloomfield, MI (US); David C McKinnon, Bloomfield, MI (US); Glenn W Abbott, West Bloomfield, MI (US)

(73) Assignee: DaimlerChrysler Corporation, Auburn Hills, MI (US)

(**) Term: 14 Years

(21) Appl. No.: 29/201,094

(22) Filed: Mar. 10, 2004

(51) LOC (7) Cl. 12-08
(52) U.S. Cl. D12/92
(58) Field of Search D12/90-92, 86; D21/424, 433; 296/185

(56) References Cited
U.S. PATENT DOCUMENTS

D408,328 S * 4/1999 Ayoub et al. D12/92
D465,436 S * 11/2002 Dehner et al. D12/92
D476,601 S 7/2003 Stoddard et al. D12/92
D477,253 S * 7/2003 Minami et al. D12/92
D483,696 S * 12/2003 Howell et al. D12/92

* cited by examiner

Primary Examiner—Melody N. Brown
(74) Attorney, Agent, or Firm—Ralph E. Smith

(57) CLAIM

The ornamental design for an automobile body, as shown and described.

DESCRIPTION

FIG. 1 is a front perspective view of an automobile body showing our new design;
FIG. 2 is a side view thereof;
FIG. 3 is a rear perspective view thereof;
FIG. 4 is a front view thereof;
FIG. 5 is a rear view thereof;
FIG. 6 is a front perspective view of an automobile body showing a second embodiment of our new design;
FIG. 7 is a side view of FIG. 6;
FIG. 8 is a rear perspective view of FIG. 6;
FIG. 9 is a front view of FIG. 6; and,
FIG. 10 is a rear view of FIG. 6.

It will be understood that the dashed lines presented in the drawings are for illustration only, and do not form a part of the claimed design.

1 Claim, 6 Drawing Sheets

<QUERY>
(conured OR clutch OR connectability OR nmofs OR fclp OR dnsr 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>

Example of automatic identified technical term and suggestion of query expansion terms

Artificial Researcher

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

accelerator pedal
case pedal device
pedal device

Artificial Researcher

The limitations of the general Natural Language Processing tools

From A-Z, towards a patent text mining application

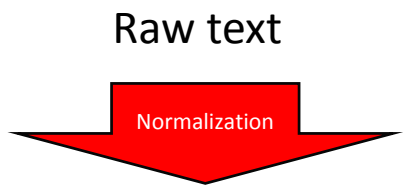
Artificial Researcher

Natural Language Processing

According to Wikipedia:
https://en.wikipedia.org/wiki/Natural_language_processing, 2019-09-02

"Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data."
…..

The different linguistic layers



PoS-tagging

Phrase chunker

Underlying layer

Example of a parser

Surface level:

The process according to claim 6, wherein the inlet temperature is between about 90°C and about 120°C.

Part of Speech:

The/DT process/NN according/VBG to/TO claim/**NN** (*VBN) 6/CD ,/, wherein/WRB the/DT inlet/NN temperature/NN is/VBZ between/IN about/RB 90/CD DEG/NNP C/NNP and/CC about/RB 120/CD DEG/NNP C./NNP

Noun Phrase Chunk:

[The/DT process/NN] according/VBG to/TO [claim/**NN** **[**(*VBN) 6/CD] ,/, wherein/WRB [the/DT inlet/NN temperature/NN] is/VBZ between/IN [about/IN 90/CD °/CD C/NN] and/CC [about/IN 120/CD °/CD C./NNP]

Constituent Information:

ROOT (FRAG (NP (DT The) (NN process)) (PP (VBG according) (PP (TO to) (NP (NP (NN claim) (CD 6)) (, ,) (SBAR (WHADVP (WRB wherein)) (S (NP (DT the) (NN inlet) (NN temperature)) (VP (VBZ is) (PP (IN between) (NP (NP (NP (QP (RB about) (CD 90))) (NP (NNP DEG) (NNP C))) (CC and) (NP (NP (QP (RB about) (CD 120))) (NP (NNP DEG) (NNP C.)))))))))))

Sentence marked with a type syntactic relations (**Typed dependencies**)

det(process-2, The-1) root(ROOT-0, process-2) dep(process-2, according-3) pcomp(according-3, to-4) pobj(to-4, claim-5) num(claim-5, 6-6) advmod(is-12, wherein-8) det(temperature-11, the-9) nn(temperature-11, inlet-10) nsubj(is-12, temperature-11) rcmod(claim-5, is-12) prep(is-12, between-13) quantmod(90-15, about-14) pobj(between-13, 90-15) nn(C-17, DEG-16) dep(90-15, C-17) cc(90-15, and-18) quantmod(120-20, about-19) conj(90-15, 120-20) nn(C.-22, DEG-21) dep(120-20, C.-22)

Pre-processing: Token detection

- Token (i.e. words, letter strings, **digits**)
  - cat, dogs, U2, padd-227
- Rhetorical structure of a discourse (e.g. commas, punctuations, digits, etc.)
  - **Digits** – numeration structure of text
  - Punctuations
    - Part of acronyms, digits marker and sentence boundaries
  - Commas
    - Clause binder: While she was cooking, her friend arrived
    - Numeration binder: mixtures of saturated hydrocarbon compounds, alicyclic hydrocarbons, aromatic hydrocarbons, etc.
    - Part of Chemical compound: 2,5-bis amidinophenyl

Pre-processing: Sentence detection

```
CELLULAR COMMUNICATIONS INC. sold 1,5500,000
common shares at $21.75 each yesterday, according
to lead underwriter L.F. Rothschild & Co.
```

Example No.	Regular expression	Correct	Errors	Ambiguities of full stop
1	[A-Za-z]\.	1,323	30	14
2	[AZa-z]\.([A-Za-z0-9]\.)+	626	0	63
3	[A-Z][bcdfgj-np-tvxz]+\.	1,397	33	26
Totals		3,876	63	103 [1]

Examples from Grefenstette et al (1994)

Definition of a sentence will differ from different domains, example from the patent text genre:

1. What is claimed is: 1. A control and communication system for a light-duty combustion engine, comprising: a circuit card; an ignition circuit carried by the circuit card and configured to control an ignition timing of the engine; and a short range wireless communication circuit carried by the circuit card.

Artificial Researcher

Language Complexity: *challenge* for general NLP

- Language dependent features
  - Word formation in English
    - suffixes
      - (e.g. to dry versus a method of dry***ing***)
    - compound noun
      - (e.g. floppy disk, air flow, blood cell, bus slot card)
  - Complex syntactic construction.

Source and Test data, *issue*

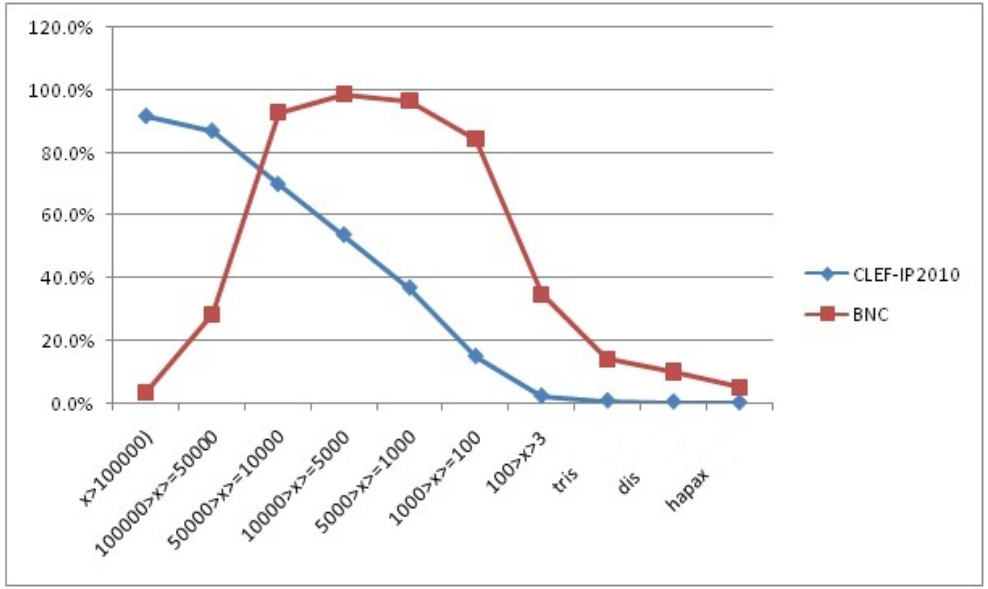
Artificial Researcher

The Penn Treebank (source data for NLP tools)

Description¤	Tagged·for·PoS·(tokens)¤	¤
Dept.·Of·Energy¤	231,404¤	¤
Dow·Jones·Newswire·stories¤	3,065,776¤	¤
Dept.·Of·Agriculture·bulletins¤	78,555¤	¤
Library·of·America·text¤	105,652¤	¤
MUC-3·messages¤	111,828¤	¤
IBM·Manual·sentences·¤	89,121¤	¤
WBUR·radio·transcripts¤	11,589¤	¤
ATIS·sentences¤	19,832¤	¤
Brown·Corpus·retagged¤	1,172,041¤	¤
Total¤	4,885,798¤	¤

Table·2:·(Marcus·et·al·1993·p.·327)¶

WordNet coverage (the large LR – manually constructed)



British National Corpus (BNC)

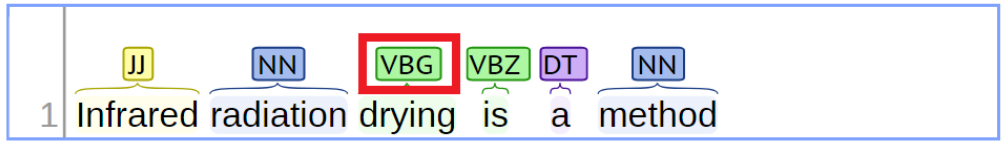
	PTB	MAREC_US_2500
Table	0	638
sentence	46 665	462 912
digits	67 690	1844824
noise	35 112	1674961
token type	41 311	1888838
tokens	947 139	10531164

	Sentence Length		Term frequency	
	PTB	MAREC_US_2500	PTB	MAREC_US_2500
average	23	43	22	55
median	20	70	2	1
max	173	10217	33 001	39858
average per document	50	5733	423	5733

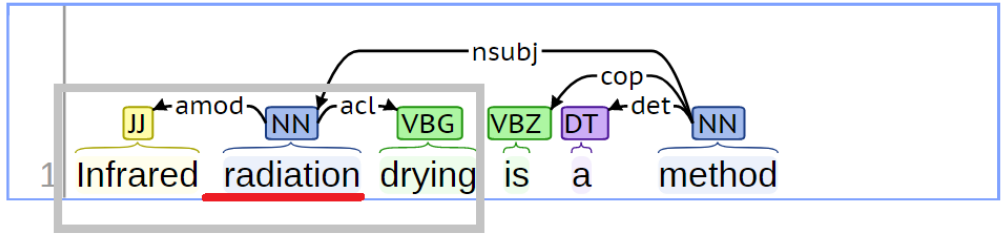
A example of the limitation of the general NLP tools

- NLP identifies noun phrases
  - Source (news text) versus Target data (patent text)

Part-of-Speech:



Basic Dependencies:



  - *Verb participles were discovered to be erroneous in patent text*

Artificial Researcher

Trick, is the know-how

Application 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*.

Application 2: Automatic Terminology Extraction

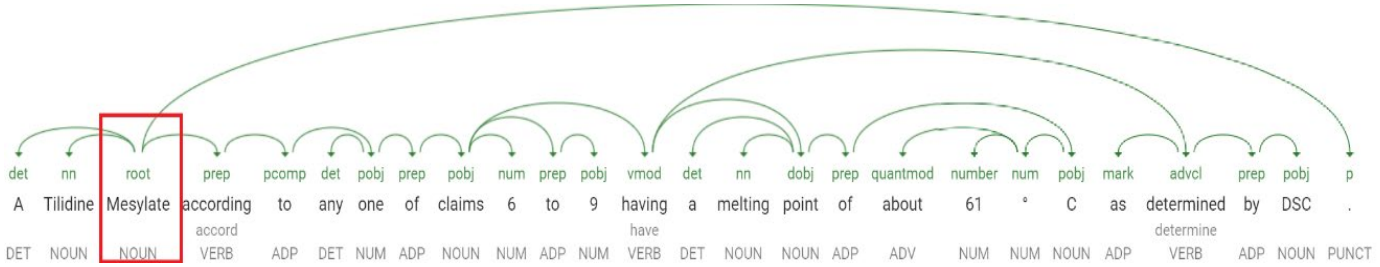
Every local **bus slot card** willing to master the bus will have to mimic 030, so it appears the 040-to-030 cycles translation adapter will always be in between the CPU and the local bus, no matter be it 040 or 060

Infrared radiation drying is a method used to process food.

Application 1: Quantity linking

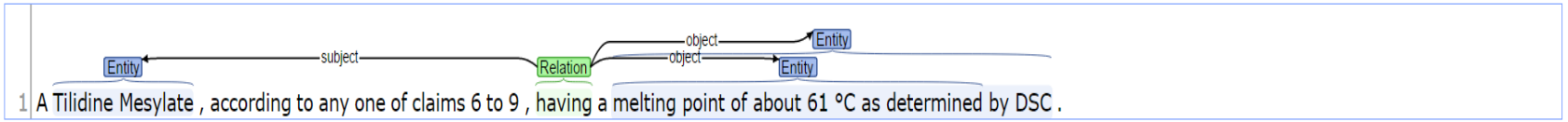
Ground truth:
Subject: *Tilidine Mesylate* Predicate: *having* object: *melting point of about 61°C*

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:

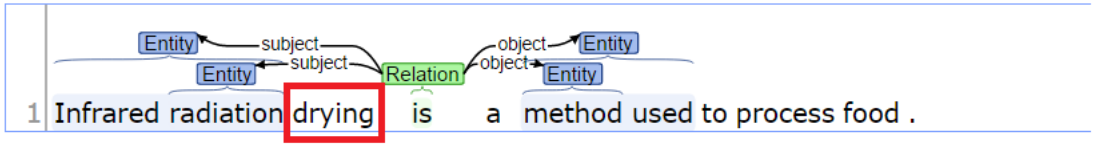


Application 2: Automatic Terminology Extraction

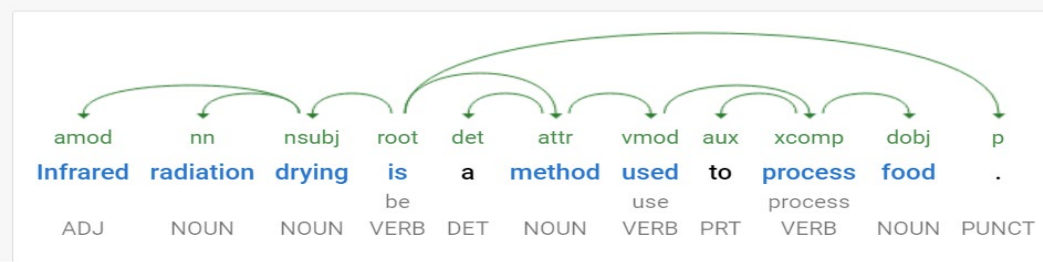
Ground truth: Infrared radiation drying

Stanford
Infrared radiation

Open IE:



Google
Infrared radiation drying

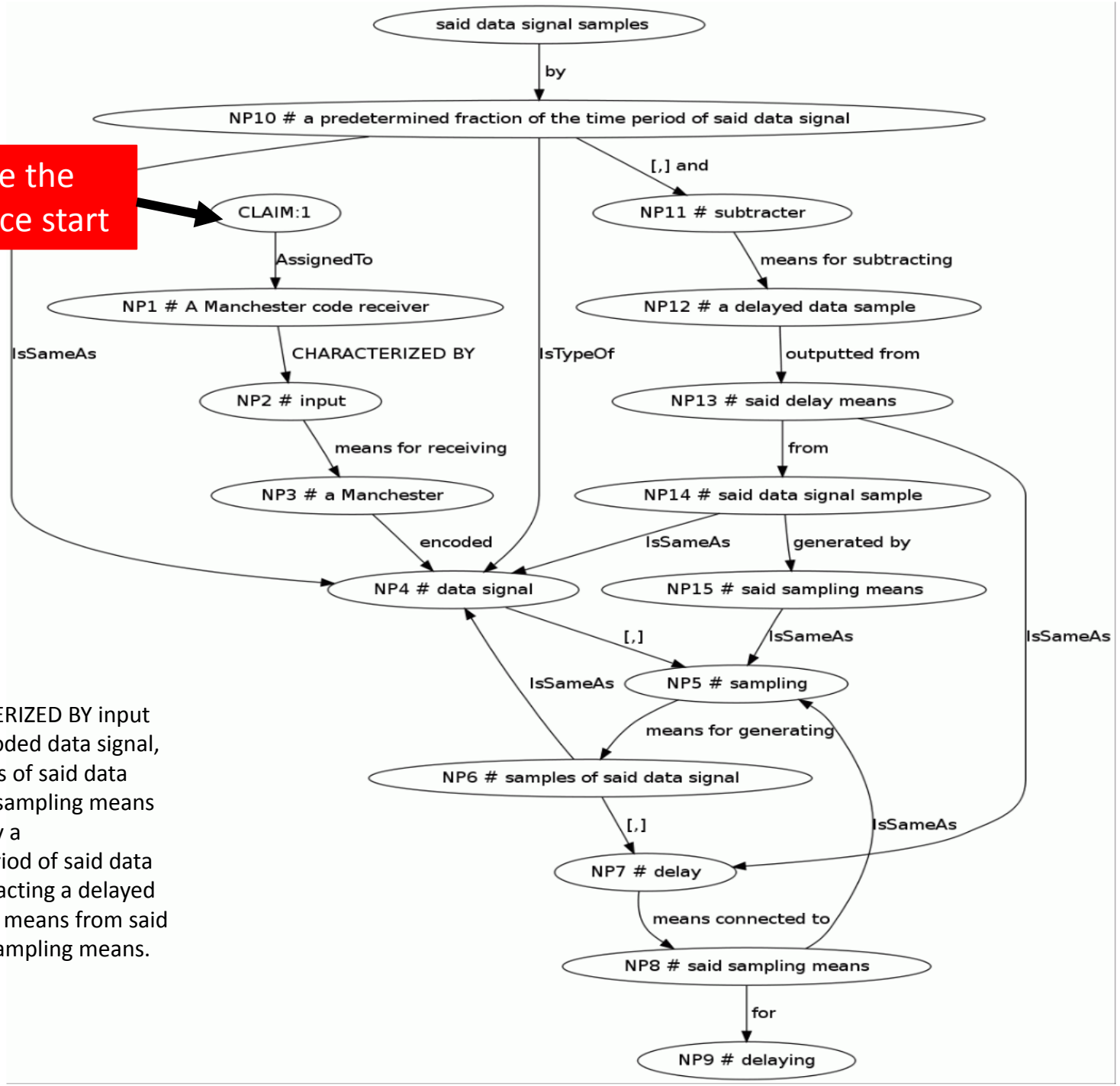


Artificial Researcher

Dependency Claim Graph

From A-Z, towards a patent text mining application

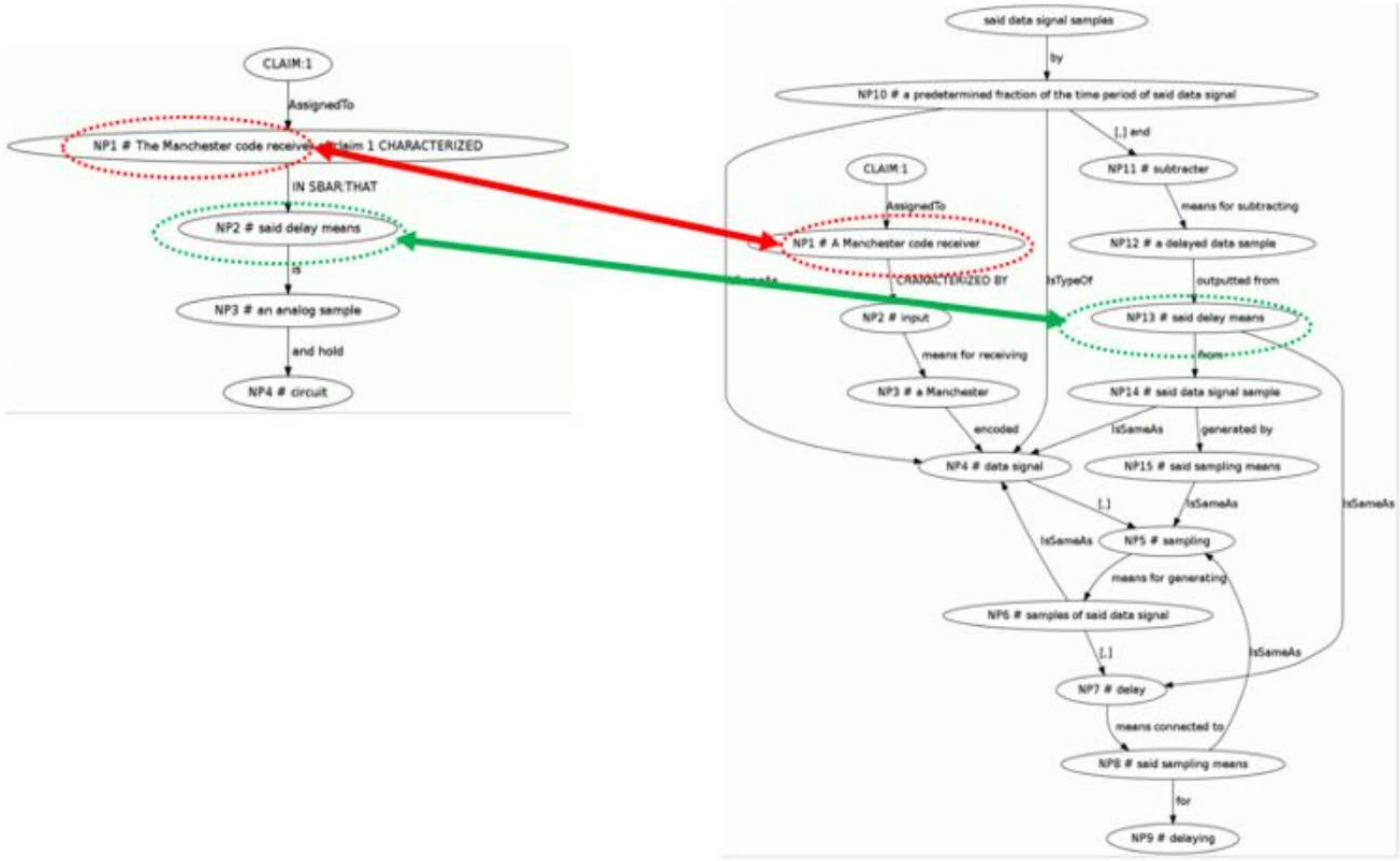
Here is where the claim sentence start



A Manchester code receiver CHARACTERIZED BY input means for receiving a Manchester encoded data signal, sampling means for generating samples of said data signal, delay means connected to said sampling means for delaying said data signal samples by a predetermined fraction of the time period of said data signal, and subtracter means for subtracting a delayed data sample outputted from said delay means from said data signal sample generated by said sampling means.

1990-09-19 EP-0289237-A3

Link internal and external relations



<NLP>**[A/DT method/NN]** **of/IN generating/VBG** **[illumination/NN characteristic/JJ data/NNS]** around/IN [an/DT image/NN display/NN device/NN] ,/, comprising/VBG :/: making/VBG [predetermined/JJ illumination/NN characteristic/JJ data/NNS] around/IN [the/DT image/NN display/NN device/NN] into/IN [a/DT type/NN block/NN indicating/VBG information/NN] on/IN **[a/DT type/NN] of/IN [illumination/NN]** ,/, [the/DT information/NN] on/IN **[the/DT type/NN] of/IN [illumination/NN]** comprising/VBG **[at/IN least/JJS one/CD] of/IN [a/DT color/NN temperature/NN] of/IN [illumination/NN] and/CC [a/DT coordinate/JJ value/NN]** in/IN [chromaticity/NN] coordinates/VBZ of/IN [illumination/NN] ;/: and/CC making/VBG [the/DT predetermined/JJ illumination/NN characteristic/JJ data/NNS] into/IN [an/DT illuminance/NN block/NN indicating/VBG information/NN] on/IN **[the/DT illuminance/NN] of/IN [illumination/NN**] ,/, [the/DT information/NN] on/IN **[the/DT illuminance/NN] of/IN [illumination/NN]** being/VBG [a/DT numerical/JJ illuminance/NN value/NN] [which/WDT] is/VBZ represented/VBN in/IN **[the/DT units/NNS] of/IN [Lux/FW]** ./.</NLP>

<DomainNLP>[A/DT method/NN of/IN generating/VBG illumination/NN characteristic/JJ data/NNS] around/IN [an/DT image/NN display/NN device/NN] ,/, comprising/VBG :/: making/VBG [predetermined/JJ illumination/NN characteristic/JJ data/NNS] around/IN [the/DT image/NN display/NN device/NN] into/IN [a/DT type/NN block/NN indicating/VBG information/NN] on/IN [a/DT type/NN of/IN illumination/NN] ,/, [the/DT information/NN] on/IN [the/DT type/NN of/IN illumination/NN] comprising/VBG [at/IN least/JJS one/CD of/IN a/DT color/NN temperature/NN of/IN illumination/NN NP-COORDINATOR:and/CC a/DT coordinate/JJ value/NN] in/IN [chromaticity/NN] ***coordinates/VBZ of/IN*** [illumination/NN] ;/: and/CC making/VBG [the/DT predetermined/JJ illumination/NN characteristic/JJ data/NNS] into/IN [an/DT illuminance/NN block/NN indicating/VBG information/NN] on/IN [the/DT illuminance/NN of/IN illumination/NN] ,/, [the/DT information/NN] on/IN [the/DT illuminance/NN of/IN illumination/NN] being/VBG [a/DT numerical/JJ illuminance/NN value/NN] SBAR:which/WDT is/VBZ represented/VBN in/IN [the/DT units/NNS of/IN Lux/FW] ./.</DomainNLP>

SENTENCE

A method of generating illumination characteristic data around an image display device, comprising: making predetermined illumination characteristic data around the image display device into a type block indicating information on a type of illumination, the information on the type of illumination comprising at least one of a color temperature of illumination and a coordinate value in chromaticity coordinates of illumination; and making the predetermined illumination characteristic data into an illuminance block indicating information on the illuminance of illumination, the information on the illuminance of illumination being a numerical illuminance value which is represented in the units of Lux.

We need to targeting the English Noun Phrase

Artificial Researcher

Rule	Original NP Sequence	Modified NP Sequence	Modifiying
"said" as an article	said/VBD [supercritical/JJ fluid/NN]	[said/VBD supercritical/JJ fluid/NN].	PoS-tagger
preposition within the preamble phrase	[The/DT soccer/NN shoe/NN] of/IN [claim/NN 4/CD]	[The/DT soccer/NN shoe/NN of/IN claim/NN 4/CD]	Chunker
include present participle	[A/DT method/NN] of/IN fabricating/VBG [a/DT semiconductor/NN device/NN]	[A/DT method/NN of/IN fabricating/VBG a/DT semiconductor/NN device/NN]	Chunker
infinitive verb tagged as NN	[said/VBD laser/NN radiation/NN] to/TO [exit/NN] [said/VBD exit/NN system/NN]	[said/VBD laser/NN radiation/NN] to/TO exit/VB [said/VBD exit/NN system/NN].	PoS-tagger
include digits into the NP	NP [The/DT method/NN of/IN any/DT of/IN claims/NNS] [12/CD to/TO 16/CD]	[The/DT method/NN of/IN any/DT of/IN claims/NNS 12/CD to/TO 16/CD]	PoS-tagger
list of NPs	in [the/DT group/NN] consisting/VBG of/IN [a/DT photoresist/NN] ,/, [a/DT photoresist/NN residue/NN] ,/, and/CC [a/DT combination/NN]	into [the/DT group/NN] consisting/VBG of/IN [a/DT photoresist/NN ,/, a/DT photoresist/NN residue/NN ,/, and/CC a/DT combination/NN]	Claims discourse adaptation specific rules
	A sub rule to 7,		
	Identifying, transition phrases listing sub clauses as seen in figure 2.B		

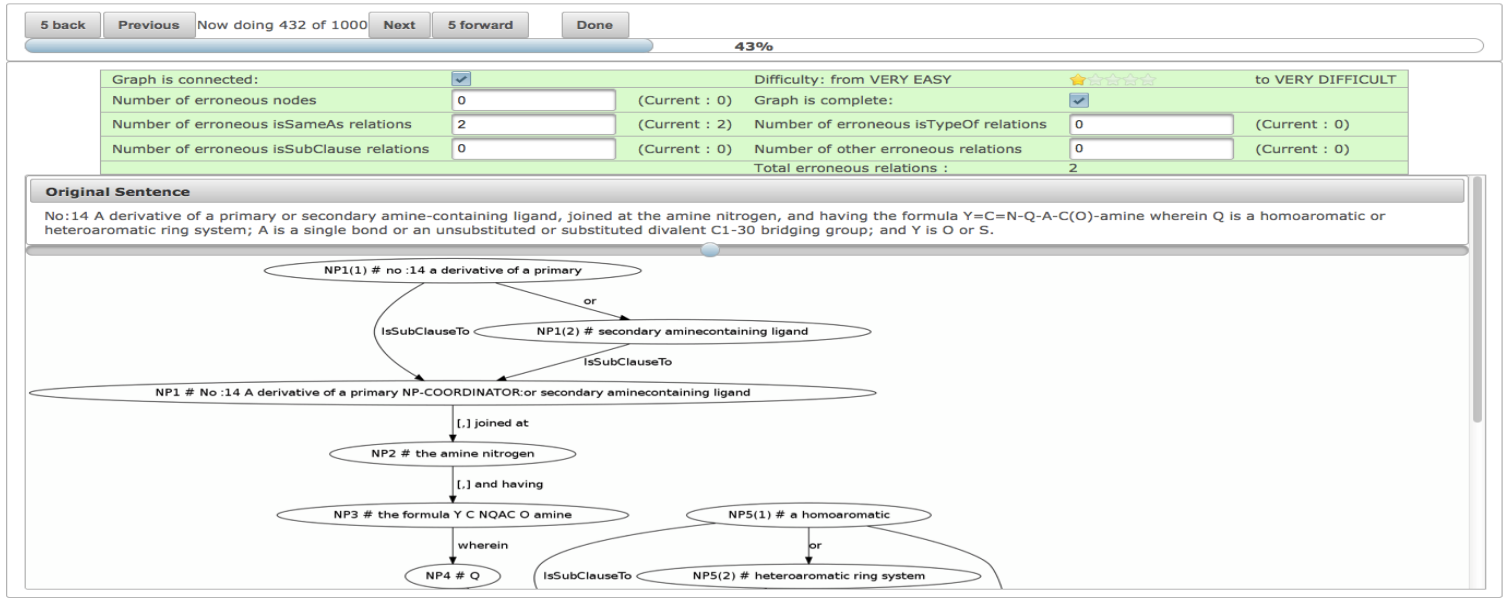
Artificial Researcher

Experiment

- CLEF-IP2012 EN 35 topics
  - 600 sentences (three assessors per sentence)

- Assessors
  - Experts (3) vs non-experts (14)
  - We defined seven parameters we ask the user to assess for each graph:
    - graph is complete,
    - graph is connected,
    - number of erroneous nodes,
    - number of erroneous IsSameAs relations,
    - number of erroneous IsSubClauseTo relation,
    - number of erroneous IsTypeOf relations,
    - number of other erroneous relations.

Evaluation:
Inter-annotation agreement



Assessor Pair	No of sentences	Connected Graphs	Erroneous Nodes	Erroneous IsSameAs	Erroneous IsTypeOf	Erroneous IsSubClauseTo	Erroneous Other Relations	Complete graphs	Graph Difficulty
Non-expert vs Expert	182	98.35	68.13	87.91	97.80	96.15	69.78	84.62	26.37
Expert vs Expert	193	97.41	61.14	84.97	97.93	98.45	64.77	74.09	56.48

Results, for different IPC Section

Only Experts

IPC	No of Sentences	Erroneous Nodes	Erroneous IsSameAs	Erroneous IsTypeOf	Erroneous IsSubClause	Erroneous Other Relations	Complete Graph	Connected Graph	Difficulty
A	182	0.05 (0.08)	0.03 (0.07)	0 (0.01)	0 (0.02)	0.04 (0.08)	0.83 (0.34)	0.99 (0.11)	1.73 (1.1)
B	158	0.11 (0.13)	0.02 (0.05)	0 (0.01)	0 (0.02)	0.08 (0.11)	0.81 (0.35)	0.98 (0.12)	1.71 (1.1)
C	165	0.09 (0.12)	0.02 (0.05)	0 (0.02)	0 (0)	0.06 (0.11)	0.8 (0.35)	0.97 (0.16)	1.69 (1.08)
D	27	0.03 (0.06)	0.01 (0.03)	0 (0.01)	0 (0)	0.02 (0.04)	0.89 (0.28)	1 (0)	1.41 (0.91)
F	27	0.08 (0.08)	0.01 (0.05)	0 (0)	0 (0)	0.04 (0.05)	0.72 (0.37)	0.98 (0.09)	1.63 (0.67)
G	99	0.13 (0.13)	0.02 (0.05)	0 (0.02)	0 (0)	0.1 (0.12)	0.81 (0.35)	0.97 (0.16)	1.81 (1.17)
H	62	0.11 (0.11)	0.04 (0.07)	0 (0.02)	0 (0)	0.07 (0.09)	0.73 (0.41)	0.92 (0.27)	2.17 (1.39)
Total	720	0.09 (0.12)	0.02 (0.06)	0 (0.01)	0 (0.01)	0.06 (0.1)	0.81 (0.35)	0.97 (0.15)	1.75 (1.12)

Artificial Researcher

Automatic Terminology Extraction

From A-Z, towards a patent text mining application

What is a Multi Word Term and what is not?
Depends on who you are asking?

Artificial Researcher

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

Artificial Researcher

Automatic Terminology Extractions

- Finding Termhoodness among phrases
  - State-of-the-art: C-Value (Frantzi et al 2000)
  - 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

Artificial Researcher

Experiment

- All sentences in the corpus containing the candidate terms need to be PoS tagged and chunked
  - 40,149,317 parsed sentences
  - 5 months processing time

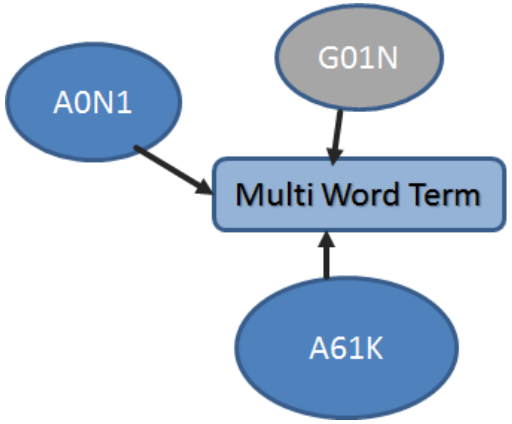
- Random sample of 637 phrases.
  - 222 negative, 451 positive
  - Manually assessed

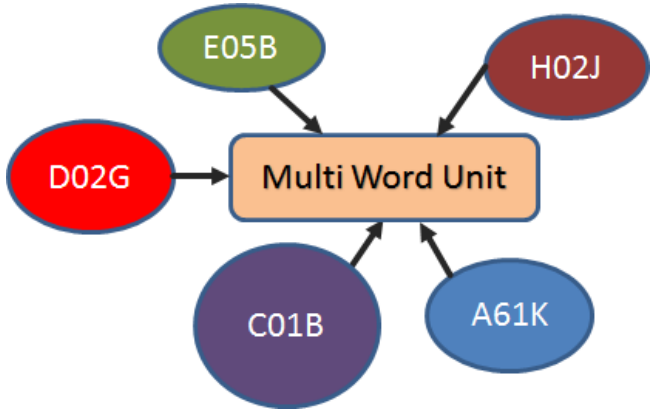
- Tested 13 different features
    - Syntax, phrase length, C-value, *IPC-distribution-values*, document frequency, mutual information

Domain knowledge: IPC-distributional-values

- Our assumption

Phrases having a homogenous distribution of IPC codes will reflect the termhoodness compared to phrases with heterogeneous distribution





Result: Learning Termhoodness

Artificial Researcher

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	MI	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	

IPC-distributional-values
Using **Random Forest Regression** model obtained F1 score of 0.845 in accuracy
(Sample data of 4,000)

Artificial Researcher

Domain ontology populations

From A-Z, towards a patent text mining application

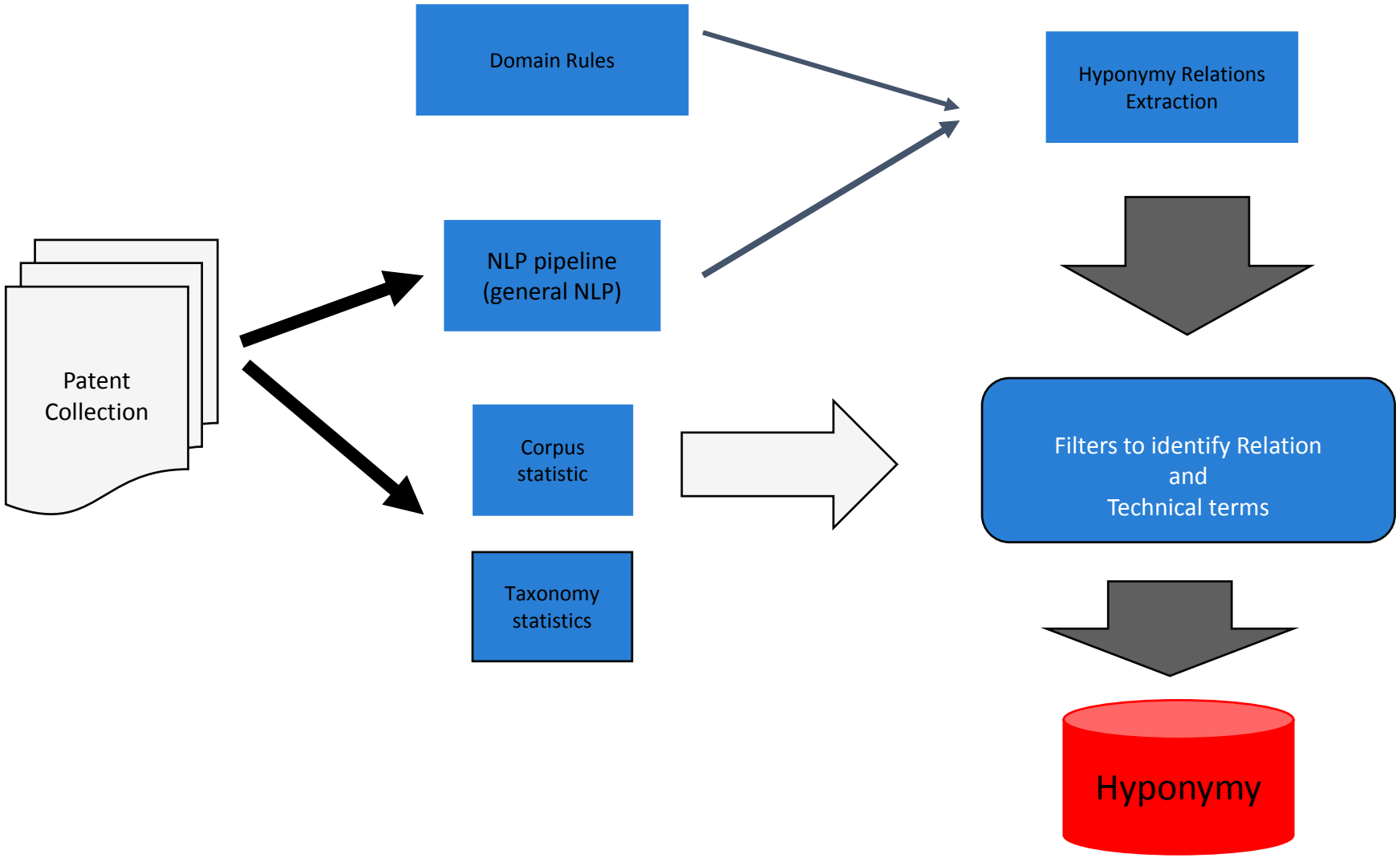
Artificial Researcher

Ontology

- …description basic categories and there relation
- Why would someone want to develop an ontology?

  - To share common understanding of the structure of information among people or software agents
  - To enable reuse of domain knowledge
  - To make domain assumptions explicit
  - To analyze domain knowledge
  - To make a Browsable search aid

Automatic Knowledgebase Ontology



Extracting Lexico- Semantic relation with different techniques

- Embedding identifies similarities between different words
  - **Word:** underwear **Similar to:** underpants , undergarment, panties, sportswear, undergarments, underclothes
  - **Word:** strength **Similar to:** strengths, strength, toughness, stronger, sfrength

(Extracted 0.6 threshold based upon Rekabsaz et al 2016 and Rekabsaz et al 2017)

- ***NLP technical term relation between multi words term***
  - **Multi word term:** *synthetic fibers* **Similar to:** *polyester fibers*
  - **Word:** fabric **Similar to:** *non cellulose fibers* , *melt-extruding thermoplastic synthetic resin fiber*, *nonwoven fabric*, *woven fabric* (different type of fabrics)

With a wider semantic definition of the hyponym property

Include both ‘part of’ and ‘member of’ in the definition:

“… an expression A is a hyponym of an expression B iff the meaning of B is part of the meaning of A and A is subordinated of B. In addition to the meaning of B, the meaning of A must contain further specifications, rendering the meaning of A, the hyponym, more specific than the meaning of B.”

(Löbner 2002)

Hypernym a word with a broad meaning constituting a category

Hyponym a word of more specific meaning than a general or superordinate term applicable to it.

Artificial Researcher

Distributional Semantic models

- Gives attributional relations in a given sample
- The usage define the meaning of a word?
- Linguistic hypothesis behind the popular usage statement

  “You shall know a word by the company it keeps”

  (Firth 1957 p. 11)

- … is this true for all meaning of a word?

Artificial Researcher

Assessment word2vec similarity candidates

The 0.6 threshold (Rekabsaz et al 2016 and Rekabsaz et al 2017)

The related words		Lexico-semantic relations				Other relations						Similarity according to the w2v model trained on CLEF-IP
word_1	word_2	hypernym	hyponym	synonym	antonym	Spelling variant	word forms	Closeness* (context)	Other relation	Same word	No relation	Cosine
court	court	0	0	0	0	0	0	0	0	1	0	1,0000
march	july	0	0	0	0	0	0	0	1	0	0	0,9124
log	logs	0	0	0	0	0	1	0	0	0	0	0,7181
hard	soft	0	0	0	1	0	0	0	0	0	0	0,7509
feline	cat	1	0	0	0	0	0	0	0	0	0	
close	proximity	0	0	1	0	0	0	0	0	0	0	0,7395
certain	however	0	0	0	0	0	0	0	0	0	1	0,7101
display	lcd	0	0	0	0	0	0	1	0	0	0	0,8296
light	hght	0	0	0	0	1	0	0	0	0	0	0,7021
patient	physician	0	0	0	0	0	0	0	1	0	0	0,7344
perfume	perfiime	0	0	0	0	1	0	0	0	0	0	0,7317
patient	patients	0	0	0	0	1	1	0	0	0	0	0,7596
light	illumination	0	0	1	0	0	0	0	0	0	0	0,7057
chair	furniture	0	1	0	0	0	0	0	0	0	0	

*closeness: occurring in the same text context e.g. part of a multi word term

The month march, or to march

The related words		Lexico-semantic relations				Other relations						Similarity according to the w2v model trained on CLEF-IP
word_1	word_2	hypernym	hyponym	synonym	antonym	Spelling variant	Word forms	closeness	Other relation	Same word	No relation	Cosine
march	march	0	0	0	0	0	0	0	0	1	0	1,0000
march	april	0	0	0	0	0	0	0	1	0	0	0,9280
march	august	0	0	0	0	0	0	0	1	0	0	0,9244
march	february	0	0	0	0	0	0	0	1	0	0	0,9240
march	september	0	0	0	0	0	0	0	1	0	0	0,9202
march	october	0	0	0	0	0	0	0	1	0	0	0,9185
march	november	0	0	0	0	0	0	0	1	0	0	0,9176
march	july	0	0	0	0	0	0	0	1	0	0	0,9124
march	january	0	0	0	0	0	0	0	1	0	0	0,9121
march	june	0	0	0	0	0	0	0	1	0	0	0,9087
march	december	0	0	0	0	0	0	0	1	0	0	0,8844
march	filed	0	0	0	0	0	0	0	0	0	1	0,7487
march	entitled	0	0	0	0	0	0	0	0	0	1	0,7477
march	jan	0	0	0	0	0	0	0	1	0	0	0,7356
march	published	0	0	0	0	0	0	0	1	0	1	0,7293
march	feb	0	0	0	0	0	0	0	1	0	0	0,7261
march	dated	0	0	0	0	0	0	0	0	0	1	0,7260

address

Artificial Researcher

The related words		Lexico-semantic relations				Other relations						Similarity according to the w2v model trained on CLEF-IP
word_1	word_2	hypernym	hyponym	synonym	antonym	Spelling variant	Word forms	closeness context	Other relation	Same word	No relation	Cosine
address	address	0	0	0	0	0	0	0	0	0	1	1
address	addresses	0	0	0	0	0	1	0	0	0	0	0,918
address	register	0	0	0	0	0	0	1	0	0	0	0,771
address	registers	0	0	0	0	0	0	1	0	0	0	0,757
address	memory	0	0	0	0	0	0	1	0	0	0	0,755
address	addr	0	0	0	0	0	0	0	0	0	1	0,753
address	accessed	0	0	0	0	0	0	1	0	0	0	0,746
address	addressing	0	0	0	0	0	0	1	0	0	0	0,744
address	byte	0	0	0	0	0	0	1	0	0	0	0,731
address	write	0	0	0	0	0	0	1	0	0	0	0,726
address	accesses	0	0	0	0	0	0	1	0	0	0	0,726
address	logical	0	0	0	0	0	0	1	0	0	0	0,718
address	written	0	0	0	0	0	0	1	0	0	0	0,706
address	fetch	0	0	0	0	0	0	1	0	0	0	0,702
address	bytes	0	0	0	0	0	0	1	0	0	0	0,701
address	destination	1	0	0	0	0	0	0	0	0	0	0,700

More Examples

The related words		Lexico-semantic relations				Other relations						Similarity according to the w2v model trained on CLEF-IP
word_1	word_2	hypernym	hyponym	synonym	antonym	Spelling variant	Word forms	close context	Other relation	Same word	No relation	Cosine
bus	Bus	0	0	0	0	0	0	0	0	1	0	1
bus	Buses	0	0	0	0	0	1	0	0	0	0	0,8456
bus	Busses	0	0	0	0	1	1	0	0	0	0	0,8360
bus	memory	0	0	0	0	0	0	0	1	0	0	0,7023
spring	spring	0	0	0	0	0	0	0	0	1	0	1
spring	springs	0		0	0	0	1	0	0	0	0	0,8654
spring	resilient	0	0	1	0	0	0	0	0	0	0	0,7260
spring	resiliently	0	0	1	0	0	0	0	0	0	0	0,7151
spring	urges	0	0	0	0	0	0	0	0	0	1	0,7119
spring	urging	0	0	0	0	0	0	0	0	0	1	0,7103
table	table	0	0	0	0	0	0	0	0	1	0	1
table	tables	0	0	0	0	0	1	0	0	0	0	0,8684
table	results	0	0	0	0	0	0	0	0	0	1	0,7539
mouse	mouse	0	0	0	0	0	0	0	0	1	0	1
hive	hive	0	0	0	0	0	0	0	0	0	1	1

Artificial Researcher

Our approach

	Patent	MedIR	MathIR	CLEF paper	Brown
Domain Rules	92,702	1,643,254	48,922	3,698	762
Simple Rules	135,550	2,084,529	70,822	5,748	950
No Rules	135,946	2,252,056	73,472	6,164	944

NLP adaptation methods

- No rules (NoRules) was used to modifying the NLP pipeline analyses
- Three rules (SimpleRules) addressing observed errors among sentence fitted the LSP patterns.
- Domain rules, (DomainRules) here we applied the simple rules (2) and the rules.

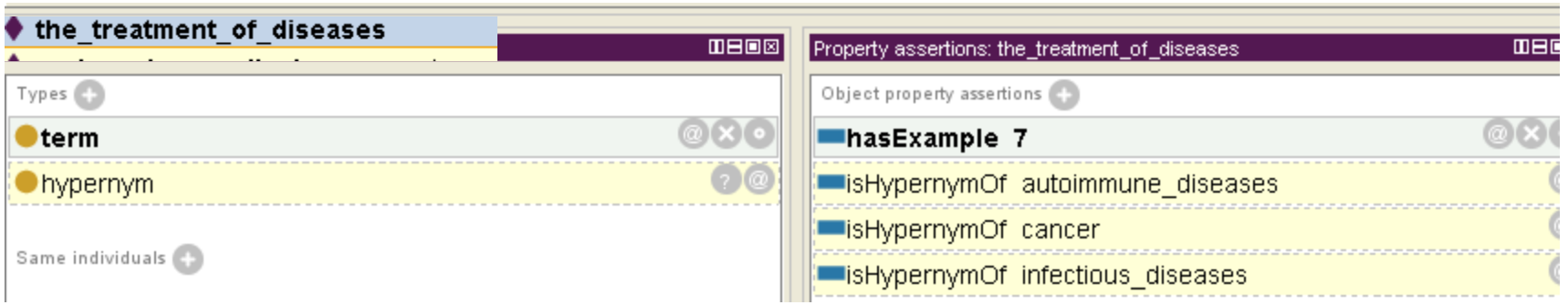
Hyponymy lexical relation extraction using Lexico-syntactic patterns

Example sentences		LSP
1	…work such author as Herrick, Goldsmith, and Shakespeare	such NP as {NP, }* {(or|and)} NP
2	Even then, we would trail behind other European Community member, such as Germany, France and Italy	
3	Bruises, wounds, broken bones or other injuries	NP{, NP}*{,} or other NP
4	Temples, treasuries, and other important civic buildings	NP{, NP}*{,} and other NP
5	All common-law countries, including Canada and England	NP{,} including {NP,}* {or|and} NP
6	… most European countries, especially France, England, and Spain	NP{,} especially {NP,}* {or|and} NP

(Hearst 1992)

PATENT

"The novel conjugate molecules are provided for the manufacture of a medicament for gene therapy , apoptosis , or for the treatment of diseases such as cancer , autoimmune diseases or infectious diseases "



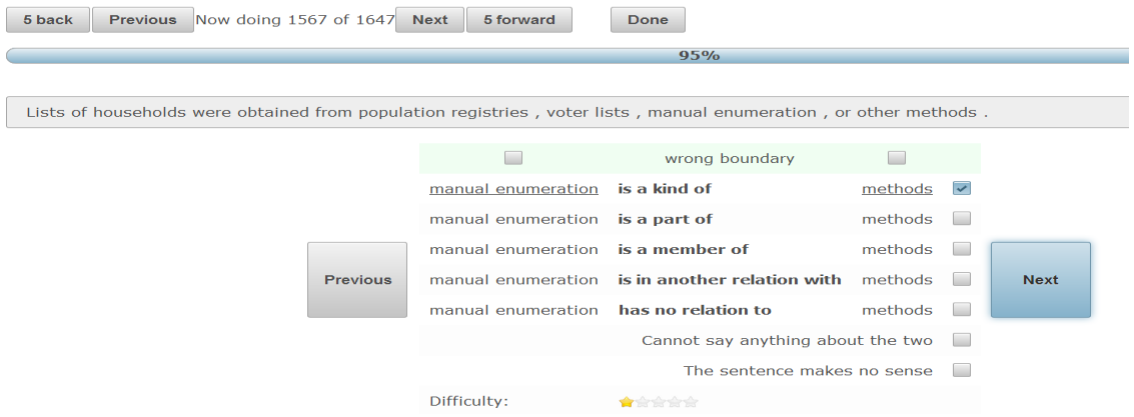
BROWN CORPUS

"Long-lived carbon-14 from the fusion process would cause four million embryonic , neonatal or childhood deaths and stillbirths over the next 20 generations , and between 200,000 and one million human beings now living would have their lives cut short by radiation-produced diseases such as leukemia"

Experiment & Evaluation

- For the evaluation only a smaller set was sampled out (1,647 instances) for manual assessment, approximately 100 instances per data collection and method.
  - one instance correspond to one relation extracted from a sentences

- Three groups: linguist, expert and non-expert.



Inter-annotator agreement:
different assessment groups

- The inter-annotation agreement for identifying relations range between 81% and 88%
- The inter-annotation agreement decreases for wrong NP boundary identifications

	MathIR		Brown	CLEFpaper
	Linguist vs Expert	Linguist vs None Linguist	Linguist vs None Linguist	Linguist+Domain knowlege vs Expert
Relations	85%	81%	83%	88%
No relation	68%	72%	72%	75%
Cannot tell	86%	77%	83%	89%
Makes no sense	90%	89%	80%	93%
hypernymBoundaryWrong	64%	67%	83%	67%
hyponymBoundaryWrong	62%	67%	85%	82%

Artificial Researcher

Results: All made from each sample

Number of positive extraction in relation to all extraction made for each sample and method

Group: Linguist	DomainRules	NoRules	SimpleRules
Brown	39%	40%	40%
MedIR	52%	33%	54%
MathIR	44%	66%	33%
CLEFpaper	50%	47%	56%
Patent	64%	71%	81%

The table displays the percentage of all examined sentences matching the LSP patterns where a positive and correct extraction was identified. For three out of five data set the method SimpleRules was preferred.

Combining NLP & Distributional Semantics

Artificial Researcher

Embedding identifies similarities between different words

- Underwear **similar to** underpants , undergarment, panties, underclothes
- Strength **similar to** strengths, strength, toughness, stronger, sfrength

(Rekabsaz et al 2016 and Rekabsaz et al 2017)

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

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

- 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*

Joined Similarity

- Does “Network lan” and “communication link” have (hyponymy) relation? Yes

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

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

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

Automatic Query Formulation and Expansion

Artificial Researcher

(12) **United States Design Patent** (10) **Patent No.: US D500,000 S**
Dyson et al. (45) **Date of Patent: ** Dec. 21, 2004**

(54) **AUTOMOBILE BODY**

(75) Inventors: **Andrew P Dyson**, West Bloomfield, MI (US); **Joseph S Dehner**, Bloomfield, MI (US); **David C McKinnon**, Bloomfield, MI (US); **Glenn W Abbott**, West Bloomfield, MI (US)

(73) Assignee: **DaimlerChrysler Corporation**, Auburn Hills, MI (US)

(**) Term: **14 Years**

(21) Appl. No.: **29/201,094**

(22) Filed: **Mar. 10, 2004**

(51) **LOC (7) Cl.** **12-08**
(52) **U.S. Cl.** **D12/92**
(58) **Field of Search** D12/90-92, 86; D21/424, 433; 296/185

(56) **References Cited**
U.S. PATENT DOCUMENTS

D408,328 S * 4/1999 Ayoub et al. D12/92
D463,436 S * 11/2002 Dehner et al. D12/92
D478,601 S 7/2003 Stoddard et al. D12/92
D477,253 S * 7/2003 Minami et al. D12/92
D483,696 S * 12/2003 Howell et al. D12/92

* cited by examiner

Primary Examiner—Melody N. Brown
(74) *Attorney, Agent, or Firm*—Ralph E. Smith

(57) **CLAIM**

The ornamental design for an automobile body, as shown and described.

DESCRIPTION

FIG. **1** is a front perspective view of an automobile body showing our new design;
FIG. **2** is a side view thereof;
FIG. **3** is a rear perspective view thereof;
FIG. **4** is a front view thereof;
FIG. **5** is a rear view thereof;
FIG. **6** is a front perspective view of an automobile body showing a second embodiment of our new design;
FIG. **7** is a side view of FIG. **6**;
FIG. **8** is a rear perspective view of FIG. **6**;
FIG. **9** is a front view of FIG. **6**; and,
FIG. **10** is a rear view of FIG. **6**.

It will be understood that the dashed lines presented in the drawings are for illustration only, and do not form a part of the claimed design.

1 Claim, 8 Drawing Sheets

Example of automatic query generation

<QUERY>
(conured OR clutch OR connectability OR nmofs OR fclp OR dnsr 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 "dri
OR "clutch connection command" OR "rear whee
path" OR "output rotational speed" OR "input ro
"detected parameter" OR "generation load torqu
"determination occurrence" OR "four-wheel driv
proceed" OR "wheel speed sensor" OR "output s
"backlash elimination" OR "drive mode switch"
range" OR "transition time" OR "wheel speed"
connection" OR "motor torque" OR "generator l
"high rate" OR "electric motor" OR "throttle ope
force" OR "connected state" OR "previous equat
"prescribed rotational speed difference" OR "12-
"disconnected state" OR "electric clutch" OR "fo
</QUERY>

Automatic query expansion terms

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

accelerator pedal
case pedal device
pedal device

Claims (1)

1. The ornamental design for an automobile body, as shown and described.

Domain Knowledge makes AI *smart*

Artificial Researcher

User request
Patent
Document

Normalization Component

Query Generation

Domain NLP Pipeline

Learning
Termhoodness

Query Expansion

Document index

Passage
Index

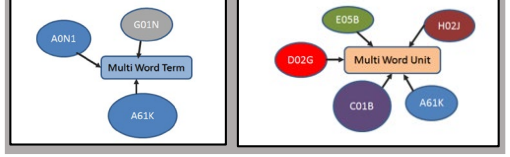
Post Rank
MERGED

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

*Distributional (DS)
models word2vec

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

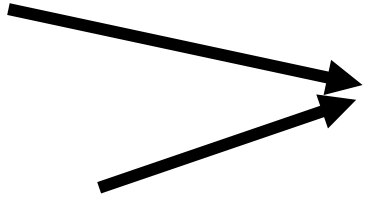
IPC-distributional-values



Hyponymy

Experiment

- CLEF-IP 2013 Collection (~3 M)
  - English topics (50)
  - Patent document represent a 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)
```

- Solr Lucene (4.7.2)
  - Select handler
- Query Generation
  - Query length
  - NLP and Statistical
  - Text section (claims or entire document)
  - Four Technical terms filters
  - Query expansion using NLP and Word Embedding
- Baseline, log(tf)*IDF (Cetintas et al 2012)

Patent Passage Retrieval CLEF-IP 2013
(1000 Passages per topic, max 100 doc)

	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

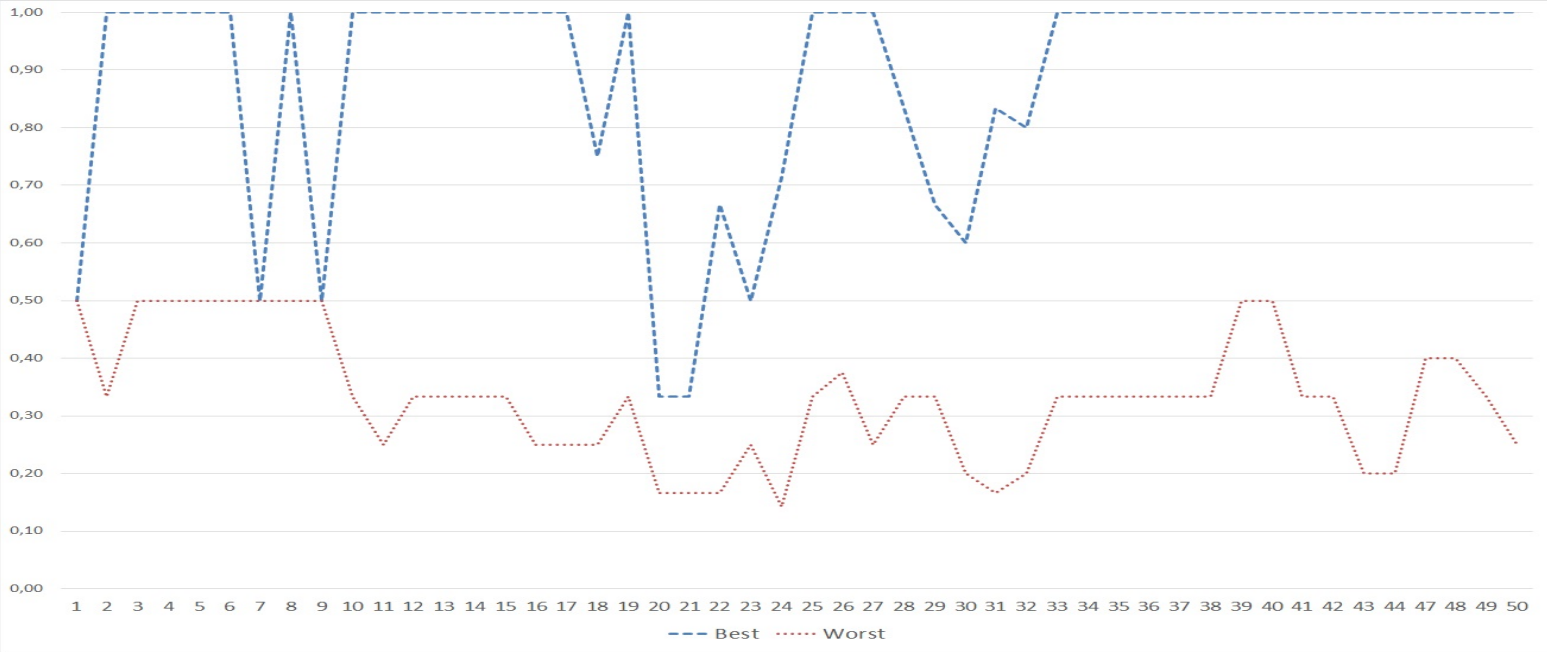
Patent Passage (Paragraph) Retrieval CLEF-IP 2013
Results with Automatic Query Expansion

Artificial Researcher

Main method	AQE	PRES	Recall	MAP	MAP (P)	Precision (P)
NLP1	Hyper_Sem5	0,563	0,653	0,271	0,106	0,207
NLP1	HyperHypo_Sem5	0,558	0,649	0,269	0,104	0,204
NLP1	Hyper_Sem15	0,554	0,634	0,273	0,109	0,205
NLP1	HyperHypo_Sem15	0,549	0,633	0,270	0,102	0,202
NLP1	Hyper_Sem10	0,548	0,628	0,266	0,105	0,203
NLP1		0,544	0,631	0,285	0,112	0,218
NLP1	Seed Ontology	0,486	0,564	0,266	0,098	0,194

- Hyper_SemN: Expansion with only hypernym relations, the top N(5,10,15) most similar words
- Hypo_SemN: Expansion with only hyponym relations, the top N(5,10,15) most similar words
- HyperHypo_SemN: Expansion with hyponym and hypernym relations, the top N(5,10,15) most similar words

What can we learn in terms of Recall?



- For 26 topics we achieve a recall of 1 with at least for one of the QF methods presented in this experiment
- *However there are significant limitation in doing ML since 50 topics are extracted form 37 patent documents*

Artificial Researcher

Conclusion

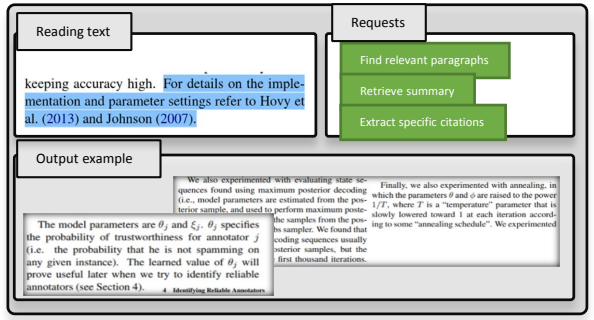
- *A successful text mining solution does not only focus on developing the technology or the best deep learning algorithm, it is much more complex.*
  - *My PhD research shows it is as important to know how to customise the text mining solution to the language, the domain, and the users need*

- **Language Complexity**
  - Word formation of new words are particular important for the patent text genre.
- **Domain Complexity**
  - Multi word terms
- **Task Complexity**
  - Information need, retrieve relevant paragraphs and not just documents

Two up coming projects

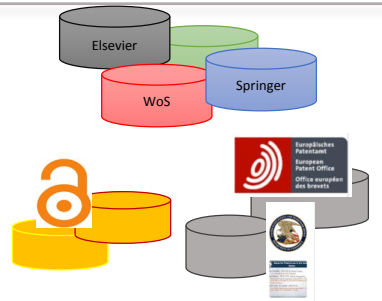
Artificial Researcher

Artificial Researcher in Science



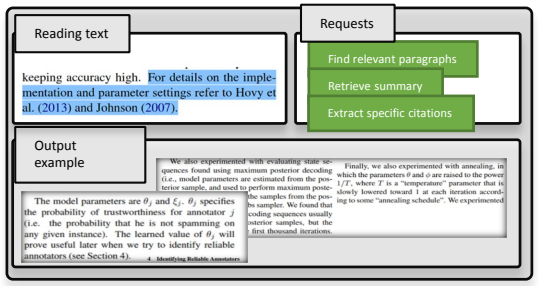
Artificial Researcher

TU WIEN Bibliothek
Your knowledge hub



vienna business agency
A service offered by the City of Vienna

Artificial Researcher in Open Access



Artificial Researcher

Bundesministerium Verkehr, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

austria wirtschafts service aws

Cross Genre retrieval

Domain knowledge is the key

Artificial Researcher

User request

Query Generation & Expansion

Domain NLP Pipeline

Learning Termhoodness

Document index

Passage Index

Post Rank MERGED

GUI

API

Active feedback from users as eLearning platform

Artificial Researcher

Questions?

Please assess the sample examples and send your assessments to andersson@ifs.tuwien.ac.at, we will put up the statistics on our web page http://www.ifs.tuwien.ac.at/patentsemtech/

Go online and give feedback with this survey

https://forms.gle/PiYYYgx27E8mGGzf6

Artificial Researcher

References

- Grefenstette G. and Tapanainen P. (1994) What is a word, what is a sentence?:problems of Tokenisation. Rank Xerox Research Centre,
- Firth J.R. .(1957) A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis, pages 1–32.
- Rekabsaz, N., Lupu, M., Hanbury, A., and Zuccon, G. (2016). Generalizing translation models in the probabilistic relevance framework. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, pages 711–720.
- Rekabsaz N., Lupu M., Baklanov A., Hanbury A., Dur A, and Andersson L. (2017). Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In Proc. of ACL. Association for Computational Linguistics, Stroudsburg, PA, USA.
- Löbner, S. (2002). Understanding Semantics. London
- Frantzi K., Ananiadou S., and Mima H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. Internat. Journal on Digital Libraries.
- Cetintas S. and Si L.. (2012) Effective query generation and post processing strategies for prior art patent search. J. AM. Soc Info. Tec..
- Hearst A. M. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics - Volume 2 (COLING '92), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 539-545. DOI: https://doi.org/10.3115/992133.992154
- Marcus M. P, Santorini B., and Marcinkiewicz M. A.(1993) Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19(2):313–330.