

Usage of Controlled Vocabularies in Industry

Artem Revenko

Director Research @ Semantic Web Company

SD LLOD 2022, Cercedilla

Terminology for this talk

[Controlled Vocabulary](#) mandate the use of predefined, authorised terms.

[Terminology](#) is a group of specialized terms and respective meanings.

[Taxonomy](#) is a scheme of (hierarchical) classification, in which things are organized into groups or types.

[Thesauri](#) consist of 1) a list of terms, 2) the relationship amongst the terms, 3) lexical variants of terms (synonyms, etc.).

[Simple Knowledge Organization System \(SKOS\)](#) is a W3C recommendation for representation of thesauri, classification schemes, taxonomies, or any other type of controlled vocabulary. Terms are called **Concepts** in SKOS.

Get2Know SWC

Semantic Web Company (SWC) and PoolParty



SWC is developer / vendor of
PoolParty Semantic Suite

Most complete and secure
**Semantic middleware /
Low-code AI platform** on
the global market

W3C standards compliant



ISO 27001:2013
certified

First release in 2009

Current version **8.2**

On-premises or
cloud-based



Over **150** customers
world-wide



Semantic AI:

Fusion of graphs,
NLP, and machine
learning



Named as Visionary
in **Gartner's Magic
Quadrant** for Metadata
Management Systems
2019 and 2020



KMWorld listed
PoolParty as
**Trend-Setting
Product** 2015 - 2021
and in the **AI 50** list of
companies in 2020

PoolParty Platform—components and features



TEXT MINING & NATURAL LANGUAGE PROCESING

GRAPH AUTOMATION

SEMANTIC AI APPLICATIONS

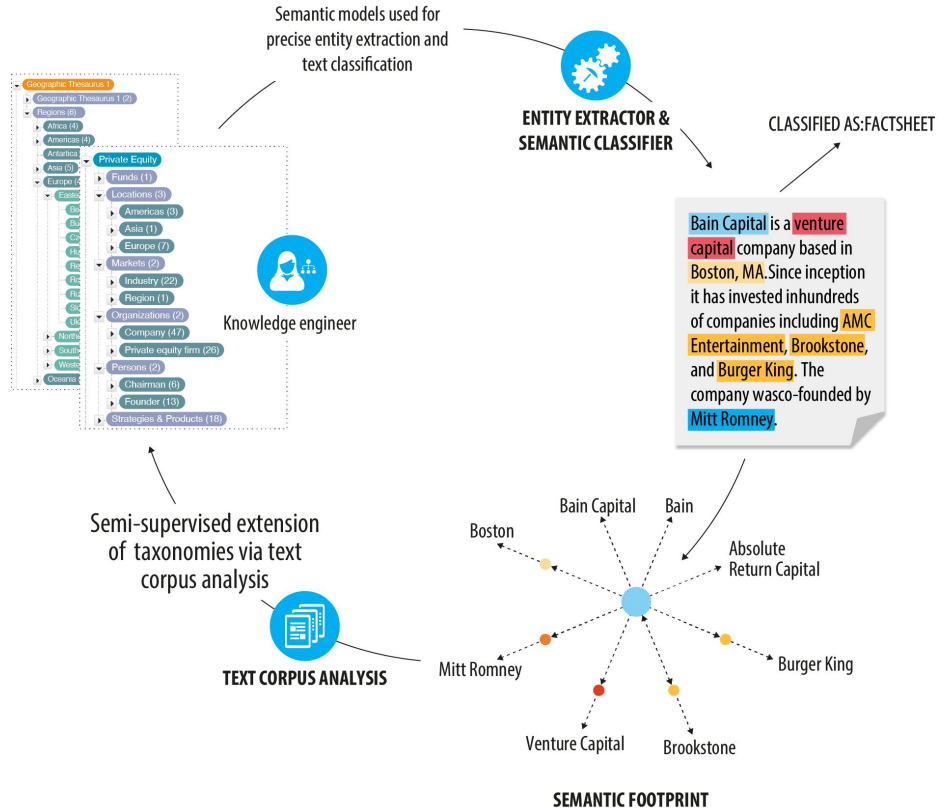
GRAPH MANAGEMENT



LOD Source: DeDBPedia

② The pyrolysis (or devolatilization) process is the thermal decomposition of materials at elevated temperatures in an inert atmosphere. It involves a change of chemical composition. Pyrolysis is most commonly used in the treatment of organic materials. It is one of the processes involved in charring wood. In general, pyrolysis of organic substances produces volatile products and leaves char, a carbon-rich, solid residue. Extreme pyrolysis, which leaves mostly carbon as the residue, is called carbonization.

Learning from Text



Customer Use Cases

Types of Use Cases

1. Controlled Vocabulary (CV) as a Service
 - a. KG as a Service
2. Enterprise Search / Information Retrieval
 - a. Data Integration: interlink structured and unstructured data
3. Similarity
4. Matching
5. Question Answering
6. Information Extraction

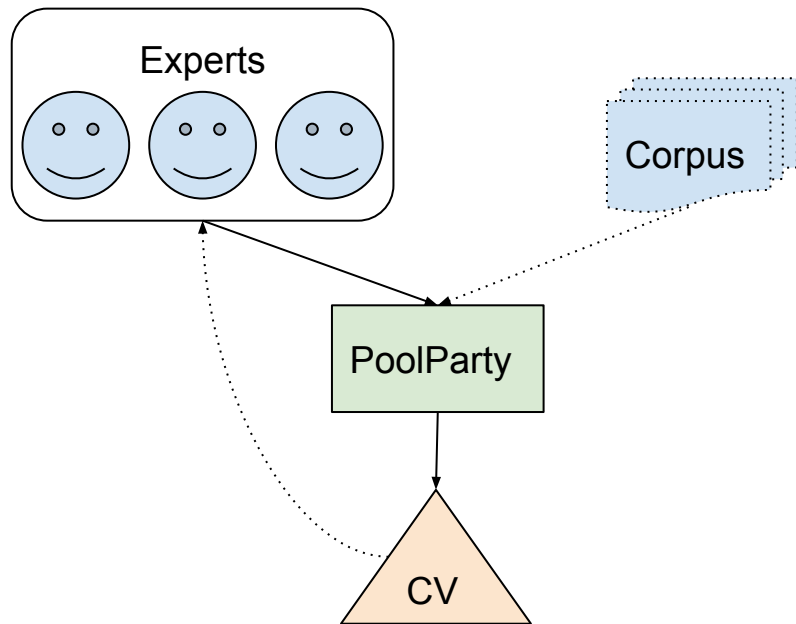
CV as a Service: Task

Given:

1. Experts
2. *Optionally*: Corpus

Expected Output:

1. Controlled Vocabulary
2. Agreement on CV by experts
3. *Optionally*: Process to update CV



CV as a Service: Usage Scenarios

Goals:

1. Make CV available
 - a. Via APIs
 - b. Via FEs
2. Manual tagging, autocompletion
3. Establish agreement on terms
 - a. Provide aspect-specific variants (language, domains, etc.)
4. KG as the final source of truth
 - a. Support corporate website
5. Versioning
6. Data transformation

Retail Industry:

Our customer is one of the world's largest furniture companies

CHALLENGE

- Inconsistencies and duplication of effort between regional units
- Product recommendations are too broad, include products irrelevant in a context
- Much domain knowledge is unstructured and implicit and difficult to access

SOLUTION

- Dynamically generated topic pages showing products and related information in their context
- Creation of a product knowledge graph
- Use of multilingual controlled vocabularies

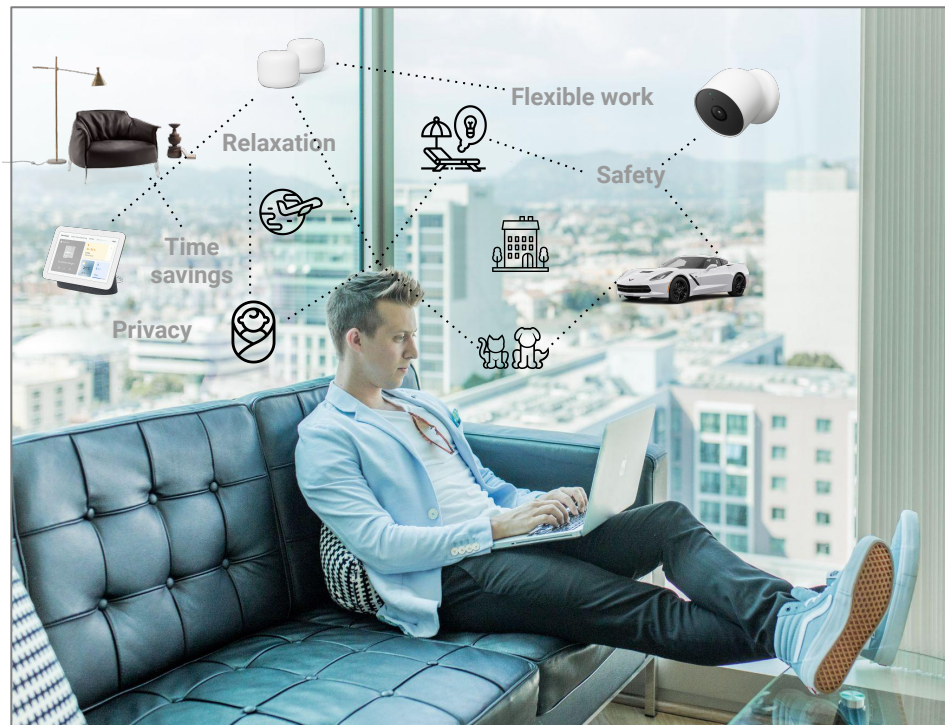
BENEFIT

- ✓ Better customer experience on the website
- ✓ More appropriate informational material and better targeted selling opportunities
- ✓ Stronger customer loyalty, building extended knowledge around the product and developing usage contexts

Knowledge graphs have been around for a long time

Knowledge graphs have long existed in every company, but they are in our heads, are largely analog and rarely already digitized.

Only by digitizing them, by making them explicit, can they be connected to the digital world and scale.



Inspired by **Adam Keresztes** (Product Owner, IKEA Knowledge Graph)

CV as a Service: Research

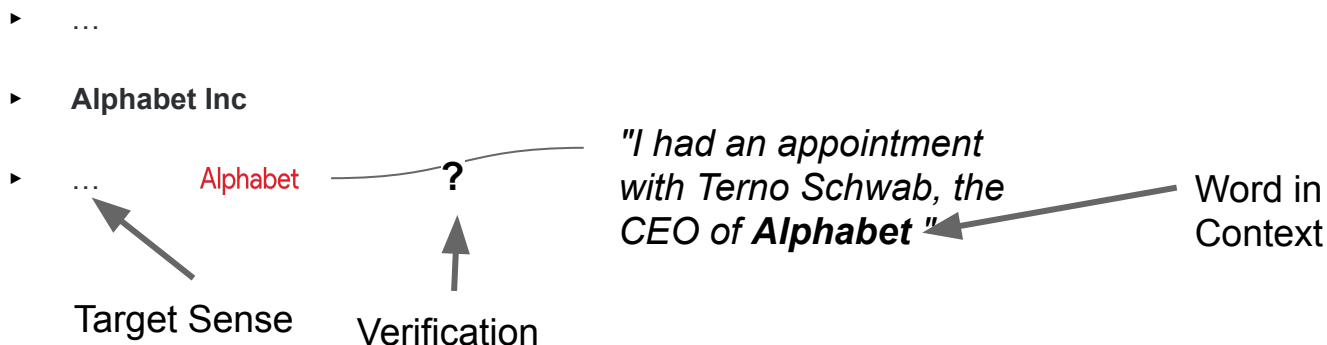
Related research tasks:

1. CV Induction (from Text)
2. CV Alignment
 - a. to open LOD
 - b. to other external vocabs
 - c. between different corporate CVs
3. Word Sense Disambiguation (WSD) for Machine Translation
4. Crowd-sourcing

Research: WSD

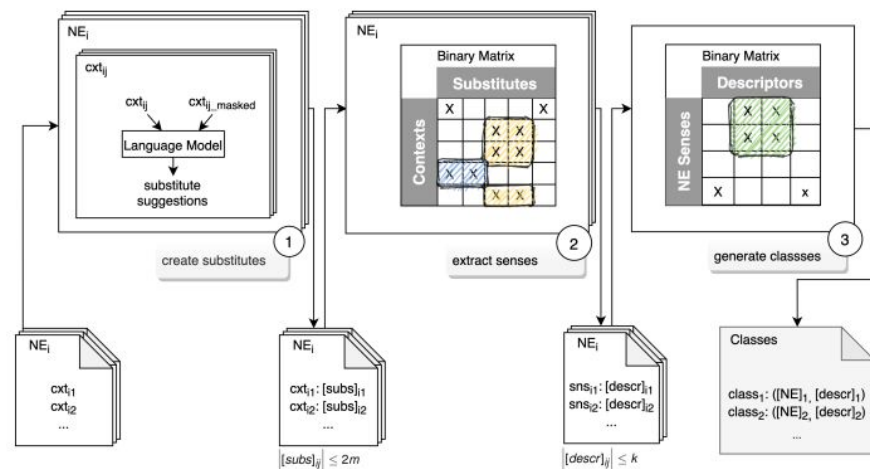
New task formulation: Target Sense Verification

- Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. [WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context](#). In *EACL 2021*
- Blogs
 - a. [Is Word Sense Disambiguation outdated?](#)
 - b. [Label unstructured data using Enterprise Knowledge Graphs 3](#)
- Repository: <https://github.com/semantic-web-company/wic-tsv>



Research: CV Induction

- Word Embeddings for synonym and hypernym discovery
 - OntoLex-FrAC module:
<https://acoli-repo.github.io/ontolex-frac/>
- Saffron: <https://github.com/insight-centre/saffron>
- Induction of senses and classes:
 - Bourgonje, P., Breit, A., Khvalchik, M., Mireles, V., Schneider, J. M., Revenko, A., & Rehm, G. (2020, January). Automatic induction of named entity classes from legal text corpora. In *ASLD@ ISWC*.
 - Revenko, A. and Mireles, V., 2019, August. The Use of Class Assertions and Hypernyms to Induce and Disambiguate Word Senses. In *International Conference on Database and Expert Systems Applications* (pp. 172-181). Springer, Cham.



Research: Crowd-sourcing

- Promoting Financial Awareness and Stability (PROFIT): <https://cordis.europa.eu/project/id/687895>
 - **The PROFIT Platform is designed as a solution catering to the exact need of action enhancement for greater financial awareness and capabilities of EU citizen.**
 - Revenko, A., Sabou, M., Ahmeti, A., & Schauer, M. (2018). Crowd-Sourced Knowledge Graph Extension: A Belief Revision Based Approach. HCOMP.

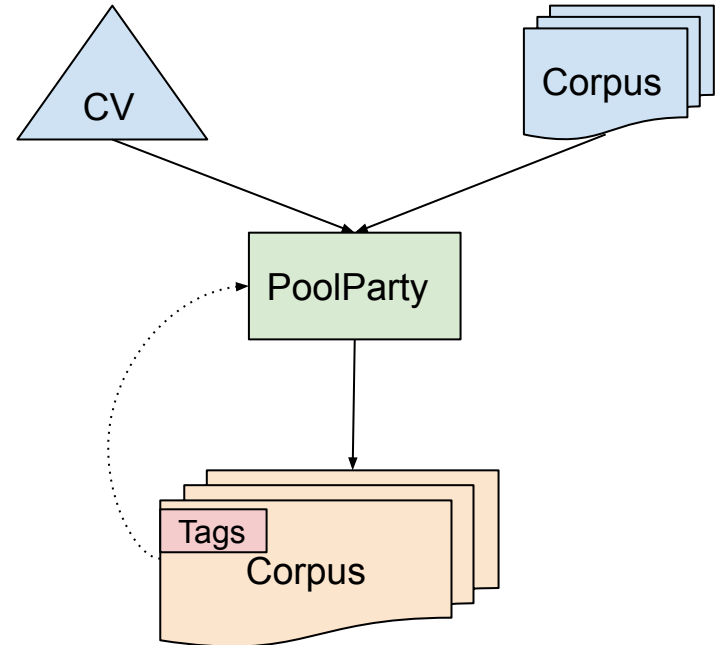
Enterprise search: Task

Given:

1. Corpus
2. Controlled Vocabulary

Expected Output:

1. Tagged Corpus
2. *Optionally*: Re-tagging procedure when Corpus / CV was updated



Enterprise search: Usage Scenarios

Goals:

1. Concept-based search
2. Full-text search with (linguistic) variants
 - a. Inflections
 - b. Synonyms
 - c. Multilingual Search
3. Graph-based navigation
4. Clustering / topic modelling

Industrial Automation:

Our client is a leading provider of **industrial automation** and **test solutions**.

CHALLENGE

- Manage shift handover documentation.
- Shift handover in the heavy industry is a very hazardous activity that can lead to accidents.
- The documentation written for the shift handover is unstructured and usually produced but never consumed.

SOLUTION

- A semantic search and recommender system over the shift handover documentation.
- Create a taxonomy and ontology to cover the topic of interest, reusing international standards for modelling the knowledge graph.

BENEFIT

- ✓ Records are easy to find, analyze and maintain.
- ✓ Based on analytics, it is possible to create measures to improve shift handover, a potentially hazardous activity.

Analysis of shift handover reports

Dashboard

Latest shifts with detected safety hazards discovered

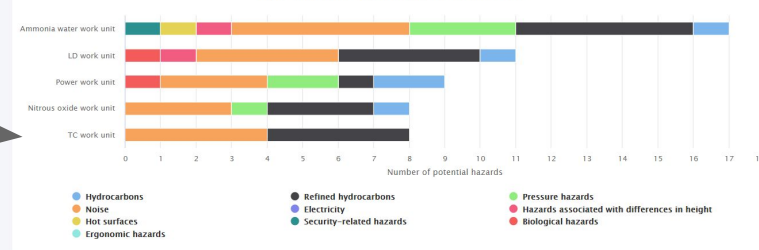
Top 5 work units in terms of number of hazards

Latest safety hazards discovered

| Work center | Unit | Start time | End time | Severity | Operations |
|--------------------------|---------------------------|---------------------|---------------------|----------|------------------------------|
| Main product work center | Hydrogen work unit | 2020-07-30 19:00:00 | 2020-07-31 06:59:00 | High | Show details |
| Utility work center | The 1st Power work unit | 2020-07-30 07:00:00 | 2020-07-30 18:59:00 | Low | Show details |
| Main product work center | Carbon monoxide work unit | 2020-07-29 19:00:00 | 2020-07-30 06:59:00 | Low | Show details |
| Utility work center | The 1st Power work unit | 2020-07-29 07:00:00 | 2020-07-29 18:59:00 | Low | Show details |
| Main product work center | Nitrous oxide work unit | 2020-07-28 19:00:00 | 2020-07-29 06:59:00 | Medium | Show details |

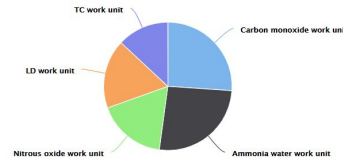
Click here to show the detail screen for the shift report

Top 5 work units with most potential hazards



Highest hazard severity associated to that shift

Top 5 work units with most potential equipment failures



Top 5 work units in terms of number of equipment failures

Enterprise search: Research

Related research tasks:

1. Query expansion
2. Tagging
3. Metadata quality
4. Classification & Clustering
5. WSD
6. Named Entity Recognition (NER)

Research: Re-tagging

- Sometimes documents and thesauri change quickly. How can we re-tag at least daily?
 - Use Case: <https://www.coronawhy.org/literature-review-demo> “over 1,000,000 scholarly articles”
 - Algorithms: [Aho-Corasick](#), [Finite-state transducers](#)
- Generating inflections
 - Thierry Declerck, Melanie Siegel, and Stefania Racioppa. 2019. [Using OntoLex-Lemon for Representing and Interlinking German Multiword Expressions in OdeNet and MMORPH](#). In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 22–29, Florence, Italy. Association for Computational Linguistics.

Research: MD quality

- ADEQUATE project
 - Knap, T. (2017, October). Towards Odalic, a Semantic Table Interpretation Tool in the ADEQUATE Project. In LD4IE@ ISWC (pp. 26-37).
 - Neumaier, Sebastian, Lörinc Thurnay, Thomas J. Lampoltshammer and Tomás Knap. "Search, Filter, Fork, and Link Open Data: The **ADEQUATE platform**: data- and community-driven quality improvements." Companion Proceedings of the The Web Conference 2018 (2018).
- Realising the Social Sciences and Humanities for the European Open Science Cloud (SSHOC) project:
<https://sshopencloud.eu/>



Research: Classification & Clustering

- Mireles, V., & Revenko, A. (2017). Evolution of Semantically Identified Topics. In HybridSemStats@ ISWC.
 - The dataset we analyze is a financial news data set. The news come from a single source (Bloomberg news) and are made available. From the original dataset, we took articles between January 2009 and November 2013, which total 447,145 documents.
 - SKOS concepts enable better detection and smooth transitions thanks to synonyms and hierarchies

| Football 6 Weeks | Korea 5 Weeks | Drought 5 Weeks | Arab Spring 8 Weeks | Fukushima 6 Weeks |
|---------------------|------------------|--------------------|------------------------|----------------------|
| World | Koreans | Wheat | Libya | Plants |
| Sport event | Korean | Crops | International | Nuclear energy |
| South Africa | South Korea | Drought | Nation | Manufacturing plant |
| Football | North Korea | Soybean | Foreign | Electricity |
| African | South Korean | World | Arabs | Cooling |
| Matching | South Koreans | Rice | Arab | Earthquake |
| Coaching | Officials | Food price | Egypt | Greenhouse gas emis. |
| South African | Nation | Nation | Air | Nuclear power plant |
| South Africans | Island | Egypt | West Asia | Nuclear fuel |
| Brazil | Foreign | Department | Tunisia | Order |
| Argentina | World | Australia | Officials | Process |
| Netherlands | Chinese | International | African | Officials |
| Mexico | Fire | Province | Industrial action | Fire |

Research: NER

Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors (Prêt-à-LLOD):
<https://pret-a-llod.eu/>



Legal Knowledge Graph for Multilingual Compliance Services (Lynx): <https://lynx-project.eu/>



- Thesaurus-based training set generation:
 - Karampatakis, S., Dimitriadis, A., Revenko, A., & Blaschke, C. (2020, May). Training NER Models: Knowledge Graphs in the Loop. In European Semantic Web Conference (pp. 135-139). Springer, Cham.
- Fine-grained NET:
 - Revenko, A., Breit, A., Mireles, V., Moreno-Schneider, J., Sageder, C., & Karampatakis, S. (2021). Annotating Entities with Fine-Grained Types in Austrian Court Decisions. In Further with Knowledge Graphs (pp. 139-153). IOS Press.

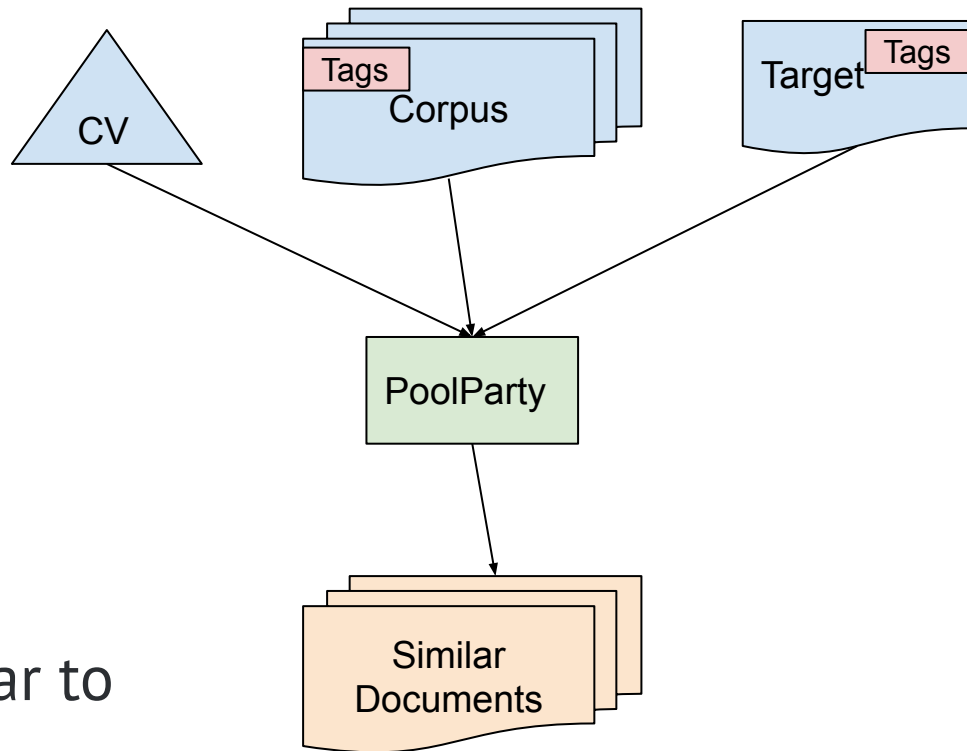
Similarity: Task

Given:

1. Thesaurus / Ontology
2. Annotated Documents / Objects
3. Target Document

Expected Output:

Documents that are similar to the target



Similarity: Usage Scenarios

Goals:

1. Find similar/duplicate documents: preparing a new project
2. Find similar resources: recommend similar clothes
3. Overcome cold-start problem in recommenders
4. Take the background structure (ontology / thesaurus) into account
 - a. Enable user control over the importance of the structure

Pharmaceutical Industry:

Our customer is one of the
top 20 companies in the
pharmaceutical industry.

CHALLENGE

- Required information to successfully complete a drug approval is diverse and scattered across numerous sources.
- The right contacts to successfully carry out a drug approval are often not easy to find, even within the corporation.

SOLUTION

- Graph-based text mining and recommender system embedded in target systems via PowerTagging
- Use of the PoolParty to efficiently maintain comprehensive knowledge graphs for the pharmaceutical industry.

BENEFIT

- ✓ Risks of becoming entangled in inconsistencies during a drug approval process are minimized.
- ✓ The time to successfully complete a drug approval is significantly reduced.

Did Dr Lehner do any drug research on polysorbates like ps 20 or ps 40?

Submit question

Be specific and imagine asking a question to another person.

Matched Tags:

PS 20

PS 40

Polysorbate

Drugs

Lehnert, Dr., Dirk

ps22

Add tag



Similar questions

P.7 Container Closure System:

It should be specified to which monograph do the Rubber Stopper part comply.

It is unclear whether the MCB was tested for bovine and/or porcine viruses by in vitro assay. If no testing was performed the sponsor should discuss this point in respect to bovine serum albumin which was used as supplement of the SCB cryopreservation medium.

Please clarify the source of origin for Enzalutamide Cap 40mg (e.g. commercial product in UK)

Associated Documents

Considerations for the Use of Polysorbates in Biopharmaceuticals

The degradation of polysorbates 20 and 80 and its potential impact on the stability of biotherapeutics

A Rapid High-Sensitivity Reversed-Phase Ultra High Performance Liquid Chromatography Mass Spectrometry Method for Assessing Polysorbate 20 Degradation in Protein Therapeutics

Associated experts



Lehnert, Dirk, Dr.

PS 20 Antibiotics
Machine Learning



Beverly Neal, Dr.

PS 80 Antibiotics Drugs



Keith Ramos, Dr.

DNA Drugs
Small Molecules Bioprinting



Did Dr Lehner do any drug research on polysorbates like ps 20 or ps 40?

Be specific and imagine

Matched Tags: PS

Similar questions

P.7 Container C
It should be spe

It is unclear whe
performed the s
supplement of th

Please clarify th

Associated Docum

Considerations

The degradation

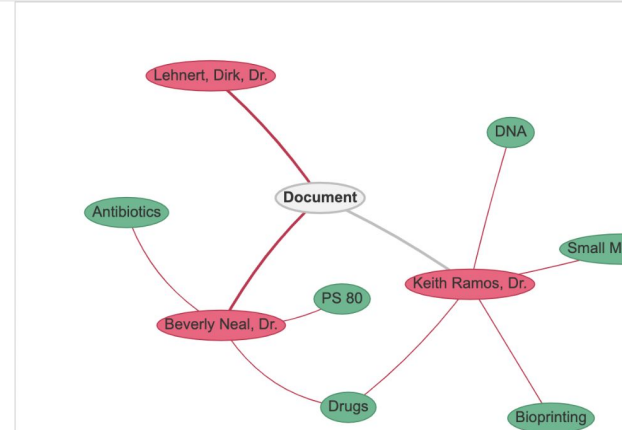
A Rapid High-Se
Method for Assessing Polysorbate 20 Degradation in Protein Therapeutics

A Rapid High-Sensitivity Reversed-Phase Ultra High Performance Liquid Chromatography Mass Spectrometry Method for Assessing Polysorbate 20 Degradation in Protein Therapeutics

Polysorbate 20 (PS20), a widely used surfactant in protein therapeutics, has been reported to undergo hydrolytic degradation during product storage, causing the release of free fatty acids. The accumulation of free fatty acids in protein therapeutics was found to result in the formation of particles due to their limited aqueous solubility at 2°C-8°C. Quantitation of free fatty acids originating from PS20 degradation is thus important during bioprocess optimization and stability testing in formulation development to ensure optimum PS20 stability as well as product and process consistency in final drug products. This work reports the development of a simple and robust, high-throughput, reversed-phase ultra high performance liquid chromatography mass spectrometry method for high-sensitivity quantitation of lauric acid and myristic acid by using isotope-labeled fatty acid internal standards. The high sensitivity (<100 ng/mL for lauric acid) and suitable precision (intermediate precision relative standard deviation of 11%) of this method enable accurate detection of lauric acid produced from the degradation of less than 1% of PS20 in a 0.2-mg/mL formulation. Using accelerated thermal stability testing, this method identifies processes that exhibit fast PS20 degradation within only days and consequently allows faster iterative optimization of the process.

Show Relations

Show full document



Similarity: Research

- Document similarity measures with SKOS concept:
 - Schneider, J.M., Rehm, G., Montiel-Ponsoda, E., Rodríguez-Doncel, V., Martín-Chozas, P., Navas-Loro, M., Kaltenböck, M., Revenko, A., Karampatakis, S., Sageder, C. and Gracia, J., 2022. **Lynx**: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain. Information Systems, 106, p.101966.
 - Usage of concepts from CV improves the robustness of identifying similar / relevant documents in a corpus

Matching: Task

Given:

1. Annotated Objects
2. Thesaurus / Ontology
3. Target Object

Output:

Other objects (usually of a different type) that match to the target

Matching: Usage Scenarios

1. Given a project description find best matching experts
2. Given equipment find standards that apply to it
3. Given employee profiles find best educational courses

Management consultancy:
Our customer is **one of the most prestigious employers in the industry.**

CHALLENGE

- The main challenge that our client faces is to find the right people for the right projects.
- The current approach is comparatively slow, unpredictable, and only takes into account experience from past projects in an unsystematic way.

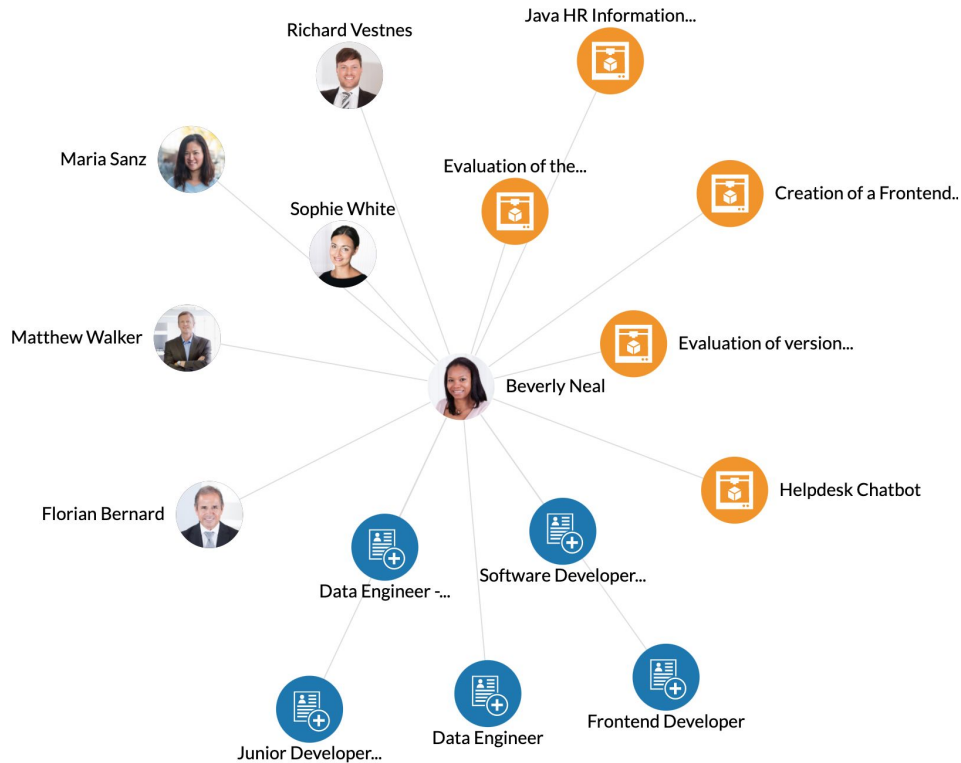
SOLUTION

- Use of PoolParty Taxonomy Management to curate skills taxonomies systematically
- Use of PoolParty Extractor in conjunction with skills taxonomies
- Highly accurate recommender system
- Improved person search system

BENEFIT

- ✓ Increased competitiveness through better reuse of existing knowledge assets
- ✓ Improved analysis capabilities to identify existing gaps in the knowledge base
- ✓ Identification of missing skills in the company: Support for strategic HR planning

Demo: Semantic Search & Matchmaking



The **HR Recommender** is a semantic matchmaking tool based on a knowledge graph. It is designed to connect employees with their coworkers, show them relevant projects, and let them know about interesting career opportunities within their organization.

[HR Recommender](#)

Matching: Approach

We apply similarity measures together with curated rules to find the best matching

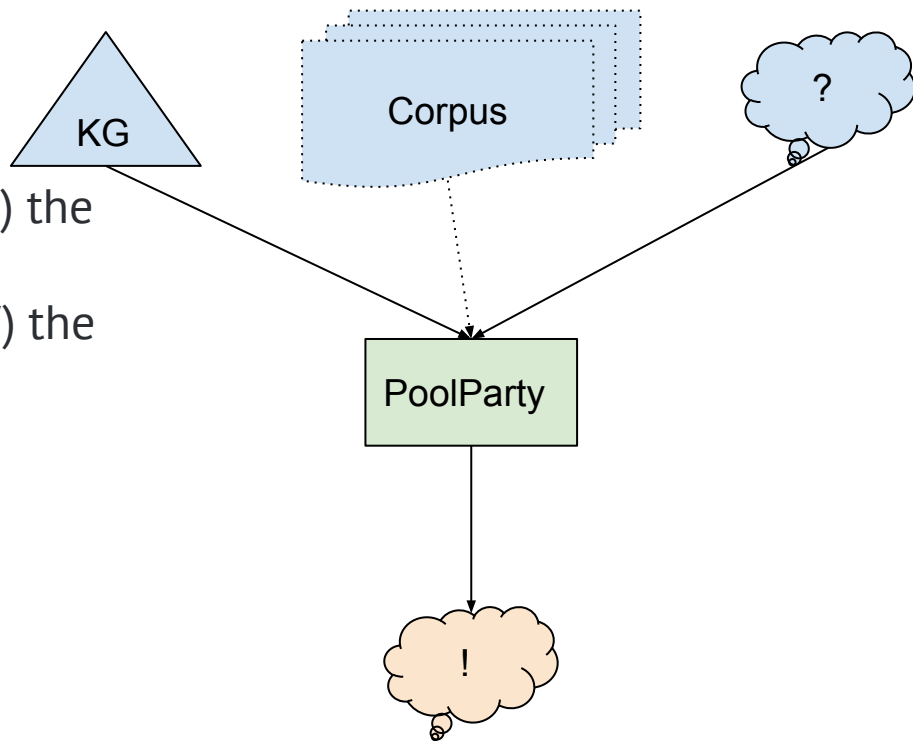
Question Answering: Task

Given:

1. Question in natural language
2. Knowledge Graph containing (part of) the answer
3. *Optionally*: Corpus containing (part of) the answer

Output:

1. Answer in natural language
2. *Optionally*: Supporting sources (KG + documents)



Question Answering: QAnswer

<https://qanswer.poolparty.biz/>

poolparty & QAnswer

who can use sql

Go

How to Use

Contact

82 %

SPARQL LIST

DID YOU MEAN

Is this the right answer? ☐ Yes ☐ No

/ Programming / SQL

/ type / Employee

TABLE

LIST

IMAGES

First

Previous

1

2

3

Next

Last

| Programming | description | Languages | type | age | gross salary | image |
|--|---|-------------------------|--------------------------|-----|--------------|---|
| Anna Sørensen <ul style="list-style-type: none">SPARQLSQL | I graduated from Ghent University with a degree in Internet Technologies. I work for JPMorgan Chase ... | English | Employee | 34 | 120000 | https://image.poolparty.biz/hr/istock/AnnaSørensen.jpg |
| | I studied Media Science at ... | | | | | |

Question Answering: Research

- Polylingual Hybrid Question Answering (PORQUE) project: <https://www.porque-project.eu/>
 - PORQUE will develop and evaluate an extensible platform for polylingual question answering by relying on knowledge graphs and text.
- QA for KG completion:
 - Khvalchik, M., Blaschke, C. and Revenko, A., 2019, August. Question Formulation and Question Answering for Knowledge Graph Completion. In International Conference on Database and Expert Systems Applications (pp. 166-171). Springer, Cham.

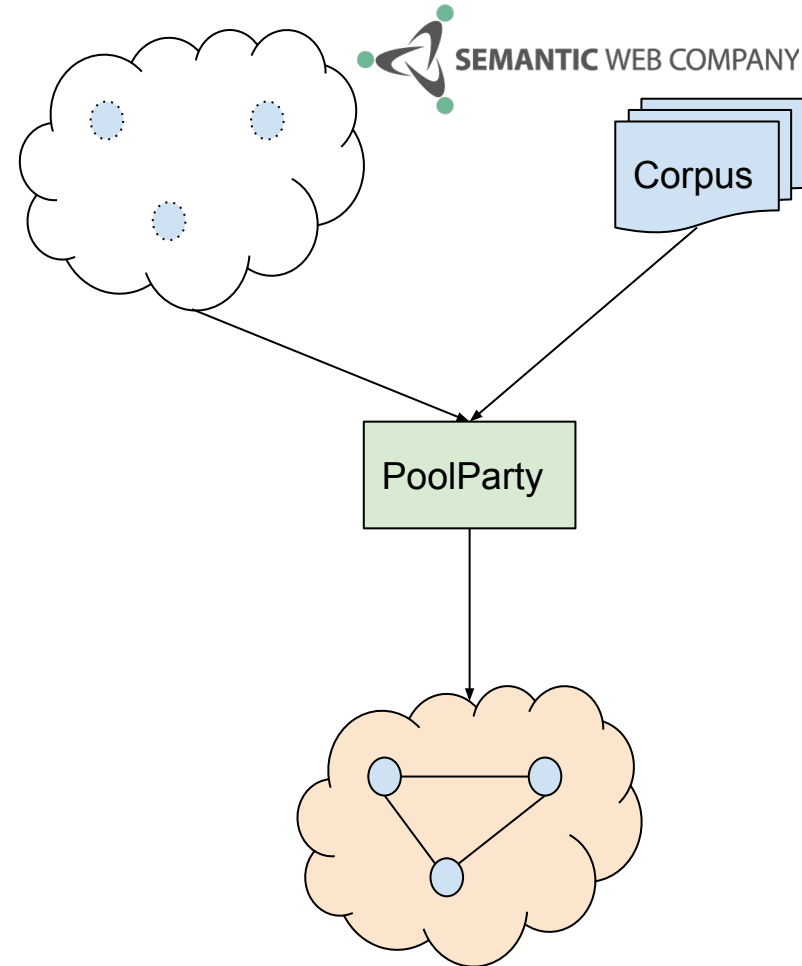
Information Extraction: Task

Given:

1. Ontology (T-Box)
2. Corpus

Output:

1. Extracted structured information (KG) complying with the ontology



Information Extraction: Usage Scenarios

1. Find various clauses and agents in contract
2. Find problems and subjects in support tickets
3. Find legal relations, agents in laws
4. Find parties and dates in agreements

Information Extraction: Contract Intelligence



Filter Documents by Rule

- ☒ Applicable law
- ☒ Confidentiality
- ☒ Insurance
- ☐ Termination
- ☐ Warranty

Applied rules

Rule **CON-1** was applied because concepts **Party**, **not communicate** and **confidential information** were found in the same context.

Matching Documents



Cloud-Service-Agreement-24-7.docx

hide details



Applicable law

Confidentiality

Insurance

Relevant sentence 1 of 18



GDB will not be responsible for any use, disclosure, modification or deletion of Customer Content resulting from any such access by third party program providers or for the interoperability of such third party programs with the Services.

3.4 Except as otherwise expressly set forth in Customer order for certain Cloud Services offerings (e.g., a private cloud hosted at Customer facility), Customer acknowledge that GDB has no delivery obligation for GDB Programs and will not ship copies of such programs to Customer as part of the Services.

3.5 The Services may contain or require the use of Separately Licensed Third Party Technology.

Customer is responsible for complying with the Separate Terms specified by GDB that govern Customer's use of Separately Licensed Third Party Technology.

GDB may provide certain notices to Customer in the Service Specifications, Program Documentation, readme or notice files in connection with such Separately Licensed Third Party Technology.

The third party owner, author or provider of such Separately Licensed Third Party Technology retains all ownership and intellectual property rights in and to such Separately Licensed Third Party Technology.

3.6 As part of certain Cloud Services offerings, GDB may provide Customer with access to Third Party Content within the Services Environment.



Cook Island - Conditions-for-a-Consultancy-Agreement.docx

show details



Information Extraction: Research

- Ontology-Based Artificial Intelligence in Environmental Sector (OBARIS) project:
<https://www.obaris.org/>
 - The FFG-funded OBARIS project (Ontology-Based Artificial Intelligence in Environmental Sector) aims to advance the status quo in the area of auditable semantic artificial intelligence systems (SAISs), by investigating both conceptual aspects of these systems as well as a technology stack that facilitates transposing these system types into concrete settings.
- Martín-Chozas, P. and Revenko, A., Thesaurus Enhanced Extraction of Hohfeld's Relations from Spanish Labour Law, DeepOntoNLP@ESWC 2021



Semantic Web Machine Learning Systems

Systematic Literature Review

Combining Machine Learning and Semantic Web - A Systematic Mapping Study

Anna Breit ^{a,*}, Laura Waltersdorfer ^b, Fajar J. Ekaputra ^b, Marta Sabou ^b, Artem Revenko ^a,
Andreas Ekelhart ^c, Jan Portisch ^d, Andreea Iana ^d, Heiko Paulheim ^d, Frank van Harmelen ^e and
Annette ten Teije ^e

The FFG-funded OBARIS project (Ontology-Based Artificial Intelligence in Environmental Sector) aims to advance the status quo in the area of auditable semantic artificial intelligence systems.

What is the Literature Review about?

What type of content?

- ▶ Semantic Web Machine Learning Systems (SWeMLS)
 - ▷ Usage of Semantic Web resource(s)
 - ▷ Usage of ML model(s)
 - ▷ Solving a specific task by combining SW and ML

What type of papers?

- ▶ From 2010 - 2020
- ▶ Peer-reviewed

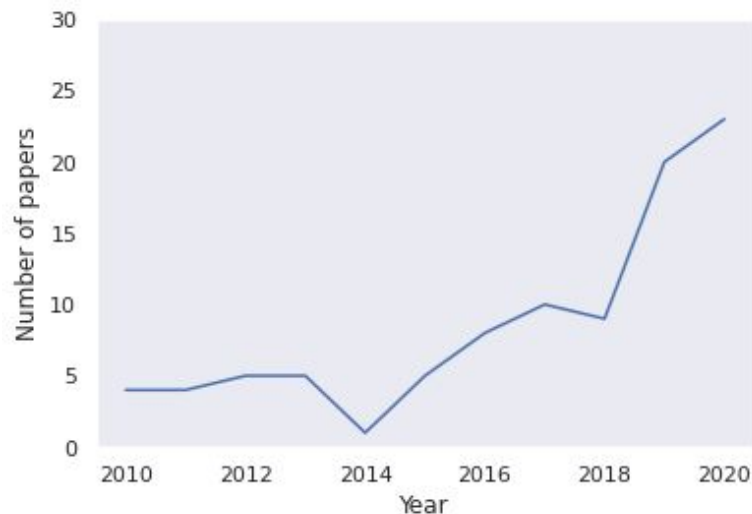
Initial search results: 2707

After strict filtering: 477

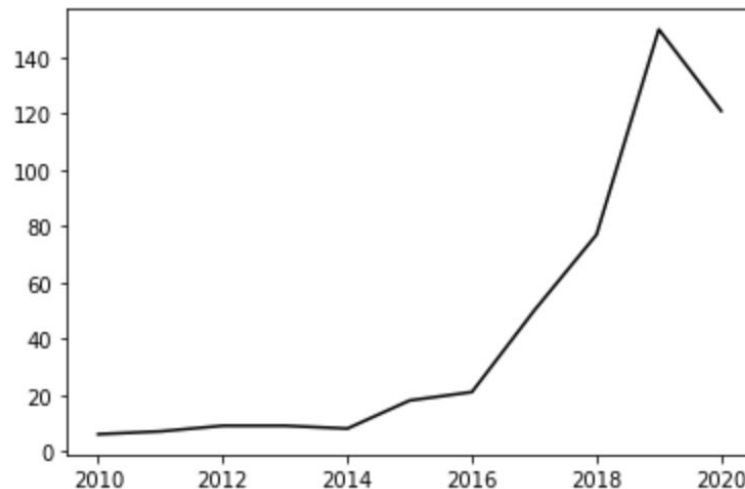
Filter: papers that use either labels (with their synonyms) or hierarchical relations.

Number of papers: Overall 94

Our Filter



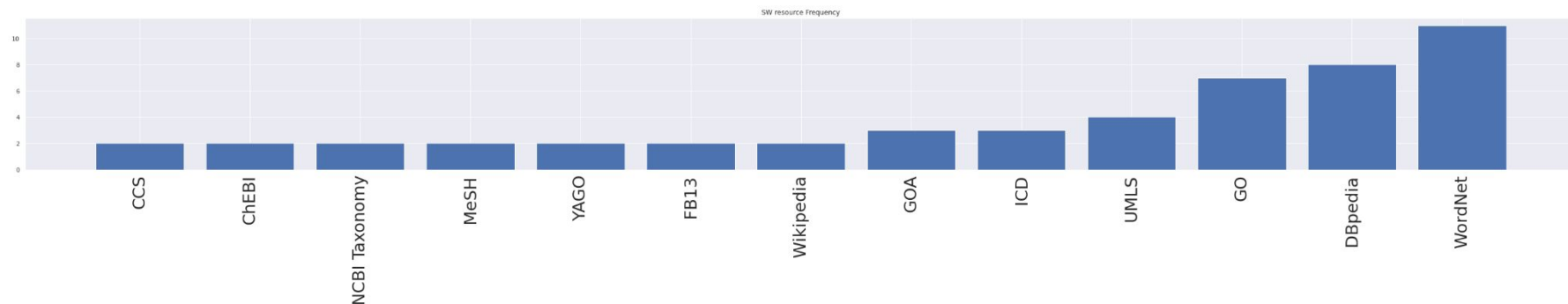
All SWeMLS papers



Semantic Web Resources

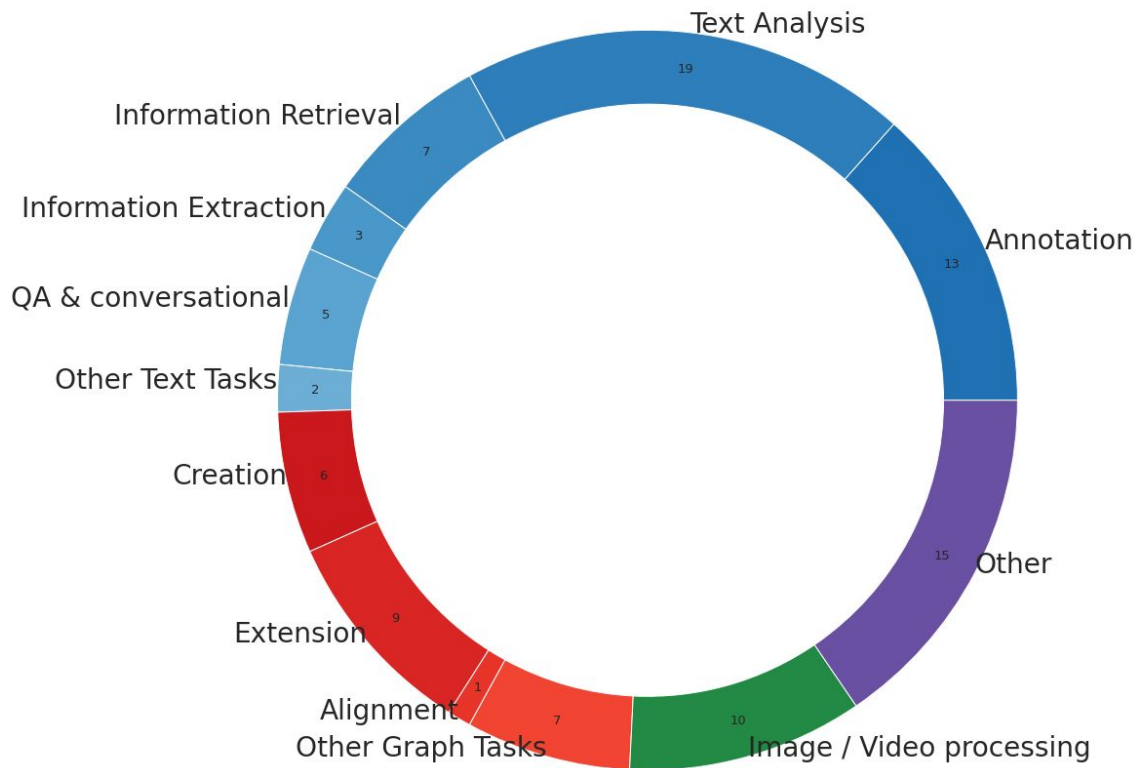
WordNet & DBpedia are most popular resources

Gene Ontology is most popular biological resource



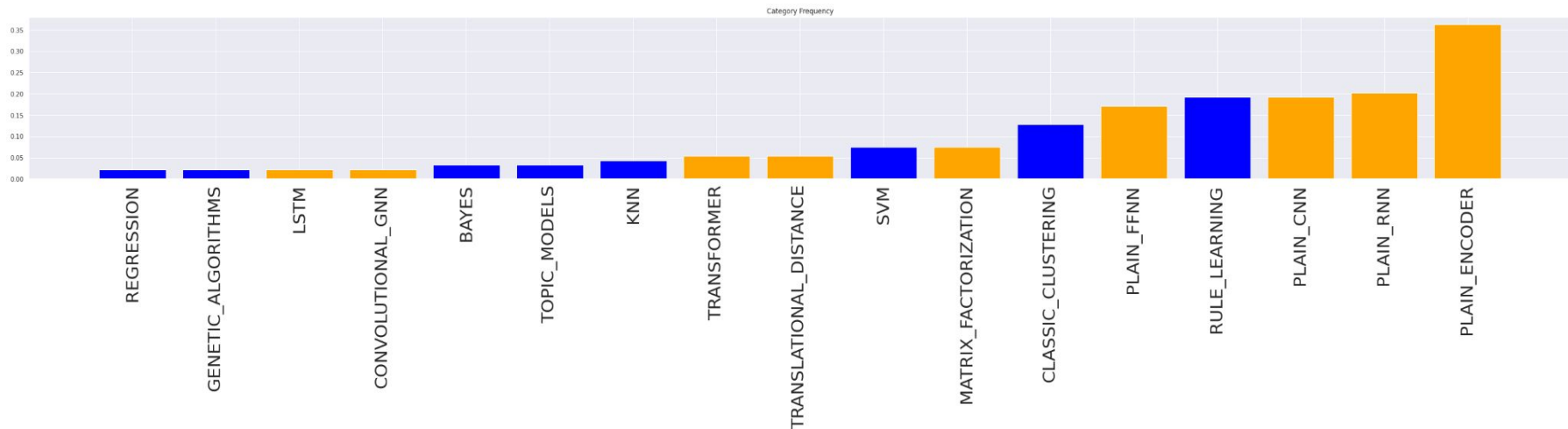
Tasks

Most tasks are related to NLP (blue), but Image & Video and Other are more than 25%!



ML models

Deep Networks (yellow) are prevalent.
Most popular encoder (~80%) is word2vec.



Why SWC?

Our Research and Product



- Combination of applied industrial projects and cutting-edge research
- Collaboration with top international scientists in frames of funded national and European projects
- Participation in top-ranked scientific conference and industrial fora
- Co-organizing [SEMANTiCS](#) conference
- Development of broadly used tool for CV management
- Large international outreach

Thank You!

Vacancies@SWC:

<https://semantic-web.com/work/>



Contact:

artem.revenko@semantic-web.com

{twitter, medium}@revenkoartem