Semantic annotation and expansion for keyword queries part 1, tutorial

Ranka Stanković
ranka@rgf.rs
University of Belgrade, Serbia

http://rgf.rs/ranka1.pptx

COST Action IC1302

2nd KEYSTONE Training School

Keyword search in Big Linked Data

University of Belgrade, Serbia

University of Belgrade is the oldest State university of the Republic of Serbia

The origin of the University of Belgrade can be tracked down to the beginning of the 19th century, when Dositej Obradović founded the College in 1808.

http://bg.ac.rs/en

keystone

University of Belgrade, Serbia

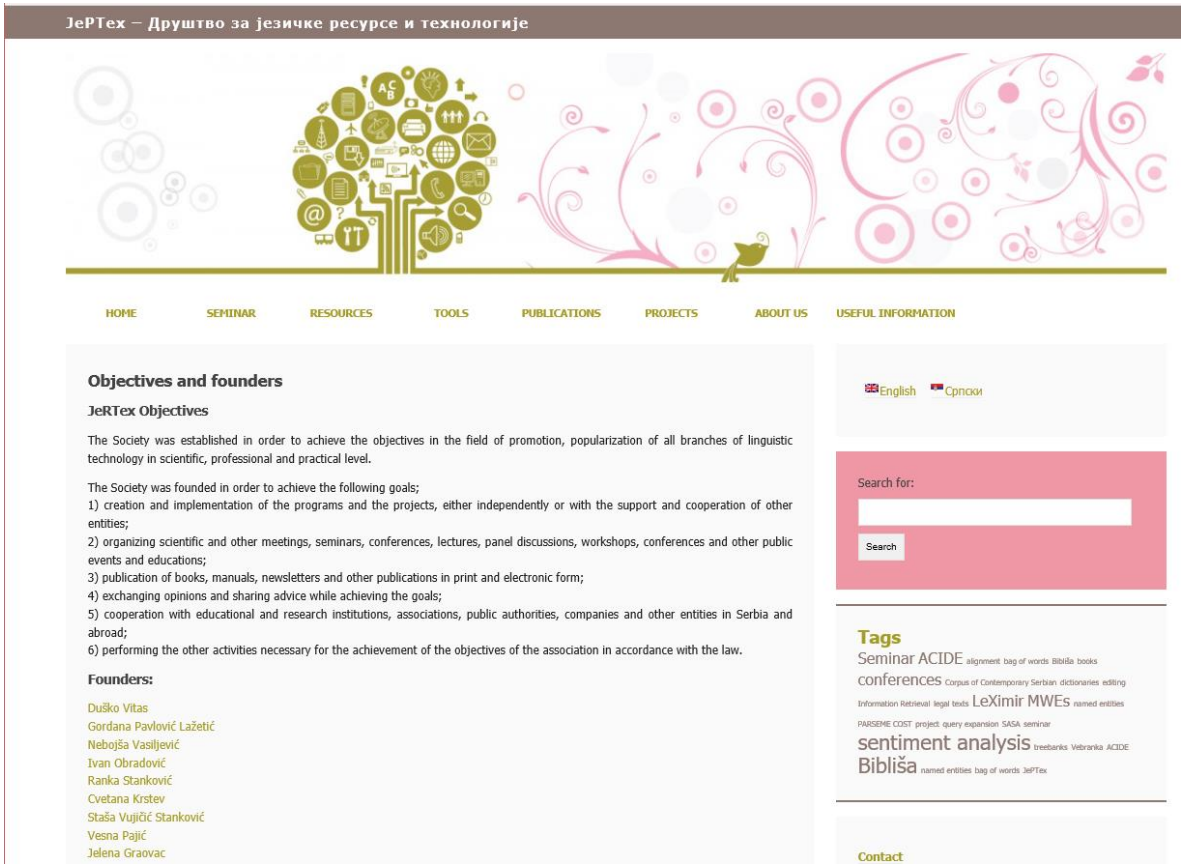
- 31 Faculties
- 11 scientific research institutes
- 6 University Centers
- University Library
- 90.000 students and 2.650 teachers

- More than 7000 staff;
- More than 320 modules;
- More than 1700 students enrolled at master and PhD programs;
- More than 350.000 graduated students, 23.000 MSc and 13.500 PhD fellows

keystone

Jerteh – Society for language resources and tools

Society was established in order to achieve the objectives in the field of promotion, popularization of all branches of linguistic technology in scientific, professional and practical level.



keystone

Main topics

- Introductory definitions
  - Information need & Information Access
  - Query & Semantic annotation
- A review of ontology based query expansion
- A keyword-based semantic retrieval approach
- Interaction Between Automatic Annotation and Query Expansion
- Query by (lexical) pattern
- Examples of application on parallel and domain specific (math) corpora

Motivation

- Enhancing the search results in large archives is a concern shared by many research community
- The improvement can come from two directions:
  - enhancing the annotations or
  - enhancing the search mechanism.
- Both directions are active research area's.

Information Access

Definition [Wikipedia]

- Area of research at the intersection of Informatics, Information Science, Information Security, Language Technology, Computer Science, and Library Science.
- The objective [...] is to simplify and make it more effective for human users to access and further process large and unwieldy amounts of data and information.

Technologies

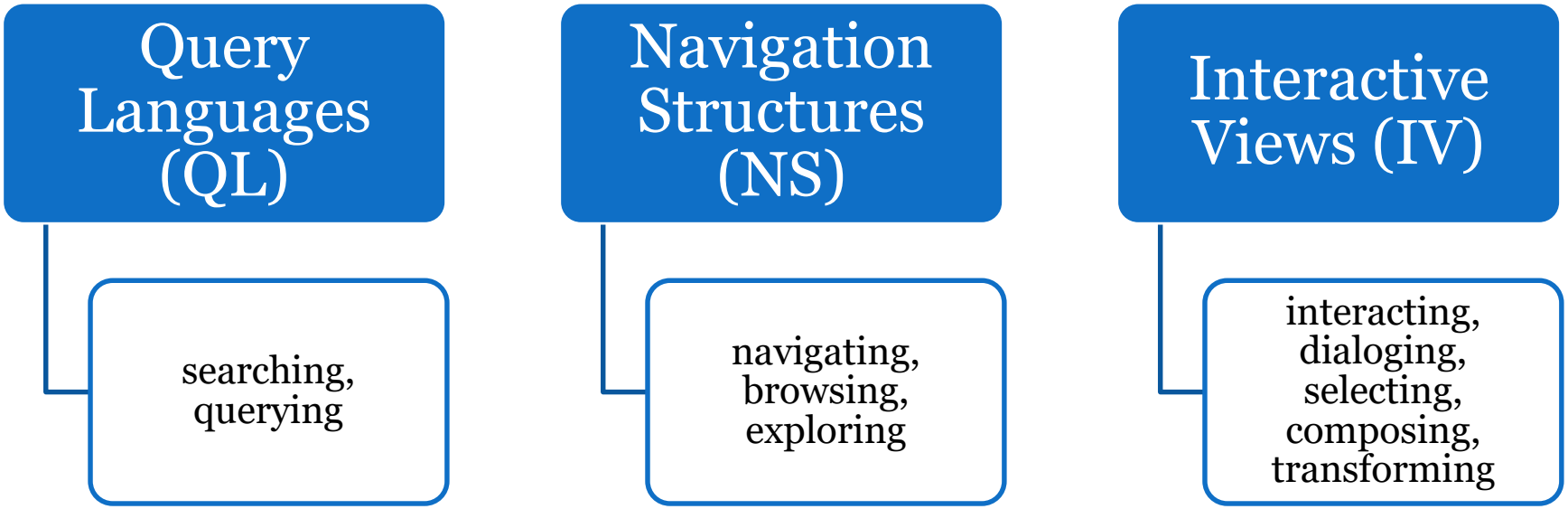
- Information Retrieval, Text Mining,
- Machine Translation, and Text Categorisation.

Related terms

- search, information retrieval, searching,
- navigation, data exploration, quering, ...

Information Access

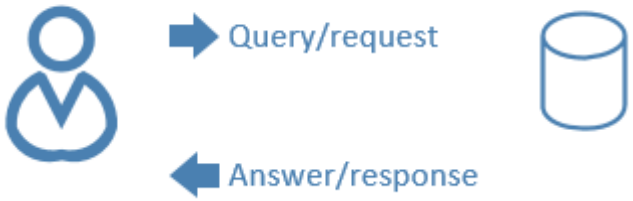
Existing approaches



keystone

Query Languages

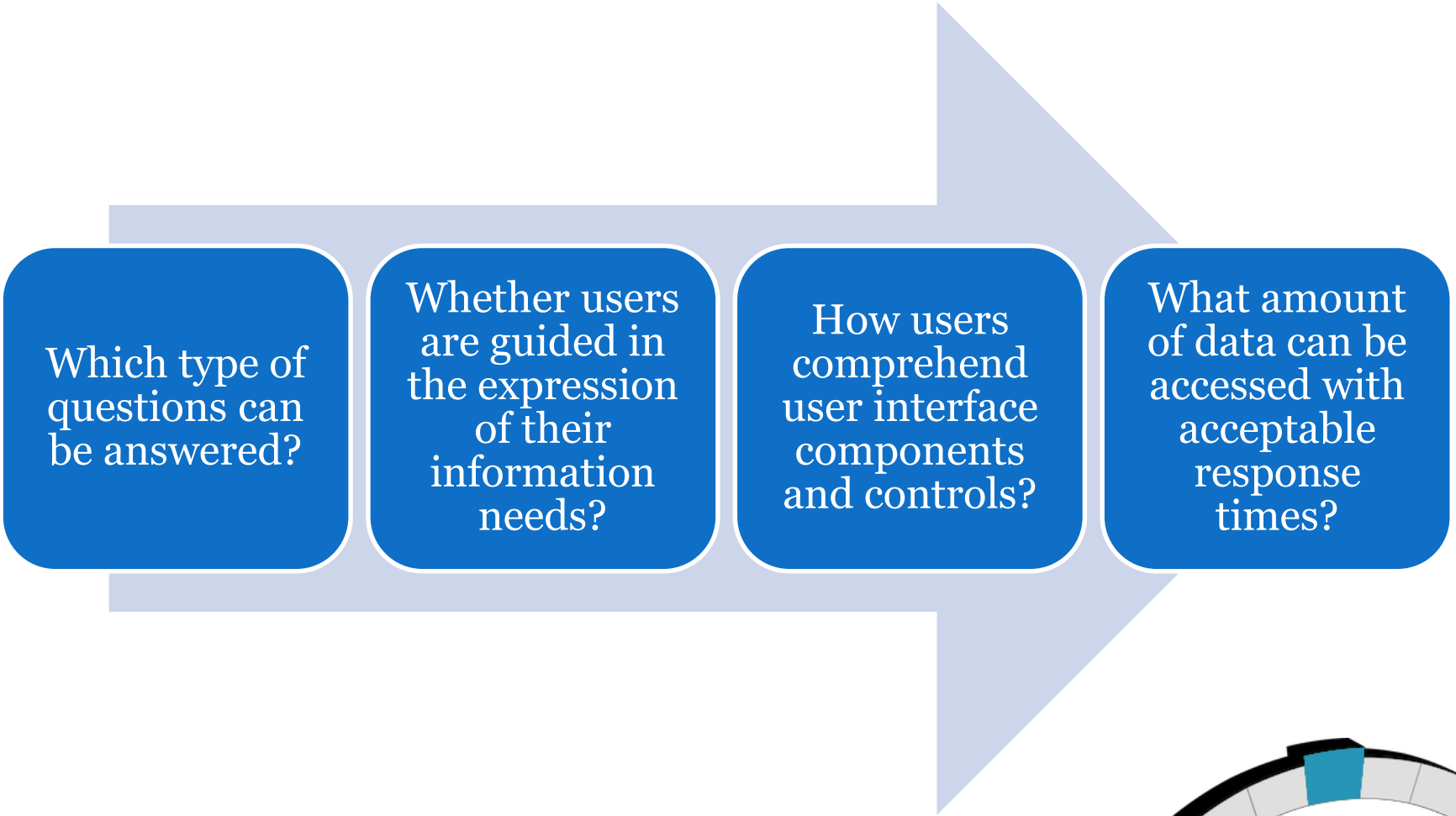
- The user provides some input query, and the system returns some answers to the query



- information retrieval (IR) search: keywords, forms
- corpus query languages: CQP, CQL, regular expressions
- formal query languages: SQL, XQuery, SPARQL
- natural language interfaces (NLI): IBM Watson

keystone

Query Languages



keystone

Information need

- An incomplete query disrupt a search engine from satisfying the user's information need.
- The query provided by the user is often unstructured and incomplete.
- We need some representation which can express the user's information need.

What is Query Expansion?

Query Expansion

- adding search terms to a user's keyword list

The goal is

- to improve precision and/or recall.

Simple example

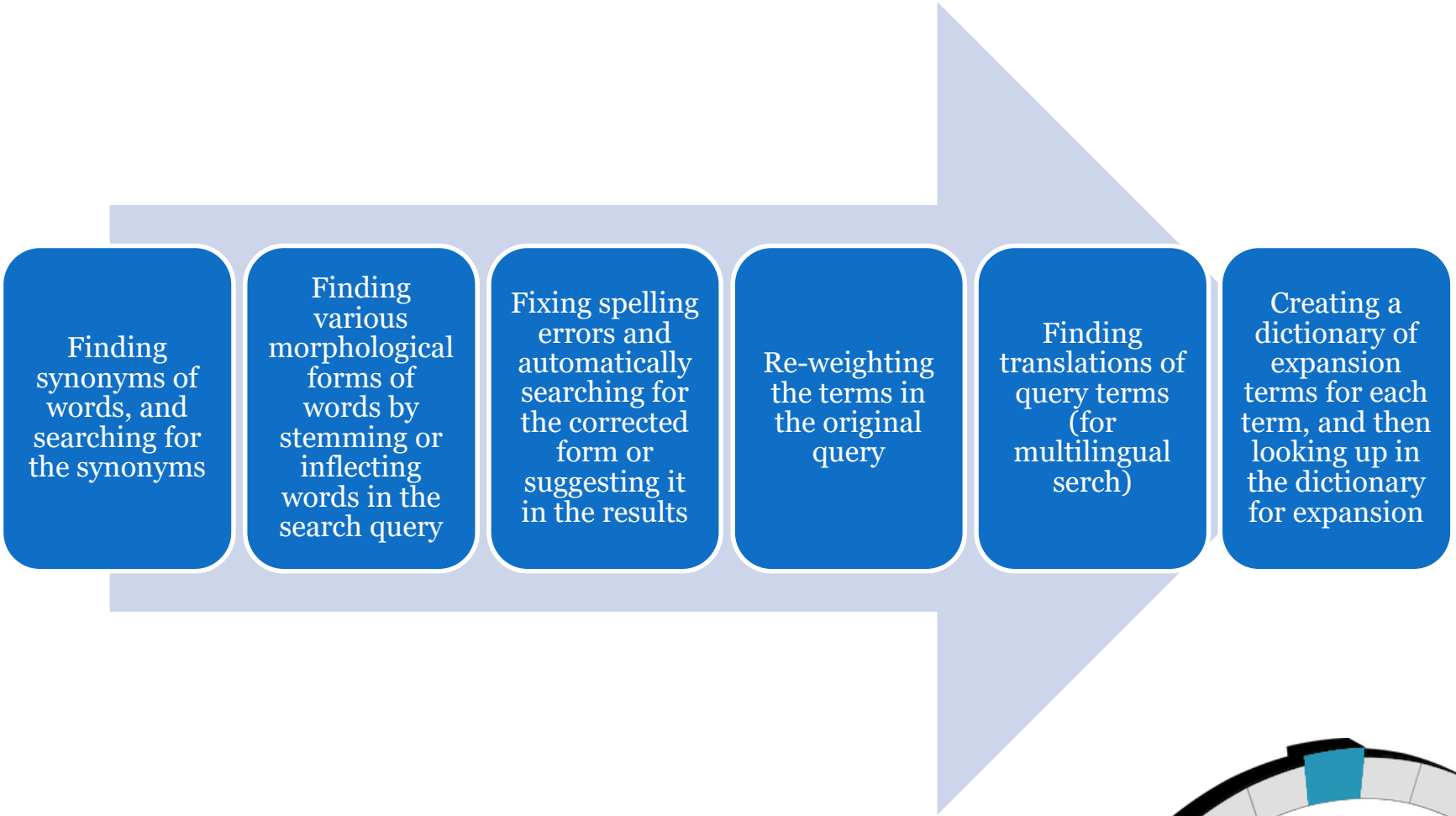
- User Query: "car"
- Expanded Query: "car cars automobile automobiles auto" etc…

keystone

Why Query Expansion?

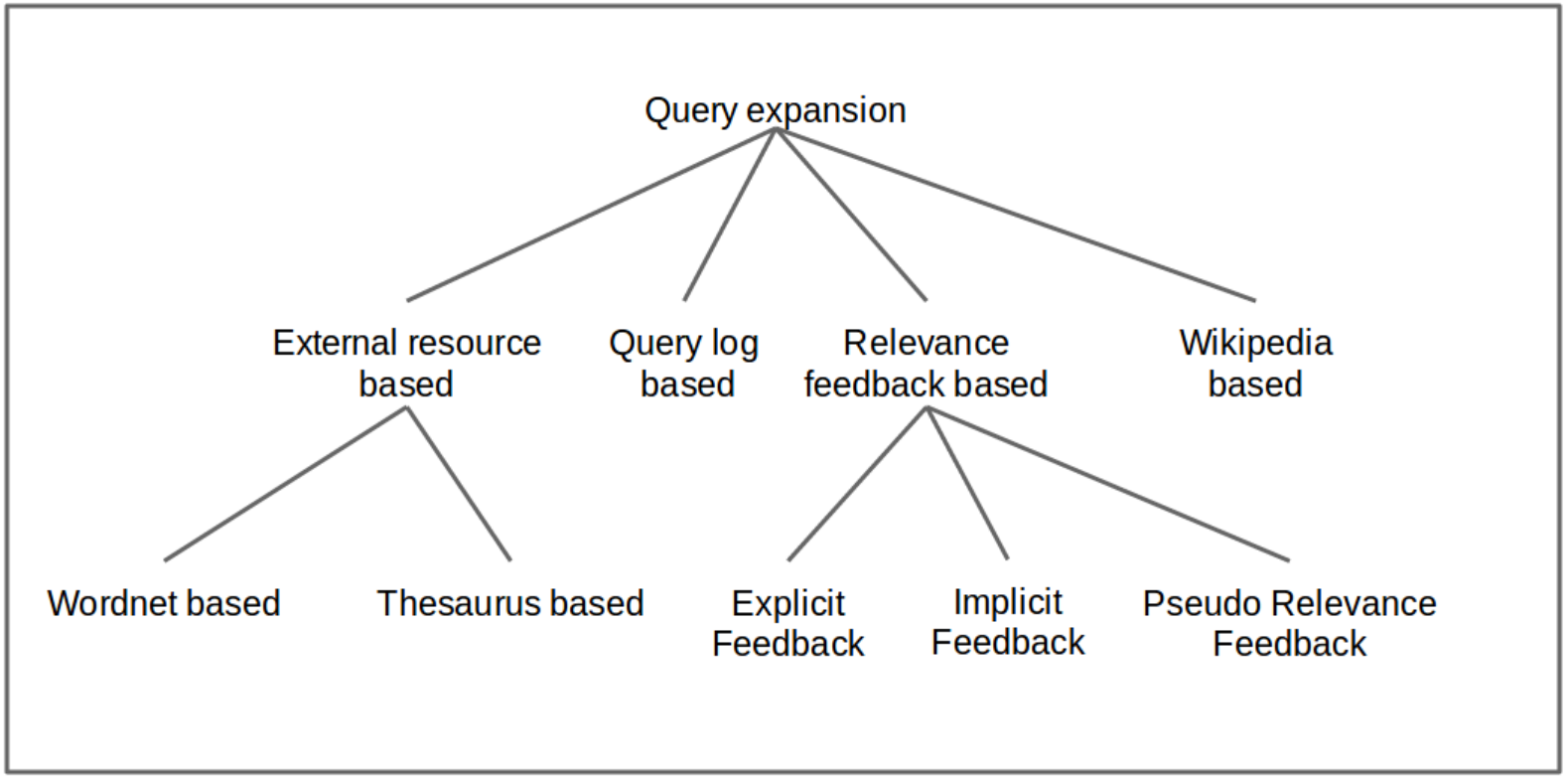
- Query expansion is very important on the web.
- The amount of information on the web is always increasing.
- Search engine users follow specific trends with their searches.
  - 2-3 words
  - Broad search term
  - Do not like to expand their queries either through refining search terms or using Boolean operators

Techniques of query expansion



keystone

Query expansion techniques



Ashish Kankaria, **Query Expansion techniques**
Indian Institute of Technology Bombay, Mumbai

keystone

External resource based query expansion

The query is expanded using some external resource like WordNet, lexical dictionaries or thesaurus.

These resources are built manually, they contain mappings of the terms to their relevant-related terms.

Techniques involve looking up in such resources and adding the related terms to query.

keystone

Thesaurus based expansion

A thesaurus

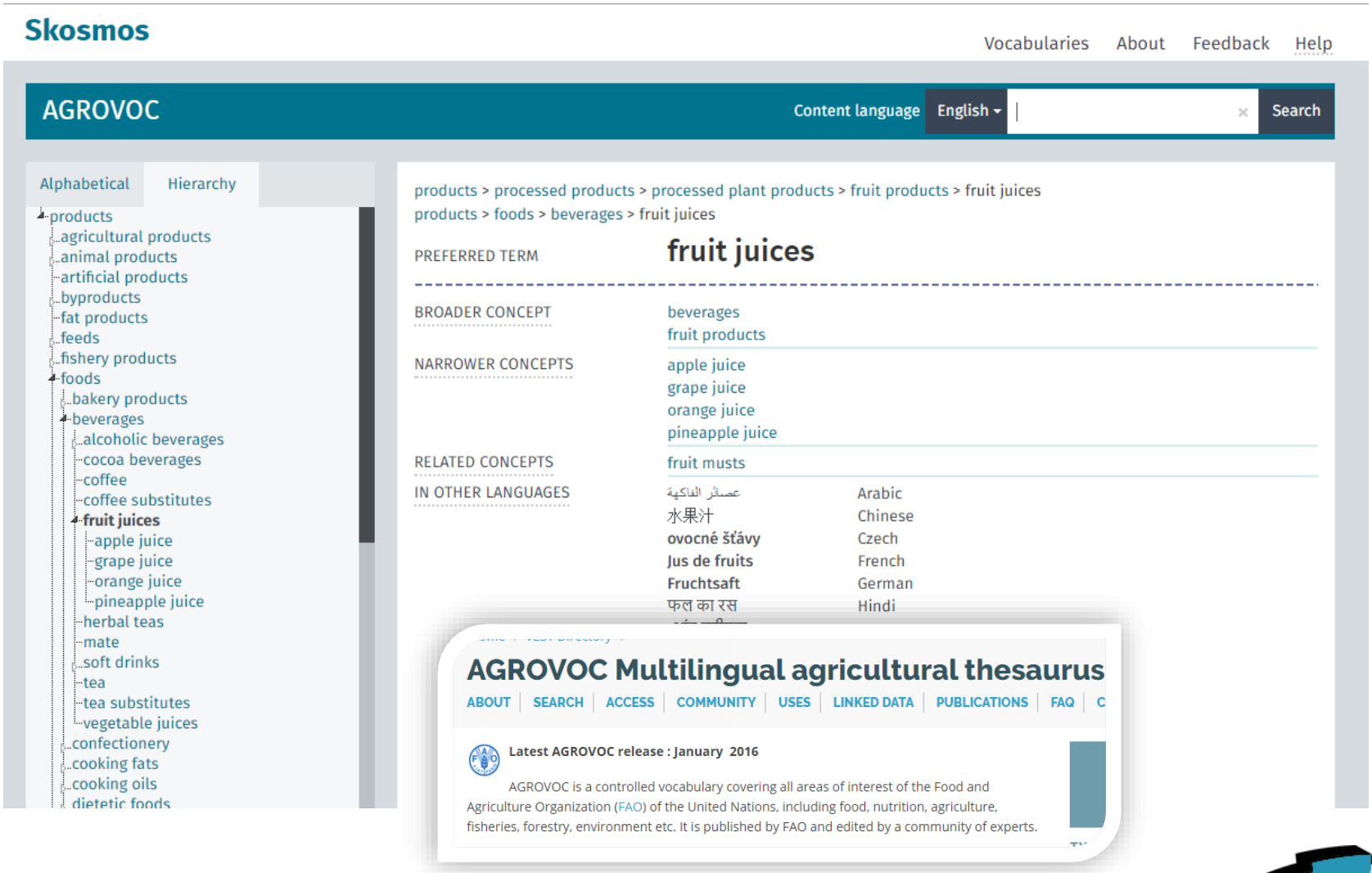
- is a data structure that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words
- is used to expand the query terms and all the connected words of query terms are added to query.

Thesaurus based system

- have been explored and put to use by many organizations.
- example is Unified Medical Language System (UMLS) used with MedLine for querying the bio medical research literature.
- Controlled vocabulary contains similar terms for each bio medical concept.

A thesaurus based query expansion system works well only for rich domain specific thesaurus.

http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus



keystone

WordNet based expansion

- WordNet is a lexical database for multiple languages.
- The similar terms from multiple languages are connected via synsets (set of senses).
- WordNet can be used to fetch related term for a particular term in multiple languages and can help in satisfying user's information need.
- Expansion can use synonyms and/or hypernyms and/or hypernyms.

http://wordnetweb.princeton.edu/perl/webwn

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: car [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

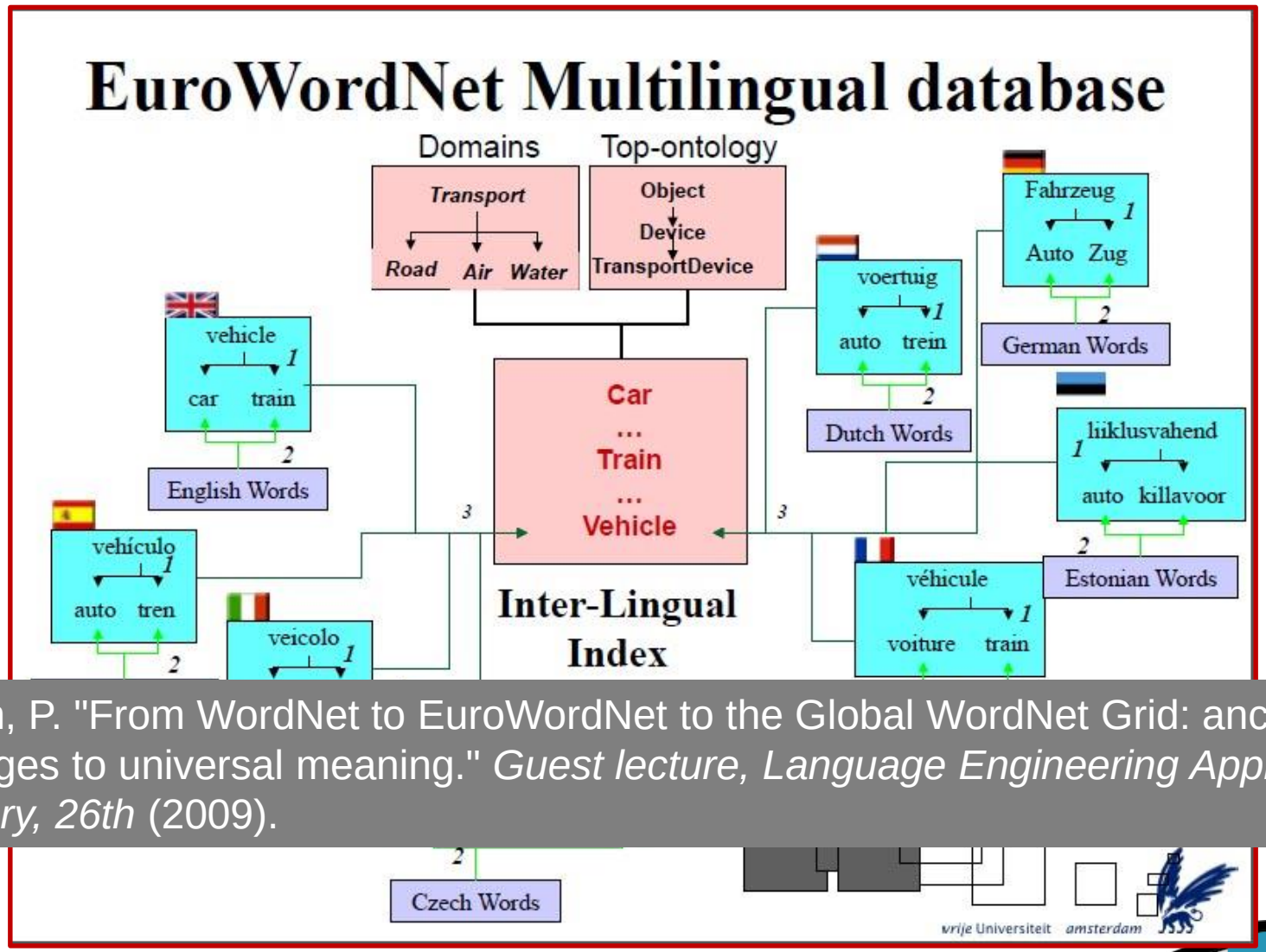
- S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) **car**, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) **car**, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) **car**, elevator car (where passengers ride up and down) *"the car was on the top floor"*
- S: (n) cable car, **car** (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*

keystone

Global WordNet

- Global WordNet Association - http://globalwordnet.org/
- A free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.
  - Organizes GWA Conferences – 8 conferences up to now
- Global WordNet Grid - which is being build around a shared set of concepts used in many wordnet projects.
  - List of all wordnets in the world (contact persons, licences etc. http://globalwordnet.org/wordnets-in-the-world/)

keystone



Vossen, P. "From WordNet to EuroWordNet to the Global WordNet Grid: anchoring languages to universal meaning." *Guest lecture, Language Engineering Applications, February, 26th* (2009).

keystone

Concepts recognized by all Balkan languages

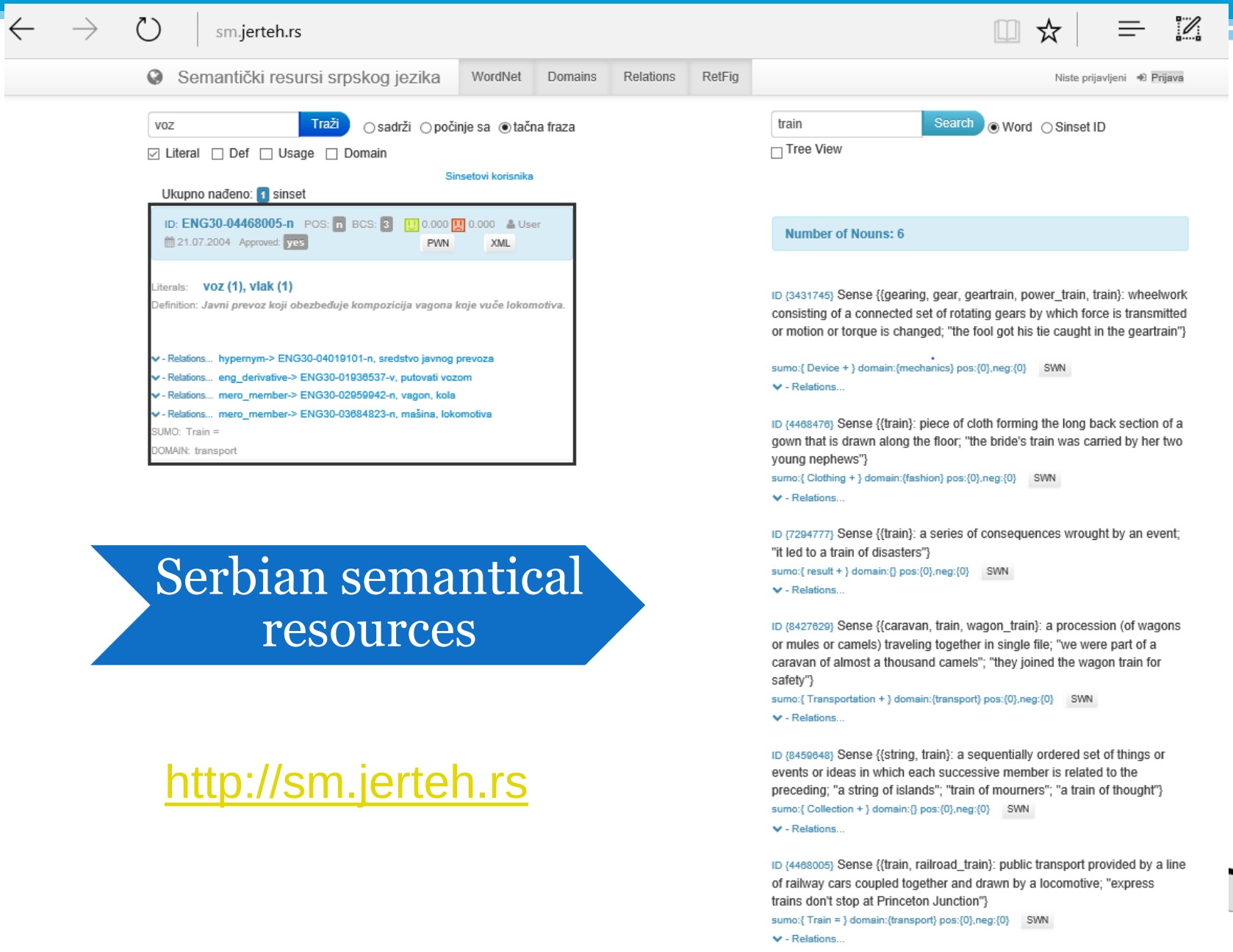
Bulgarian	кадаиф	халва
Greek	καντα͏ΐφι	χαλβάς
Romanian	cataif	halva
Serbian	кадаиф	алва
Turkish	kadayıf	kağıt helva

Cvetana Krstev, Ivan Obradović, Duško Vitas, "Developing Balkan specific concepts within BalkaNet - a multilingual database of semantic networks", in Proceedings of the 5th International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL 2006, Sofia, Bulgaria, October, 2006, eds. Svetla Koeva and Mila Dimitrova-Vulchanova, pp. 94-98, The Department of Computational Linguistics, Institute of Bulgarian Language, 2006

keystone

Usage of wordnets

- Improve recall of textual based analysis:
  - Query → Index
    - Synonyms: commence → begin
    - Hypernyms: taxi → car
    - Hyponyms: car → taxi
    - Meronyms: trunk → elephant
    - Lexical entailments: used a gun → shot
  - Inferencing:
    - what things can be used for transport?
  - Expressions in language generation and translation:
    - alternative words and paraphrases



Serbian semantical resources

http://sm.jerteh.rs

keystone

WordNet XML representation

```
<SYNSET>
   <ID>ENG30-04468005-n</ID>
   <POS>n</POS>
   <SYNONYM>
        <LITERAL>train<SENSE>1</SENSE></LITERAL>
        <LITERAL>railroad train<SENSE>1</SENSE></LITERAL>
   </SYNONYM>
   <ILR><TYPE>hypernym</TYPE>ENG30-04019101-n</ILR>
   <ILR><TYPE>eng_derivative</TYPE>ENG30-01936537-v</ILR>
   <DEF>public transport provided by a line of railway cars coupled together
   and drawn by a locomotive</DEF>
   <USAGE>express trains don't stop at Princeton Junction</USAGE>
   <BCS>3</BCS>
   <DOMAIN>transport</DOMAIN>
   <SUMO>Train<TYPE>=</TYPE></SUMO>
   <RILR>ENG30-01936537-v<TYPE>eng_derivative</TYPE></RILR>
   <RILR>ENG30-02859729-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-02959942-n<TYPE>holo_member</TYPE></RILR>
   <RILR>ENG30-02971579-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-03394480-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-03541393-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-03684823-n<TYPE>holo_member</TYPE></RILR>
   <RILR>ENG30-03711044-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-03896233-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-04103918-n<TYPE>category_domain</TYPE></RILR>
   <RILR>ENG30-04334504-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-04349306-n<TYPE>hypernym</TYPE></RILR>
   <RILR>ENG30-10403876-n<TYPE>category_domain</TYPE></RILR>
   <RILR>ENG30-10647745-n<TYPE>category_domain</TYPE></RILR>
</SYNSET>
```

```
<SYNSET>
   <ID>ENG30-04468005-n</ID>
   <POS>n</POS>
   <SYNONYM>
        <LITERAL>voz<SENSE>1</SENSE><LNOTE>nema</LNOTE></LITERAL>
        <LITERAL>vlak<SENSE>1</SENSE><LNOTE>nema</LNOTE></LITERAL>
   </SYNONYM>
   <DEF>Javni prevoz koji obezbeđuje kompozicija vagona koje vuče lokomotiva.
   </DEF>
   <BCS>3</BCS>
   <ILR>ENG30-04019101-n<TYPE>hypernym</TYPE></ILR>
   <ILR>ENG30-01936537-v<TYPE>eng_derivative</TYPE></ILR>
   <ILR>ENG30-02959942-n<TYPE>mero_member</TYPE></ILR>
   <ILR>ENG30-03684823-n<TYPE>mero_member</TYPE></ILR>
   <NL>yes</NL>
   <STAMP>User 21/07/2004 00:00:00</STAMP>
   <SUMO>Train<TYPE>=</TYPE></SUMO>
   <SENTIMENT><POSITIVE>0.00000</POSITIVE><NEGATIVE>0.00000
   </NEGATIVE></SENTIMENT>
   <DOMAIN>transport</DOMAIN>
   <RILR>ENG30-02959942-n<TYPE>holo_member</TYPE></RILR>
   <RILR>ENG30-03684823-n<TYPE>holo_member</TYPE></RILR>
   <RILR>ENG30-04019101-n<TYPE>hyponym</TYPE></RILR>
</SYNSET>
```

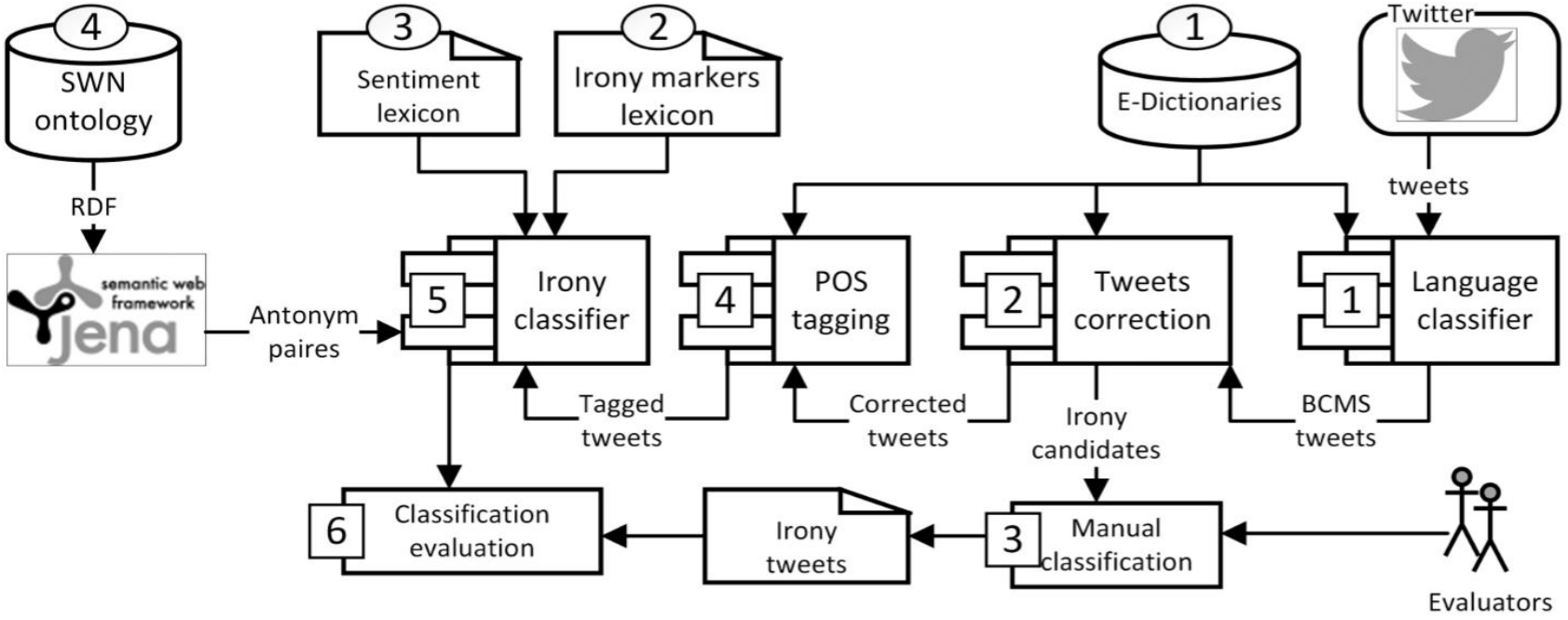

keystone

Using WordNet Knowledge for Irony Classification

- A language dependent model for classification of statements into ironic and non-ironic.
  - language resources: morphological dictionaries, sentiment lexicon, lexicon of markers and a WordNet based ontology.
  - features: antonymous pairs obtained using the reasoning rules over the SrpWN (R), antonymous pairs in which one member has positive sentiment polarity (PPR), polarity of positive sentiment words (PSP), ordered sequence of sentiment tags (OSA), POS tags and irony markers (M).
- Evaluation on a collection of tweets that had been manually annotated according to irony.
- The collection of tweets is in the Serbian language (or Bosnian/Croatian/Montenegrin).
- The best achieved accuracy of the developed classifier acc = 86.1% was achieved with the set of 5 features — (PPR, PSP, POS, OSA, M).

keystone

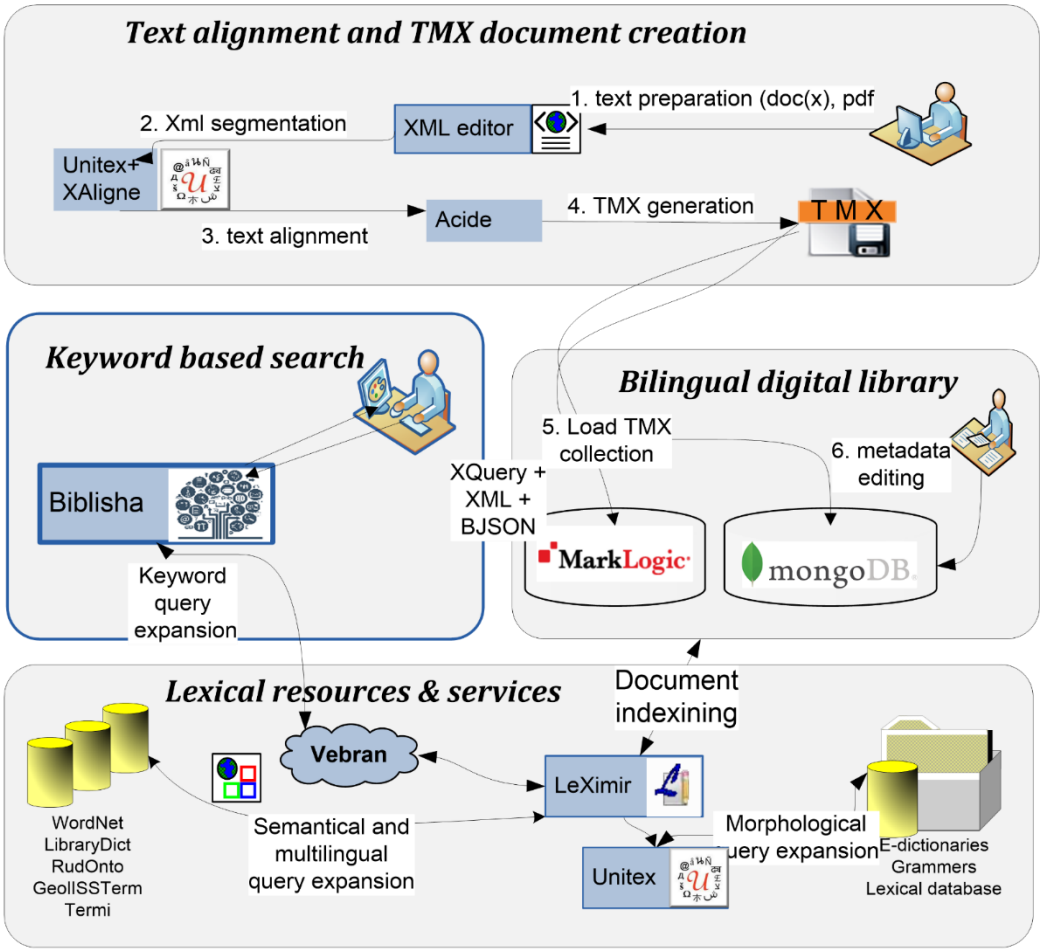
Architecture of the ironic tweets classifier



HrTal2016, accepted, Dubrovnik September 2016
Mladenovic M., Krstev C., Mitrovic. J. Stankovic R.
„Using WordNet Knowledge for Irony Classification “

keystone

Bilingual digital library search demo in hands-on session



keystone

Query logs based expansion

- Query logs are maintained by search engine in order to analyze the behavior of the user while interacting with search engine.
- Query logs can be used to analyze the user's preference and adds corresponding terms to query.
- Method fails when user search something which is not related to earlier searches.
- List of all the documents visited for a particular query can be stored for further use.
- It can be used to learn associations by combining evidence from various lexical sources like WordNet.

Relevance feedback based expansion

Process

- execute the initial query on collection and extract top k documents.
- use ranked document to improve the performance of retrieval.

Assumption

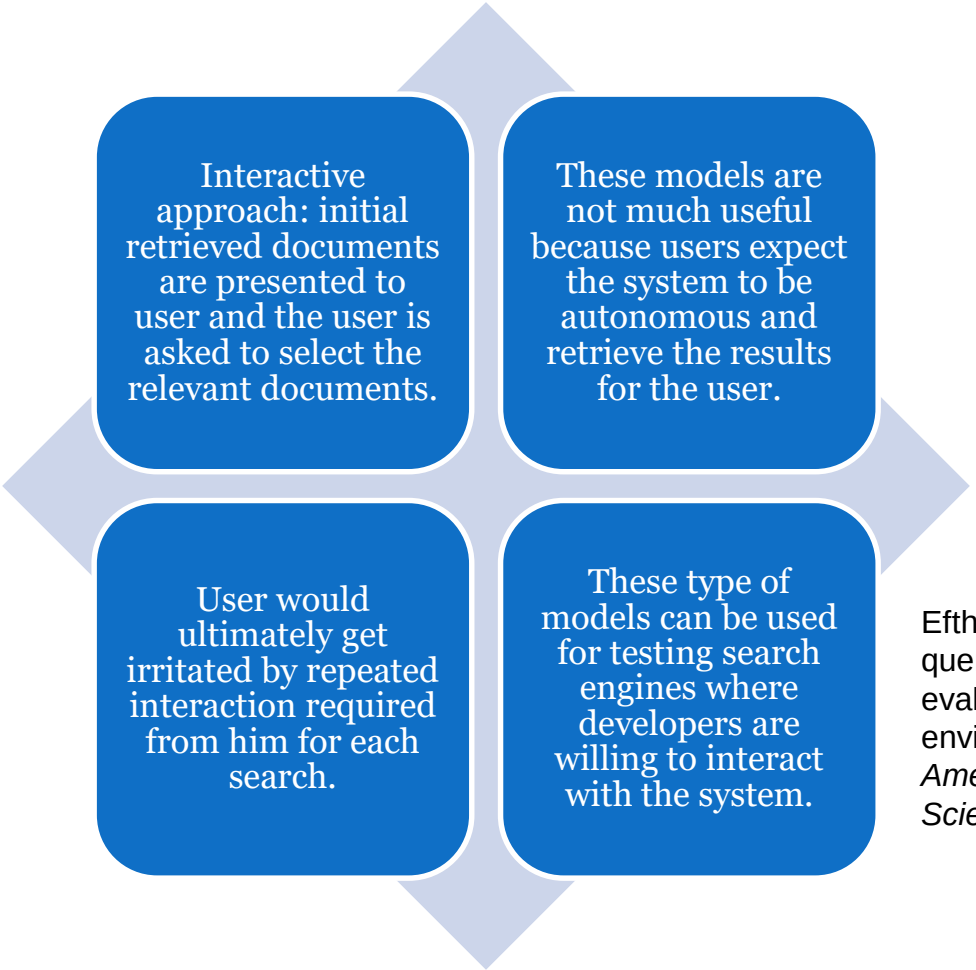
- initial retrieved documents are relevant and thus can be used to extract expansion terms.
- the initial retrieval algorithm of search engine is good.

Model types

- Explicit feedback from user
- Implicit feedback
- Pseudo Relevance Feedback (PRF)

keystone

Explicit feedback from user



Efthimiadis, Efthimis N. "Interactive query expansion: a user-based evaluation in a relevance feedback environment." *Journal of the American Society for Information Science* 51.11 (2000): 989-1003.

keystone

Implicit feedback

- User's feedback is inferred by the system.
- The feedback can be inferred from user's behavior like:
  - The pages which user opens for reading, or
  - pages on which user clicks once the results are displayed back to the user
  - Time spent on page

Pseudo Relevance Feedback (PRF)

- Initial query is fired and top k results are obtained.
- Important terms, mostly based on co-occurrence, from these documents are extracted and added to query.
- Expanded query is re fired to retrieve final set of documents which are made available to the user.

Process

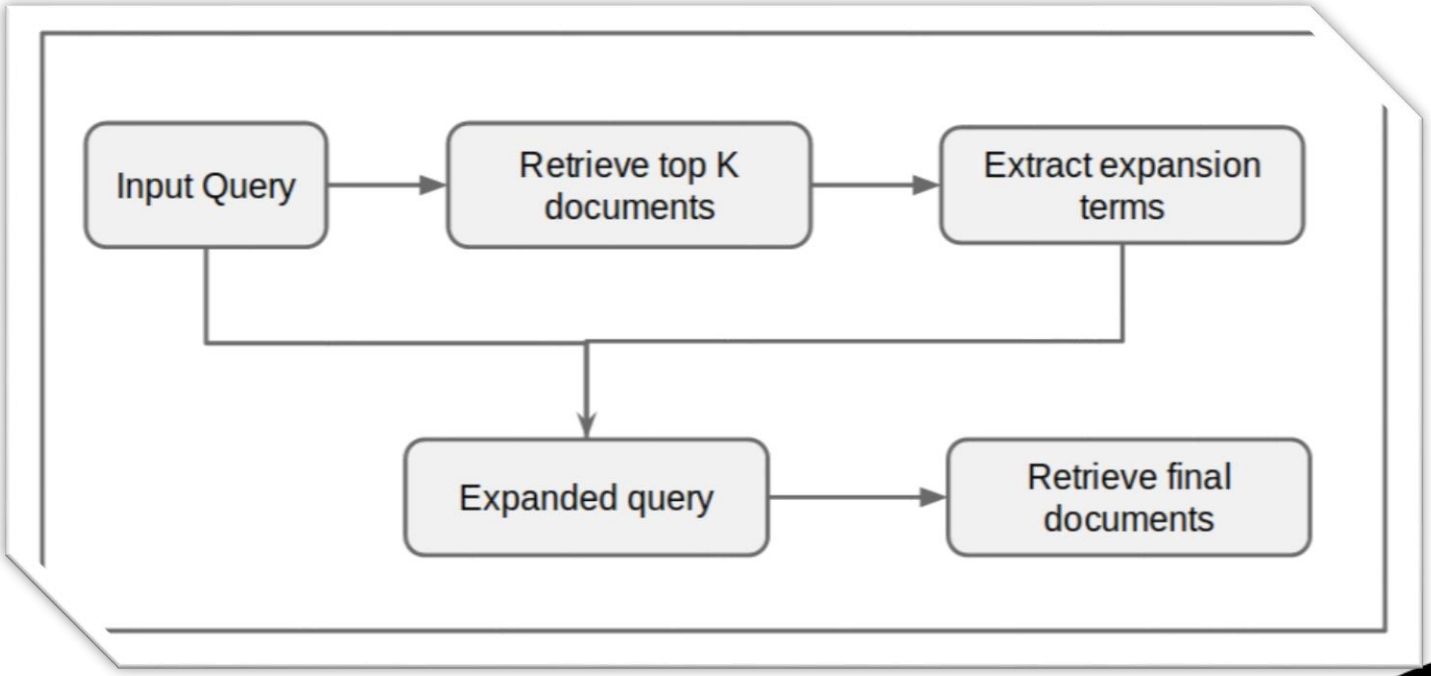
- The relevancy of expansion terms depends upon the initial retrieved documents.
- Pseudo relevance feedback captures the important terms only based on co-occurrence (not enough for correctness of results)
- Semantic and lexical properties of word should be considered.
- Feedback is independent of the user there is a chance of topic drift.

Note

keystone

Pseudo Relevance Feedback (PRF)

- Typical work flow for PRF (Blind Feedback) based information retrieval system



keystone

Pseudo Relevance Feedback (PRF)

- Automates the manual part of relevance feedback and has the advantage that assessors are not required.
- Successfully applied in various IR frameworks and has been proved to improve precision and recall of search engines.
- Improvement of PRF by -
  - refining relevant document set
  - refining the expansion terms from PFR
  - using selective query expansion
  - varying the importance of documents

PFR improvements

- terms in the document which are nearer to the query terms are assigned more weight.
- use of an assisting language (AL) to improve the performance retrieval of search engine.
  - translate the query to an assisting language
  - perform PFR twice, once for in QL and other in AL
  - Merge the expansion terms obtained from both the PRF instances using translation and retrieve the documents for expanded query.
  - multilingual PRF using English as AL for French, German, Hungarian,...

PFR improvements

- Takes into account the structure of the documents while assigning priorities to the expansion terms.
- The intuition behind the idea is that a term that occurs in title section of a document is more important for that document than the term which occurs in the body.
- Title more compactly represents the entire document and thus it is very less probable that title will have a word which is unrelated to the document.
- Examples:
  - Wikipedia document collection: Title, Body, Infobox and Categories.
  - Project collection: Institution, Project Name, Location, Domain, Responsible person

Query expansion using Wikipedia

- Wikipedia query expansion is based on the category assignments of its articles.
  - The base query is run against a Wikipedia collection and each category is assigned a weight proportional to the number of top-ranked articles assigned to it.
  - Articles are then re-ranked based on the sum of the weights of the categories to which each belongs.
- Thesaurus can be produced from Wikipedia articles (with some irrelevant results).
- Category information
  - can be used by calculating distances between document categories and target categories.
  - has more value than link information.

Query Expansion Issues

- Two major issues
  - Which terms to include?
  - Which terms to weight more?
- Concept-Based vs. Term-Based Query Expansion
  - Is it better to expand based upon the individual terms in the query, or the overall concept of the query?
- Classes of QE
  - Manual approach - Human generated thesauri
  - Interactive Query Expansion
  - Automatic Query Expansion

Approaches to Query Expansion

- Global Analysis
  - Considers all the documents in the system.
- Local analysis
  - Uses some initially retrieved documents for expansion terms.

Another classification:

- Document-term based approach.
- Query-term based approach.
- Combined approach.

Global Analysis

- Term clustering
- Latent Semantic Indexing
- Similarity Thesauri

Disadvantages

- Corpus wide statistical analysis takes computation time.
- Cannot address term mismatch problem.

The Need For Thesauri

- Naturally assumed that pulling words from a thesauri would increase:
  - The number of documents retrieved.
  - Possibly precision.
- The car example: “car” vs. “car, auto, automobile, vehicle, sedan, etc…”
  - Which would retrieve the largest number of documents?
  - Is larger necessarily better?

Human and Automatically Generated Thesauri

Earliest work began in the 1950s.

- H.P. Luhn
- *Thesaurofacet* – detailed list of engineering terms

Largely used in

- Medicine,
- Agronomy,
- Natural science,
- Technological fields.

keystone

Drawbacks of Handcrafted Thesauri

- Cost
  - Development.
  - Maintenance.
  - Cost often outweighs benefit.
- Time
  - It often takes a long time for thesauri to develop.
  - Hard to keep up with the pace of scientific and technological development.

Automatically Generated Thesauri

- Global analysis method.
- 3 Steps.
  - Extract word co-occurrences or sintactic patterns.
  - Define word similarities.
    - Based upon word co-occurrence or lexical relationship.
  - Cluster words based upon their similarities.
- Not proven very successful.
  - As late as 1990 many industries were still using handcrafted thesauri.

Interactive Query Expansion

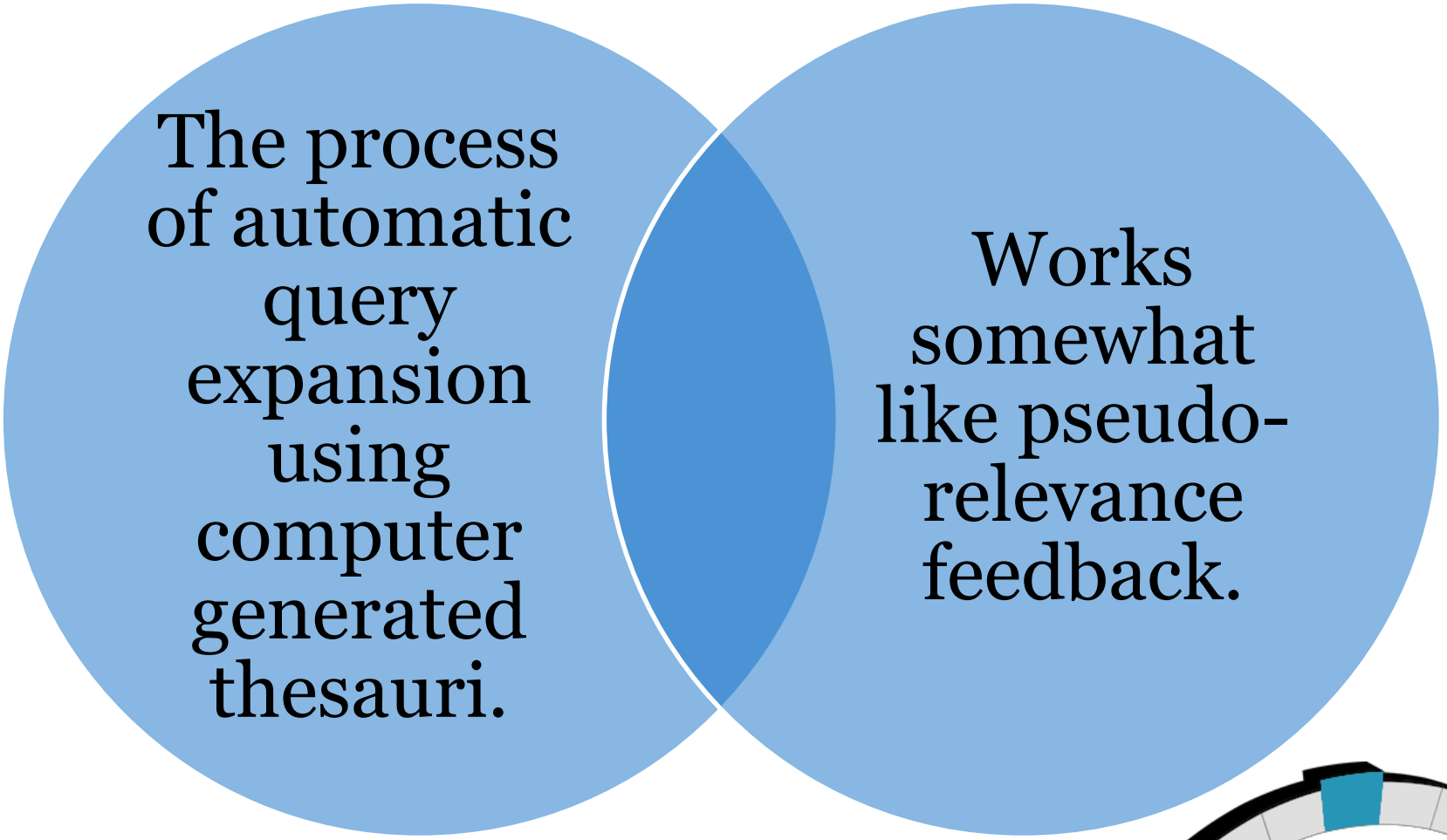
- Uses a thesaurus.
  - After initial query is submitted,
  - the system returns a list of associated and relevant
  - words derived from both the result set and a thesaurus.
- Useful, but more research is needed.

keystone

Relevance Feedback

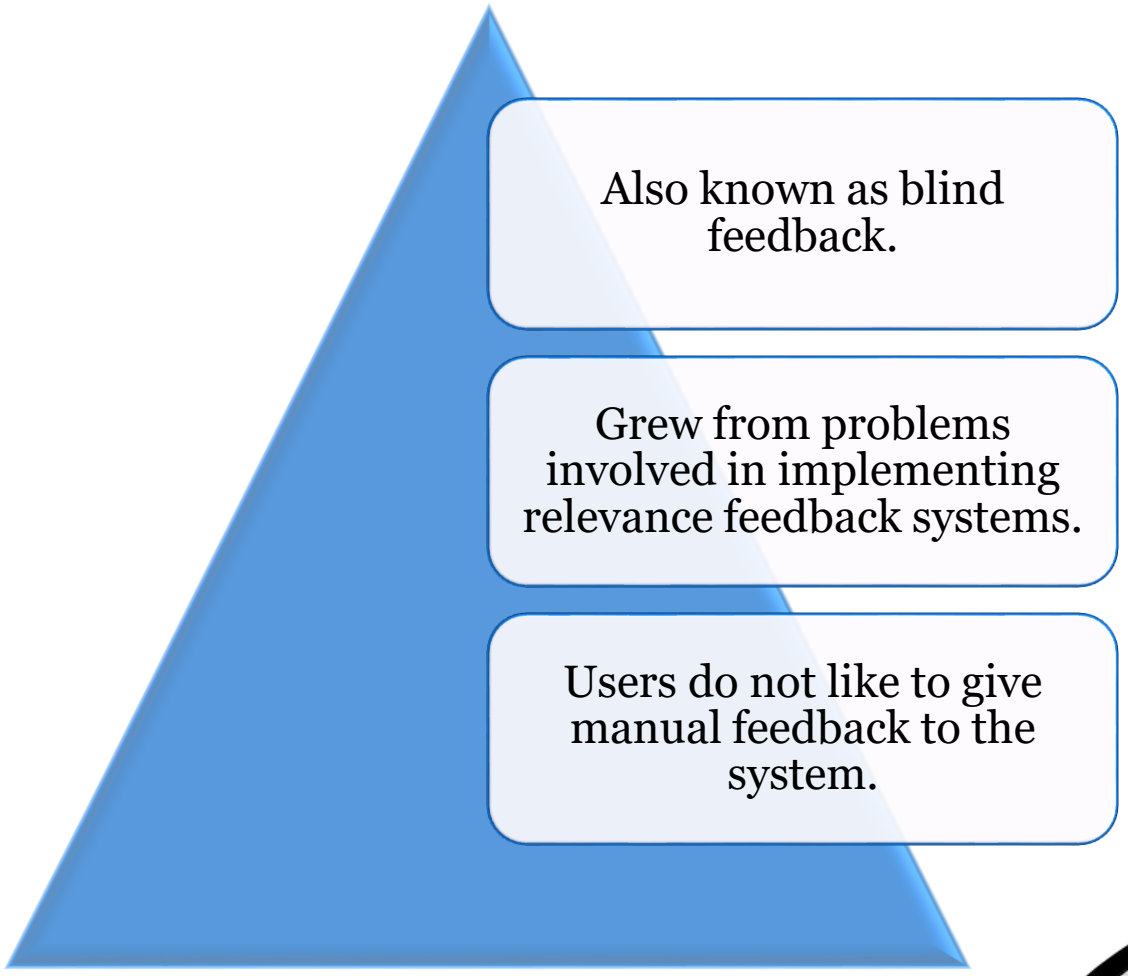
- Local analysis + interactive.
- Significant improvement in recall and precision over early query expansion work.
- Basic process as follows.
  - The user creates their initial query which returns an initial result set.
  - The user then selects a list of documents that are relevant to their search.
  - The system then re-weights and/or expands the query based upon the terms in the documents.

Automatic Query Expansion



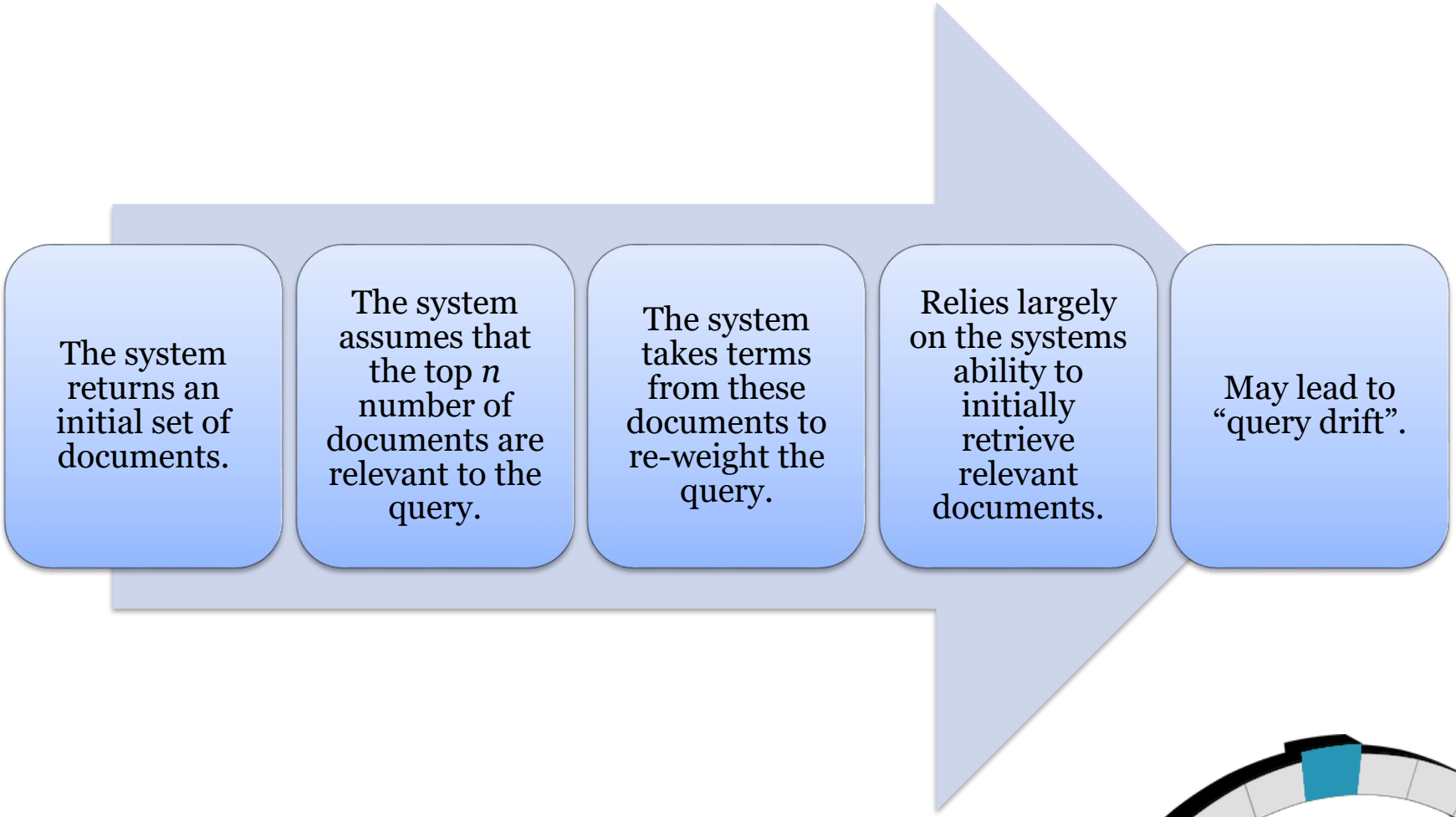
keystone

Pseudo-relevance Feedback



keystone

Pseudo-relevance Feedback Process



keystone

Concept Based Query Expansion

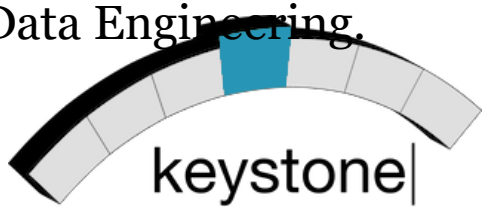
- Uses terms that are closer to the concept of query rather than individual query terms.
- Determining concept representing a query is hard.
- Mathematical approach

Qiu, Y. and Frei, H.P. 1993. Concept Based Query Expansion. Proceedings of 16th SIGIR.

Mining for Query Expansion

- Needs a log of queries fired and the corresponding documents clicked by the user.
- If a set of documents is often selected for the same queries, then the terms in this document are strongly related to terms in the queries.
- Takes advantage of user judgment implied in the logs.
- Described in the paper

Cui, H.; Wen, J.R.; Nie, J.Y; and Ma, W.Y. 2003. Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering.



Unitex

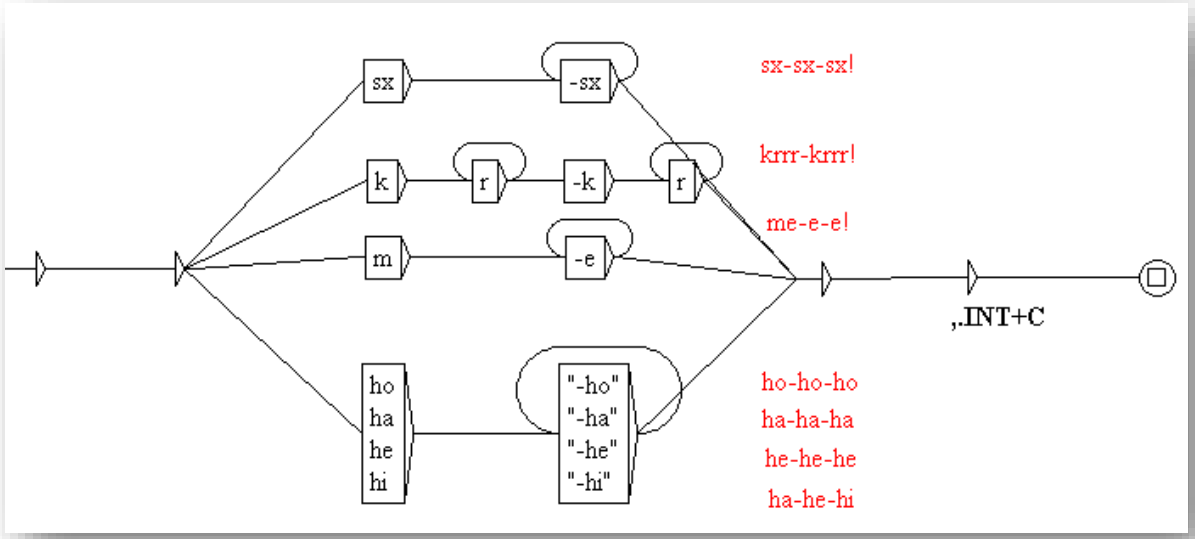
- http://www-igm.univ-mlv.fr/~unitex

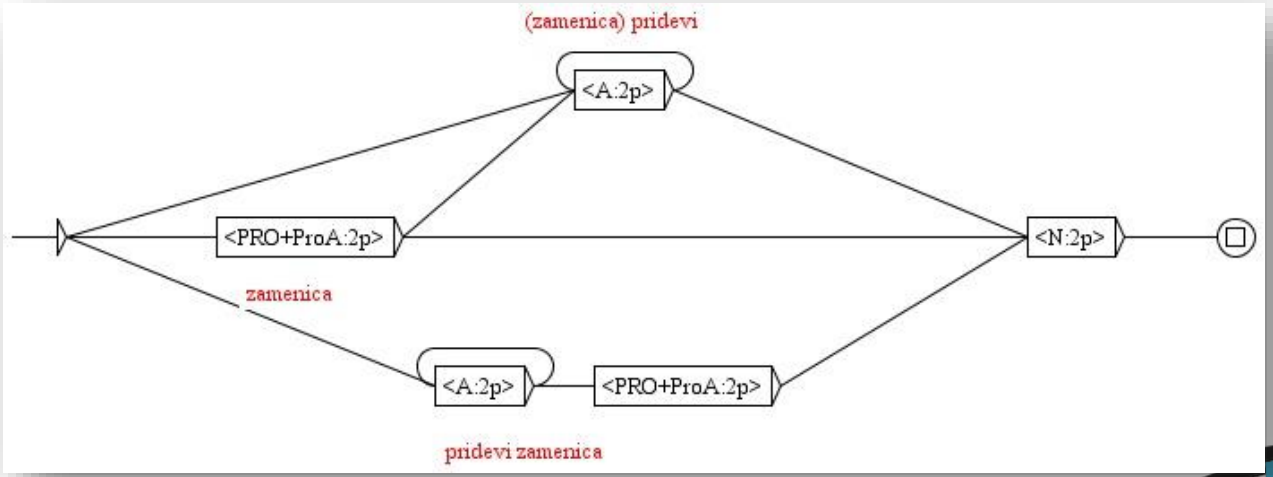
keystone

Query in Unitex

- Regular expressions
- **<to be>** - all word forms linked to lemma
- **<N>** - search for POS – all nouns
- **<DET><A><N+Hum>** - noun preceeded by determiner and adjective
- **<N+Hum>** - nouns with semantic tag Human
- **<N+NProp+Hum~Inh>** - ...not Inhabitans
- **(<A>+<PRO+ProA>) <love>**
- **(<this>+<that>) <A> <N>**

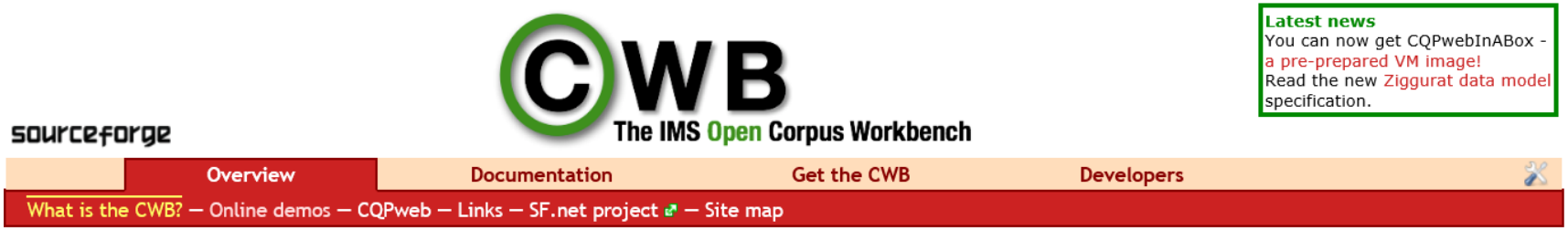
keystone


sx
-sx
k
r
-k
r
m
-e
ho
ha
he
hi
"-ho"
"-ha"
"-he"
"-hi"
sx-sx-sx!
krrr-krrr!
me-e-e!
,.INT+C
ho-ho-ho
ha-ha-ha
he-he-he
ha-he-hi


(zamenica) pridevi
<A:2p>
<PRO+ProA:2p>
<N:2p>
zamenica
<A:2p>
<PRO+ProA:2p>
pridevi zamenica

keystone

About CQPweb



http://cwb.sourceforge.net/cqpweb.php

keystone

About CQPweb

- **CQPweb** is a web-based app for **CQP** query processor.
- CQPweb is designed to replicate the user-interface of BNCweb tool, which also uses CQP as a back-end.
- Unlike BNCweb, CQPweb can be used with *any* corpus.
- CQPweb is especially suitable for students, non-linguists
- CQPweb can be used in three ways.
  - Via a public server. There are many of these out there; the one run by Andrew Hardie, CQPweb's main developer, is https://cqpweb.lancs.ac.uk.
  - By getting a copy of the code and installing it directly on your own computer
  - By downloading CQPwebInABox, a Virtual PC which has CQPweb pre-installed (with two sample corpora included!)

Terminology extraction

- Terminology mining, term extraction, term recognition, or glossary extraction, is a subtask of information extraction.
- The goal of terminology extraction is to automatically extract relevant terms from a given corpus.
  - Used in topic-driven web crawlers, web services, recommender systems, etc.
  - Essential to the language industry
  - Used for conceptualizing a knowledge domain or for supporting the creation of a domain ontology or a terminology base
  - Used for semantic similarity, knowledge management, human translation and machine translation, etc.
- One of the first steps to model the knowledge domain is to collect a vocabulary of domain-relevant terms, as linguistic view of domain concepts.
- Automatic term extraction includes
  - linguistic processing (part of speech tagging) to extract candidates, i.e. noun phrases, NPs (e.g. credit card", adjective-NPs "local tourist information office", and prepositional-NPs "board of directors").
  - filtering the candidate list using statistical and machine learning methods.

Ranka Stanković, **Cvetana Krstev**, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac, "Rule-based Automatic Multi-word Term Extraction and Lemmatization", *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari *et al.*, ISBN 978-2-9517408-9-1.

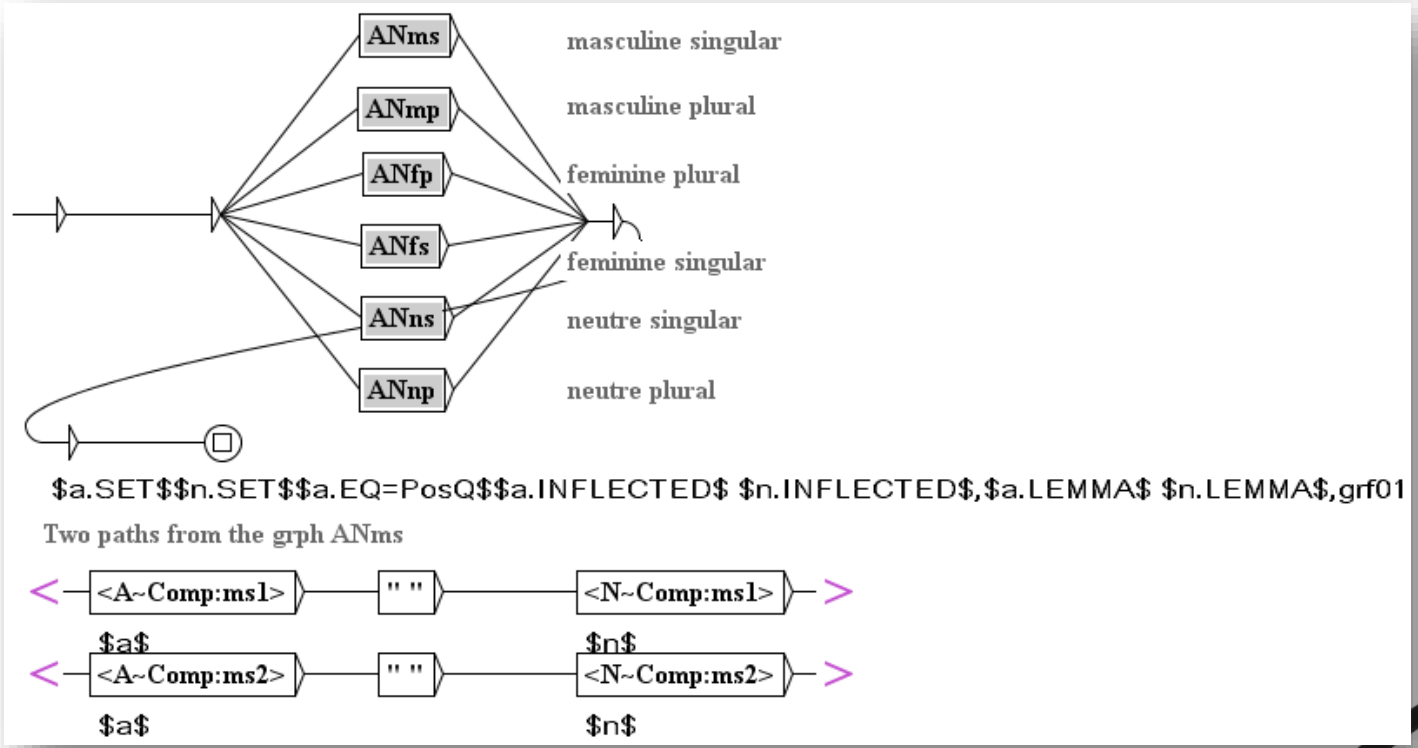
keystone

Examples of term extraction

- AXN – an adjective followed by a noun; the adjective and the noun have to agree in all four grammatical categories; e.g. zemni gas ‘natural gas’.
- AXAXN – a noun preceded by two adjectives that agree with it in gender, number, case and animate-ness; e.g. površinski istražni radovi ‘surface exploration works’.
- NNgiPrepNp - a noun followed by a noun in the genitive case and a prepositional phrase (as in case 4b); e.g. priprema ležišta za otkopavanje ‘deposit preparation for mining’.
- NNgiNgiNgi - a noun followed by three nouns/adjectives in the genitive case; e.g. istraživanje ležišta mineralnih sirovina ‘exploration of mineral deposits’.
- NprepNpNgi - a noun followed by a prepositional phrase; e.g. bakar sa primesama zlata ‘copper with a sprinkling of gold’.
- 2XAXN - an adjective followed by a noun that agrees in all four grammatical categories and preceded by a word that does not inflect in the MWU; e.g. magmatsko-eruptivni masiv ‘magmatic-igneous massif’.

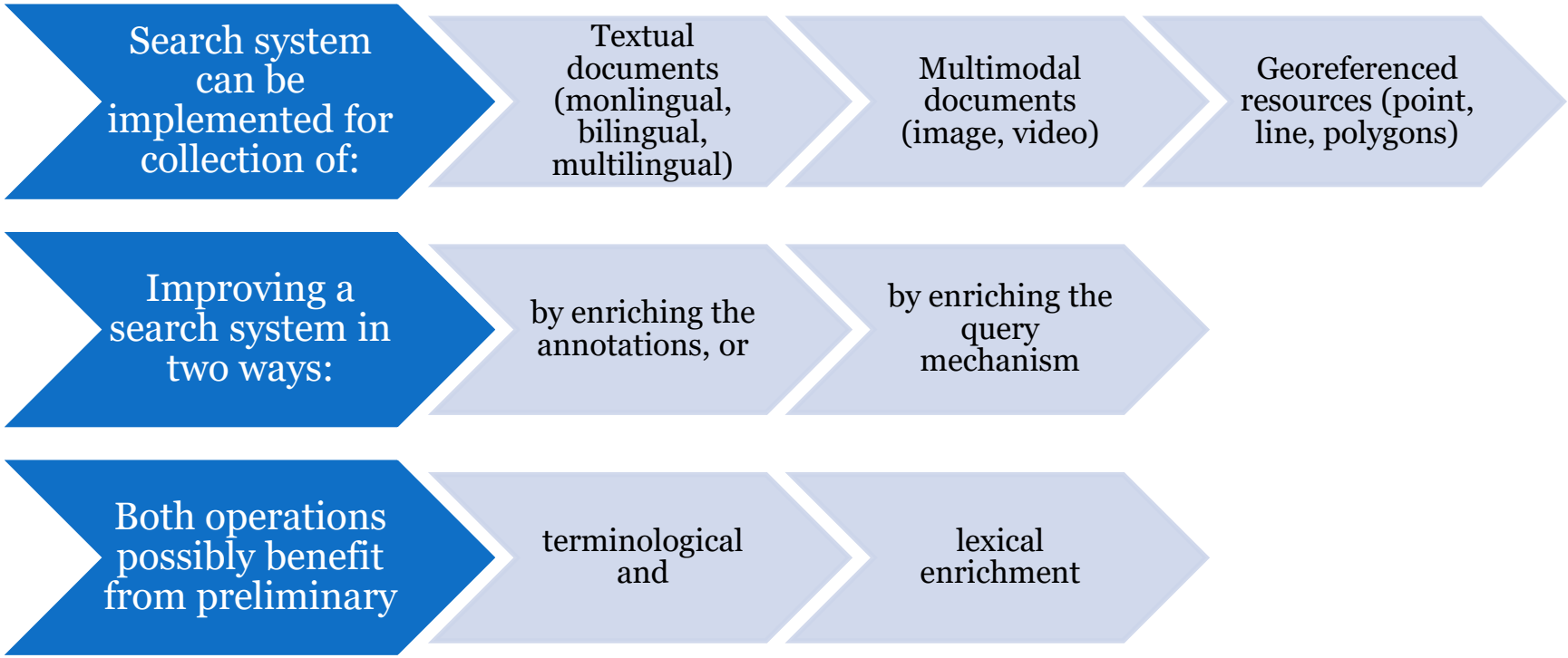
FST for extraction of MWUs of type AXN

- Two paths from one of the subgraphs that illustrate the agreement between adjectives and nouns
- Dictionary variable used for FST output in the form $a.LEMMA$ retrieves a lemma of recognized word form a thus performing the simple word lemmatization



keystone

Interaction Between Automatic Annotation and Query Expansion



Veronique Malais, Laura Hollink, and Luit Gazendam, The Interaction Between Automatic Annotation and Query Expansion: a retrieval experiment on a large cultural heritage archive, SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/

keystone

Enriching the annotations

- Manual annotation leads to a low number of keywords per document and improvement can be in
  - Facilitating manual creation of annotations
  - Creation of semi-automatic annotations
  - Automatically created annotations.
- Tools can be used for semi-automatic semantic annotation, extracted from text resources.
- Automatically generated annotations seldom reach the quality level of manual annotations.

Annotation and Query expansion

- Retrieval of not only documents that match the query concept, but also documents that are annotated with concepts that are related to the query.
- Ontology based query expansion
- What is the effect of query expansion in the context of automatic annotation?
- Is query expansion still beneficial when applied to lower-quality automatic annotations?
- And is it still necessary if a larger number of annotations is generated?
- Case study:
  1. Compute a baseline by querying a corpus of hand-made metadata.
  2. Query the automatically generated annotations of the same corpus.
  3. Query the hand-made metadata using query expansion.
  4. Query the automatically generated annotations using query expansion.

Creating annotations

- The term “annotation” implies, very generally speaking, to attach data to some other piece of data.
- Applied to different domains
  - Document annotations
  - Semantic Wikis
  - Semantic Blogs
  - Tagging
- Annotations create a relationship between URIs and build up a network of data.
- The Semantic Web is about shared terminology, achieved through consistent use of URIs.

http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf

Document annotations

Annotations can be

- manual (performed by one or more people),
- semi-automatic (based on automatic suggestions),
- or fully automatic.

Manual annotation tools

- allow users to add annotations to web pages or other resources, and share these with others.
- An example annotation would relate the text “Santiago de Compostela” to an ontology, identifying it as a city, as a capital of Galicia.

Automatic tools

- can perform similar annotations (such as named-entity recognition) without manual intervention.

Semantic Wikis

(Regular) Wikis

- Are collaborative hypertext authoring environments for collaborative writing and editing
- enable users to describe resources in natural language,

Semantic Wikis

- allow users to make formal descriptions of resources by annotating the pages that represent those resources.
- enable users to additionally describe resources in a formal language.

Adding metadata

- to ordinary Wiki content improves
- retrieval, information exchange, and knowledge reuse

E. Oren. SemperWiki: a semantic personal Wiki. In SemDesk in ISWC. 2005.

keystone

Semantic Blogs

Blogs (or weblogs)

- are online journals or diaries, usualy individual posts, created and presented in reverse chronological order.

An annotation in blogs

- a statement about a post or a category
- For example: classify posts with categories like “sports”, “cinema” or “Novak Djokovic”

Semantic Blogging

- annotations are extended, and allow association on an ontological basis.

keystone

Tagging

- Tagging systems (e.g. Flickr, Facebook, Linkedin,…) allow users add tags to a web resources
- Tags express some unspecified relation between the resource and whatever the term refers to.
- Token (words) tagging in text
  - to connect with gramatical features
  - to (related) concepts
  - to other language(s) equivalents

keystone

Approaches

- Annotation approaches types:
  - completeness of the result (i.e. how well does it capture the real-world situation) and
  - commitment to the result (i.e. usability, understanding).
- For example,
  - tags require little effort and result in high commitment (through the collaborative tagging), but
  - they have a low completeness (one can not make complex statements about the real world, but only assign shallow tags).

Types of annotations

- Informal annotations,
- Formal annotations,
  - that have formally defined constituents and are thus machine-readable, and
- Ontological annotations,
  - that have formally defined constituents and use only ontological terms that are socially accepted and understood.

keystone

Oren, Eyal, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. "What are semantic annotations." Relatório técnico. DERI Galway 9 (2006): 62.

Definition 1 (Annotation). *An annotation A is a tuple (a_s, a_p, a_o, a_c), where a_s is the subject of the annotation (the annotated data) a_o is the object of the annotation (the annotating data) a_p is the predicate (the annotation relation) that defines the type of relationship between a_s and a_o, and a_c is the context in which the annotation is made.*

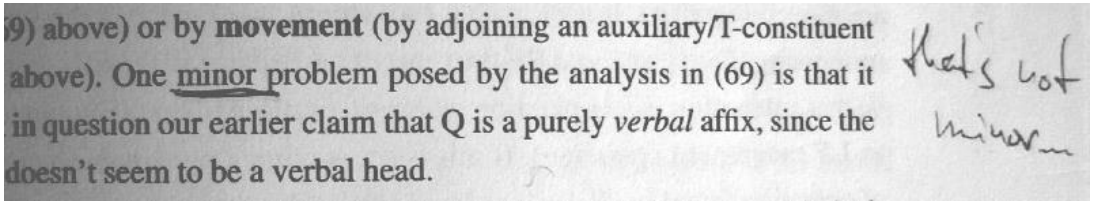
Definition 2 (Formal annotation). *A formal annotation A_f is an annotation A, where the subject a_s is a URI, the predicate a_p is a URI, the object a_o is a URI or a formal and the context a_c is a URI.*

Definition 3 (Ontological annotation). *A ontological annotation A_s is a formal annotation A_f, where the predicate a_p and the context a_c are an (arbitrarily complex) ontological term, and the object a_o conforms[10] to an ontological definition of a_p.*

keystone

Examples

- Informal



- Formal

```
<http://papers.org/minimalism#minor>
 <disagree> "that's not minor!".
```

- Ontological

```
<http://papers.org/minimalism#minor>
  ibis:con
    [ rdf:type ibis:Argument;
      rdf:label "that's not minor!" ].
```


keystone

How to classify annotations

- Association
  - way an annotation is associated with the annotated resource - whether the annotation is embedded in the annotated resource, or references the resource externally.
- Subject granularity
  - granularity of the annotation subject: e.g. is the annotation about a document, a section inside a document, a sentence, or a word?
- Representation distinction
  - whether the tool distinguishes annotations about documents from annotations of the concept described in or otherwise related to the document?
- Terminology reuse (“heterogeneity” , “interoperability”)
  - whether an annotation is self-confined with its own terminology, or whether an annotation uses terms from (one or more) existing ontologies, and are thus interoperable and understandable for others.
- Object type (“annotation form”)
  - type of annotation object: is it a literal or textual object, a structural object (including a hyperlink to another page), or an ontological object?
- Context context of the anno
  - when was it made, by whom, and within what scope: the annotation could for example be temporally scoped (it is only valid in 2016) or spatially scoped (it is only valid in Spain).
  - If the annotation is not about a document, then the context could also be the document the annotation is derived from

Named entity recognition

- **Named entity recognizers** identify proper names in documents, and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like.
- In the sentence:
  - **Italy**'s business world was rocked by the announcement **last Thursday** that **Mr. Verdi** would leave his job as vice-president of **Music Masters of Milan, Inc** to become operations director of **Arthur Andersen**.
- 'Italy' would be identified as a place, 'last Thursday' as a date, 'Verdi' as a person, 'Music Masters of Milan, Inc' and 'Arthur Andersen' as companies.
- Some would consider recognition of 'Milan' as a place, and identifying 'Arthur Andersen' as a person as an error in this context.

Slides about NE compiled from from Cvetana Krstev presentation Named Entities

Named entities and coreferences

- MUC defined a coreference task as linking together multiple expressions that refer to a given entity.
- In the context of information extraction, the role of coreference annotation is to ensure that information associated with multiple mentions of an entity can be collected together.
- For instance,
  - **<coref id='100'>**International Business Machines **</coref>**
  - **<coref id='101' type='ident' ref='100'>**IBM**</coref>**
- The acronym **IBM** refers to the identical notion as the phrase **International Business Machines**.

Some examples of **ENAMAX** tags/3

Country name is a part of a name of an organization:

- **<ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.</ENAMEX>**

Country name is not a part of a name of an orgnaization

- **<ENAMEX TYPE="ORGANIZATION">Hyundai, Inc.</ENAMEX> of <ENAMEX TYPE="LOCATION">Korea</ENAMEX>**

City name is not a part of a name of an university

- **<ENAMEX TYPE="ORGANIZATION">University of California</ENAMEX> in <ENAMEX TYPE="LOCATION">Los Angeles</ENAMEX>**

Compound expressions in which place names are separated by a comma are to be tagged as separate instances of **LOCATION**

- **<ENAMEX TYPE="LOCATION">Kaohsiung</ENAMEX>, <ENAMEX TYPE="LOCATION">Taiwan</ENAMEX>**

Some examples of TIMEX tags

Time

- **<TIMEX TYPE="TIME">twelve o'clock noon</TIMEX>**
- **<TIMEX TYPE="TIME">5 p.m. EST</TIMEX>**

Date

- **<TIMEX TYPE="DATE">January 1990</TIMEX>**
- **<TIMEX TYPE="DATE">fiscal 1989</TIMEX>**
- **the <TIMEX TYPE="DATE">autumn</TIMEX> report (?)**
- **<TIMEX TYPE="DATE">third quarter of 1991</TIMEX>**
- **<TIMEX TYPE="DATE">the fourth quarter ended Sept. 30</TIMEX>**

Some examples of NUMEX tag

monetary expressions:

- **<NUMEX TYPE="MONEY">20 million New Pesos</NUMEX>**
- **<NUMEX TYPE="MONEY">$42.1 million</NUMEX>**
- **<NUMEX TYPE="MONEY">million-dollar</NUMEX> conferences**

percentage

- **<NUMEX TYPE="PERCENT">15 pct</NUMEX>**

Named Entity Categories and TEI

- One chapter of TEI (Text Encoding Initiative) guidelines is dedicated to named entities:
  - **P5: Guidelines for Electronic Text Encoding and Interchange**
  - Chapter 13: **Names, Dates, People, and Places**
- Elements and their attributes are described in this chapter that can be used when a special TEI module is included **namesdates** – without it only basic elements can be used, for instance for names those are **name** and **rs**.

Person names in TEI

- <persName>
  - <surname>
  - <forename>
  - <roleName>
  - <addName>
  - <nameLink>
  - <genName>

- Examples:

```
<persName key="DUDO1">
  <roleName type="honorific" full="abb">Mme
  </roleName>
  <nameLink>de la</nameLink>
  <surname>Rochefoucault</surname>
</persName>
<persName>
  <forename>Charles</forename>
  <genName>II</genName>
</persName>
```

Geopolitical names in TEI

- <placeName>
  - <district>
  - <settlment>
  - <region>
  - <country>
  - <bloc>

- Examples:

```
<placeName key="LSEA1">
  <country type="nation">Laos</country>,
  <bloc type="sub-continent">Southeast
Asia</bloc>
</placeName>
<placeName>

<settlement type="city">Rochester</settlement>,
  <region type="state">New York</region>
</placeName>
```

Organization names in TEI

- **<orgName>** - Examples:
  - **About a year back, a question of considerable interest was agitated in the <orgName key="PAS1" type="voluntary"> <placeName key="PEN">Pennsyla.</placeName> Abolition Society</orgName>....**
  - **A spokesman from <orgName type="regional"> <orgName type="acronym">IBM</orgName> <country type="acronym">UK</country> </orgName> said ...**

Problems with NER

- Many referring expressions are proper names and may therefore exhibit initial capital letters in English (and many other European languages), e.g., **John Smith**, **Thomson Corporation** and **Los Angeles**.
- The presence of an initial capital does not guarantee that one is dealing with part of a name, since initial capitalization is also used:
  - at the start of sentences,
  - Variables in mathematics, chemical symbols, **X-rays**,...
  - Acronyms that are not named entites (**FC** – for football club)
  - Acronyms in short messages: **OMG** (Oh, my God), etc.
- Also, for some named entities no initial capital letter is used, e.g. **eBay**.

NER system for Serbian

- Entities that are tagged belong to classes:
  - **Person names** (full names and distinguished person names) their titles, roles and functions, if present, preceding or following them;
  - **Geopolitical names** – countries and settlments – **geographic names** – water bodies and oronyms.
  - **Organization names** – including names of political parties.
  - **Number expressions – monetary, measurements, count, percentage**
  - **Time expressions – dates, times of day, periods and frequencies,** absolute and relative

General resources used for the Serbian NER

- Comprehensive morphological e-dictionaries of Serbian in DELA/DELAF format:
  - simple words,
  - Multi-word names;
- including:
  - general lexica,
  - geographic names,
  - personal names,
  - encyclopedic knowledge (in development).
- Dictionary entries are provided with elaborate semantic markers.

Examples of Dictionary Entries

- Geographic names:
  - **Dunav,N+NProp+Top+Hyd** (Danub is a proper name, geographic notion, hydronym)
  - **Atlanski okean,N+NProp+Top+Hyd** (Atlantic Ocean)
- Geopolitical names:
  - **Madrid,N+NProp+Top+Gr** (**Madrid** – a proper name, city)
  - **Španija,N+NProp+Top+Dr** (Spain– a proper name, country)
- Organizations:
  - **Atinska novinska agencija,N+NProp+Org+Acr=ANA**
- Person names:
  - **Venizelos,N+NProp+Hum+Last+Cel** (a last name of a famous person)
  - **Riga od Fere,N+NProp+Hum+Last+Cel** (a full name of a famous person)

The general approach – rule-based supported by lexical resources

- Use of dictionaries
- Use of local grammars to specify the context
  - For rejecting false recognitions
  - For accepting false rejections
- A task: recognition and tagging of **hydronyms** (water bodies) in Serbian newspaper texts.
- Problems: hydronyms are ambiguous with:
  - other geographic names: ***Bosna*** – a river and a region.
  - personal names: ***Una*** – a river and a feminine name, ***Sava*** – a river and a masculine name
  - Common nouns: ***Kupa*** – a river, and but als a form of a noun ***kup***, a verb ***kupati***...

The first solution

- We use as a text a small collection of news dealing with recent floods in Serbia in 2014 named ***Poplave*** (~10.000 simple words)
- For retrieving names of water bodies we use a pattern:
  - **<N+NProp+Top+Hyd>**
- All names of water bodies (recorded in e-dictionaries) but also a number of false recognitions:
  - **Oko** – a preposition 'around' and a form of a name **Oka**
- 89 matches / 7 false matches

keystone

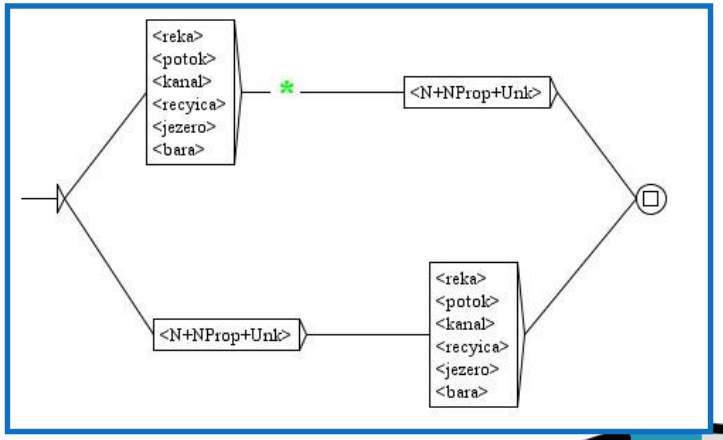
The improvement

- Use local grammar that take into consideration only simple word hydronym names that are not ambiguous with other proper or common names.
- Recognize MWU hydronym names in dictionaries (usually they are not ambiguous)
- Use context that have (river, lake, on the bank of..., hydropower on...)
- Use context that have („has flooded“...)
- Check context –if it is in list with already recognized woter body names.

keystone

The last improvement

- We try to retrieve some more entries – even those uppercase unknown words but with an obligatory key word following or preceding it
  - 76 matches in a collection **Poplave** / no false recognitions / 82 correct hydronym names
  - differences from a previous recognition
  - Example: *Tulovska reka,* (reka) *Lugomir*



Cvetana Krstev, Ivan Obradović, Miloš Utvić, Duško Vitas, "A system for named entity recognition based on local grammars", J Logic Computation 24(2), pp. 473-489, 2014, Oxford Journals, doi:10.1093/logcom/exs079, first published online February 19, 2013

Прелиминарна техноекономска студија услова и могућности експлоатације лежишта угља Черевић Нови Сад, Фрушка Гора, Рударски институт Београд. Ову студију је требало урадити на основу података из Елабората о прорачуну резерви угља Ц2 категорије између Беочина и Баноштора. Институт за грађевинарство Суботица, Миливој Макар, дипл. инж. руд. Беочин. Истраживано подручје налази се на северним падинама Фрушке Горе, између Беочина на истоку и Баноштора на западу. Угљени слојеви се даље настављају према западу до Корушке. Студија о хидрогеолошким истраживањима у зони између Параћина и Главице у циљу отварања новог изворишта (I фаза). Београд Др Војислав Томић, доцент, Невен Крешић, дипл.инж.

Прелиминарна техноекономска студија услова и могућности експлоатације лежишта угља <top.gr>Черевић</top.gr> <top.gr>Нови Сад</top.gr>, <top.geo>Фрушка Гора</top.geo>, <org>Рударски институт <top.gr>Београд</top.gr></org>.
Ову студију је требало урадити на основу података из Елабората о прорачуну резерви угља Ц<amount.exact>2 категорије</amount.exact> између <top.gr>Беочина</top.gr> и ... <pers><persName.full>Др Војислав Томић</persName.full><role>, доцент</role></pers>, <persName.full>Невен Крешић</persName.full>

keystone

Tools

- Manual annotation tools: Annotea, OntoMat, COHSE, WebAno,...
- (semi-)automatic annotation GATE, Unitex, NooJ,...
- Statistical tagging: treetagger, Stanford POS tagger ,....
- Statistical natural language processing and corpus-based computational linguistics:
  - An annotated list of resources
    - Tools: Machine Translation, POS Taggers, NP chunking, Sequence models, Parsers, Semantic Parsers/SRL, NER, Coreference, Language models, Concordances, Summarization, ...
    - Corpora: Large collections, Particular languages, Treebanks, Discourse, WSD, Literature, Acquisition
    - SGML/XML
    - Dictionaries, Lexical/morphological resources
    - Courses, Syllabi, and other Educational Resources
    - Mailing lists
    - Other stuff on the Web: General, IR, IE/Wrappers, People, Societies

Semantic Annotation of Texts with RDF Graph Contexts

- The semantic annotation of texts consists in extracting semantic relations between domain relevant terms in texts.
- Several studies address the problem of capturing complex relations from texts.
- They combine statistical and linguistic analyses.
- In the semantic annotation generation, the aim is to identify existing relations, belonging to the domain ontology, within instances in texts and to complete them with the description of the domain concepts related by these identified relations.

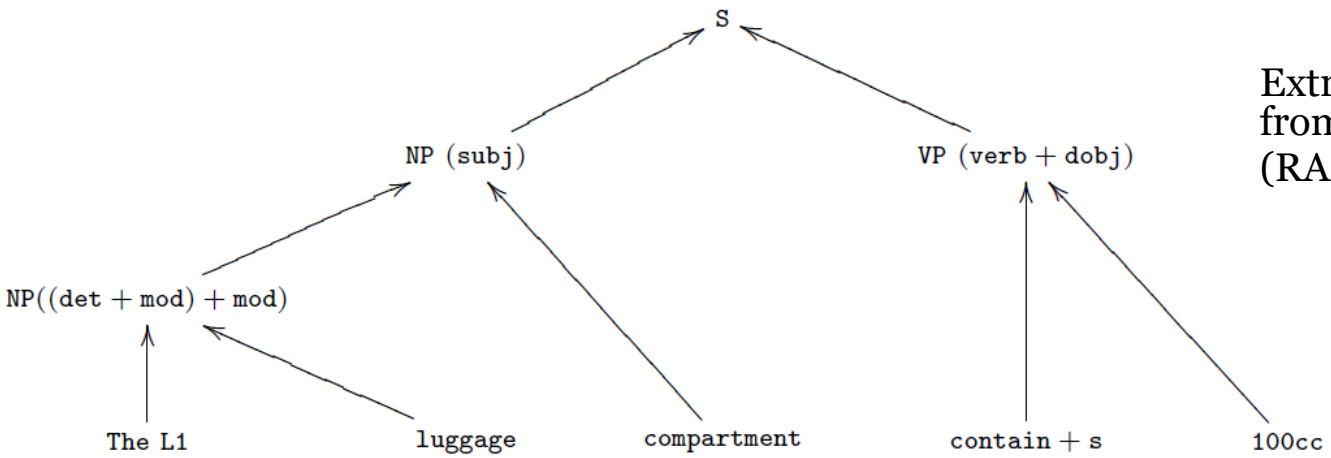
Cherfi, H., Corby, O., Faron-Zucker, C., Khelif, K., & Nguyen, M. T. (2008, July). Semantic Annotation of Texts with RDF Graph Contexts. In ICCS Supplement (pp. 75-82).

Semantic Annotation of Texts with RDF Graph Contexts

- Mapping between grammatical elements of each sentence in the analysed text and the corresponding entities in the dedicated-domain ontology.
  - the detection of relations described in a domain ontology,
  - the detection of terms linked by the identified relations based on term linguistic roles (subject, object, etc.) in the sentence, and
  - he generation of a corresponding annotation of the analysed text.
- Distinguishing between the ontological level and the instance level when linking a term in the text to the ontology: a term is identified to an instance of a concept rather than to the concept itself;
- Enriching the extracted instances of conceptual relations with contextual knowledge.
- Corese semantic search engine: RDF graph-based knowledge representation, SPARQL anf RDF contextual metadata (contexts).

keystone

NLP-Driven Semantic Annotation of Texts



Extraction of relations from texts
(RASP for english)

The L1 luggage compartment contains 100cc.	
RASP syntactic tree analysis	RDF annotation
("S" ("NP" ("NP" "The" "L1") "luggage" "compartment") ("VP" "contain::s" ("NP" "100cc")) ".")	<spro:Luggage_compartment rdf:about="#L1"> <spro:contain> <spro:Capacity rdf:about="#100cc" /> </spro:contain> </spro:Luggage_compartment>

Mapping of grammatical constituents to RDF triples (RASP output » RDF triples)

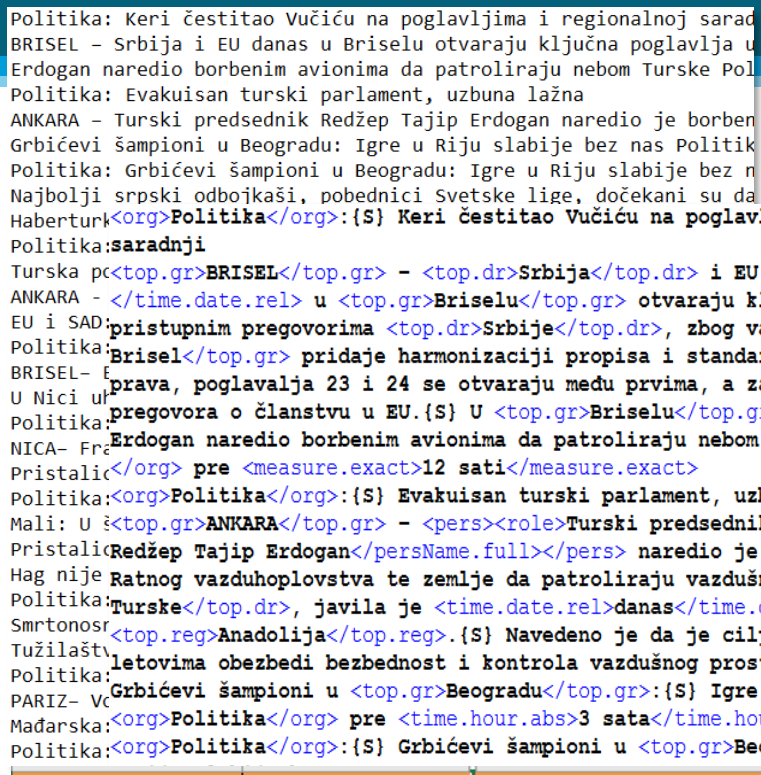
keystone

NLP-Driven Semantic Annotation of Texts

- S – V – O (sentence in active form)
  - O – V – S (sentence in passive form)
  - subordinate phrases “independent” from the main sentence and rhetorical relations
  - ambiguous subject/object constituents

- constitutes are major problem and may lead to a deadlocks when querying with SPARQL
  - “RDF graph context” with recursive capability for more expressive structures

keystone



<org>**Politika**</org>:{S} **Keri čestitao Vučiću na poglavljima i regionalnoj saradnji**

<top.gr>**BRISEL**</top.gr> **–** <top.dr>**Srbija**</top.dr> **i EU** <time.date.rel>**danas**</time.date.rel> **u** <top.gr>**Briselu**</top.gr> **otvaraju ključna poglavlja u pristupnim pregovorima** <top.dr>**Srbije**</top.dr>**, zbog važnosti koju** <top.gr>**Brisel**</top.gr> **pridaje harmonizaciji propisa i standarda u oblasti vladavine prava, poglavlja 23 i 24 se otvaraju među prvima, a zatvaraju tek pred kraj pregovora o članstvu u EU.**{S} **U** <top.gr>**Briselu**</top.gr>**...** »

Erdogan naredio borbenim avionima da patroliraju nebom Turske <org>**Politika**</org> **pre** <measure.exact>**12 sati**</measure.exact>

<org>**Politika**</org>:{S} **Evakuisan turski parlament, uzbuna lažna**

<top.gr>**ANKARA**</top.gr> **–** <pers><role>**Turski predsednik**</role><persName.full>**Redžep Tajip Erdogan**</persName.full></pers> **naredio je borbenim avionima Ratnog vazduhoplovstva te zemlje da patroliraju vazdušnim prostorom** <top.dr>**Turske**</top.dr>**, javila je** <time.date.rel>**danas**</time.date.rel> **agencija** <top.reg>**Anadolija**</top.reg>.{S} **Navedeno je da je cilj da se patrolnim letovima obezbedi bezbednost i kontrola vazdušnog prostora, prenosi TAS..**

Grbićevi šampioni u <top.gr>**Beogradu**</top.gr>:{S} **Igre u Riju slabije bez**

<org>**Politika**</org> **pre** <time.hour.abs>**3 sata**</time.hour.abs>

<org>**Politika**</org>:{S} **Grbićevi šampioni u** <top.gr>**Beogradu**</top.gr>:{S}

CategoryID	SubCategID	Caption in Engish	Frequency
person	pers	Who is it?	20
person	function	What does he/she do?	5
organization	organization	What is it?	25
location	location	At which place/where?	42
count	measure	How much?	7
count	money	How much money?	0
count	amount	How much of what?	4
count	percent	What percentage?	0
time	date	Which day?	10
time	hour	What time?	5
time	duration	How long does it take?	0
time	frequency	How many times repeated	0

CategID	SubCategID	Caption in Engis	Ran	Tagged Named Entity	Frequenc
person	pers	Who is it?	1	Dade Vujasinović Smrtonosne p	2
person	pers	Who is it?	2	Stefanović	2
person	pers	Who is it?	3	Vujasinović	2
person	pers	Who is it?	4	Vujanović	2
person	pers	Who is it?	5	Redlep Tajip Erdogan	1
person	pers	Who is it?	6	Nikola Tesla	1
person	pers	Who is it?	7	Nikole Grbića	1
person	pers	Who is it?	8	Akin Ozturk	1
person	pers	Who is it?	9	Sinišu Malog	1
person	pers	Who is it?	10	Radoslava Dada	1
person	function	What does he/she do?	1	Turski predsednik	1
person	function	What does he/she do?	2	selektora	1
person	function	What does he/she do?	3	bivši komandant turskog ratnog	1
person	function	What does he/she do?	4	gradonačelnika	1
person	function	What does he/she do?	5	savetnik mađarskog premijera z	1
organizatior	organization	What is it?	1	Politika	18
organizatior	organization	What is it?	2	EU Politika	1
organizatior	organization	What is it?	3	Evropska unija	1
organizatior	organization	What is it?	4	NATO	1
organizatior	organization	What is it?	5	Skupštinom grada	1
organizatior	organization	What is it?	6	Više javno tulilaštvo u Beogradu	1
organizatior	organization	What is it?	7	Tulilaštvo	1
organizatior	organization	What is it?	8	IS Politika	1
location	location	At which place/where?	1	Nici	5
location	location	At which place/where?	2	SAD	3
location	location	At which place/where?	3	Beogradu	3
location	location	At which place/where?	4	Beograd	3
location	location	At which place/where?	5	Turske	2
location	location	At which place/where?	6	Turska	2
location	location	At which place/where?	7	Mađarska	2
location	location	At which place/where?	8	BRISEL	2
location	location	At which place/where?	9	Briselu	2
location	location	At which place/where?	10	ANKARA	2
count	measure	How much?	1	11 sati	1
count	measure	How much?	2	12 sati	1
count	measure	How much?	3	40 minuta	1
count	measure	How much?	4	6 sati	1
count	measure	How much?	5	5 sati	1
count	measure	How much?	6	7 sati	1
count	measure	How much?	7	9 sati	1
count	amount	How much of what?	1	dvoje albanskih drlavljana	1
count	amount	How much of what?	2	84 osobe	1
count	amount	How much of what?	3	50 osoba	1
count	amount	How much of what?	4	devet ljudi	1
time	date	Which day?	1	danas	7
time	date	Which day?	2	četvrtak	1
time	date	Which day?	3	juče	1
time	date	Which day?	4	prošle nedelje	1
time	hour	What time?	1	3 sata	1
time	hour	What time?	2	4 sata	1
time	hour	What time?	3	2 sata	1
time	hour	What time?	4	uveče	1
time	hour	What time?	5	jutros	1

keystone

Extending Full Text Search Engine for Mathematical Content

- Index and search for mathematical content on the WWW using full text search engine
- Linearization, transformation rules, generalisation rules and ordering algorithm simplify the complex and highly symbolic mathematical structures into linear structures with well-defined symbols

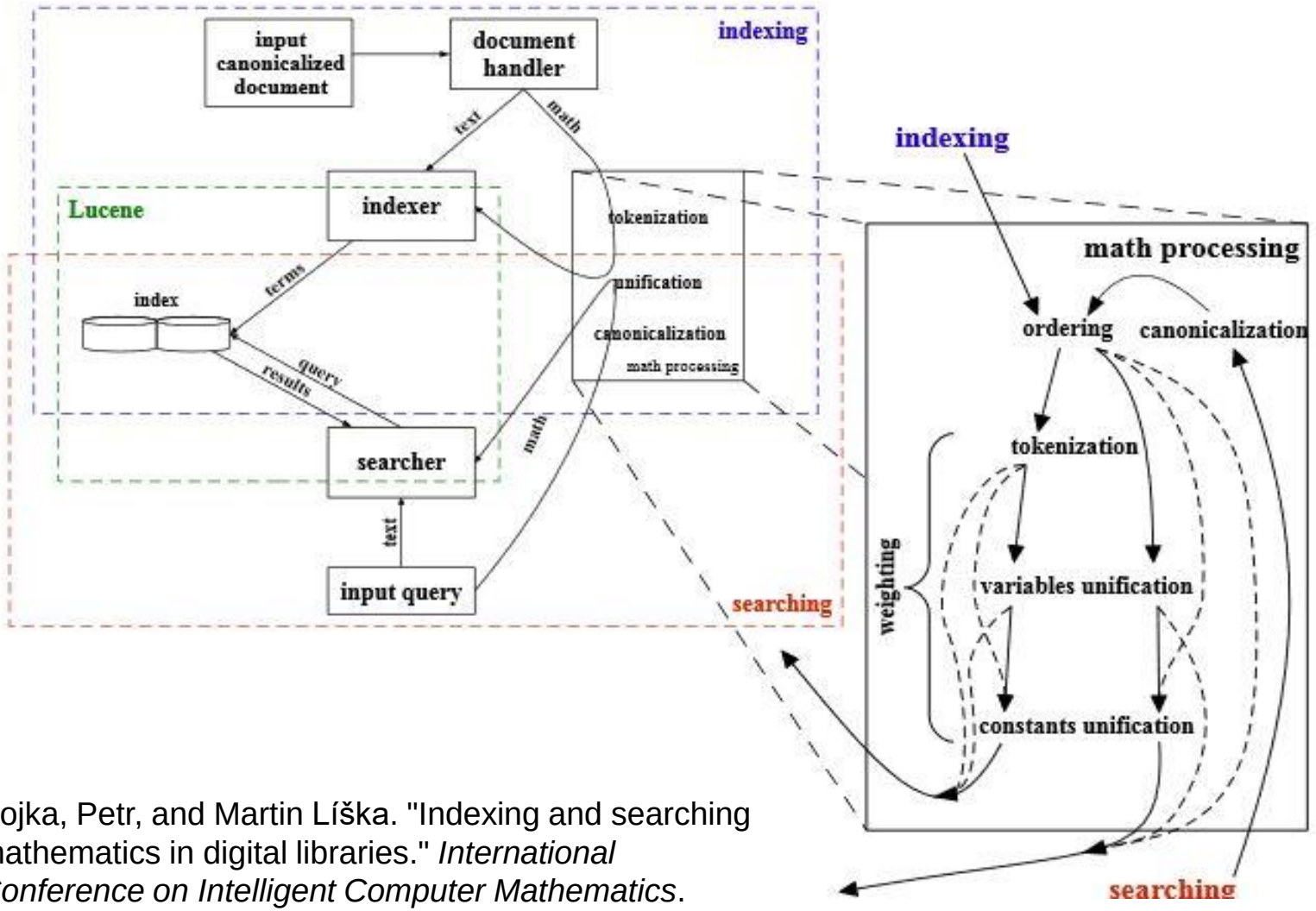
1. *Partial evaluation:* $7 + a + 5 \xrightarrow{(\text{converted to})} 12 + a$
2. *Approximate numerical constants:* $5.82 \doteq 6$
3. *Remove brackets using distributivity:* $a * (b + c) \rightarrow a * b + a * c$
4. *Multiply tokens:* $\frac{a+b}{2} * \Pi \rightarrow \frac{\Pi a + \Pi b}{2}$
5. *Assign each numerator its own denominator:* $\frac{\Pi a + \Pi b}{2} \rightarrow \frac{\Pi a}{2} + \frac{\Pi b}{2}$
6. *Replace constants with* const *symbol:*
   $74 + a^2 + b^2 \rightarrow \text{const} + a^{\text{const}} + b^{\text{const}}$
7. *Replace unknown constants, variables with* id *symbol:*
   $a^2 - b^2 + 2bc \rightarrow id_1^2 - id_1^2 + 2id_1id_2$
   $\text{or} \rightarrow id_1^2 - id_2^2 + 2id_1id_2 \ldots$

keystone

Extending Full Text Search Engine for Mathematical Content

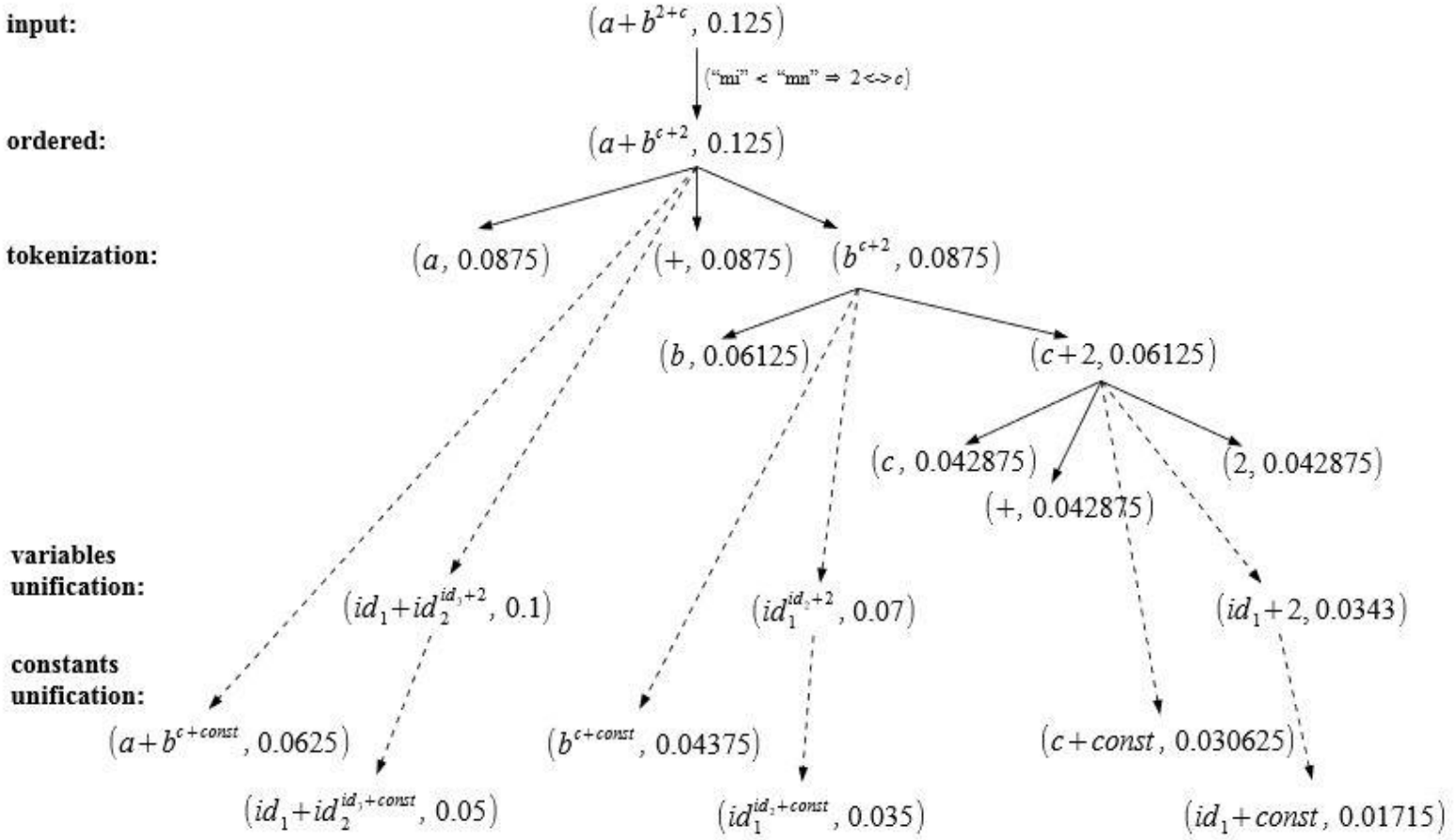
- Searching phase is the only user interactive phase of a search engine
- User enters a query (in LaTex) which is executed and the results are displayed.
- This includes several steps:
  - 1) query parsing,
  - 2) mapping query operators to supported internal constructs,
  - 3) finding all words/phrases from the query,
  - 4) evaluating the logic of the query and collecting suitable documents,
  - 5) sorting them according to their rank,
  - 6) displaying the result list.
- The mathematical extension is part of 1), 2) and 6).

keystone



Sojka, Petr, and Martin Líška. "Indexing and searching mathematics in digital libraries." *International Conference on Intelligent Computer Mathematics*. Springer Berlin Heidelberg, 2011.

keystone


input:
$(a+b^{2+c}, 0.125)$
("mi" < "mn" ⇒ 2<->c)
ordered:
$(a+b^{c+2}, 0.125)$
tokenization:
$(a, 0.0875)$
$(+, 0.0875)$
$(b^{c+2}, 0.0875)$
$(b, 0.06125)$
$(c+2, 0.06125)$
$(c, 0.042875)$
$(+, 0.042875)$
$(2, 0.042875)$
variables unification:
$(id_1+id_2^{id_3+2}, 0.1)$
$(id_1^{id_2+2}, 0.07)$
$(id_1+2, 0.0343)$
constants unification:
$(a+b^{c+const}, 0.0625)$
$(b^{c+const}, 0.04375)$
$(c+const, 0.030625)$
$(id_1+id_2^{id_3+const}, 0.05)$
$(id_1^{id_2+const}, 0.035)$
$(id_1+const, 0.01715)$

keystone

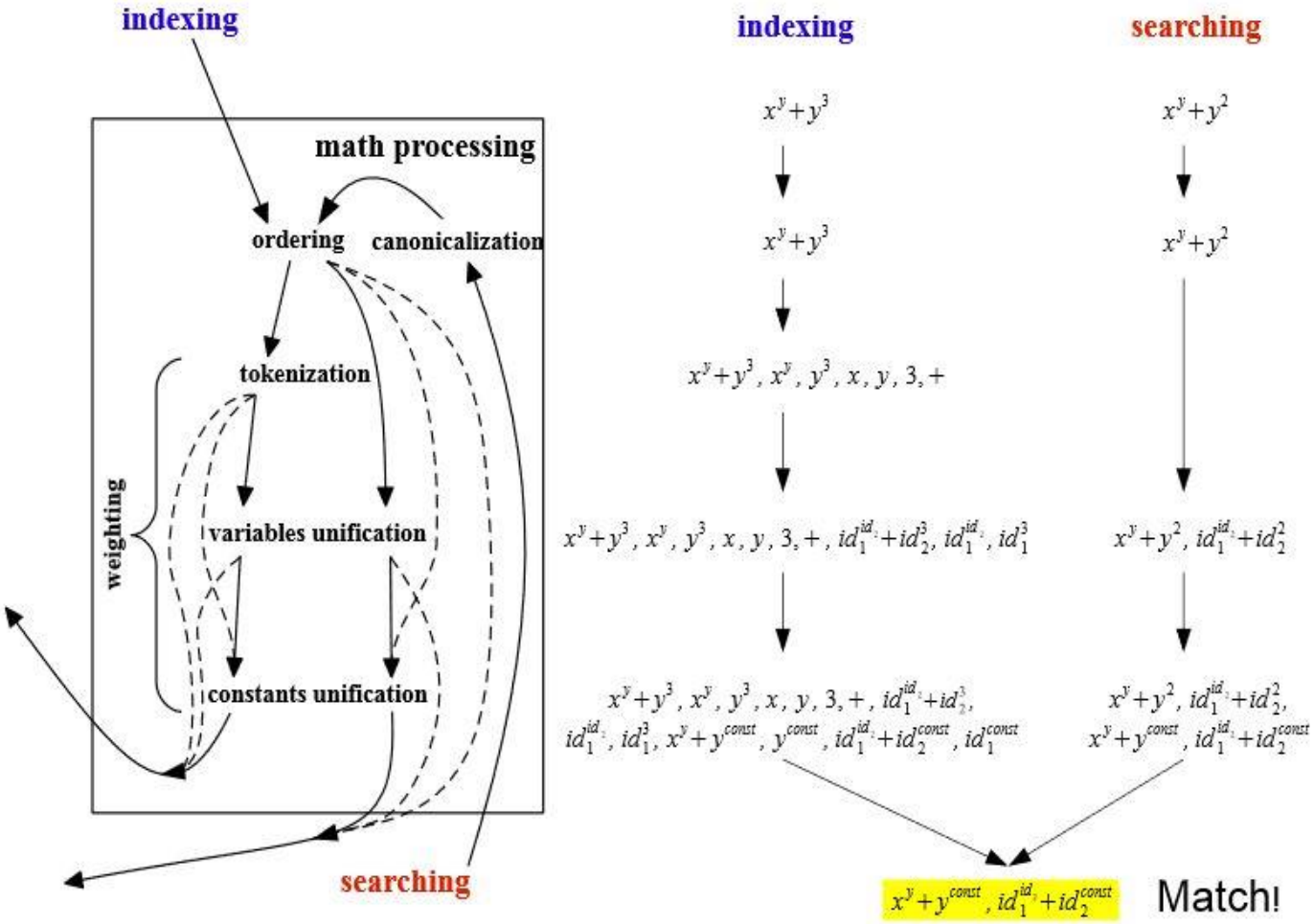


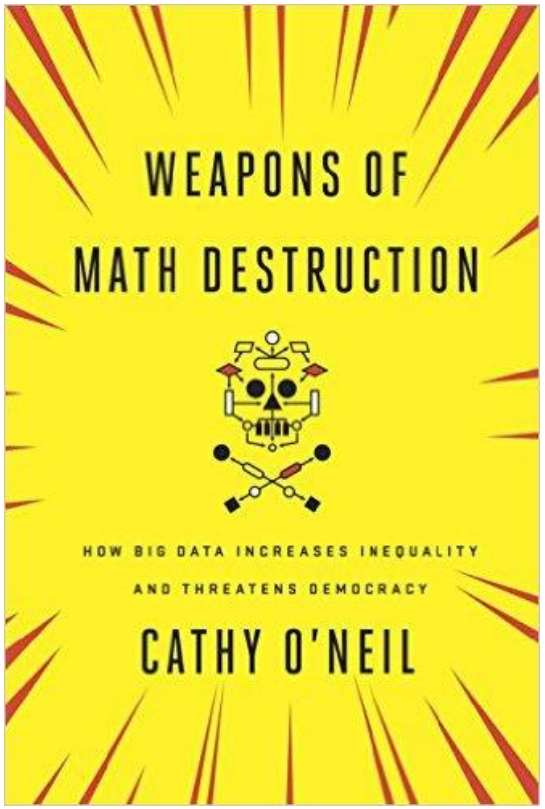
Figure 3: Math-aware search in MIaS

keystone

System	Input documents	Internal representation	Approach	α-eq.	Query language	Queries	Indexing core
MathDex	HTML, TEX/LATEX, Word, PDF	Presentation MathML (as strings)	syntactic	✗	?	text, math, mixed	Apache Lucene
LeActiveMath	OMDoc, OpenMath	OpenMath (as string)	syntactic	✗	OpenMath (palette editor)	text, math, mixed	Apache Lucene
LATEXSearch	LATEX	LATEX (as string)	syntactic	✗	LATEX	titles, math, DOI	?
MathWeb Search	Presentation MathML, Content MathML, OpenMath	Content MathML, OpenMath (substitution trees)	semantic	✔	QMath, LATEX, Mathematica, Maxima, Maple, Yacas styles (palette editor)	text, math, mixed	Apache Lucene (for text only)
EgoMath	Presentation MathML, Content MathML, PDF	Presentation MathML trees (as strings)	mixed	✔	LATEX	text, math, mixed	EgoThor
MIaS	any (well-formed) MathML	Canonical Presentation MathML trees (as compacted strings)	math tree similarity/ normalization	✔	AMS-LATEX or MathML	text, math, mixed	Apache Lucene/ Solr

- EuDML – European digital mathematics library https://eudml.org/
- WebMIaS is a web interface for Math Indexer and Searcher (MIaS) math aware searching engine, https://mir.fi.muni.cz/webmias-demo/
- Enhancing Searching of Mathematics, http://www.dessci.com/en/reference/searching/
- WolframAlfa

Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy



Cathy O'Neil

“Far too many algorithms in use today are being used as weapons against populations, whether they are consumers, workers, prisoners, or teachers.
'll talk about a few which I consider the worst kind - and which I call Weapons of Math Destuction - namely, those that are opaque, widespread, and powerful enough to cause tremendous destruction through feedback loops. I will also discuss suggestions for data scientists, policy makers, and the public for how to combat them.”

keystone

For more details see

- Attardi, G., S. Di Marco and F. Sebastiani. 1998. Automated Generation of Category-Specific Thesauri for Interactive Query Expansion.
- Grefenstette, G. 1992. Use of Syntactic Context to Produce Term Association Lists for Text Retrieval. In *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark*, ed. N. Belkin, P. Ingwersen and A. M. Pesjtersen: pp. 89-97. New York: ACM Press.
- Ide, E. 1971. New Experiments in Relevance Feedback. In G. Salton. *The SMART Retrieval System: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Qiu, Y., 1993. Concept Based Query Expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval.*
- Schutze, H. and J. Pederson. 1997. A Cooccurance-based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management* 33, no. 3: pp. 307-318.
- Walker, D. 2001. Query Expansion Using Thesauri.

keystone

Thank you for your attention

Hvala na pažnji

Хвала на пажњи

keystone