Automatic Text Analysis by Artificial Intelligence

Dunja Mladenić[2] and Marko Grobelnik
Jožef Stefan Institute, Artificial Intelligence Laboratory, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: dunja.mladenic@ijs.si, marko.grobelnik@ijs.si, http://ailab.ijs.si/,

[2]Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

Keywords: text mining, natural language processing, machine learning, data visualization, big data

Received: November 7, 2012

Text is one of the traditional ways of communication between people. With the growing availability of text data in electronic form, handling and analysis of text by means of computers gained popularity. Handling text data with machine learning methods brought interesting challenges to the area that got further extended by incorporation of some natural language specifics. As the methods were capable of addressing more complex problems related to text data, the expectations become bigger calling for more sophisticated methods, in particular a combination of methods from different research areas including information retrieval, machine learning, statistical data analysis, data mining, natural language processing, semantic technologies. Automatic text analysis become an integral part of many systems, pushing boundaries of research capabilities towards what one can refer to as an artificial intelligence dream - never ending learning from text aiming at mimicking ways of human learning. The paper presents development of text analysis research in Slovenian that we have been personally involved in, pointing out interesting research problems that have been and are still addressed by the research, example tasks that have been addressed and some challenges on the way.

Povzetek: V članku je predstavljen razvoj raziskav analize besedil z metodami umetne inteligence v Sloveniji.

1 Introduction

Word expressed as a sound is known as a fundamental phenomena in creation of our world. “Every element of the universe is in a constant state of vibration manifested to us as light, sounds and energy. The human senses perceive only a fraction of the infinite range of vibration, so it is difficult to comprehend that the Word mentioned in the Bible is actually the totality of vibration which underlines and sustains the creation.” [1] . Written words are one of the traditional ways of communication over space and time. “Communication is a gift to know. Communication is a gift to understand. Communication is a gift to realize.” [2] . To understand what has been communicated by some text is not always easy. Automatic text analysis can often contribute to understand the text, to gaining knowledge from the data provided in textual form, to realize the underlying facts that have been communicated via the text.

As electronic media become widely used, the amount of texts in electronic form has grown rapidly and is still growing. While these texts are primarily aiming at human readers, it is not uncommon to use computer programs to manipulate texts. Text handling by computer programs has a wide range of usage from enabling text editing, storing and indexing text for searching and retrieval, ranking documents, classifying documents, extracting information and knowledge, question answering, etc.

In this paper we present development of artificial intelligence research in Slovenia related to handling of text data that we have been personally involved in. The paper points out some interesting research problems that have been and are still addressed, listing some example tasks that have been addressed in our group and some challenges on the way. We conclude by providing discussion and some direction for future research on automatic text analysis.

2 Handling text data

In the 1990s handling of text data by machine learning techniques was inspired mainly by information retrieval, where machine learning methods were used primarily for classification of documents regarding their relevance to a given query (as an alternative to the information retrieval ranking methods). At that time, machine learning was also applied for personalized information delivery on text, such as, learning to filter relevant Netnews, suggesting potentially relevant hyperlinks on Web documents [3] , [4] , browsing the Web [5] , powering intelligent agents [6] . As texts (documents, Web pages, news articles) are often manually labeled by some topic category (e.g., a news on acquisitions, a Web page on artificial intelligence), this is a natural area for applying machine learning methods to train a classifier for topic classification. The problem is far from being a trivial application of machine learning methods to a new domain. The number of classes may get much larger than what was usual at the time for machine learning methods to handle, requiring a careful handling of efficient classifier construction [7] and pruning the space of classifiers to be consulted at the classification of a new example [8] .

Using words as features is, in such a setting, a common way of representing text documents so that machine learning methods can be applied on them. As each word from the vocabulary is assigned a feature with its value being based on the frequency of the word in a document, the number of features easily got to several tens of thousands. Moreover, one can think about some more sophisticated features beyond single words, such as sequences of words [9] , additionally increasing the feature space. This requires careful handling of the problem including efficient feature selection [10] .

Even though many relevant problems can and have been addressed at the level of documents using machine learning methods and at the level of sentences and words using natural language processing methods, there is still a way to go towards automatically obtaining knowledge from text to be used for ontology extension and reasoning. Extracting knowledge form the text and representing it in logical forms means that a computer can reason on it, provide hopefully some interesting insights and propose new conclusions. One of the earlier attempts included information extraction from Web pages using manually constructed wrappers and forming rules connecting the extracted information [11] . A step towards extraction of knowledge for ontology generation is presented in [12] , where natural language processing is used in combination with semantic technologies. While there are a number of similar efforts in direction of knowledge extraction from text, the problem of obtaining logical statements corresponding to some text remains open.

In general different methods from the area of artificial intelligence can be used for obtaining knowledge from text [13] ranging from classification and clustering, to association rule construction and visualization.

3 Example tasks

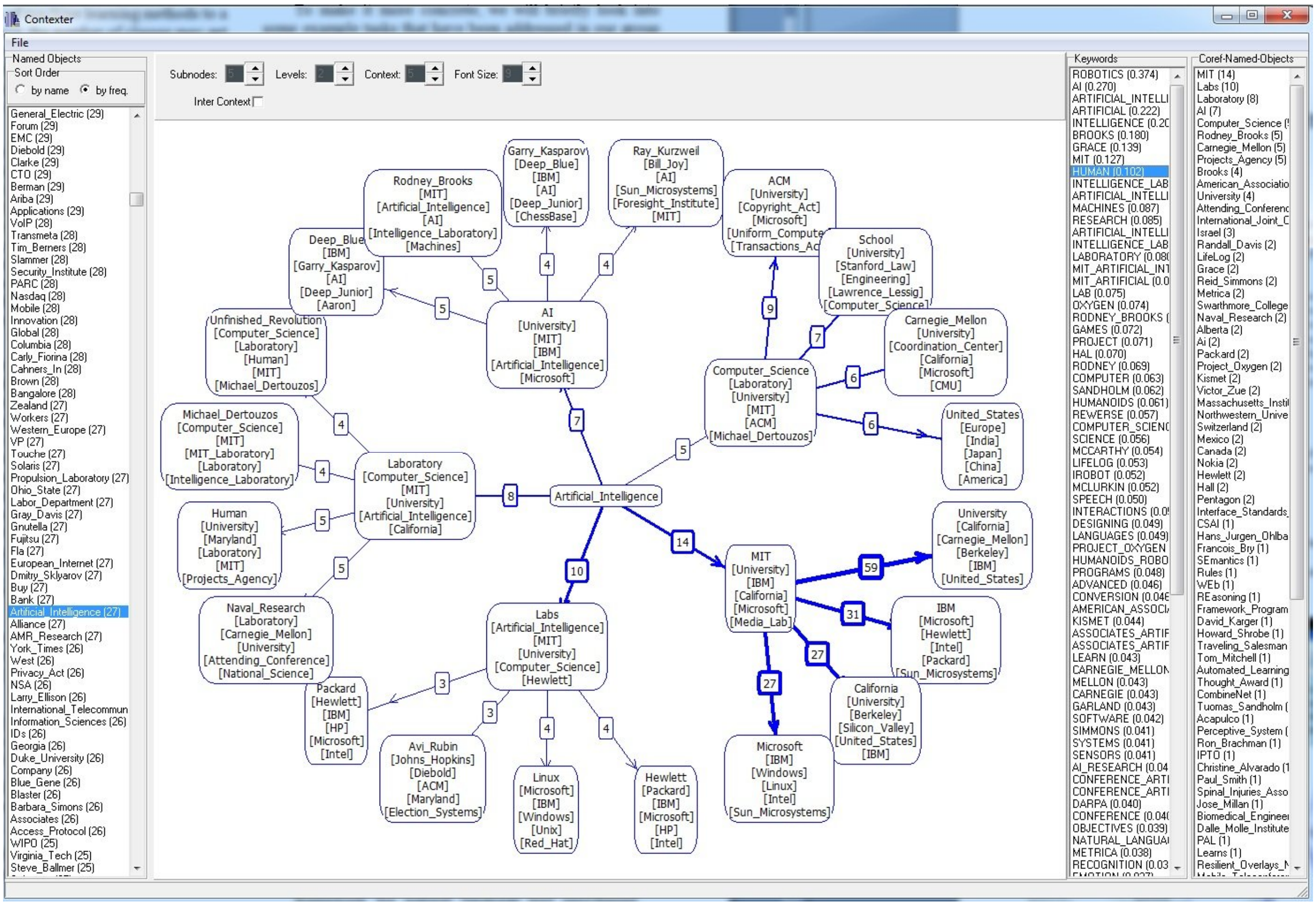
When we talk about applying artificial intelligence methods on text data, what we have in mind is a whole range of methods and problems that in some way involve analysis of text data. Many of these problems have been addressed in the area of Text Mining. For the definition of text mining we have adapted the definition of Data Mining, so we can say that text mining is about finding interesting regularities in large text data, where interesting means: non-trivial, hidden, previously unknown and potentially useful. Looking from the linguistic and semantic technologies perspective, text mining can be defined as finding semantic and abstract information from the surface form of text.

To be more concrete, we will briefly look into some example tasks that have been addressed in our group during the last twenty years by applying artificial intelligence methods on text data. These include:

- visualization of text available in news articles, visualization of named entities over time, visualization of document corpus, visualization of Web pages;
- triplet extraction from text, document representation using semantic graph, document summarization;
- text enrichment, contextual question answering;
- semi-automatic ontology construction from document corpus, ontology extension;
- knowledge extraction from text, text mining combined with social network analysis.

Visualization of text data available in news articles can be based on named entity extraction, as news are usually mentioning some named entities (e.g., people, countries, organizations) putting them in some relation. Visualization of news articles as proposed in [14] is based on extracting named entities from news articles and representing them in a graph (name entities being vertices connected if they appear in the same news article). The graph of entities is enriched with contextual information in the form of characteristic keywords and name entities related to the entity in the focus. Operations for browsing a graph are implemented to be efficient enabling interactive user experience with a quick capturing of large amounts of information present in the original text. Figure 1 shows the user interface on ACM Technology News consisting of 11000 article abstracts.

Named entities that have time information associated to them can be related to each other on a time scale. An approach relating people, places, organizations and events extracted from Wikipedia and linking them on a time scale is proposed in [15] . Relevant Wikipedia pages are identified by categorizing the articles as containing people, places or organizations. Then a timeline if generated linking the named entities and extracting events and their time frame.


Contexter
File
Subnodes: 5
Levels: 2
Context: 5
Font Size: 9
Inter Context □
Rodney_Brooks [MIT] [Artificial_Intelligence] [AI] [Intelligence_Laboratory] [Machines]
Garry_Kasparov [Deep_Blue] [IBM] [AI] [Deep_Junior] [ChessBase]
Ray_Kurzweil [Bill_Joy] [AI] [Sun_Microsystems] [Foresight_Institute] [MIT]
ACM [University] [Copyright_Act] [Microsoft] [Uniform_Compute [Transactions_Ac
School [University] [Stanford_Law] [Engineering] [Lawrence_Lessig] [Computer_Science]
Deep_Blue [IBM] [Garry_Kasparov] [AI] [Deep_Junior] [Aaron]
Unfinished_Revolution [Computer_Science] [Laboratory] [Human] [MIT] [Michael_Dertouzos]
AI [University] [MIT] [IBM] [Artificial_Intelligence] [Microsoft]
Carnegie_Mellon [University] [Coordination_Center] [California] [Microsoft] [CMU]
Computer_Science [Laboratory] [University] [MIT] [ACM] [Michael_Dertouzos]
Michael_Dertouzos [Computer_Science] [MIT] [MIT_Laboratory] [Laboratory] [Intelligence_Laboratory]
United_States [Europe] [India] [Japan] [China] [America]
Laboratory [Computer_Science] [MIT] [University] [Artificial_Intelligence] [California]
Artificial_Intelligence
Human [University] [Maryland] [Laboratory] [MIT] [Projects_Agency]
University [California] [Carnegie_Mellon] [Berkeley] [IBM] [United_States]
MIT [University] [IBM] [California] [Microsoft] [Media_Lab]
Naval_Research [Laboratory] [Carnegie_Mellon] [University] [Attending_Conference] [National_Science]
Labs [Artificial_Intelligence] [MIT] [University] [Computer_Science] [Hewlett]
IBM [Microsoft] [Hewlett] [Intel] [Packard] [Sun_Microsystems]
Packard [Hewlett] [IBM] [HP] [Microsoft] [Intel]
California [University] [Berkeley] [Silicon_Valley] [United_States] [IBM]
Avi_Rubin [Johns_Hopkins] [Diebold] [ACM] [Maryland] [Election_Systems]
Linux [Microsoft] [IBM] [Windows] [Unix] [Red_Hat]
Hewlett [Packard] [IBM] [Microsoft] [HP] [Intel]
Microsoft [IBM] [Windows] [Linux] [Intel] [Sun_Microsystems]
5
4
4
5
9
7
6
6
4
4
7
8
5
5
5
14
59
31
10
27
27
3
3
4
4

Named Objects
Sort Order
○ by name ● by freq.

General_Electric (29)
Forum (29)
EMC (29)
Diebold (29)
Clarke (29)
CTO (29)
Berman (29)
Ariba (29)
Applications (29)
VoIP (28)
Transmeta (28)
Tim_Berners (28)
Slammer (28)
Security_Institute (28)
PARC (28)
Nasdaq (28)
Mobile (28)
Innovation (28)
Global (28)
Columbia (28)
Carly_Fiorina (28)
Cahners_In (28)
Brown (28)
Bangalore (28)
Zealand (27)
Workers (27)
Western_Europe (27)
VP (27)
Touche (27)
Solaris (27)
Propulsion_Laboratory (27)
Ohio_State (27)
Labor_Department (27)
Gray_Davis (27)
Gnutella (27)
Fujitsu (27)
Fla (27)
European_Internet (27)
Dmitry_Sklyarov (27)
Buy (27)
Bank (27)
Artificial_Intelligence (27)
Alliance (27)
AMR_Research (27)
York_Times (26)
West (26)
Privacy_Act (26)
NSA (26)
Larry_Ellison (26)
International_Telecommun
Information_Sciences (26)
IDs (26)
Georgia (26)
Duke_University (26)
Company (26)
Blue_Gene (26)
Blaster (26)
Barbara_Simons (26)
Associates (26)
Access_Protocol (26)
WIPO (25)
Virginia_Tech (25)
Steve_Ballmer (25)

Keywords

ROBOTICS (0.374)
AI (0.270)
ARTIFICIAL_INTELLI
ARTIFICIAL (0.222)
INTELLIGENCE (0.20
BROOKS (0.180)
GRACE (0.139)
MIT (0.127)
HUMAN (0.102)
INTELLIGENCE_LAB
ARTIFICIAL_INTELLI
MACHINES (0.087)
RESEARCH (0.085)
ARTIFICIAL_INTELLI
INTELLIGENCE_LAB
LABORATORY (0.08
MIT_ARTIFICIAL_INT
MIT_ARTIFICIAL (0.0
LAB (0.075)
OXYGEN (0.074)
RODNEY_BROOKS (
GAMES (0.072)
PROJECT (0.071)
HAL (0.070)
RODNEY (0.069)
COMPUTER (0.063)
SANDHOLM (0.062)
HUMANOIDS (0.061)
REWERSE (0.057)
COMPUTER_SCIENC
SCIENCE (0.056)
MCCARTHY (0.054)
LIFELOG (0.053)
IROBOT (0.052)
MCLURKIN (0.052)
SPEECH (0.050)
INTERACTIONS (0.0
DESIGNING (0.049)
LANGUAGES (0.049)
PROJECT_OXYGEN
HUMANOIDS_ROBO
PROGRAMS (0.048)
ADVANCED (0.046)
CONVERSION (0.04
AMERICAN_ASSOCI
KISMET (0.044)
ASSOCIATES_ARTIF
ASSOCIATES_ARTIF
LEARN (0.043)
CARNEGIE_MELLON
MELLON (0.043)
CARNEGIE (0.043)
GARLAND (0.043)
SOFTWARE (0.042)
SIMMONS (0.041)
SYSTEMS (0.041)
SENSORS (0.041)
AI_RESEARCH (0.04
CONFERENCE_ARTI
CONFERENCE_ARTI
DARPA (0.040)
CONFERENCE (0.04
OBJECTIVES (0.039)
NATURAL_LANGUA
METRICA (0.038)
RECOGNITION (0.03

Coref-Named-Objects

MIT (14)
Labs (10)
Laboratory (8)
AI (7)
Computer_Science (
Rodney_Brooks (5)
Carnegie_Mellon (5)
Projects_Agency (5)
Brooks (4)
American_Associatio
University (4)
Attending_Conferenc
International_Joint_C
Israel (3)
Randall_Davis (2)
LifeLog (2)
Grace (2)
Reid_Simmons (2)
Metrica (2)
Swarthmore_College
Naval_Research (2)
Alberta (2)
Ai (2)
Packard (2)
Project_Oxygen (2)
Kismet (2)
Victor_Zue (2)
Massachusetts_Instit
Northwestern_Unive
Switzerland (2)
Mexico (2)
Canada (2)
Nokia (2)
Hewlett (2)
Hall (2)
Pentagon (2)
Interface_Standards_
CSAI (1)
Hans_Jurgen_Ohlba
Francois_Bry (1)
SEmantics (1)
Rules (1)
WEb (1)
REasoning (1)
Framework_Program
David_Karger (1)
Howard_Shrobe (1)
Traveling_Salesman
Tom_Mitchell (1)
Automated_Learning
Thought_Award (1)
CombineNet (1)
Tuomas_Sandholm (
Acapulco (1)
Perceptive_System (
Ron_Brachman (1)
IPTO (1)
Christine_Alvarado (1
Paul_Smith (1)
Spinal_Injuries_Asso
Jose_Millan (1)
Biomedical_Enginee
Dalle_Molle_Institute
PAL (1)
Learns (1)
Resilient_Overlays_N

Figure 1: *Contexter* graphical interface for browsing/visualizing the name-entity network. Showing context of named entity *Artificial Intelligence* (eg., occurring 5 times in the news collection with *Computer Science* that occurs 6 times with *Carnegie Mellon*. Among the most important keywords for the news where *Artificial Intelligence* occurs, we can see *robotics, AI, human, machines* etc.

General document corpus can be also visualized using clustering methods on text data [16] . Document corpus visualization can be further used in **Semi-automatic construction of topic ontology** using machine learning to cluster document, to map documents onto some existing ontologies, to suggest concept naming [17] .

In addition to addressing problems that focus on handling documents as the main units, it is also relevant to split texts into smaller units, such as, paragraphs, sentences, words or even characters. In this way one can **annotate text** on different levels of granularity including topic category of the whole document, extraction of facts mentioned in the text, named entity extraction and resolution (into some ontology such as, DBPedia, OpenCyc). Figure 3 shows an example output for a homepage annotations produced by Enrycher [18] , where the identified named entities are linked to concepts in DBPedia, OpenCyc and GeoNames. In addition, text of the homepage is assigned several topic categories from Open directory (by machine learning methods an efficient text classifier is constructed from Open directory). Enrycher also produces a semantic graph of text and extracts interesting statements from it, such as Dunja Mladenic is Enwise expert. Text annotations has been also used for **enhancing visualization of web pages** [19] by transforming the page to semantic graph, using machine learning to rank the triplets enabling page visualization as a semantic graph of a chosen size (see Figure 2).

Extracting triplets from text [20] involves some more or less sophisticated natural language processing to extract what would be considered as subject – predicate – object triplets from sentences. Even though the original approach uses parsing of a sentence to get its logical form (extracting subject-predicate-object) [21] , reasonable results have been achieved by using predefined patterns, such as noun phrase – verb phrase – noun phrase to extract triplets [18] . The extracted triplets can be also generalized to a kind of templates [22] , such as, country – borders – country that can be further used to extend an ontology or to extract information from text.

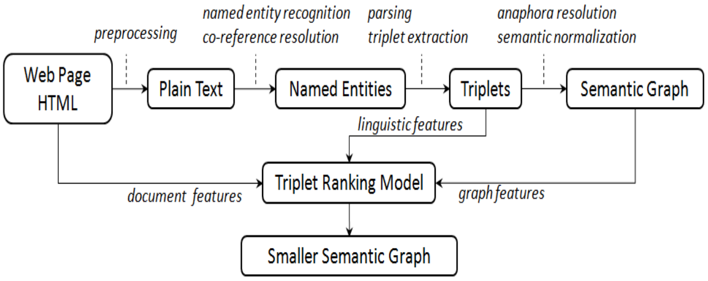
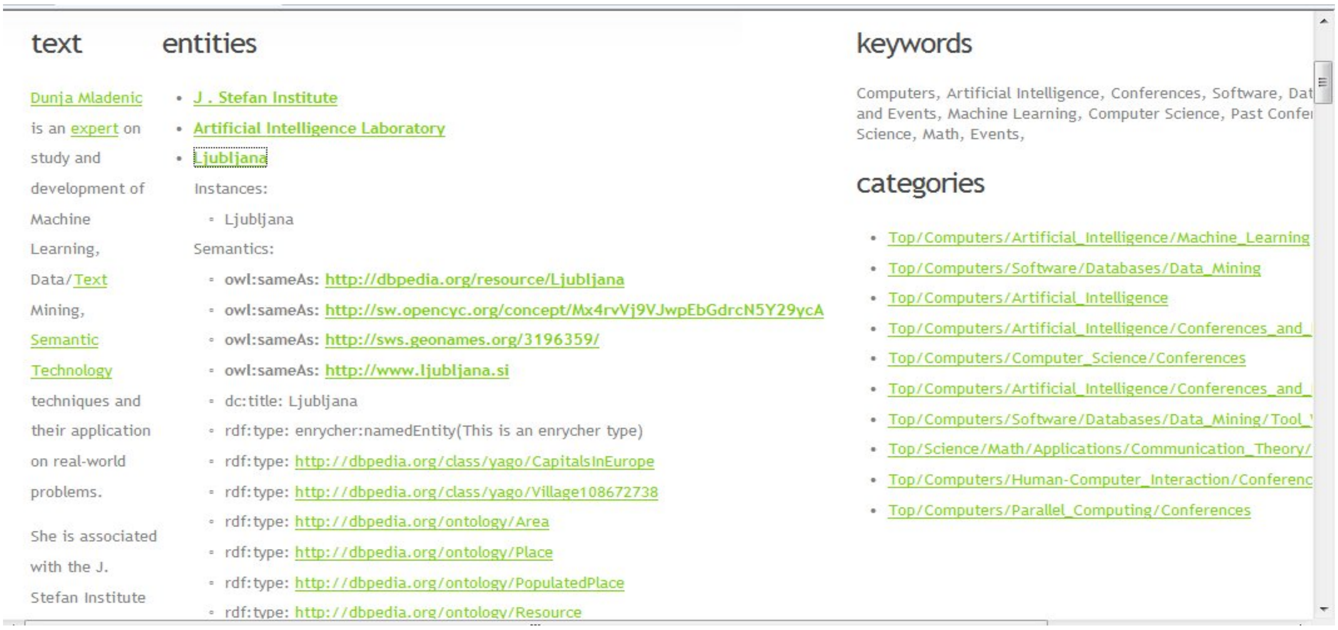


Figure 2: Steps in enhanced web page visualization.



text

Dunja Mladenic is an expert on study and development of Machine Learning, Data/Text Mining, Semantic Technology techniques and their application on real-world problems.

She is associated with the J. Stefan Institute

entities

- J . Stefan Institute
- Artificial Intelligence Laboratory
- Ljubljana

  Instances:
  - Ljubljana

  Semantics:
  - owl:sameAs: http://dbpedia.org/resource/Ljubljana
  - owl:sameAs: http://sw.opencyc.org/concept/Mx4rvVj9VJwpEbGdrcN5Y29ycA
  - owl:sameAs: http://sws.geonames.org/3196359/
  - owl:sameAs: http://www.ljubljana.si
  - dc:title: Ljubljana
  - rdf:type: enrycher:namedEntity(This is an enrycher type)
  - rdf:type: http://dbpedia.org/class/yago/CapitalsInEurope
  - rdf:type: http://dbpedia.org/class/yago/Village108672738
  - rdf:type: http://dbpedia.org/ontology/Area
  - rdf:type: http://dbpedia.org/ontology/Place
  - rdf:type: http://dbpedia.org/ontology/PopulatedPlace
  - rdf:type: http://dbpedia.org/ontology/Resource

keywords

Computers, Artificial Intelligence, Conferences, Software, Dat
and Events, Machine Learning, Computer Science, Past Confe
Science, Math, Events,

categories

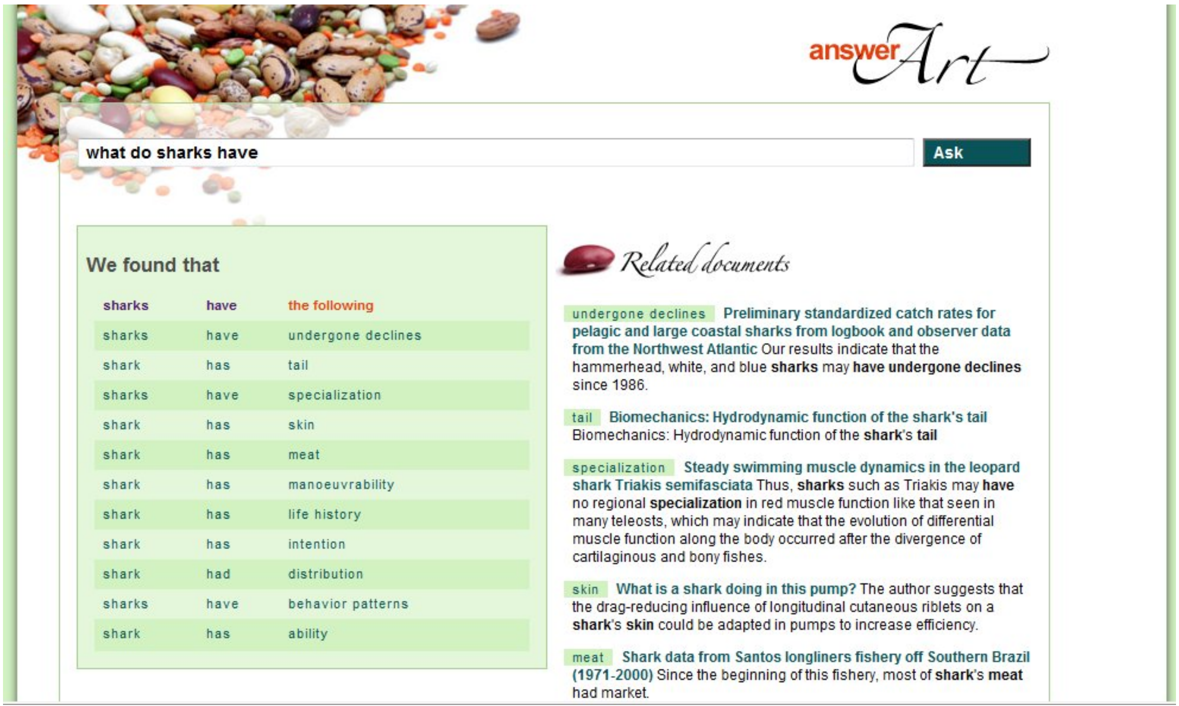
- Top/Computers/Artificial_Intelligence/Machine_Learning
- Top/Computers/Software/Databases/Data_Mining
- Top/Computers/Artificial_Intelligence
- Top/Computers/Artificial_Intelligence/Conferences_and_
- Top/Computers/Computer_Science/Conferences
- Top/Computers/Artificial_Intelligence/Conferences_and_
- Top/Computers/Software/Databases/Data_Mining/Tool_
- Top/Science/Math/Applications/Communication_Theory/
- Top/Computers/Human-Computer_Interaction/Conferenc
- Top/Computers/Parallel_Computing/Conferences

Figure 3: Enrycher providing text annotation on a text of homepage of Dunja Mladenić, linking entities to existing ontologies eg., Ljubljana is linked to concepts in DBPedia, Opencyc, GeoNames. The page is annotated by keywords (computers, artificial intelligence, etc.) and by topic categories from Open directory (eg., Top/Computer_Science/ Artificial_Intelligence/Machine_Learning).

Document summarization aims at construction of a shorter version of the original document. It can be performed using different approaches, one of them as proposed in [21] is based on extracting triplets from text to obtain semantic structure of a document feeding features to a machine learning classifier trained to classify triplets for being included in the document summary or not.

Question answering can be also based on triplet extraction [23] , [24] . The whole document collection used for finding the answers is transformed into a collection of triplets and the question transformed into triplet is matched against the collection of triplets. Figure 4 shows interface of Answer Art system responding on the question "What do sharks have?" by listing answers in the form of triplets (eg., sharks have tail) and enabling the user to access the related documents that were used to obtain the listed answers.

Ontology construction and extension is usually performed entirely manually or semi-automatically by



We found that

sharks	have	the following
sharks	have	undergone declines
shark	has	tail
sharks	have	specialization
shark	has	skin
shark	has	meat
shark	has	manoeuvrability
shark	has	life history
shark	has	intention
shark	had	distribution
sharks	have	behavior patterns
shark	has	ability

Related documents

undergone declines **Preliminary standardized catch rates for pelagic and large coastal sharks from logbook and observer data from the Northwest Atlantic** Our results indicate that the hammerhead, white, and blue **sharks** may **have undergone declines** since 1986.

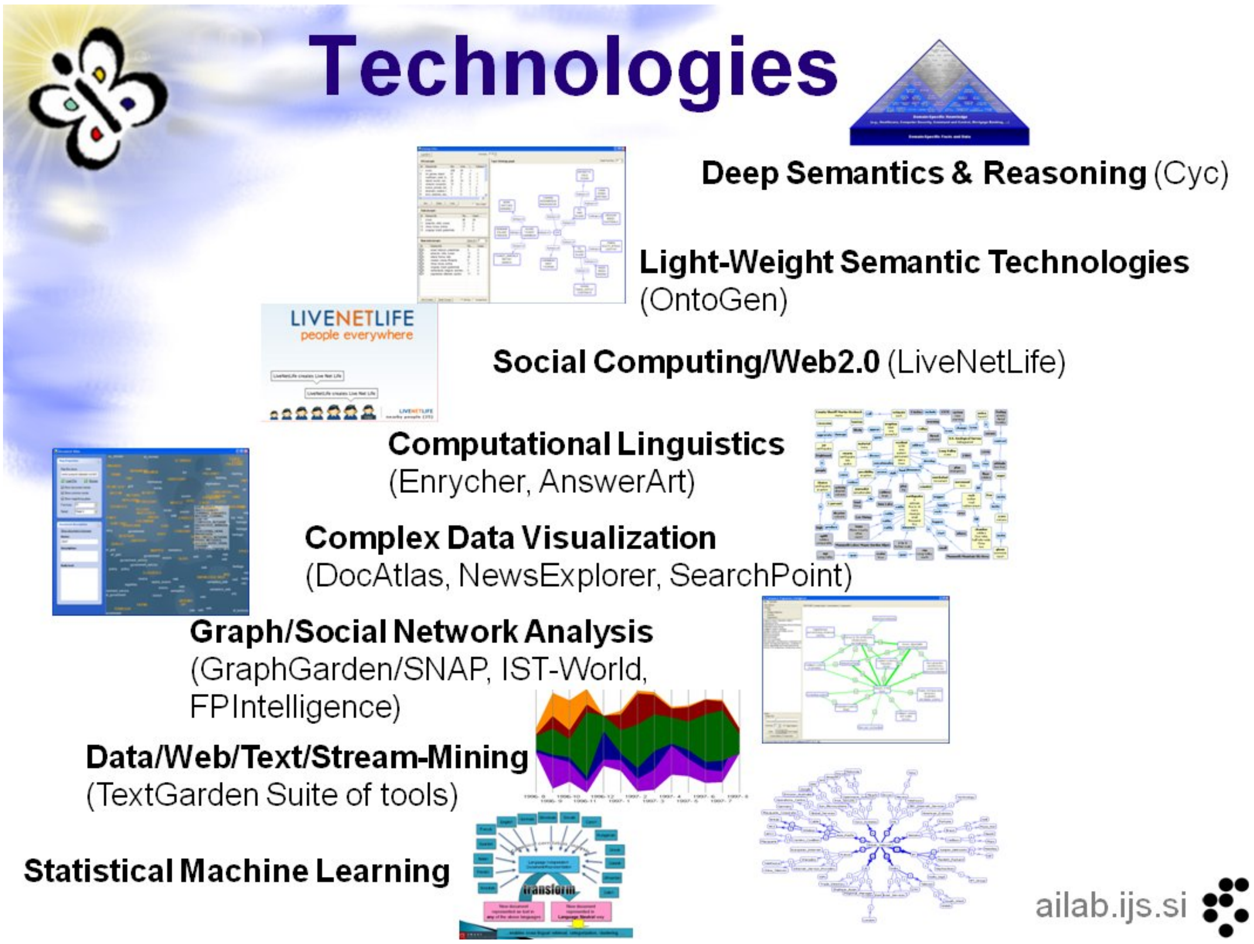
tail **Biomechanics: Hydrodynamic function of the shark's tail** Biomechanics: Hydrodynamic function of the **shark**'s **tail**

specialization **Steady swimming muscle dynamics in the leopard shark Triakis semifasciata** Thus, **sharks** such as Triakis may **have** no regional **specialization** in red muscle function like that seen in many teleosts, which may indicate that the evolution of differential muscle function along the body occurred after the divergence of cartilaginous and bony fishes.

skin **What is a shark doing in this pump?** The author suggests that the drag-reducing influence of longitudinal cutaneous riblets on a **shark**'s **skin** could be adapted in pumps to increase efficiency.

meat **Shark data from Santos longliners fishery off Southern Brazil (1971-2000)** Since the beginning of this fishery, most of **shark**'s **meat** had market.

Figure 4 *Answer Art* on question "*What do sharks have?*", based on documents on fisheries & aquaculture and ASFA ontology lists that *sharks undergo decline, have tail, have specialization, have skin, have meat* etc.



Technologies

Deep Semantics & Reasoning (Cyc)

Light-Weight Semantic Technologies (OntoGen)

Social Computing/Web2.0 (LiveNetLife)

Computational Linguistics (Enrycher, AnswerArt)

Complex Data Visualization (DocAtlas, NewsExplorer, SearchPoint)

Graph/Social Network Analysis (GraphGarden/SNAP, IST-World, FPIntelligence)

Data/Web/Text/Stream-Mining (TextGarden Suite of tools)

Statistical Machine Learning

ailab.ijs.si

Figure 5: Diagram showing different kind of technologies involving text data developed by Artificial Intelligence Laboratory at J. Stefan Institute.

applying some methods from artificial intelligence [25] . Annotation of text by the concepts of an existing ontology, as for instance used in Figure 3, is limited by the concepts that already exist in the ontology, unless we extend the existing ontology. Novel methodology for semi-automatic ontology extension aggregating the elements of text mining and user-dialog approaches is proposed in [26] . The domain of interest for ontology extension is defined by keywords and a glossary of relevant terms with term descriptions. Collaborative manual ontology extension can be supported by analysis of the dynamics of ontology changes over time. This can be especially useful when dealing with larger ontologies, where a number of editors having different expertise contribute to different parts of the ontology.

Methods from **social network analysis** can be used to gain some insights into editors' interaction with the ontology, their expertise and the ontology changes. An example approach proposed in [27] enables visualization of ontology concepts through the view of editors interacting with the concepts. Social network analysis in combination with analysis of text data can provide insights into research collaboration between institutions and countries, as proposed in [28] , where collaborations on European research projects is addressed.

Semantic technologies have been successfully applied to visualization of temporal data [30] and to support the users in dealing with information overload [31] . It was also recognized that by means of semantic technologies context of the data and the user may be used to support the user's personal productivity [32] . An approach to analysis of communication between individuals inside an organization using semantic technologies is proposed in [29] . The data on communication activity is first cleaned and transformed into a set of transactions reflecting the communication out of which a graph is constructed. The graph of transactions is represented as a matrix and fed into an tool for semi-automatic ontology construction. Out of the communication activity data an institutional ontology is constructed showing communities and important players inside the institution (eg., key people that are often involved in communication, isolated groups, well connected groups, etc.).

4 Discussion and future directions

Different Artificial Intelligence methods have been successfully applied on text data addressing a number of relevant problems. Figure 5 shows some of the technologies and the associated prototypes we have developed in our group at J. Stefan Institute ranging from statistical machine learning and data/web/text mining, to analysis of social networks and graphs, complex data visualization, computational linguistics, social computing, light-weighted semantic technologies and deep semantics with reasoning. As the methods in general become more sophisticated, the problems become more complex and researchers are constantly facing new challenges.

As an example, we can point out the fact that each text we have been handling is written is some natural language. The majority of artificial intelligence approaches focus on a single language, some handle multiple languages and other work in a cross-lingual setting adding to the complexity of the tasks and opening new challenges, as for instance in multilingual document retrieval [33] and multilingual sentiment analysis [34] . There are a number of open research challenges related to developing linguistic resources for different languages and covering multilingual and cross-lingual settings.

Another important direction of research is ensuring scalability of approaches as it is becoming common to deal with large data, also referred to as Big Data. Digging for knowledge is big data is very common goal, hoping that we will avoid traps of just noticing statistical artifacts instead of real, true phenomenon we are interested in revealing form the data. "…truth is simple, straight and with a smile. You don't have to remember it. You have to say it, you have to know it and then you have to live it."[2] .

Acknowledgements

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under a number of since 2000, the most recent being PASCAL2 (ICT-216886-NoE), RENDER (ICT-257790-STREP) and XLike (ICT-STREP-288342).

References

[1] Bhajan, Y. The Aquarian Teacher, pp.66, KRI, 2003.

[2] Bhajan, Y. Conscious Communication, KRI, 2006.

[3] Mladenić, D. Personal WebWatcher: Implementation and Design, *Technical Report IJS-DP-7472*, J. Stefan Institute, Slovenia, 1996.

[4] Joachims, T., Mladenić, D. Browsing-assistant, tour guides und adaptive WWW-server. *KI Journal, Künstl. Intell. (Oldenbourg)*, 1998, vol. 3, pp. 23-29.

[5] Mladenić, D. Web browsing using machine learning on text data. In *Intelligent exploration of the web*,111, New York; Heidelberg: Physica-Verlag, 2002, pp. 288-303.

[6] Mladenić, D. Text-learning and related intelligent agents : a survey. *IEEE intelligent systems and their applications*, 1999, vol. 14, pp. 44-54.

[7] Grobelnik, M, Mladenić, D. Simple classification into large topic ontology of web documents. *CIT. J. Comput. Inf. Technol.*, 2005, vol. 13, pp. 279-285.

[8] Mladenić, D. Turning Yahoo into an automatic Web-page classifier. In *Proceedings ECAI-1998*. Chichester [etc.]: John Wiley & Sons, 1998, pp. 473-474.

[9] Mladenić, D., Grobelnik, M. Word sequences as features in text-learning. In *Proceedings of the seventh Electrotechnik and Computer Science Conference ERK-1998, 1998*, Ljubljana: IEEE Region 8, Slovenian section IEEE, 1998, pp. 145-148.

[10] Mladenić, D., Grobelnik, M. Feature selection on hierarchy of web documents. *Decision support systems journal*, 2003, vol. 35, pp. 45-87.

[11] Ghani, R., Jones, R., Mladenić, D., Nigam, K., Slattery, S.. Data mining on symbolic knowledge extracted from the Web. In Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *KDD-2000 working notes : workshop on text mining,* Boston, MA, USA., 2000, pp. 29-36.

[12] Baxter, D., Klimt, B., Grobelnik, M, Schneider, D.I., Witbrock, M.J., Mladenić, D. Capturing document semantics for ontology generation and document summarization. In *Semantic knowledge management : integrating ontology management, knowledge discovery, and human language technologies*. Berlin; Heidelberg: Springer, cop. 2009, pp. 141-154.

[13] Grobelnik, M, Mladenić, D. Automated knowledge discovery in advanced knowledge management. Journal of knowledge management, 2005a, vol. 9, pp. 132-149.

[14] Grobelnik, M, Mladenić, D. Visualization of news articles. *Informatica (Ljublj.)*, 2004, 28:4, pp. 375-380.

[15] Bhole, A., Fortuna, B., Grobelnik, M, Mladenić, D. Extracting named entities and relating them over time based on Wikipedia. *Informatica (Ljublj.)*, 2007, 31:4, pp. 463-468.

[16] Fortuna, B., Mladenić, D., Grobelnik, M. Visualization of text document corpus. *Informatica (Ljublj.)*, 2005, 29:4, pp. 497-502.

[17] Fortuna, B., Grobelnik, M, Mladenić, D. OntoGen: semi-automatic ontology editor. *Lect. notes comput. sci.*, 2007, vol. 4558, pp. 309-318.

[18] Štajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenić, D., Grobelnik, M. A service oriented framework for natural language text enrichment. *Informatica (Ljublj.)*, 2010, 34:3, pp. 307-313.

[19] Dali, L., Rusu, D., Mladenić, D. Enhanced web page content visualization with Firefox. *Lect. notes comput. sci.*, 2009, lNAI 5782, pp. 718-721.

[20] Rusu, D., Fortuna, B., Grobelnik, M, Mladenić, D. Semantic graphs derived from triplets with application in document summarization. *Informatica (Ljublj.)*, 2009, 33:3, pp. 357-362.

[21] Leskovec, J., Grobelnik, M., Milic-Frayling, N., Learning Sub-structures of Document Semantic Graphs for Document Summarization, In *Proceedings of LinkKDD 2004* Workshop at KDD International conf.

[22] Sipoš, R., Mladenić, D., Grobelnik, M., Brank, J. Modeling common real-world relations using triples

extracted from n-grams, In the *Proc. of The Semantic Web Fourth Asian Conference, ASWC 2009*, Lecture Notes in Computer Science, 2009, 5926, pp. 16-30.

[23] Dali, L., Rusu, D., Fortuna, B., Mladenić, D., Grobelnik, M. *Question answering based on semantic graphs. WWW-2009 Workshop on Semantic Search 2009.*

[24] Bradeško, L., Dali, L., Fortuna, B., Grobelnik, M, Mladenić, D., Novalija, I., Pajntar, B. Contextualized question answering, In Proc. of ITI-2010.

[25] Novalija, I., Mladenić, D., Bradeško, L. OntoPlus : text-driven ontology extension using ontology content, structure and co-occurrence information. *Knowedge.-based systems*, 2011, 24:8, pp. 1261-1276.

[26] Novalija, I., Mladenić, D. Ontology extension towards analysis of business news. *Informatica (Ljublj.)*, 2010, 34:4, pp. 517-522.

[27] Tomašev, N., Mladenić, D. Social network analysis of ontology edit logs. *CIT. J. Comput. Inf. Technol.*, 2010, 18:2, pp. 191-200.

[28] Grobelnik, M, Mladenić, D. Analysis of a database of research projects using text mining and link analysis. In *Data mining and decision support : integration and collaboration*, Boston; Dordrecht; London: Kluwer Academic Publishers, 2003, pp. 157-166.

[29] Grobelnik, M, Mladenić, D., Fortuna, B.. Semantic technology for capturing communication inside an organisation. *IEEE internet computing*, 2009, 13:4, pp. 59-66.

[30] Fortuna, B., Mladenić, D., Grobelnik, M. Visualization of temporal semantic spaces. In *Semantic knowledge management : integrating ontology management, knowledge discovery, and human language technologies*. Berlin; Heidelberg: Springer, cop. 2009, pp. 155-169.

[31] Simperl, E., Thurlow, I., Warren, P., Dengler, F., Davies, J., Grobelnik, M, Mladenić, D., Gomez-Perez, J.M., Ruiz Moreno, C. Overcoming information overload in the enterprise : the active approach. *IEEE internet computing*, 2010, 14:6, pp. 39-46.

[32] Dolinšek, I., Grobelnik, M, Mladenić, D. Managing and understanding context. In *Context and semantics for knowledge management: technologies for personal productivity*. Heidelberg: Springer, 2011, pp. 91-106

[33] Rupnik, J., Muhič, A., Škraba, P. Multilingual Document Retrieval Through Hub Languages, In: *Proceedings of the Fifteenth International Multiconference Information Society 2012*. Ljubljana: Institut Jožef Stefan, 2012.

[34] Štajner, T., Novalija, I., Mladenić, D. Informal sentiment analysis in multiple domains for English and Spanish, In: *Proceedings of the Fifteenth International Multiconference Information Society 2012*. Ljubljana: Institut Jožef Stefan, 2012.