ARTIFICIAL INTELLIGENCE HANDLING TEXT DATA

Dunja Mladenić, Marko Grobelnik
Artificial Intelligence Laboratory, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
e-mail: dunja.mladenic@ijs.si, marko.grobelnik@ijs.si

ABSTRACT

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 got bigger calling for a combination of methods from different research areas including information retrieval, machine learning, statistical data analysis, data mining, natural language processing, semantic technologies. *Nowadays* automatic text analysis is an integral part of many systems, pushing boundaries of research capabilities towards artificial intelligence dream on 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.

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. 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 editing, storing and indexing to searching and retrieving, ranking, classifying, extracting information and knowledge, question answering, etc.

In this paper we present development of artificial intelligence research in Slovenia involving handling of text data that we have been personally involved in, pointing out 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.

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 [2], [3], browsing the Web [4], powering intelligent agents [5]. 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 topics. However, the problem is far from being a trivial application of machine learning methods to a new domain, as, for instance, 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 [8] and pruning the space of promising classifiers to be consulted for classification of a new example [7].

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 [6], additionally increasing the feature space. This requires careful handling of the problem including efficient feature selection [9].

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 [14]. A

step towards extraction of knowledge for ontology generation is presented in [15], 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 [16] 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 from Usama Fayad, 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 make it 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.

Visualization of text data as given in news articles [10] 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. Furthermore, the named entities extracted from news or some other text that has time information associated to it can be related over time [11]. General document corpus can be also visualized using clustering methods on text data [12]. 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 [13].

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) [22]. The text annotations can be also used for **enhancing visualization of text**, for instance on web pages [23].

Extracting triplets from text [17] 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 is by parsing the sentence to get its logical form giving subject-predicate-object [21], reasonable results have been achieved by using predefined patterns, such as noun phrase-verb phrase-noun phrase [22]. The extracted triplets can be also generalized to a kind of templates [20], such as, country – borders – country that can be further used to extend an ontology or to extract information from text.

Document summarization addresses automatic construction of a shorter version of the original text document. It can be performed using different approaches, one of them 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 [21].

Question answering [18], [19], such as, "where do tigers live?", can be also based on triplet extraction where the whole document collection used for finding the answers is transformed into a collection of triplets to enable matching with the questions.

Ontology construction and extension is usually performed entirely manually or semi-automatically by applying some methods from artificial intelligence [24]. Annotation of text by the concepts from an existing ontology is limited by the concepts that already exist in the ontology, unless we extend the ontology in which case it is desirable to focus on a domain of interest [25]. When dealing with larger ontologies, a number of editors having different expertise contribute to different parts. Data analysis can be used to gain some insights into their interaction with the ontology, their expertise and the ontology changes by means of social network analysis and visualization [26].

Social network analysis in combination with analysis of text data can provide insights into research collaboration between institutions and countries [27], while semantic technologies have been successfully applied to analyze communication between individuals inside an organization [28], in visualization of temporal data[30]and to support the users in dealing with information overload [29]. It was also recognized that context of the data and the user may be relevant for the addressed problem [33].

4 DISCUSSION

Different Artificial Intelligence methods have been successfully applied on text data addressing a number of relevant problems. Figure 1 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 [31] and multilingual sentiment analysis [32]. There are a number of open research challenges related to developing linguistic resources for different languages and covering multilingual and cross-lingual settings.

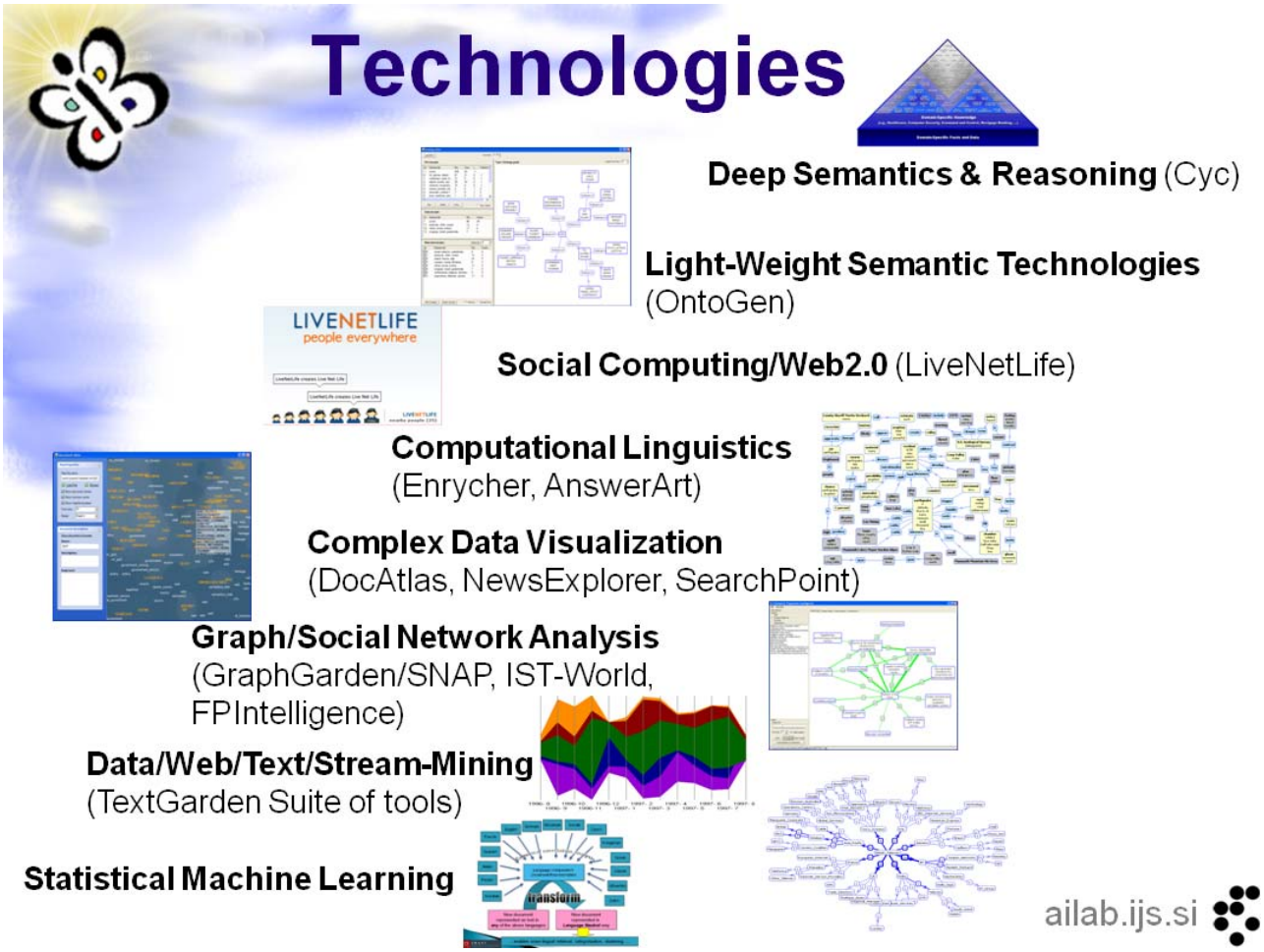

Figure 1: *Diagram showing different kind of technologies involving text data developed by artificial intelligence laboratory.*

ACKNOWLEDGMENTS

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

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

[4] 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.

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

[6] 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.

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

[8] Grobelnik, M, Mladenić, D. Simple classification into

large topic ontology of web documents. *CIT. J. Comput. Inf. Technol.*, 2005, vol. 13, pp. 279-285.

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

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

[11] 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.

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

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

[14] 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.

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

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

[17] 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.

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

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

[20] Sipoš, R., Mladenić, D., Grobelnik, M., Brank, J. Modeling common real-word 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.

[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] 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.

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

[24] 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.

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

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

[27] 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.

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

[29] 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.

[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] 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.

[32] 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.

[33] 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