Using Text Processing in a Multimedia Environment. UAIC Activities in MUCKE Project

Adrian Iftene
Faculty of Computer Science,
"Alexandru Ion Cuza" University of Iasi
General Berthelot, No. 16
adiftene@info.uaic.ro

ABSTRACT
This paper presents main activities of UAIC ("Alexandru Ioan Cuza" University) team from the MUCKE project. MUCKE addressed the stream of multimedia social data with new and reliable knowledge extraction models designed for multilingual and multimodal data shared on social networks. Credibility models for multimedia streams are a novel topic, and constituted the main scientific contribution of the project. UAIC group was involved in the main tasks of the project: building the data collection, text processing, diversification in image retrieval and data credibility.

Author Keywords
Image retrieval; text processing; diversification; YAGO; Flickr

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces. H.3.2. Information Storage and Retrieval: Information Storage.

General Terms
Human Factors; Design.

INTRODUCTION
MUCKE (Multimedia and User Credibility Knowledge Extraction) project of type ERA-NET CHIST-ERA started in 2012 and finished at the end of 2015. MUCKE departed from current knowledge extraction models, which are mainly quantitative, by giving a high importance to the quality of the processed data, in order to protect the user from an avalanche of equally topically relevant data. MUCKE came with two central innovations: automatic user credibility estimation for multimedia streams and adaptive multimedia concept similarity. Adaptive multimedia concept similarity departed from existing models by creating a semantic representation of the underlying corpora and assigning a probabilistic framework to them.

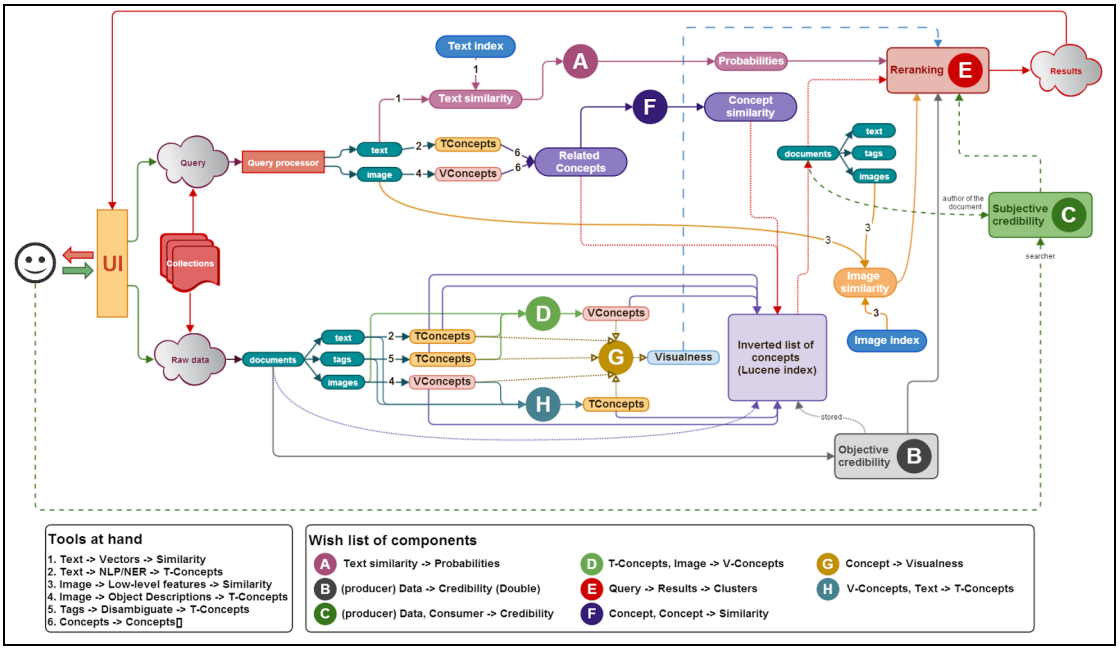

Figure 1. MUCKE Architecture [1].

Although a lot of research efforts were directed towards automatic information mining and important results are reported in different domains, the topic remains complicated with a huge potential. Important challenges arise from the heterogeneous character of raw data, from the scalability of processing methods and from the reliability of extracted knowledge.

Heterogeneity comes in different forms: nature of the documents (text, image, speech, movie), diversity of the languages used in text documents, or language particularities for different data sources. Scalability remains an the most important issue for multimedia streams, whose processing with a good quality and exploitation under real time constraints are still problematic. While it is difficult for the results of automatic techniques to match the quality of manually created resources, the latter imply huge investments when dealing with large-scale data.

Figure 1 shows an overview of the MUCKE framework, covering how documents are processed, concepts extracted and indexed, similarity computed based on concepts, text and images, and how credibility is estimated and fed into the re-ranking process to and how credibility is estimated and fed into the re-ranking process to improve the final set of results. In the rest of the paper, we will see the main tasks in which UAIC group was involved.

NEW DATA COLLECTION

The complex nature of the project objectives required the mobilization of different multimedia data sources in order to mine all necessary user-related knowledge. With over 6 billion of photo uploads, Flickr (flickr.com) is one of the largest photo repositories on the Web and constituted our main source of visual information. Consortium members downloaded images and textual metadata from 1,000,000 Flickr users.

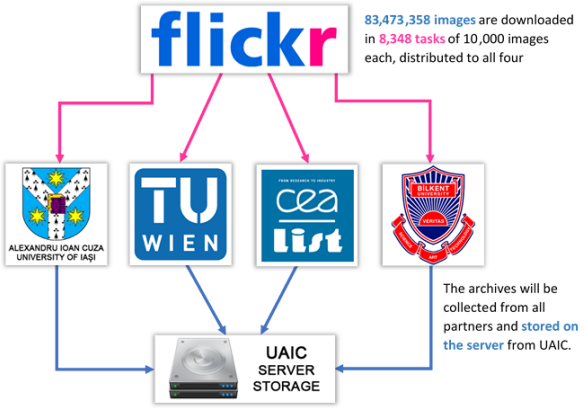

Figure 2. How the download process works.

Given that Flickr contains a mix of personal and social relevant data, focus was on downloading the latter type, which is useful for information extraction tasks. UAIC coordinated the data collection effort but, in order to speed up the process, it was distributed among all partners (Figure 2).

During the download process, some statistics were provided in the form of two charts: one for ongoing and one for completed tasks. These charts were dynamically updated, thus enabling us to check our tasks and see the overall status of the download process at any time (See Figure 3).

At the end of downloading process, the MUCKE corpus contained metadata and images based on Wikipedia concepts that are often used to annotated Flickr images. Wikipedia concepts are ranked using the number of corresponding Flickr images which is divided by the log of incoming Wikipedia links in order to penalize very common concepts. In the end, the top 200 concepts with the highest frequency of occurrence in Flickr are represented in the current corpus.

At the end of MUCKE project, we collected over 80 million images and their associated metadata that have been downloaded mainly from the Flickr database.

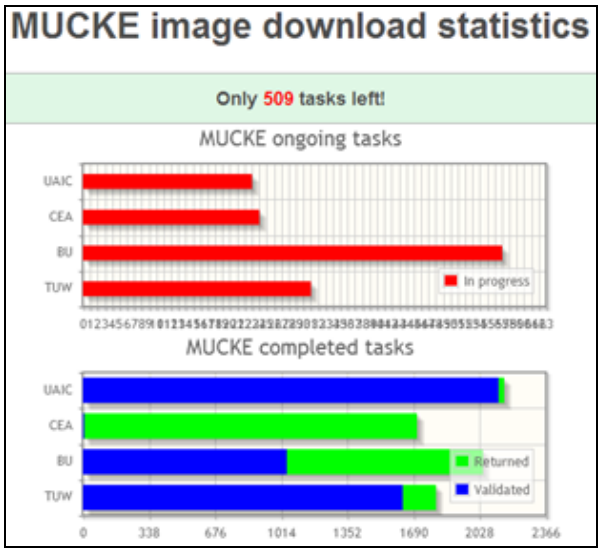

Figure 3. Image download statistics during the download process.

BASIC TECHNIQUES FOR TEXT PROCESSING

Before any keywords or query semantic analysis is performed, some shallow linguistic processing steps are commonly required, such as lemmatization, Part-of-Speech (POS) tagging, anaphora resolution, named entity identification. The aim of text processing was to help us to map images metadata to Wikipedia concepts.

Lemmatization

Lemmatization identifies the root word beyond the inflected forms, which is necessary since most concepts from Wikipedia contain root words. This allows us access to both linguistic resources such as WordNet [28] (for word senses

and semantic relations) and to other linguistic processing tools using root word lexicons or patterns. For all shallow processing steps we identified free existing tools, for English, French, German and Romanian [16, 17].

POS-Tagging

POS-tagging is another important pre-processing step, identifying the parts of speech of the words in the target document. Usually POS-taggers also add morpho-syntactic annotations to the words. This processing step is essential for any deeper analysis, since it conveys some data regarding the grammatical relations between the words, usually correlated to semantic relations. For lemmatization and POS-tagging we used a web-service [25] found to perform best for the Romanian language [26] used in UAIC's experiments, but all European languages have equivalent quality tools available.

Identifying Noun-Phrases and Named Entity

Identifying Noun-Phrases and Named Entity is another step relevant for our project, because it helps us to identify Wikipedia articles (in fact Wikipedia concepts). For Romanian, we used tools from [8, 9, 14 and 18].

Additionally, for named entity recognition we used web service described in [25], with two rules that allow us to transform proper names into named entities. The rules were these: (1) many successive capitalized words and marked by POS-Tagger as proper names are grouped into a single entity, (2) many capitalized words separated by linking words such as "din", "de", etc. (in En: from, of) were also grouped into one entity (for example "Venus din Milo" (in En: Venus from Milo), "Camera Națională de Pensii Publice" (in En: National Chamber of Public Pensions), etc.). We validated these entities using Wikipedia or our external resources (an entity is considered valid if it has a corresponding Wikipedia page or it exists in our external resources).

For named entities classification, we looked at the words in the neighborhood of the entity in order to make the classification. Thus, in the situations where we found in the text expressions such as "Palatul Roznovaru" (in En: Roznovaru Palace) or "munții Rodnei" (in En: the Rodnei Mountains), we considered the corresponding type "palace" or "mountain" for them. In the other situations, we used our external resources, where we have entities of the type location, organization, people and other.

Anaphora resolution

Anaphora Resolution identifies semantic identity between different parts (usually NPs) of a text. We did the anaphora resolution both in the original text and in the Wikipedia articles [12]:

(1) In the original text we started from the classification of the named entity. For instance, after we classified "Munții Rodnei" as "munte", all the appearances of the word "munte" (En: mountain) (that were not followed by the word "Rodnei") have been replaced by the expression "Munții Rodnei".

(2) At the level of Wikipedia articles, we consider that the first paragraph refers to the presented concept from the current Wikipedia page. Thus, we considered that all expressions such as: "A fost inaugurat" (in En: was inaugurated), "Este conceput" (in En: is designed/conceived), "Este considerat" (in En: is considered), etc. refer to the main concept, described in that Wikipedia page.

ADVANCED TECHNIQUES USED FOR TEXT PROCESSING

Query reformulation

Query reformulation provides techniques of improving the quality of search results by extending or replacing parts of the original query. This process is very similar to a required step in a question answering system as described in [15].

Here, we apply two approaches:

(1) a global technique, which analyses the body of the query in order to discover word relationships (synonyms, homonyms or other morphological forms from WordNet), to remove stop words ("a", "un", "la", "pentru", (English: the, a, at, for), etc.), to remove wh- words ("cine", "ce", "de ce", "unde", (English: who, what, why, where), etc.) and to correct any spelling errors;

(2) local feedback which implies the analysis of the results returned by the initial query, leading to re-weighting the terms of the query and relating it with entities and relationships originating from the target ontology. Further discussion on expanding search queries using Yago and Wikipedia can be found in [13].

Another experiment carried out at UAIC with the goal of augmenting the existing query using the information extracted from large data resources such as Wikipedia and Freebase [5]. The system uses a POSTagger for the Romanian language, afterwards it identifies and classifies the named entities. For these entities, the external resources mentioned above are used and concepts are identified (an entity that can appear in the text in various forms) and the relations between them. Further details about this experiment can be found in [7].

Diversification with YAGO

From all collected data for MUCKE, we selected around 30.000 images (a small collection), with aim to perform several processing tasks at both textual (on associated metadata) and image level and retrieve the results in a diversified way. Over small collection, we built a system which allowed users to retrieve multimedia content [11].

To improve a search query we first looked for the relevant words in the query in the results provided by a text-processing module. The text processing module is used to process on one hand, the images associated metadata and, on the other hand, the user queries. Standard tools are used for POS-tagging [25], lemma identification [26] and named entity identification [9]. After the images associated metadata is processed, the image collection is indexed with Lucene [21]. In order to achieve diversification in the results set, the system incorporates a query expansion module that makes use of the YAGO ontology [29] (see Figure 4).

Yago ontology comprises well known knowledge about the world [10]. It contains information extracted from Wikipedia [27] and other sources like WordNet [28] and GeoNames [6] and it is structured in elements called entities (persons, cities, etc.) and facts about these entities (which person worked in which domain, etc.).

For example, with Yago we are able to replace in a query like "tennis player on court", where we have two entities ("tennis player" and "court"), the entity "tennis player" with instances like "Roger Federer", "Rafael Nadal", "Andy Murray", etc. Thus, instead of performing a single search with the initial query, we perform several searches with the new queries, and in the end we combine the obtained partial results in a final result set.

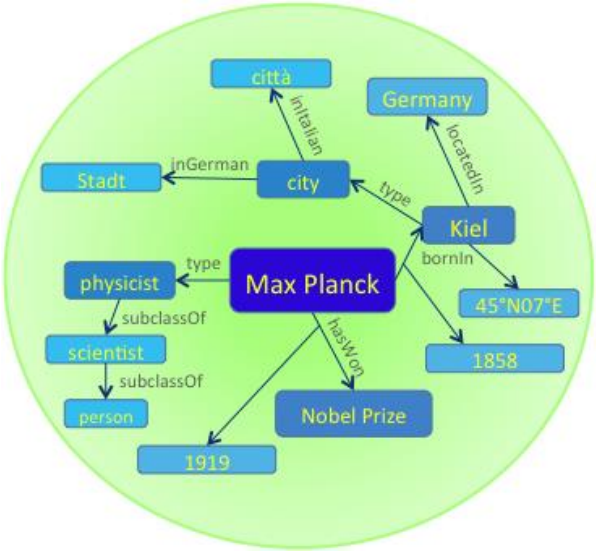

Figure 4. YAGO Ontology [29].

In Figures 5 and 6 are presented results obtained with "*tennis player on court*" query in our application and in Google.

How we can see, results offered by our system contains both concepts ("tennis player" and "court"), while in the Google are cases when concept "tennis player" is missing.

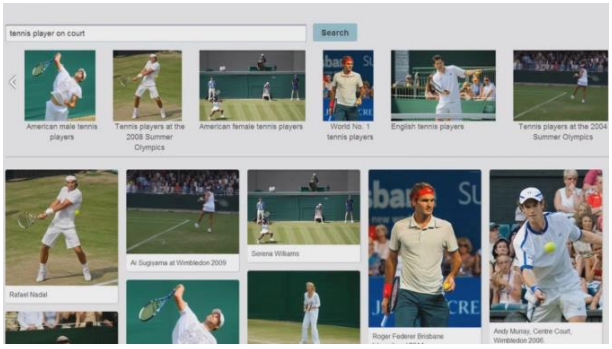

Figure 5. Results for query "tennis player on court" in our application.

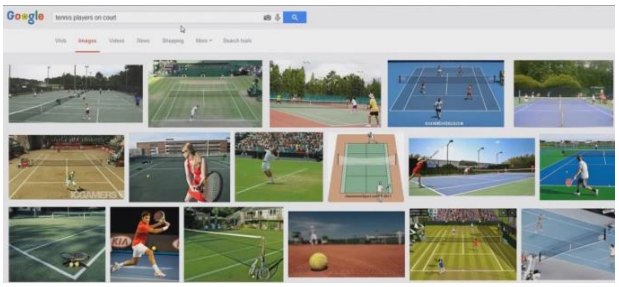

Figure 6. Results for query "tennis player on court" in Google application.

AUTOMATIC IMAGE ANNOTATION

The aim of this component is to add relevant keywords to an image without annotation [18]. In order to do this, the first step was to create a collection of images that was annotated by human annotators, while the second step was to expand this collection of images performing search on the Internet using keywords associated to the initial collection of annotated images. Currently, for a new picture, we can identify similar images in our collection of images and based on the keywords associated with them, we can determine what keywords characterize this new image.

Creation of gold collection with annotated images

The initial collection of images consisted of 100 images, from different areas (art, furniture, sport, other, etc.). The images were selected by six human experts and then were manually annotated by human annotators. Some of the images have words in their visual content to see how this can influence the process of annotation. In the process of annotating, the only criterion was to write keywords in the Romanian language, criterion that was established from the beginning.

Comparing the keywords entered by users for the same picture, it was seen that there were small differences among the words entered, most of them were from the same lexical

family or they were synonyms. Performing an analysis on what users have annotated, it can be said that their tendency was to introduce, on average, 3.41 keywords per image, with a minimum of 2 keywords for an image and a maximum of 12 keywords for an image. Looking further into the keywords that they have entered, it can be said that most users have opted for simple words and not phrases.

After completing this step, we increased the initial collection as it follows: for each image we added 10 new images to our collection, thus increasing the image collection to 1,000 images. For this, we searched for Google images using lists of keywords associated with each image. For the first 10 results, we initially associated the list of keywords used in the search process, followed by a process of verification, corrections, additions to this list, this process was done with human annotators [19].

Reverse Image Search

This module uses the 1,000 collection of images with related keyword lists obtained at the previous step. Regarding this collection, we know that the list contains relevant keywords associated with images.

The main purpose of this module is to generate a list of keywords that characterize an image given by the user. This is done using the LIRE library [20] (Lucene Image REtrieval) [22], which compares the new image with the images from our collection. It establishes a set of 20 images most closely to the new inserted image in terms of texture and color. After that, starting from the 20 lists of keywords associated to these images, we use lemmatization, the synonymy relation and the processing of expressions. In the end, using keywords order in the lists and their frequency, we decide the list keywords for new image.

The conclusion that can be drawn from the analysis of obtained results: the more images we have in our collection of annotated images, the more chances of finding similar images.

CONCLUSIONS

In present, it is a growing interest in hosting multimedia information and looking for such content. Large companies like Flickr, Google, Bing, Yahoo, and Microsoft have such platforms available to users both through its own search pages, and through APIs available to application developers. In order to create collection of images we needed in the project, we used the Flickr network (for images) and Wikipedia encyclopedia (to establish concepts). For Flickr API, we created a crawler which allowed us to save locally relevant information: both images and associated metadata. Since downloading the massive amount of information required a huge effort on our part, this activity was implemented with all partners involved in the project MUCKE.

We could also see that text processing techniques and related resources can make a significant contribution to improving quality of the image retrieval task. The principal methods of doing this come either from improving the way it is used query (extension of search terms, finding semantic relations between terms in the query and use clues from the image, use multilingualism to expand the scope of search) or on improving description of an image. For this, we have identified and used several tools for text processing, such as POS taggers, tools for lemmatization, for anaphora resolution, for identification of name entities, etc.). We also used semantic resources such as WordNet, Yago, Wikipedia, GeoNames and Freebase. The quantitative evaluation was performed for all experiments described with good results, more details are available in [7, 13 and 18].

Our algorithms, presented in this paper, have been successfully used in the Plant Identification task from CLEF [3], in Image CLEF evaluation campaigns (in Scalable Concept Image Annotation Challenge task and Plant Identification task) and in MediaEval benchmark [2, 4, 9, 12, 23 and 24].

ACKNOWLEDGMENT

We want to thank to UAIC team, to bachelor and master students from Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iasi, which was involve in the all stages of this project development. This work is partially supported by POC-A1-A1.2.3-G-2015 program, as part of the PrivateSky project (P_40_371/13/01.09.2016).

REFERENCES

1. Bierig, R., Șerban, C., Siriteanu, A., Lupu, M, and Hanbury, A. A System Framework for Concept- and Credibility-Based Multimedia Retrieval. In *ICMR 2014 Proceedings of International Conference on Multimedia Retrieval*, Glasgow, Scotland, (2014), 543-550.
2. Calfa, A., Silion, D., Bursuc, A. C., Acatrinei, C. P., Lupu, R. I., Cozma, A. E., Pădurariu, C. and Iftene, A. Using Textual and Visual Processing in Scalable Concept Image Annotation Challenge. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum - ImageCLEF2015*, 1391, 8-11 September 2015, Toulouse, France, (2015).
3. CLEF 2017: http://clef2017.clef-initiative.eu/ (Last time accessed on 30 June, 2017)
4. Cristea, A.G., Savoaia, M. M., Martac, M. A., Pătraș, I. C., Scutaru, A. O., Covrig, C. E. and Iftene, A. Using Machine Learning Techniques, Textual and Visual Processing in Scalable Concept Image Annotation Challenge. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum - ImageCLEF2016*, Evora, Portugal, (2016).
5. Freebase: https://www.freebase.com (Last time accessed on 30 June, 2017)
6. GeoNames: http://www.geonames.org/ (Last time accessed on 30 June, 2017)

7. Gherasim, L. M. and Iftene, A. Extracting Background Knowledge about World from Text. In *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, Craiova, 18-19 September 2014, ISSN 1843-911X, (2014), 199-208.
8. Gînscă, A. L., Boroș, E., Iftene, A., Trandabăț, D., Toader, M., Corîci, M., Perez, C. A. and Cristea, D. Sentimatrix - Multilingual Sentiment Analysis Service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (ACL-WASSA2011), Portland, Oregon, USA, (2011).
9. Gînscă, A. L., Popescu, A., Lupu, M., Iftene, A. and Kanellos, I. Evaluating User Image Tagging Credibility. *Experimental IR meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, 9283, Publisher Springer International Publishing. In Proceedings of 6th International Conference of the CLEF Association, CLEF'15 Toulouse, France, September 8-11, (2015), 41-52.
10. Hoffart, J., Suchanek, F., Berberich, K. and Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Elsevier, Artificial Intelligence*, 194, (2013), 28-61.
11. Iftene, A. and Alboaie, L. Diversification in an image retrieval system based on text and image processing. Computer Science Journal of Moldova, 22 (3) (66), (2014), 1-10.
12. Iftene, A., Moruz, A. and Ignat, E. Using Anaphora resolution in a Question Answering system for Machine Reading Evaluation. *Notebook Paper for the CLEF 2013 LABs Workshop - QA4MRE*, 23-26 September, Valencia, Spain, (2013).
13. Iftene, A., Sirițeanu, A. and Petic, M. How to Do Diversification in an Image Retrieval System. In *Proceedings of the 10th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Craiova, 18-19 September 2014, (2014), 153-162.
14. Iftene, A., Trandabăţ, D. and Pistol, I. Grammar-based automatic extraction of definitions and applications for Romanian. In *Proceedings of RANLP workshop" Natural Language Processing and Knowledge Representation for eLearning environments*, (2007), 978-954.
15. Iftene, A., Trandabăţ, D., Moruz, A., Pistol, I., Husarciuc, M. and Cristea, D. Question answering on English and Romanian languages. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer Berlin Heidelberg, (2009), 229-236.
16. Iftene, A., Trandabăț, D., Moruz, A.-M., Pistol, I., Husarciuc, M. and Cristea, D. Question Answering on English and Romanian Languages. In Multilingual Information Access Evaluation, Vol. I Text Retrieval Experiments, Springer, Heidelberg, CLEF 2009, LNCS 6241, Part I, (2010), 229-236.
17. Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A. M., Husarciuc, M. and Cristea, D. UAIC participation at QA@ CLEF2008. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer Berlin Heidelberg, (2008), 385-392.
18. Laic, A. and Iftene, A. Automatic Image Annotation. In *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, Craiova, 18-19 September 2014, ISSN 1843-911X, (2014), 143-152.
19. Laic, A., Gherasim, L. M. and Iftene, A. Expanding a gold collection of images using the Flickr network. In *Proceedings of the Conference on Mathematical Foundations of Informatics MFOI'2016*, July 25-29, 2016, Chisinau, Republic of Moldova, (2016), 241-250.
20. LIRE: http://www.semanticmetadata.net/lire/ (Last time accessed on 30 June, 2017)
21. Lucene: http://lucene.apache.org/ (Last time accessed on 30 June, 2017)
22. Lux, M. and Marques, O. Visual Information Retrieval using Java and LIRE. *Synthesis Lectures on Information Concepts, Retrieval*, and S, Morgan & Claypool Publishers, 112 pages, (2013).
23. MediaEval Benchmark: http://www.multimediaeval.org/
24. Șerban, C., Sirițeanu, A., Gheorghiu, C., Iftene, A., Alboaie, L. and Breabăn, M. Combining image retrieval, metadata processing and naive Bayes classification at Plant Identification 2013. *Notebook Paper for the CLEF 2013 LABs Workshop - ImageCLEF - Plant Identification*, 23-26 September, Valencia, Spain, (2013).
25. Simionescu, R. Hybrid POS Tagger. In *Proceedings of Language Resources and Tools with Industrial Applications*, Workshop (Eurolan 2011 summerschool), (2011).
26. UAIC NLP Tools: http://nlptools.infoiasi.ro/WebPosRo/ (Last time accessed on 30 June, 2017)
27. Wikipedia: http://en.wikipedia.org/wiki/Main_Page (Last time accessed on 30 June, 2017)
28. WordNet: http://wordnet.princeton.edu/ (Last time accessed on 30 June, 2017)
29. Yago ontology: https://www.mpi-inf.mpg.de/departments/databases-and-informationsystems/research/yago-naga/yago/ (Last time accessed on 30 June, 2017)