

M U C K E

## Text Processing Module

Adrian Iftene\*, Adrian Popescu#, Ionut Pistol\*,  
Mihai Lupu+, Allan Hanbury+

\*Univ. "Al.I.Cuza", Iasi

#CEA, LIST, LVIC France

+ Vienna University of Technology, ISIS, IMP

Contacts: [adiftene@info.uaic.ro](mailto:adiftene@info.uaic.ro), [adrian.popescu@cea.fr](mailto:adrian.popescu@cea.fr)

MUCKE Project, Deliverable 2.1

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Basic techniques for text processing</b>	<b>4</b>
2.1	Lemmatisation . . . . .	4
2.2	POS-Tagging . . . . .	5
2.3	Identifying Noun-Phrases and Named Entity . . . . .	5
2.4	Anaphora resolution . . . . .	6
2.5	Query reformulation . . . . .	6
<b>3</b>	<b>Advanced Techniques for Text Processing</b>	<b>7</b>
3.1	Explicit Semantic Analysis . . . . .	7
3.1.1	ESA Adaptation for ad-hoc multilingual IR (ESA-C) . . . . .	8
3.1.2	Enrichment and retrieval framework . . . . .	10
3.1.3	Experiments . . . . .	11
3.2	Multilinguality . . . . .	13
3.2.1	ESA for Bilingual Lexicon Extraction . . . . .	13
3.3	Diversification with Yago . . . . .	15
3.4	Automatic Image Annotation . . . . .	16
3.4.1	Creation of gold collection with annotated images . . . . .	18
3.4.2	Reverse Image Search . . . . .	19
<b>4</b>	<b>Conclusions</b>	<b>20</b>

## Abstract

This document summarizes the existing tools used by partners in text processing tasks. The text tools are used to process the text inserted by user like query in our system, or to process titles, keywords, etc. for images from Flickr database. Also, we use them when we need to process and to analyse text from Wikipedia pages. What we will present below is related to Romanian, French, English and German languages.

# 1 INTRODUCTION

---

Task 2.1 involved the use of available language technologies to improve image retrieval. This involves either mainly the effort to accurately attach keyword concepts to images found in the same context. The source of the concept words can vary as text can be associated with images either as a set of original keywords found in a collection of images (eg. Flickr), a text accompanied by images (eg. Wikipedia) or if the text serves as a search query (eg. Google Image Search).

Difficulties for this task are mostly posed by two aspects of text semantics: the ambiguity of multi-sense words and the anaphoric references. Also a concern, especially when using existing resources (Flickr, Wikipedia) is the possibility of an erroneous association between a concept and an image, which can be detected and corrected using textual clues (other keywords, manual annotations). Improving the set of concepts describing and image can also be done by automatically adding additional keywords.

Another important aspect of improving image retrieval using language technologies involves the improvement and/or expansion of the search query. This can require processing of the query itself to identify search terms, finding related words and matching search terms with various keywords attached to images.

The following sections briefly describe some efforts made concerning overcoming all the above perceived difficulties, focusing on the relevant language technologies identified and used in our developments. Section 2 goes over the basic techniques for text processing available and used, while sections 3 describe the advanced techniques for text processing (Explicit Semantic Analysis and multilinguality in image retrieval, Diversification with Yago and how query expansion and reformulation can be performed otherwise). The last section contains a brief overview of the current and future work.

## 2 BASIC TECHNIQUES FOR TEXT PROCESSING

---

Before any keywords or query semantic analysis is performed, some shallow linguistic processing steps are commonly required, such as lemmatisation, Part-of-Speech (POS) tagging, anaphora resolution, named entity identification.

### 2.1 LEMMATISATION

Lemmatisation identifies the root word beyond the inflected forms, which is necessary since most dictionaries contain only root words. This allows us access to both linguistic resources such as WordNet<sup>1</sup> (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 (freely available and

---

<sup>1</sup><http://wordnet.princeton.edu/>



accurate) existing tools. When further text processing was required we developed some tools and resources which are described in the sections below.

## 2.2 POS-TAGGING

POS-tagging is another frequently mention 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 lemmatisation and POS-tagging we used a web-service [12] found to perform best for the Romanian language<sup>2</sup>, used in UAIC's experiments, but all European languages have equivalent quality tools available.

## 2.3 IDENTIFYING NOUN-PHRASES AND NAMED ENTITY

Identifying Noun-Phrases (NPs) and Named Entity (NEs) is another step found to be relevant for our purposes, since it allows us to identify the most significant content of a query (or a Wikipedia article, for example), which leads to identifying relevant keywords and accurately matching queries to images described by other keywords. In our query reformulation and expansion efforts this step allows us to focus our effort on the parts of the query which contain the most semantic content, which most often are the NPs (nouns and related words) and NEs (usually proper names). Both for NP-chunking and NE recognition useful tools were identified and used (see [3], [8] and [2]).

Additionally, for named entity recognition, based on XML file obtained with service described in [SIM11], using two rules, we transformed proper names into named entities. The rules were these: (1) many successive capitalized words and marked by POSTagger 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 Interna?ional? de Comer?" (in En: International Chamber of Commerce), 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 tried to identify the type of the entities found in the previous step. For this, we analyse the text and we looked at the words in the neighbourhood of the entity in order to make the classification. Thus, in the situations where we found in the text expressions such as "Palatul Ghica" (in En: Ghica Palace), "Muzeul Luvru" (in En: the Louvre Museum), "fluvial Sena" (in En: the Seine River), we considered the corresponding type "palace", "museum" or "river" for them. In the other situations, we used our external resources, where we have entities of the type location, organization, people and other. In addition to this, we created a list of subtypes for each

<sup>2</sup><http://nlptools.infoiasi.ro/WebPosRo/>

of the four types using the class hierarchy from DBpedia<sup>3</sup> and Schema.org<sup>4</sup>. Thus, we can identify that an entity of the type “museum” or “palace” is actually an entity of the type location.

## 2.4 ANAPHORA RESOLUTION

Anaphora Resolution (AR) identifies semantic identity between different parts (usually NPs) of a text. This is usually done by morpho-syntactic clues and closeness of candidate NPs, for further details on the system used in most of our experiments see [5]. We did the anaphora resolution both in the original text and in the Wikipedia articles. (1) In the original text we started from the classification of the named entity. Thus, after the classification of the entity in a certain class, we used it to replace the references to the class in the text by its corresponding entity. For instance, after we classified “Palatul Ghica” as “palace”, all the appearances of the word “palat” (En: palace) (that were not followed by the word “Ghica”) have been replaced by the expression “Palatul Ghica”. (2) At the level of Wikipedia articles, we took advantage of their structure and of the fact 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. The same approaches are used for the pronouns in the text.

## 2.5 QUERY REFORMULATION

Query reformulation provides a technique of improving search results by extending or replacing parts of the original query while keeping the relevance of the results. This process is very similar to a required step in a question answering system as described in [7]. We face two major issues that occur when an end user entered a query: it is not precise enough, meaning that there are too many results returned, most of them being irrelevant or it is not abstract enough, meaning that the search does not return any results at all. 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 [6]. 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<sup>5</sup>. The system uses a POSTagger for the Romanian language, afterwards it identifies and classifies the named entities. For these entities, the external

<sup>3</sup><http://dbpedia.org/About>

<sup>4</sup><http://schema.org/>

<sup>5</sup><https://www.freebase.com>

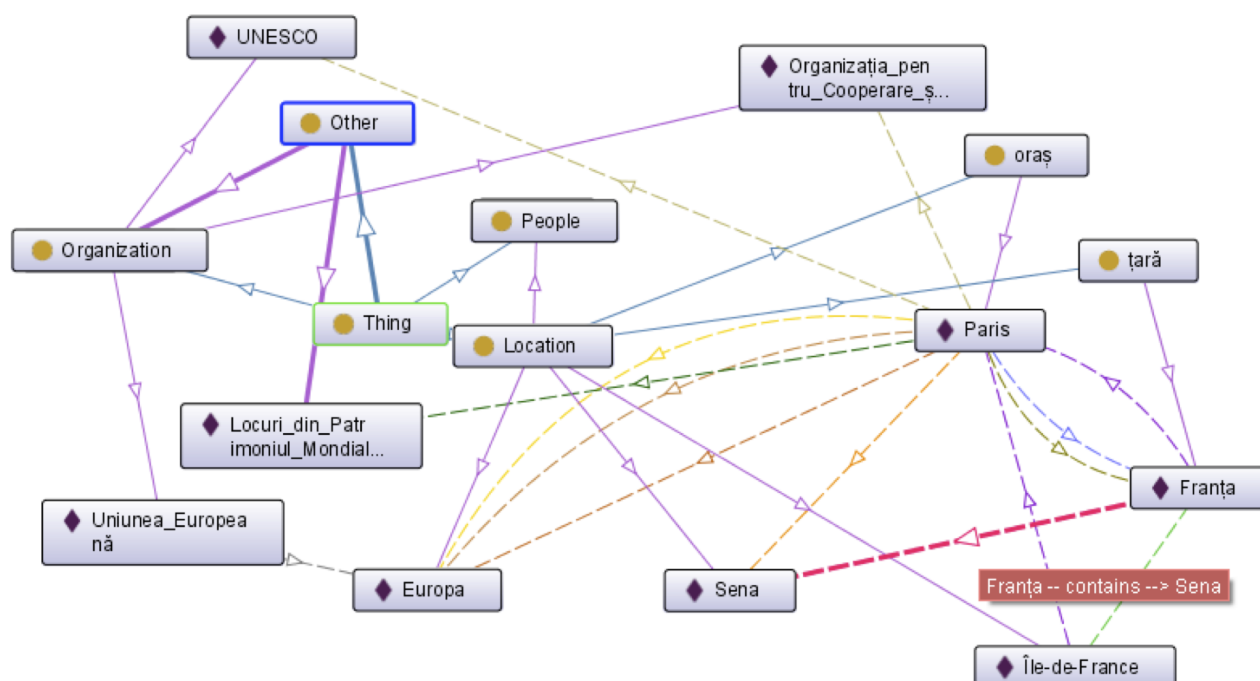


FIGURE 1: THE FINAL GRAPH FOR PARIS [2]

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 (see Figure 1).

The problems we faced started from the results returned by POSTagger that are not always accurate, but which can be improved with some post-processing. Other problems are connected to the relation identification among concepts in the text that is analysed. Sometimes, these relations are too detailed and they should be parameterised. In the future, using the analysed texts, we plan to build a resource with the features of the identified entities and with the relations among them. This resource would increase in time and it should be useful to those interested in analysing the named entities. Further details about this experiment can be found in [2].

Improving and correcting keywords attached to images can yield significant improvements to any image retrieval system, so a significant effort was carried out at UAIC in this direction.

## 3 ADVANCED TECHNIQUES FOR TEXT PROCESSING

### 3.1 EXPLICIT SEMANTIC ANALYSIS

Explicit Semantic Analysis [1] is a method that maps textual documents onto a structured semantic space. Since its introduction in 2007, ESA was successfully exploited in different natural language processing and information retrieval tasks. The success of this simple method lies in the richness and the quality of the underlying conceptual space. In the original evaluation, ESA outperformed state of the art methods in a word relatedness estimation task and different developments were

subsequently proposed. Put simply, ESA exploits classical text weighting schemes, such as TF-IDF, to model concepts from a structured resource, such as Wikipedia. A relation between words and the concepts that structure the space is established by inverting the concepts' vectorial representations. Thus, each word of the vocabulary has an associated high-dimension projection onto the concept space of the underlying resource. Finally, in order to compare two words or two documents, the representations of individual words are summed and the resulting vectors compared. In information retrieval, the most useful component of ESA is the mapping of words onto concepts that can be used for topic expansion or consolidation.

### 3.1.1 ESA ADAPTATION FOR AD-HOC MULTILINGUAL IR (ESA-C)

We proposed an ESA adaptation to information retrieval tasks that gives priority to categorical information [11]. The comparison with a classical ESA implementation showed that a significant improvement was obtained in an image retrieval setting. Moreover, the method compared favourably with other state of the art indexing and retrieval schemes. We extended the work in [11] and proposed to use ESA for query expansion and consolidation.

We proposed a version of ESA that gives a privileged role to categorical information where we used two scores to rank Wikipedia concepts:

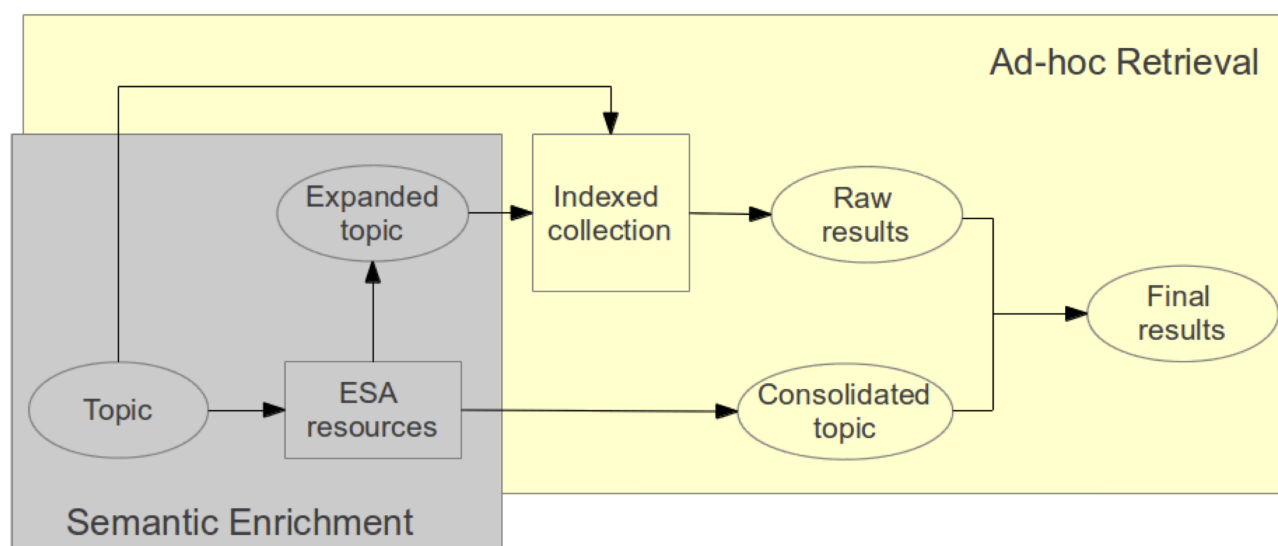
- A Boolean score that captures the number of common words between the initial topic and the words found in the categories associated to Wikipedia concepts.
- The score used in the classical ESA in order to rank concepts, based on the sum of the contributions of the individual words

Since topics are often short, ties are often obtained with the Boolean score and these are broken using the second, finer grained score. The introduction of the Boolean score has two main objectives. First, categorical information should be favoured in order to obtain concepts that are hierarchically related (i.e. isA relation) to the initial topic or to parts of it. Second, it is possible to identify which parts of the initial query an ESA related concept is related to. Table 1 shows some examples. For instance, the categories of *Tropical Fish* are *Fish stubs* and *Aquaria* and the topic would have a Boolean score of 1 (out of a maximum of 2). Similarly, *Freshwater bivalve*, the top ranked concept with classical ESA, only loosely related to the initial topic, has a Boolean score of 0 since its only category is *Bivalves*. The categorical ranking rightly gives a better position to *Tropical Fish* compared to *Freshwater bivalve* since the first concept is more closely related to *Freshwater fish*. Here we modified our ESA adaptation for IR in two directions. First, given that categorical information is often sparse, we added the words contained in the first 150 characters after the concepts name in the first paragraph of the category words.

This enrichment of the categorical space is motivated by the fact that the first paragraph of Wikipedia article is often a definition that contains salient concepts related to the target one. The limitation to words contained in string of 150 characters is useful since the first paragraph has

**TABLE 1:** TOP 10 ESA RELATED CONCEPTS FOR TOPICS FRESHWATER FISH, JEAN-JAQUES ROUSSEAU AND CROCKERY DOLL HOUSE. THE SECOND COLUMN CONTAINS RESULTS FOR CLASSICAL ESA, WHILE THE THIRD RESULTS FOR THE ADAPTED VERSION OF ESA (ESA-C) PRESENTED IN THIS SUBSECTION

<b>Topic CHiC-051 <i>Freshwater Fish</i></b>		
Rank	ESA	ESA-C
1	Freshwater bivalve	Eastern freshwater cod
2	Freshwater mollusc	Ide (fish)
3	Tropical fish	New Zealand longfin eel
4	Freshwater, Humboldt County, California	Common galaxias
5	Fish fillet processor	European perch
6	Bait fish	Green swordtail
7	Fish marketing	Rainbowfish
8	Bottom fishing	Common rudd
9	Freshwater, Isle of Wight	Spotted bass
10	Bank fishing	Common bream
<b>Topic CHiC-058 <i>Jean-Jaques Rousseau</i></b>		
Rank	ESA	ESA-C
1	Confessions (Rousseau)	Confessions (Rousseau)
2	Saint-Jean	Considerations on the Government of Poland
3	Considerations on the Government of Poland	Discourse on the Arts and Sciences
4	Eugne Rousseau (chess player)	Emile, or On Education
5	John Jacques, Baron Jacques	Essay on the Origin of Languages
6	Eugene Rousseau (saxophonist)	Discourse on Inequality
7	Jean-Jacques Henner	Letter to M. D'Alembert on Spectacles
8	Victor Rousseau	Pygmalion (Rousseau)
9	Bobby Rousseau	Julie, or the New Heloise
10	Discourse on the Arts and Sciences	Le devin du village
<b>Topic CHiC-064 <i>Crockery doll house</i></b>		
Rank	ESA	ESA-C
1	Peg wooden doll	Mabel Lucie Attwell
2	Composition doll	Bringing Up Father
3	Anatomically correct doll	The Tale of Mrs. Tiggy-Winkle
4	Bisque doll	China doll
5	Black doll	Japanese traditional dolls
6	Paper doll	Queen Mary's Dolls' House
7	Madame Alexander	Bild Lilli doll
8	Fashion doll	Vivien Greene
9	Doll	Paper Dolls (band)
10	China doll	Wall House (Elkins Park, Pennsylvania)



**FIGURE 2:** OVERVIEW OF THE SEMANTIC ENRICHMENT AND RETRIEVAL FRAMEWORK

varying length and contains information that is only loosely related to the concepts when it is long. The second modification is a concept detection that is used to produce a third score which favours articles that contain longer concepts from the initial query over other articles. At equal categorical scores, the inclusion of concept detection allows us to favour a Wikipedia concept that includes Jean-Jacques Rousseau in its text when compared to another concept that includes Jean-Jacques and Rousseau separately.

The obtained results are especially encouraging for the semantic enrichment task. The ESA-C adaptation of Explicit Semantic Analysis outperforms the original version of the method. Future work includes adding supplementary material to the conceptual space (i.e. Wikipedia corpus) in order to enrich concept descriptions and to try to cover a larger range of concepts. A second line of work involves a shift towards a more semantic representation of ESA concepts that goes beyond the current bag-of-words modelling and involves concept detection and disambiguation. A third research axis will focus on ways to predict the chances of success of automatic expansion in order to perform this task only when the topic is suited.

### 3.1.2 ENRICHMENT AND RETRIEVAL FRAMEWORK

The framework devised here was used for both semantic enrichment and ad-hoc retrieval and is summarized in Figure 2. The semantic enrichment process exploits only topic expansion with the ESA versions (ESA and ESA-C) and returns ranked lists of results using different ranking schemes.

**Semantic Topic Enrichment Framework** The purpose of the semantic enrichment process is to return a ranked list of concepts that are semantically related to the initial topic and could be used for query expansion. To test multilingual rankings, we introduced fusion methods that exploit the explicit interlingual links available in Wikipedia using either different fusion schemes based on the scores in individual languages. In all cases, the proposed enrichments are collection independent.

Only Wikipedia concepts formed of at most 4 words were retained in the final rankings. Lists of related concepts obtained with the original version of ESA and with the adapted ESA-C version are presented in Table 1.

**Ad-hoc Retrieval Framework** Within CHiC, the objective of the ad-hoc retrieval process is to return the best results possible using whatever automatic method at hand. In our approach, the target topic is first processed using ESA resources to expand and consolidate it. The initial words and the expanded concepts are then compared to the index of the collection in order to retrieve a raw list of results. The elements of this list are then compared to the consolidated version of the topic in order to obtain the final list of results. Similarly to the ranking of ESA related concepts, two similarity measures are used:

- A Boolean score to measure a coarse similarity between the initial topic or its related concepts and the documents in the collection.
- The cosine similarity is used to measure the degree of similarity between a topic and corresponding documents.

The Boolean score has a higher priority than the cosine similarity, which is used only to break ties. For multilingual runs, the process is performed for each of the languages processed and then results are combined by ranking results by decreasing scores.

### 3.1.3 EXPERIMENTS

CEA group submitted runs for both the semantic enrichment and the ad-hoc retrieval subtasks and we analyse them here. Unfortunately, this analysis is altered by the fact that an important bug in the scoring of related was discovered after the release of official results. This bug had a strong negative impact on the quality of results for all runs that exploited fusion techniques for semantic enrichment and automatic topic expansion for ad-hoc retrieval. The bug biases individual Boolean scores but the order of concepts is not affected and a comparison of ESA versions remains possible. Boolean scores of expanded concepts were overrated compared to the Boolean scores of documents found using terms from the original topic. All affected runs are indicated by a "\*" sign in the following tables.

**Semantic Enrichment** Eight runs were submitted to the semantic enrichment subtask. Even though the results for 6 out of 8 runs are biased here, there are some interesting conclusions that we can draw from Table 2. The comparison between `ceaListEnglishMonolingualOriginal` and `ceaListEnglishMonolingual` is favourable to the latter method. The original ESA implementation has significantly poorer performances compared to the adapted method introduced (P@10 0.365 vs. 0.66). The privileged role given to categories and to the first words in the concept text, coupled with concept detection in the queries have a positive impact on semantic enrichments.



**TABLE 2: SEMANTIC ENRICHMENT ACCURACY MEASURED USING P@10 OF RELEVANT AND OF RELEVANT + PARTLY RELEVANT RESULTS**

Run name	P@10	P@10 (rel + part.rel)
ceaListEnglishMonolingual	<b>0.468</b>	<b>0.66</b>
ceaListEnglishMonolingualOriginal	0.212	0.364
ceaListEnglishRankEnglish	0.34	0.56
ceaListEnglishRankMultilingual	0.3382	0.5556
ceaListEnglishBooleanEnglish *	0.228	0.436
ceaListEnglishBooleanMultilingual *	0.22	0.428
ceaListEnglishCosineEnglish *	0.076	0.164
ceaListEnglishCosineMultilingual *	0.076	0.164

None of the fusion methods proposed improves results compared to the best submitted run but this is at least in part due to the bug that affected the values of Boolean concept scores. When comparing the fusion schemes, there are no significant differences between monolingual and multilingual fusions. Since the same concepts were proposed but languages differed, this results shows that the ground truth is of high quality. The cosine-based fusion strongly degrades results, while the fusion based on ranks is closer to the original results.

Important differences occur at the topic level. For `ceaListEnglishMonolingual`, when examining CHiC-51 (freshwater fish) and CHiC-58 (Jean-Jacques Rousseau), all top 10 related concepts are at least partly related to the initial topic. Inversely, results are very poor (9 out of 10 irrelevant enrichments) for topics CHiC-64 (crockery doll houses) and CHiC-65 (sea sunset). These failures are probably due to a poor mapping of the topic in the Wikipedia corpus for CHiC-64 and to the very small number of Wikipedia concepts that cover both sea and sunset.

**Ad-hoc retrieval** For “noExpansion” runs, results are ranked first by the number of terms from the initial topic that appear in the document and then by the cosine similarity between the consolidated version of the topic and document representations. For the other runs, the Boolean score of related ESA concepts biases the results. In the end, 16 runs were submitted to the semantic enrichment subtask. Results Due to the bug that affected all the runs that involved ESA based topic expansion, it is difficult to compare runs that did not involve expansion and the others. However, it is worthwhile noticing the best submitted run, i.e. `ceaListMultilingualNoExpansion` a simple fusion of results obtained for individual languages, gave interesting results compared to monolingual runs that involved no ESA expansion.

When comparing `ceaListMultilingualOriginal` and `ceaListMultilingualFiltered`, the two multilingual runs that exploit ESA and ESA-C, obtained results are better for the second run (MAP 0.0805 vs. 0.0977). This result confirms the one obtained for semantic enrichment, where ESA-C was also superior to classical ESA. It is also in line with our findings from [11], which showed that giving a privileged role to categorical information is beneficial in an image retrieval scenario. The favourable comparison of ESA-C with ESA is also confirmed for English (MAP 0.321 vs. 0.304) and Italian (MAP 0.165 vs. 0.0222).



**TABLE 3: MAP PERFORMANCES FOR AD-HOC RETRIEVAL RUNS**

Run name	MAP
ceaListMultilingualNoExpansion	<b>0.1878</b>
ceaListFrenchNoExpansion	0.0478
ceaListFrenchFiltered *	0.0290
ceaListGermanNoExpansion	0.0631
ceaListGermanFiltered *	0.0505
ceaListMultilingualOriginal *	0.0805
ceaListMultilingualFiltered *	0.0977
ceaListDutchFiltered *	0.0377
ceaListEnglishOriginal *	0.0304
ceaListEnglishFiltered *	0.0321
ceaListItalianOriginal *	0.0165
ceaListItalianFiltered *	0.0222
ceaListNorwegianFiltered *	0.0251
ceaListPolishFiltered *	0.0109
ceaListSpanishFiltered *	0.0204
ceaListSwedishFiltered *	0.0123

## 3.2 MULTILINGUALITY

Part of the work at CEA LIST was focused on semantic topic expansion and consolidation based on Explicit Semantic Analysis (ESA) versions in different languages. The corpora used for testing the developed systems was provided by the CLEF ChiC Lab as part of the 2013 multilingual ad-hoc and multilingual semantic enrichment tasks, in which CEA LIST took part. A more detailed description of this participation can be found in [10].

A way in which multilinguality can improve an image retrieval system was considered: merging results of queries in multiple languages. The corpora used for testing the developed systems was provided by the CLEF ChiC Lab as part of the 2013 multilingual ad-hoc and multilingual semantic enrichment tasks, in which CEA LIST took part.

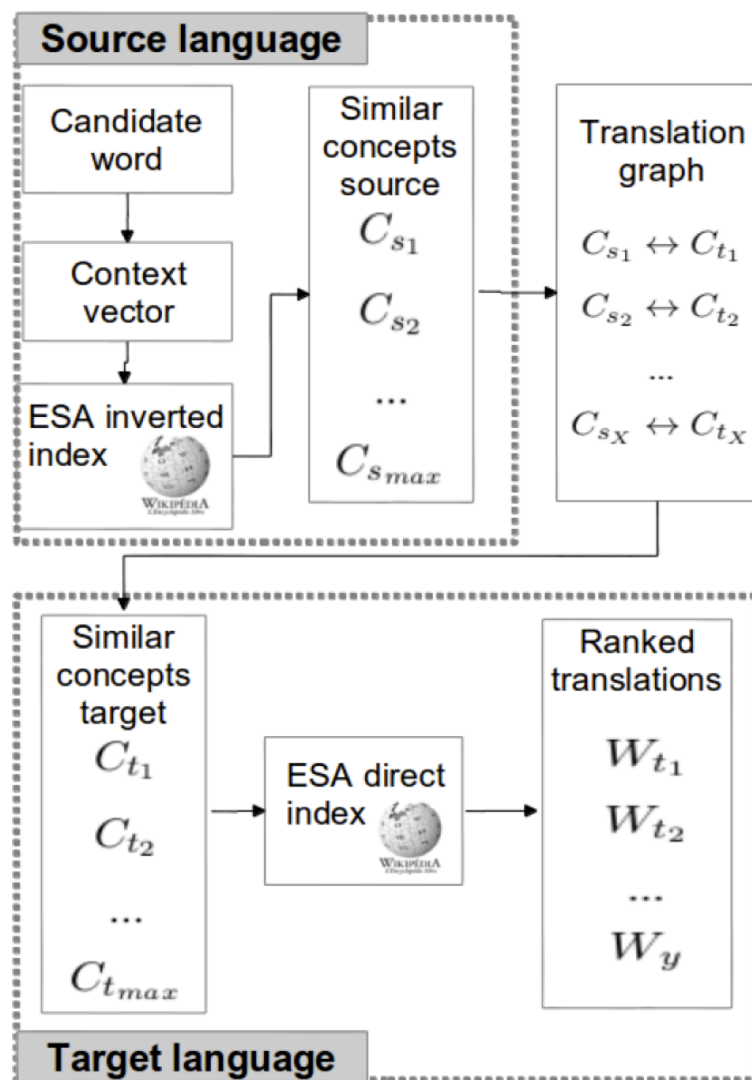
Given the strong multilingual character of the evaluation corpus, the main objectives of the experiments were to test the efficiency of semantic topic expansion and consolidation based on Explicit Semantic Analysis (ESA) versions in different languages. Another objective was multilingual fusion of results obtained in the different languages of the corpus.

### 3.2.1 ESA FOR BILINGUAL LEXICON EXTRACTION

The main objective of our approach is to devise lexicon translation methods that are easily applicable to a large number of language pairs, while preserving the overall quality of results. A subordinated objective is to exploit large scale background multilingual knowledge, such as the encyclopaedic content available in Wikipedia. As we mentioned, ESA [1] was exploited in a number of NLP tasks but not in bilingual lexicon extraction.

Figure 3 shows the overall architecture of the lexical extraction process we propose. The process is completed in the following three steps:

1. Given a word to be translated and its context vector in the source language, we derive a ranked



**FIGURE 3:** OVERVIEW OF THE EXPLICIT SEMANTIC ANALYSIS ENABLED BILINGUAL LEXICON EXTRACTION

- list of similar Wikipedia concepts (i.e. articles) using the ESA inverted index.
2. Then, a translation graph is used to retrieve the corresponding concepts in the target language.
  3. Candidate translations are found through a statistical processing of concept descriptions from the ESA direct index in the target language.

In this section, we first introduce the elements of the original formulation of ESA necessary in our approach. Then, we detail the three steps that compose the main bilingual lexicon extraction method illustrated in Figure 3. Finally, as a complement to the main method we introduce a measure for domain word specificity and present a method for extracting generic translation lexicons.

**Source Language Processing** The objective of the source language processing is to obtain a ranked list of similar Wikipedia concepts for each candidate word ( $W_{cand}$ ) in a specialized domain. To do this, a context vector is first built for each  $W_{cand}$  from a specialized monolingual corpus.

The association measure between  $W_{cand}$  and context words is obtained using the Odds-Ratio [9]. Wikipedia concepts in the source language  $C_s$  that is similar to  $W_{cand}$  and to a part of its context words are extracted and ranked.

**Translation Graph Construction** To bridge the gap between the source and target languages, a concept translation graph that enables the multilingual extension of ESA is used. This concept translation graph is extracted from the explicit translation links available in Wikipedia articles. This concept translation graph is exploited in order to connect a word's conceptual space in the source language with the corresponding conceptual space in the target language. Only a part of the articles have translations and the size of the conceptual space in the target language is usually smaller than the space in the source language. For instance, the French-English translation graph contains 940,215 while the French and English Wikipedias contain approximately 1.4 million articles, respectively 4.25 million articles as of July 2013 .

**Target Language Processing** The third step of the approach takes place in the target language. Using the translation graph, we select the 100 most similar concept translations (threshold determined empirically after preliminary experiments) from the target language and use their direct ESA representations in order to retrieve potential translations for the candidate word  $W_{cand}$  from source language. Then, these candidate translations  $WT$  are ranked.

### 3.3 DIVERSIFICATION WITH YAGO

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 [12], lemma identification [12] and named entity identification [3]. After the images associated metadata is processed, the image collection is indexed with Lucene<sup>6</sup>. In order to achieve diversification in the results set, the system incorporates a query expansion module that makes use of the Yago<sup>7</sup> ontology (see Figure 4).

The Yago ontology comprises well known knowledge about the world [4]. It contains information extracted from Wikipedia and other sources like WordNet and GeoNames 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”, the entity “tennis player” with instances like “Roger Federer”, “Rafael Nadal”, 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. In Figures 5 and 6 are presented results obtained with “tennis player” query in Google and in our application. How we can see, results offered by Google are similar from the point of view of concepts presented in

<sup>6</sup><http://lucene.apache.org>

<sup>7</sup><https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago>

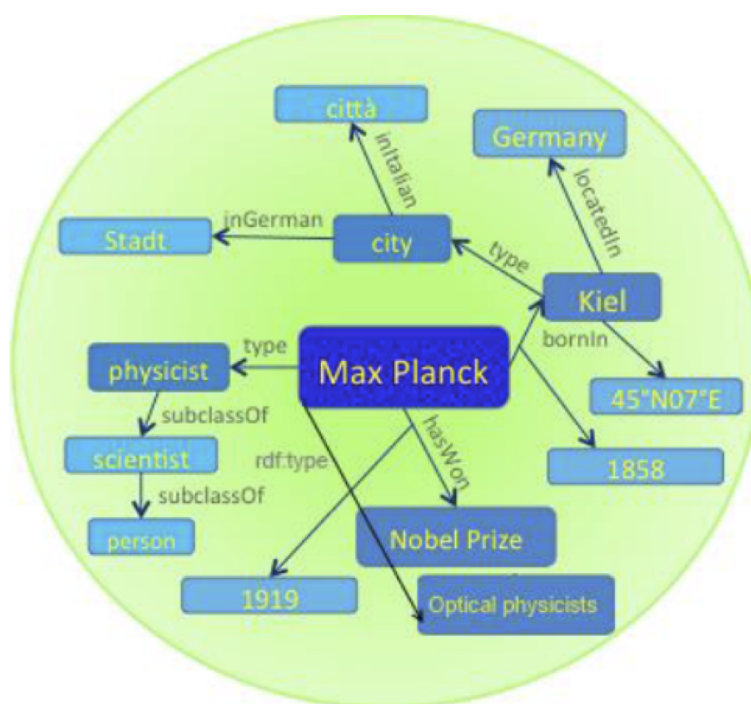


FIGURE 4: EXAMPLE FROM THE YAGO ONTOLOGY

images returned. In the case of our application are more “colours” and more concepts in comparison with results offered by Google.

Because of its structure, Yago will be used only when the text queries will match WordNet concepts that are linked by a hypernymy relationship to other Wikipedia entities, such as, person, location or organization.

To decide when to use Yago, we created a resource based on hierarchies of Wikipedia categories. For this, we started with Romanian Wikipedia which has 8 groups of categories: culture, geography, history, mathematics, society, science, technology, privacy. In turn, these categories have subcategories or links to pages directly. In the end, we obtained 8 big groups with 134 categories, which are subdivided into several subcategories and pages (hierarchical depth depends on each category and subcategory). In general, this hierarchy covers most of the concepts available for Romanian. For example, for Sport, we obtained 70 subcategories containing other subcategories and 9 pages. Going through these categories and subcategories, we built specific resources with words that signal concepts of type person, location and organization.

### 3.4 AUTOMATIC IMAGE ANNOTATION

The project that we want to present proposes that each image is connected to relevant keywords according to its content. 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



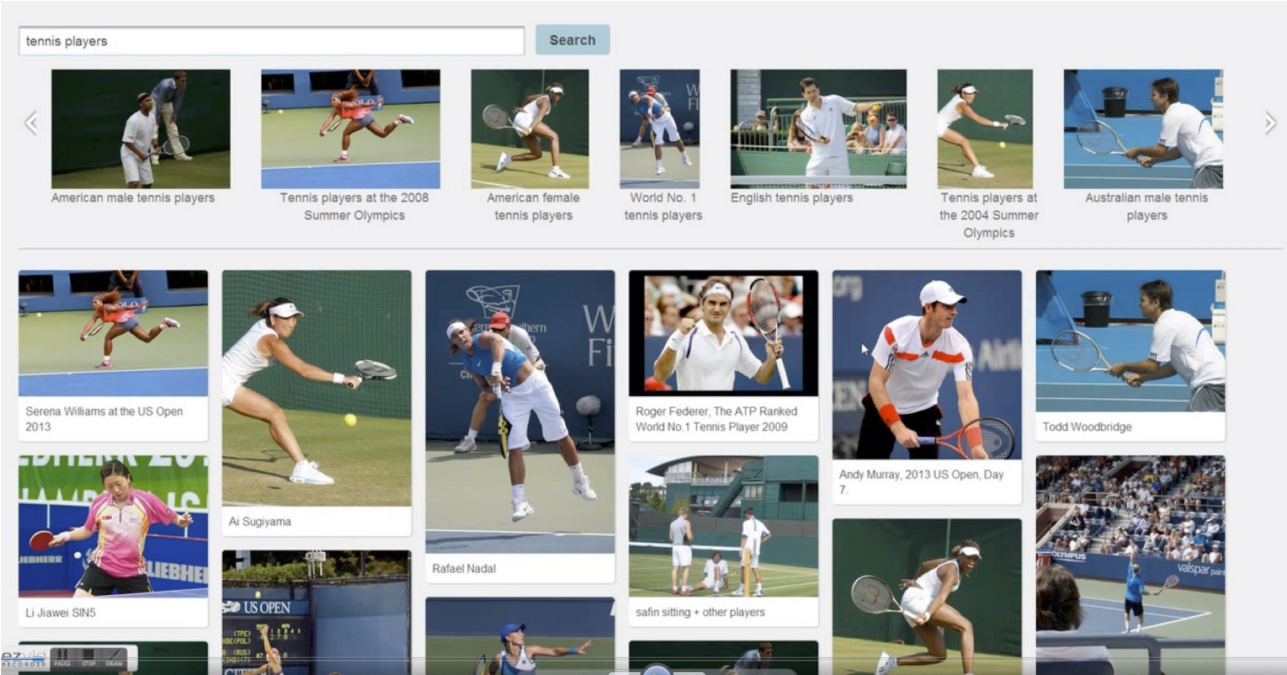


FIGURE 5: RESULTS OFFERED BY OUR APPLICATION FOR QUERY "TENNIS PLAYER"

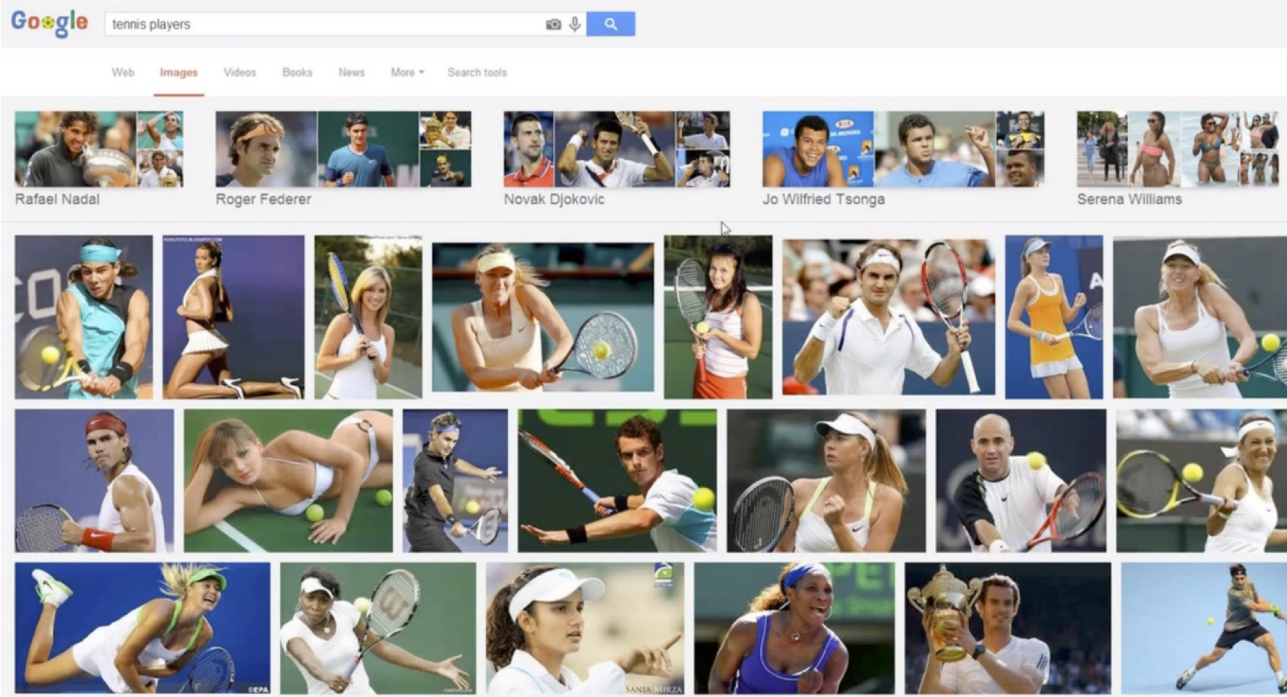


FIGURE 6: RESULTS OFFERED BY GOOGLE IMAGE SEARCH FOR QUERY "TENNIS PLAYER"

image.

### 3.4.1 CREATION OF GOLD COLLECTION WITH ANNOTATED IMAGES

The initial collection of images consisted of 100 images, from different areas. The images were categorized in the following proportions: 30% images with peoples, 15% images from nature, 20% images with animals and the remaining images were from various categories (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, there were 28 volunteers in third-year and master students of the Faculty of Computer Science from Iași. They had to annotate 100 images; 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. Each user was able to annotate how many pictures she/he wanted, but in the analysis entered only keywords by 21 users who have annotated all 100 images. In Figure 7 we can see the application used during the annotation process.

Performing an analysis on what users have annotated over a period of two weeks, 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. As a general rule, they have chosen to annotate the content of the image that quickly appears in sight. In the end, the 21 users have entered a total of 1,514 keywords for 100 images.

Furthermore, we have implemented an algorithm which, for each image, counts the frequency of lemmas of the keywords associated by users and keeps those with a frequency of at least 4. Beside frequency, we considered the relation of synonymy using the Romanian WordNet. Among all synonyms, we kept the keyword which appears more often at the users who have annotated the image.

For expressions, we used the division into component words, and then we calculated the frequency of word components based on lemma and synonymy. If all components of the expression had an occurrence frequency over 4, we decided to keep the expression and give up the words which appeared in the expression. In the end, we considered for every image a list of keywords in a descending order of frequency (of course, for frequencies over 4).

Each image initially contained around 30-40 different keywords from all users, and afterward, we applied the algorithm, the number of keywords was reduced to approximately 3-4 keywords per image. The average remained of 3.32 keywords per image, with a minimum of 1 and a maximum of 7. It can be seen that filtering was done quite rigorously.

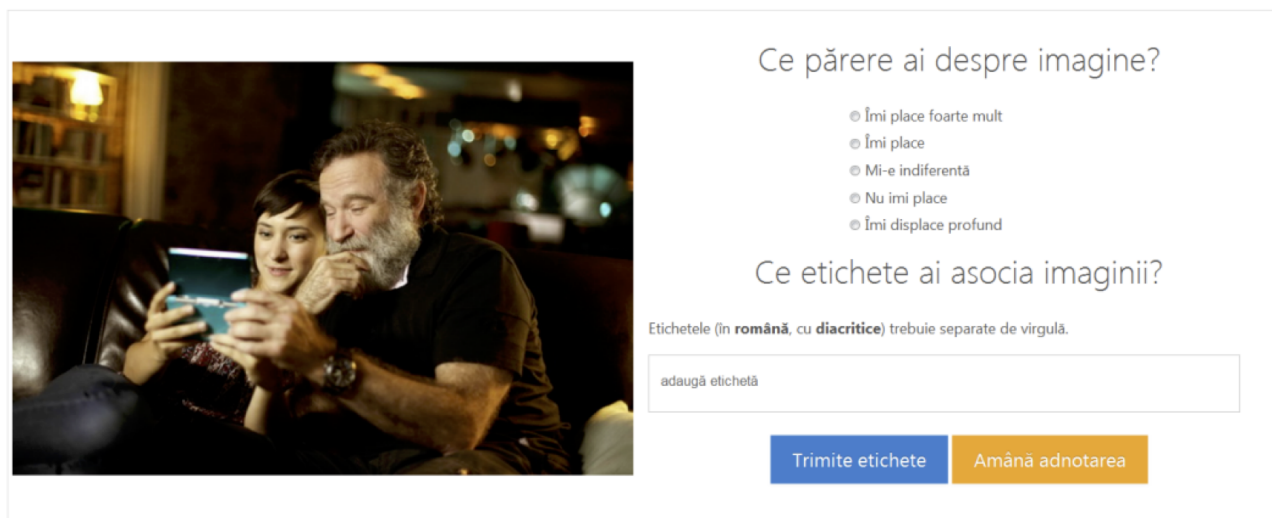
After completing this step, we increased the initial collection with 100 images as it follows. For

# Adnotare imagini

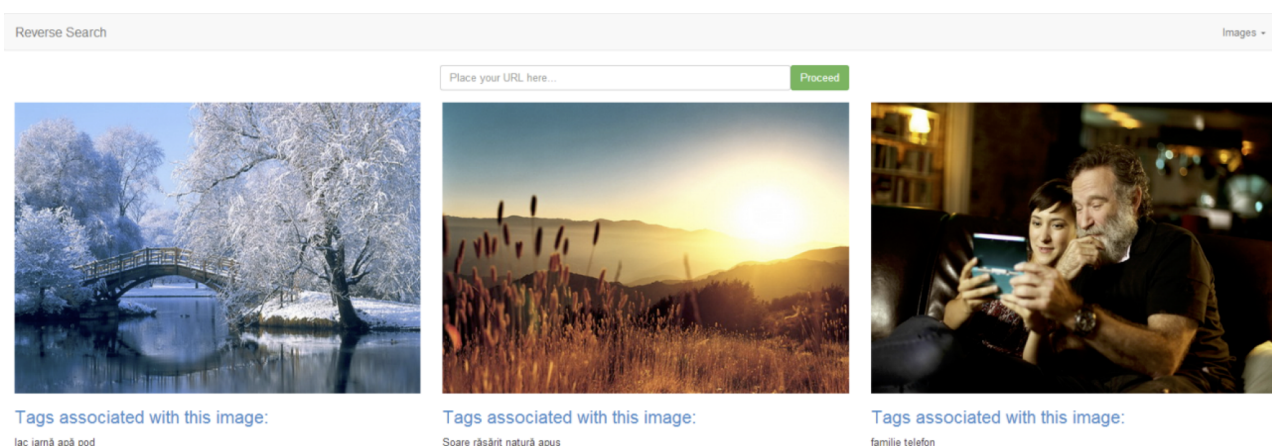
adiftene@info.uaic.ro

Ai adnotat 8 imagini până acum.

Deconectare



**FIGURE 7:** APPLICATION INTERFACE WHERE USERS CAN ANNOTATE IMAGES



**FIGURE 8:** REVERSE IMAGE SEARCH APPLICATION

each image from the initial collection, 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.

### 3.4.2 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 as it follows:





- The user introduces an URL and presses the Proceed button (See Figure 8).
- Behind the applications, we use the LIRE<sup>8</sup> library (Lucene Image REtrieval), 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.
- To establish the list of keywords that we associate with an image and their order in this list, we apply the algorithm from section above. For that, the input that we use is represented by lists of keywords from 20 similar images, and then we use lemmatization, the synonymy relation and the processing of expressions.
- An exception to the above is the case when all Euclidean distances between the new image and all images from the collection are below 0.2. This value was found experimentally and it tells us that the new image is too different in comparison with the existing images from our collection. In this case, we can't associate keywords to the new image.

In Figure 9 is the case when the application works as we wanted and it is able to build a list of relevant keywords to be associated to a new image. It can be seen that there are similar images in the database with the one introduced by the user, and the list of keywords contains relevant keywords for this image.

In Figure 9, because of the limit imposed by the Euclidean distance, the user will receive a negative response. This means that the algorithm didn't find similar images with the user image in the collection of images, and thus it is unable to create a list of keywords that characterize it. It can be seen that in the collection of images there are no similar images with the one introduced by the user (the images shown in the second row have the Euclidian distance below the limit imposed by us). We also note that these images don't contain common keywords, which may characterize the new image inserted by the user.

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

## 4 CONCLUSIONS

---

The work described in previous sections show that linguistic processing technologies and resources can have a significant contribution to improving image retrieval. The main means to do so is either by improving the way the search query is used (expanding search terms, finding semantic relations between search terms and image clues, using multilinguality to extend the scope of the search) or by improving the keyword description of an image. Valuable Natural Language Processing tools were identified and used (POS-taggers, lemmatisers, Anaphora solvers, Name Entity Recognizers, indexers and others). Also of significant value are the freely available (and of high quality) Language Resources such as the WordNet, Yago, Wikipedia, GeoNames and Freebase.

<sup>8</sup><http://www.semanticmetadata.net/lire/>



Image added by you:



Tags associated with this image:

mare nisip plajă

Image similar with the one above:



Tags associated with this image:

mare nisip plajă umbrelă



Tags associated with this image:

mare nisip palmier plajă



Tags associated with this image:

culori vesele copii

**FIGURE 9:** REVERSE IMAGE SEARCH WHEN THERE ARE SIMILAR IMAGES IN THE COLLECTION

Image added by you:



There are no tags associated with this image!

Image similar with the one above:



Tags associated with this image:

dispozitive tehnologice cloud



Tags associated with this image:

cuvințe wikipedia



Tags associated with this image:

Disney Mickey Mouse desene animate copilărie Donald personaje

**FIGURE 10:** REVERSE IMAGE SEARCH WHEN THERE ARE NOT SIMILAR IMAGES IN THE COLLECTION

Quantitative evaluation was carried out for all the described experiments with good results, more details are available in [10, 2, 6, 8]. The work carried out as part of Task 2.1 also led to successful participation to tracks in CLEF 2013 and to the development of image annotation software and a significant collection of manually annotated images.

## References

---

- [1] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence, IJCAI'07*, 2007.
- [2] L.M. Gherasim and A. Iftene. Extracting background knowledge about world from text. In *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, 2014.
- [3] A. L. Ginsca, E. Boros, A. Iftene, D. Trandabat, M. Toader, M. Corici, C. A. Perez, and D. Cristea. Sentimatrix - multilingual sentiment analysis service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 2011.
- [4] J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. In *Artificial Intelligence, vol. 194*, 2013.
- [5] A. Iftene, A. Moruz, and E. Ignat. Using anaphora resolution in a question answering system for machine reading evaluation. In *Notebook Paper for the CLEF 2013 LABs Workshop - QA4MRE*, 2013.
- [6] A. Iftene, A. Siriteanu, and M. Petic. 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"*, 2014.
- [7] A. Iftene, D. Trandabat, A. Moruz, I. C. Pistol, M. Husarciuc, and D. Cristea. Question answering on english and romanian languages. In *C. Peters et al. (Eds.): CLEF 2009, LNCS 6241, Part I (Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments)*, 2010.
- [8] A. Laic and A. Iftene. Automatic image annotation. In *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, 2014.
- [9] P. Langlais Laroche. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *23rd International Conference on Computational Linguistics (Coling2010)*, 2010.
- [10] A. Popescu. Cea list's participation at the clef chic 2013. In *Working Notes of CLEF 2013*, 2013.

- [11] A. Popescu and G. Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval*, 2011.
- [12] R. Simionescu. Hybrid pos tagger. In *Proceedings of "Language Resources and Tools with Industrial Applications" Workshop*, 2011.