CEA LIST's participation at the CLEF CHiC 2013

Adrian Popescu

CEA, LIST, Vision & Content Engineering Laboratory, Gif sur Ivette, France
adrian.popescu@cea.fr

Abstract.
For our first participation to the CLEF CHiC Lab, we submitted runs to the multilingual ad-hoc and multilingual semantic enrichment tasks. 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. ESA was adapted for the 10 languages that are best represented in the Europeana corpus. Wikipedia dumps from March 2012 were used for French and English and from March 2013 for the other languages. One problem that arises when modeling short documents, such as queries, with classical ESA vectors no information is available whether the concept is related to the entire topic only to a part of it. To overcome this problem, two adaptations of ESA (ESA-C) are Wikipedia concepts that are linked to the highest number of concepts from the original topic. In the ad-hoc task, ESA and ESA-C have two roles: to expand the topic with related concepts and to create consolidated topic models which contain the original topic words along with other related keywords. We submitted both monolingual and multilingual runs without topic expansion and using topic expansion and consolidation with classical ESA and ESA-C. The best results are obtained in a multilingual setting with no expansion and ESA-C topic consolidation. For the semantic enrichment task, we propose lists of related Wikipedia concepts using either a monolingual ranking or a voting scheme that surfaces related concepts that appear in the largest number of languages. Here, the best results are obtained in a monolingual ranking configuration that exploits ESA-C.

Keywords

Multilingual Information Retrieval, Semantic Enrichment, Explicit Semantic Analysis

1 Introduction

The CHiC evaluation lab[1] aims to explore different aspects related to the retrieval of cultural heritage content stored in digital libraries such as Europeana[2]. There are two subtasks in the multilingual task, dealing with multilingual ad-hoc retrieval and multilingual semantic enrichment. For the ad-hoc task, participants were provided with metadata in 13 different languages and were free to use any automatic method in order to return ranked lists of results for a set of 50 diversified topics. For the semantic enrichment task, a subset of 25 topics from the initial pool was provided and the objective was to return a ranked list of 10 related concepts that could be used to enrich the initial topic and might help to precise the user's information need.

[1] http://www.promise-noe.eu/chic-2013/home
[2] http://www.europeana.eu/

2 Explicit Semantic Analysis (ESA)

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.

Radinsky and al. [4] added a temporal temporal dimension to ESA vectors and showed that this addition improves the results for word relatedness.

Hassan and Mihalcea [2] introduced Salient Semantic Analysis, a variant of ESA that relies on the detection of salient concepts prior to linking words and concepts. The merits of their method are difficult to estimate since the comparison is often made with an in-house ESA implementation whose results are significantly poorer than those presented in [1].

We proposed an ESA adaptation to information retrieval tasks that gives priority to categorical information [3]. The comparison with a classical ESA implementation showed that a significant improvement was obtained in an image retrieval setting. Moreover, the method compared favorably with other state of the art indexing and retrieval schemes. Here, we extend the work in [3] and propose to use ESA for query expansion and consolidation, two operations that are explained in more details in Subsection 4.

ESA has only weak language dependence and was already deployed in several languages. Sorg and Cimiano [5] proposed an extension of the method to different languages and showed that the method is useful in cross-lingual and multilingual retrieval settings. Here we create ESA vectors in the 10 most represented languages out of the 13 present in the Europeana collection. The following languages are supported: English, German, French, Spanish, Italian, Dutch, Swedish, Norwegian, Polish and Finnish. Adaptations to different languages include detection and removal of Wikipedia disambiguation and list pages and detection of category section.

2.1 Classical Formulation of ESA

Put simply, ESA exploits classical text weighting schemes, such a tf-idf, to model concepts from a structured resource, such a 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 (see Subsection 2.2.

Classical ESA representations are well adapted for single words, since there is nothing to be done, and for long documents, since the summing operation smooths individual contributions and an accurate semantic representation of the document is obtained. However, the method has some drawbacks for documents such as retrieval topics that contain only few words (typically 2 to 4 words). Here, the smoothing of individual contributions is not sufficient because the contribution of a single word can be higher than that of the others and the obtained related concepts could be related to a part of the topic only. An illustration of this type of problem is provided in table 1 which presents the top 10 ESA concepts associated to topics *Freshwater Fish* and *Jean-Jaques Rousseau*[3]. The results from table 1 indicate that most ESA top ranked concepts are not related to the entire query. When examining results for topic CHiC-051, *Freshwater bivalve* and *Freshwater, Isle of Wight* are related to *freshwater* while *Bait fish* and *Bank fishing* are related to fish. Similarly, when examining results for topic CHiC-058, we notice that several ESA top concepts are brought up by the family name textitRousseau and have little semantic relatedness

[3] The wrong spelling of Rousseau's name was extracted from Europeana logs and provided as such for the task.

with the original topic. Concepts found for topic CHiC-064 (*crockery doll house*) are related to doll but not the other terms from the query.

We use an in-house implementation of ESA that includes only the optimization cues publicly available until recently [4]. To validate our implementation, we performed the word similarity task described in [1], with the same version of Wikipedia, and the method achieved a 0.72 correlation with human judgments (to be compared with 0.75 reported by [1]).

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, presented in subsection 2.1, while the third results for the adapted version of ESA (ESA-C) presented in subsection 2.2.

Topic CHiC-051 *Freshwater Fish*		
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
Topic CHiC-058 *Jean-Jaques Rousseau*		
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
Topic CHiC-064 *Crockery doll house*		
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)

[4] A full list was recently made public at https://github.com/faraday/wikiprep-esa/wiki/roadmap but the remaining cues were not yet integrated in our implementation.

2.2 ESA Adaptation for ad-hoc multilingual IR (ESA-C)

We already proposed a version of ESA that gives a privileged role to categorical information in [3]. There, 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 their are broke using the second, finer grained score.

The introduction of the boolean score has two main objectives. First, categorical information should be favored 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. 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 concepts 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 character is useful since the first paragraph has 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 favors articles that contain longer concepts from the initial query over other articles. At equal categorical scores, the inclusion of concept detection allows us to favor a Wikipedia concept that includes *Jean-Jacques Rousseau* in its text when compared to another concept that includes *Jean-Jacques* and *Rousseau* separately. The top related concepts obtained with ESA-C for *Freshwater fish* and *Jean-Jacques Rousseau* are presented in the third column of table 1. In both cases, the top 10 concepts are much more closely related to the initial topics compared to the use of classical ESA. The list for topic CHiC-051 contains only fish and the list for CHiC-058 includes different works of *Jean-Jacques Rousseau*. *Crockery doll house* is a specialized one, which is not well represented in Wikipedia and the retrieved concepts are still unrelated to the entire topic in a large majority of cases. This last topic illustrates one limitation of all ESA implementation, namely the poor mapping between the initial document and the knowledge included in the underlying conceptual space.

The use of the lists of related ESA concepts for both semantic enrichment and for ad-hoc retrieval is detailed in Section 4.

3 Europeana Collection Processing

We kept documents of the Europeana Collection that belong to the 10 languages processed with ESA. Separate indexes were created for each of the modeled languages. Then selected metadata associated to the following fields: “dc:title”, “dc:description”, “dc:subject”, “dc:type”, “dcterms:medium”, ”dc:date“. The retained fields were merged into a bag-of-words representation and then modeled using a tf-idf scheme. Due to the use of probabilistic models of the topics (see Subsection 4.2), the tf-idf representation of documents was subsequently transformed into a probabilistic form by dividing the weight of each word by the sum of the scores of all words in the document. Monolingual collections are stemmed using the corresponding Perl Snowball stemmer implementation.

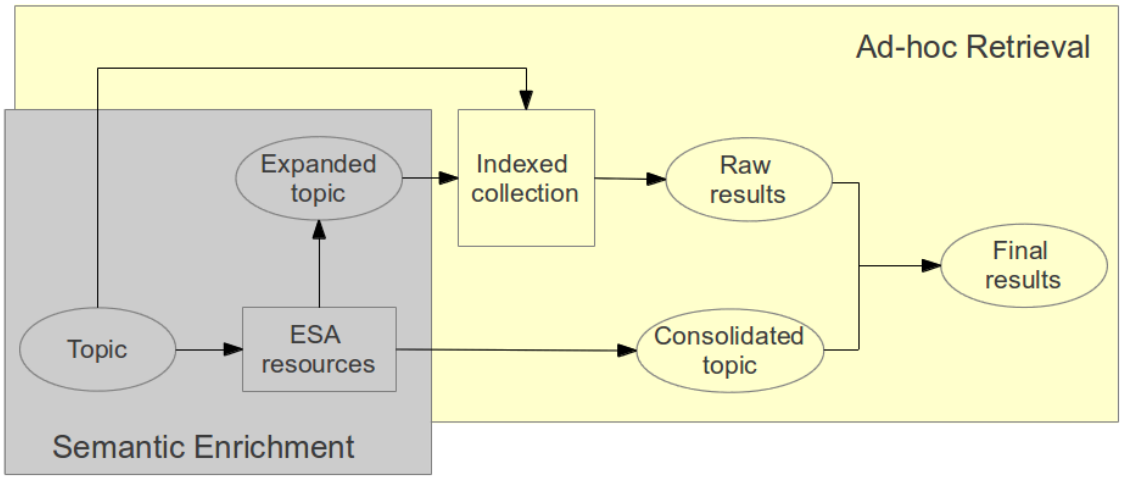

Fig. 1. Overview of the semantic enrichment and retrieval framework.

4 Enrichment and retrieval framework

The framework devised here was used for both semantic enrichment and ad-hoc retrieval and is summarized in figure 1. 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 detailed in Subsection 4.2.

4.1 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 use 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.

4.2 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.

Table 2. Top 10 words form the consolidated version of the topics for *Freshwater Fish*, *Jean-Jaques Rousseau* and *crockery doll house*. The second column contains results for classical ESA, presented in subsection 2.1 while the third column presents results for the adapted version of ESA (ESA-C) presented in subsection 2.2.

Topic CHiC-051 *Freshwater Fish*		
Rank	ESA	ESA-C
1	fish	fish
2	freshwater	freshwater
3	acquarium	galaxia
4	fillet	aquarium
5	fishery	species
6	bait	fin
7	water	water
8	species	river
9	lake	trout
10	fin	carp
Topic CHiC-058 *Jean-Jaques Rousseau*		
Rank	ESA	ESA-C
1	jacques	rousseau
2	jean	jean
3	rousseau	de
4	de	jacques
5	french	french
6	le	le
7	saint	paris
8	paris	philosopher
9	la	pygmalion
10	baptiste	pierre
Topic CHiC-064 *Crockery doll house*		
Rank	ESA	ESA-C
1	doll	doll
2	house	house
3	toy	toy
4	barbie	barbie
5	goo	mattel
6	album	album
7	mattel	goo
8	nrhp	film
9	licca	licca
10	song	song

Topic expansion is performed in a way similar to description provided in Subsection 4.1. Consolidation is a by-product of the expansion process and aims to obtain an expanded version of the initial topic that contains, in addition to the original words, other words that are semantically related to the topic but are not part of it. The words are ranked by summing their individual scores associated to the top 100 ESA related concepts and then by multiplying this sum with the log of the number of different articles in which they appear. This last score is used in order to favor words that are associated to a large number of Wikipedia concepts related to the initial articles. Up to 1000 related words are retained for each topic and a probabilistic model of the consolidated versions is obtained by dividing individual word scores by the sum of all scores. In table 2, we present top 10 words related to each topic obtained with ESA and ESA-C. A majority of the obtained words are semantically related to the initial topic, although some outliers appear. For *freshwater fish*, all top 10 words are related to the initial topic and thus useful for ranking results. In the case of *Jean-Jaques Rousseau*, there are French stop words that were not removed. For *Crockery doll house*, *nrhp* (abbreviation of National Register of Historic Places) appears since this acronym is strongly related to *house*. Similarly to the collection processing, the consolidated versions of the topics are stemmed.

5 Experiments

As we mentioned, we have 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 subsections.

5.1 Semantic Enrichment

Submitted runs The following eight runs were submitted to the semantic enrichment subtask:

- *ceaListEnglishMonolingual* - Related concepts are obtained with ESA-C. The experiment is monolingual since it only exploits the English Wikipedia version. Proposed expansions are collection independent.
- *ceaListEnglishMonolingualOriginal* - Related concepts are obtained with classical ESA. The experiment is monolingual since it only exploits the English Wikipedia version. Proposed expansions are collection independent.
- *ceaListEnglishRankEnglish* - Rank fusion for monolingual results obtained with ceaListEnglishMonolingual. The rank of the concept is obtain by averaging its ranks in different languages. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only concepts that have an English version are considered.
- *ceaListEnglishRankMultilingual* - Rank fusion for monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score in an individual language is presented.
- *ceaListEnglishBooleanEnglish ** - Fusion of boolean scores for monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only English versions of Wikipedia concepts are presented. Proposed expansions are collection independent. Only concepts that have an English version are considered.

– *ceaListEnglishBooleanMultilingual ** - Fusion of boolean scores monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score in an individual language is presented. Only concepts that have an English version are considered.
– *ceaListEnglishCosineEnglish ** - Fusion of cosine similarity scores for monolingual results obtained with ceaListEnglishMonolingualOriginal. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only English versions of Wikipedia concepts are presented. Only concepts that have an English version are considered.
– *ceaListEnglishCosineMultilingual ** - Fusion of cosine similarity scores monolingual results obtained with ceaListEnglishMonolingualOriginal. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score is presented. Only concepts that have an English version are considered.

Results Even though the results for 6 out of 8 runs are biased here, there are some interesting conclusions that we can draw from table 3. The comparison between ceaListEnglishMonolingualOriginal and ceaListEnglishMonolingual is favorable 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.

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 show 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.

Table 3. 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	**0.468**	**0.66**
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

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*.

5.2 Ad-hoc retrieval

Submitted runs 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 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. As we mentioned, multilingual fusion was performed by The following 16 runs were submitted to the semantic enrichment subtask:

- **ceaListMultilingualNoExpansion** - Multilingual run that retrieves only documents which contain at least one word from the initial topic.
- **ceaListFrenchNoExpansion** - Monolingual French run that retrieves documents which contain at least one word from the initial topic.
- **ceaListGermanNoExpansion** - Monolingual German run that retrieves documents which contain at least one word from the initial topic.
- **ceaListMultilingualOriginal *** - Multilingual run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Wikipedia concepts obtained with classical ESA.
- **ceaListMultilingualFiltered *** - Multilingual run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Wikipedia concepts obtained with ESA-C.
- **ceaListDutchFiltered *** - Monolingual Dutch run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Dutch Wikipedia concepts obtained with a version of Explicit Semantic Analysis adapted to short documents (topics).
- **ceaListEnglishFiltered *** - Monolingual English run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related English Wikipedia concepts obtained with ESA-C.
- **ceaListFrenchFiltered *** - Monolingual French run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related French Wikipedia concepts obtained with ESA-C.
- **ceaListGermanFiltered *** - Monolingual German run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related German Wikipedia concepts obtained with ESA-C.
- **ceaListItalianFiltered *** - Monolingual Italian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Italian Wikipedia concepts obtained with ESA-C.
- **ceaListNorwegianFiltered *** - Monolingual Norwegian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Norwegian Wikipedia concepts obtained with ESA-C.
- **ceaListPolishFiltered *** - Monolingual Polish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Polish Wikipedia concepts obtained with ESA-C.
- **ceaListSpanishFiltered *** - Monolingual Spanish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Spanish Wikipedia concepts obtained with ESA-C.
- **ceaListSwedishFiltered *** - Monolingual Swedish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Swedish Wikipedia concepts obtained with ESA-C.
- **ceaListEnglishOriginal *** - Monolingual English run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related English Wikipedia concepts obtained with ESA-C.
- **ceaListItalianOriginal *** - Monolingual Italian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Italian Wikipedia concepts obtained with classical ESA.

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 involved expansion and the others. However, it is worthwhile noticing the 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. O.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 [3], which showed that giving a privileged role to categorical information is beneficial in an image retrieval scenario. The favorable 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 4. MAP performances for ad-hoc retrieval runs.

Run name	MAP
ceaListMultilingualNoExpansion	**0.1878**
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

6 Conclusion

The results obtained at CLEF CHiC 2013 encourage us to pursue the investigation of enrichment and retrieval techniques for cultural heritage. The obtained results are especially encouraging for the semantic enrichment task. The ESA-C adaptation of Explicit Semantic Analysis clearly outperforms the original version of the method. The ad-hoc retrieval results were hampered by the overrated value of boolean scores and we plan to correct these experiments in order to have an accurate assessment of the influence of automatic topic expansion.

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 modeling 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.

Acknowledgments

This work was supported via Egonomy, a French "Grand Emprunt" project.

References

1. Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
2. Samer Hassan and Rada Mihalcea. Semantic relatedness using salient semantic analysis. In *AAAI*, 2011.
3. Adrian Popescu and Gregory Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval*, pages 33:1–33:8, 2011.
4. Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.
5. P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, 2012.