Chemnitz at the CHiC Evaluation Lab 2012:

Creating an Xtrieval Module for Semantic Enrichment

Jens Kürsten, Thomas Wilhelm, Daniel Richter, and Maximilian Eibl

Chemnitz University of Technology, Dept. of Computer Science, 09107 Chemnitz, Germany
{firstname.lastname}@cs.tu-chemnitz.de
http://www.tu-chemnitz.de/cs/mi

Abstract. Cultural heritage is one of the most valuable resources that describe the creative power of mankind. In this article we describe a total number of 96 experiments that have been submitted as contributions to the three subtasks of the *Cultural Heritage in CLEF* pilot evaluation lab. At the core of the majority of these experiments lies a prototype implementation for semantic enrichment based on DBpedia. The evaluation of the experiments demonstrate that semantic enrichment does not improve retrieval effectiveness in comparison to straightforward baseline experiments. The results also indicate that automatic query expansion does not improve retrieval performance for the pilot lab test collection. Further experiments are needed in order to be able to draw conclusions on whether semantic enrichment can improve retrieval results on cultural heritage collections or not.

Keywords: Ad-hoc Retrieval, Semantic Enrichment, Cultural Heritage

1 Introduction

Cultural heritage is one of the most valuable resources that describe and document human creative power. Nowadays, many different types of organisations, such as libraries, museums, and audiovisual archives, own specific collections which provide an insight into contemporary history. [1] Web portals like Europeana[1] aim to provide access to a wide range of cultural heritage collections, but a variety of challenges like different types of documents (namely text, image, audio, and video), different metadata description schemes, or different languages contribute to the complexity of the underlying retrieval system. In order to provide the user with the information that is most valuable to her or him, the *Cultural Heritage in CLEF (CHiC)* pilot evaluation lab [2] addresses these key problems by means of three types of evaluation tasks:

- Ad-hoc Retrieval Task,
- Variability Task,
- Semantic Enrichment Task.

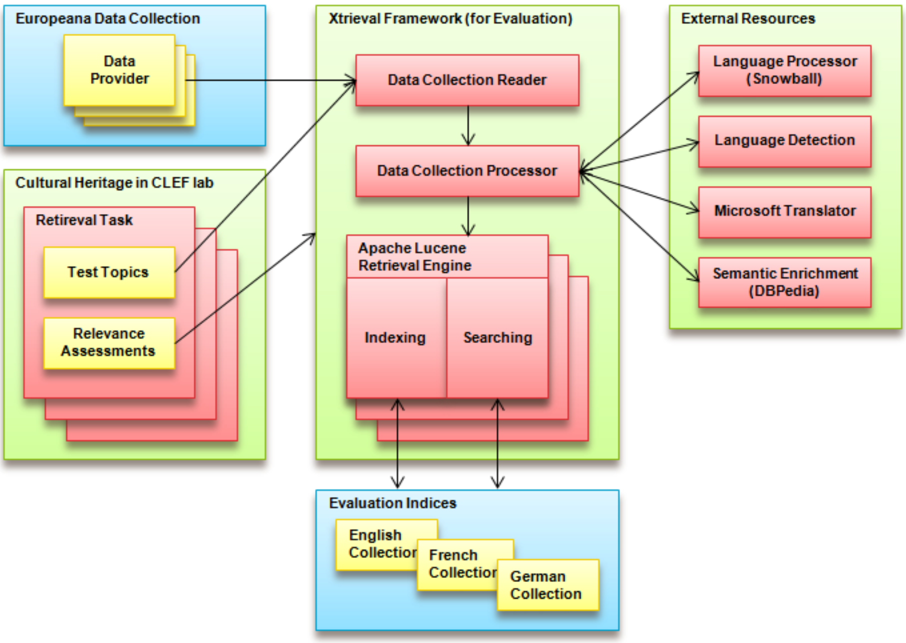
[1] http://www.europeana.eu/portal/ (retrieved on August 16th, 2012)

This contribution describes the system and the resources that have been used to tackle all of the CHiC tasks. It continues with the description of a semantic enrichment module. Then it provides an overview on the experimental set-up that was employed to approach the individual subtasks. A summary of all submitted experiments is provided subsequently. They are presented together with an analysis of the obtained results and further experiments. The final section of this article provides a review of the most important observations and resulting directions for future work on the topic.

2 System Overview

The experiments for the CHiC evaluation lab set an important milestone for the Chemnitz retrieval group: the Xtrieval framework [3] has been used for five years and a variety of retrieval tasks in the context of CLEF (see [4] for an overview of past results). Naturally, in order to design and implement the experiments for the Ad-hoc and Variability tasks the Xtrieval framework was used again. An additional module has been developed to create contextual expansions for the semantic enrichment task. Figure 1 illustrates the system architecture and the resources that were used in the experiments.

Fig. 1. Schematic overview on the system architecture and employed resources



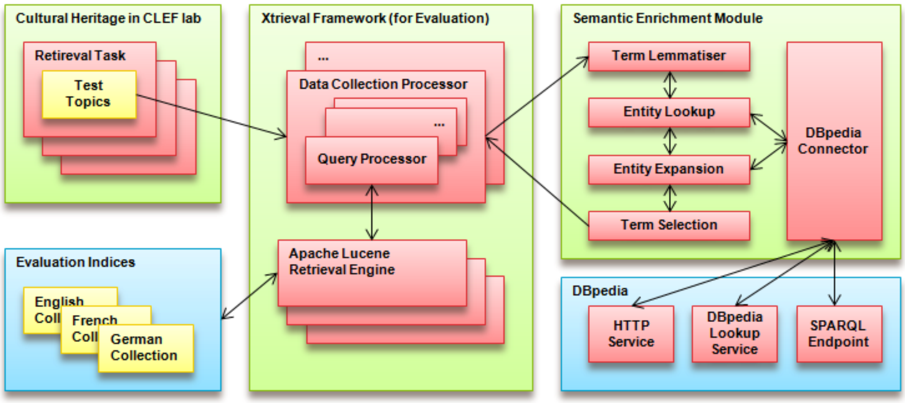
The following resources were used to prepare and to conduct the retrieval experiments:

- Apache Lucene[2] in version 3.6 as the core retrieval engine,
- The Snowball project[3] for stop word removal and stemming,
- Language detection[4] to analyse the language distribution and validity of tags,
- Microsoft Translator[5] to generate queries in collection-specific languages,
- DBpedia[6] to extract enrichment terms for short topics.

3 Xtrieval Extension for Semantic Enrichment

A specific problem in the domain of web search and cultural heritage web portals are very short queries. Retrieving documents for such queries is very difficult due to the lack of context information. One approach to address the issue is to develop sophisticated algorithms for automatic or semi-automatic semantic enrichment. For the experiments presented hereafter, a term-based enrichment module has been developed as extension for Xtrieval. This module aims to return a set of entities containing broader or more specific terms for a given query or concept. In its first prototype implementation it provides access to DBpedia resources. DBpedia can be "considered [as] the Semantic Web mirror of Wikipedia". [7] It allows the extraction of factual information from Wikipedia pages as well as the connections between pages.

Fig. 2. Abstract representation of the employed semantic enrichment



[2] http://lucene.apache.org/core/ (retrieved on August 16th, 2012)

[3] http://snowball.tartarus.org/ (retrieved on August 16th, 2012)

[4] http://code.google.com/p/language-detection/ (retrieved on August 16th, 2012)

[5] http://www.microsofttranslator.com/dev/ (retrieved on August 16th, 2012)

[6] http://dbpedia.org/ (retrieved on August 16th, 2012)

[7] http://wiki.dbpedia.org/DBpediaLive (retrieved on August 16th, 2012)

Specific features that make DBpedia attractive for the problem at hand are:

- Inter-language links, which can be used for translation,
- Disambiguation links, which help to direct a query to specific topics,
- Relationship extraction, which provides conceptually related entities or terms,
- Redirects, which allow to guess a related entity based on a given (set of) term(s),
- Entity-specific information like geo-coordinates, location names, or person data,
- Links to specific resources for term-based expansion such as article categories, category labels, general labels, or article abstracts.

Although the area of Semantic Web research created the foundations for the design and development of the semantic enrichment module for Xtrieval, the focus of this contribution lies in the experiments to approach the tasks of the CHiC evaluation lab. More information on Semantic Web technology and research as well as the details on the architecture and use cases of DBpedia can be found in [5], [6], and [7].

An easy integration into Xtrieval was one particular requirement of the semantic enrichment extension (see Figure 2). For this reason the semantic enrichment module (SEM) has to be developed in Java[8]. As the DBpedia provides web-based application programming interfaces, the extension is implemented on top of the Apache HttpComponents library[9]. Three types of interfaces to DBpedia are accessed in order to obtain semantic enrichments:

- The SPARQL endpoint[10],
- The DBpedia lookup service[11],
- The DBpedia named pages graph[12].

These interfaces are used for entity discovery (lookup service), verification (named pages graph), and expansion (SPARQL). To avoid the repeated fetching of identical resources the *DBpedia Connector* (see Figure 2, right) implements an HTTP client that supports local caching. Out of the four components of the SEM, only two (namely *Entity Lookup* and *Entity Expansion*) are employed to query the DBpedia resources.

The semantic enrichment process works as follows. In a first step the *Term Lemmatiser* transforms the given terms (i.e. the topic titles), which may or may not represent one or more named entities. This procedure includes the following steps:

- Stop word removal (restricted to the first or last terms),
- N-gram analysis for each individual term for a later comparison of similarity with discovered named entities,
- Term order alternation to discover “hidden” named entities.

[8] http://www.java.com/ (retrieved on August 16th, 2012)

[9] http://hc.apache.org/ (retrieved on August 16th, 2012)

[10] http://dbpedia.org/sparql (retrieved on August 16th, 2012)

[11] http://lookup.dbpedia.org/api/search.asmx/ (retrieved on August 16th, 2012)

[12] http://dbpedia.org/resource/ (retrieved on August 16th, 2012)

The resulting terms are treated as a stream of tokens and transferred to the *Entity Lookup* component. Here, the actual lookup via DBpedia is performed repeatedly for all possible combinations of terms and term orders, starting with the complete stream of tokens. This process is continued by removing individual terms until only a list of individual terms remains. These individual terms are also treated as entity candidates and checked via *Entity Lookup*. In case of a successful lookup, the level of the process determines the further course of action. The longer the stream of tokens that matched a DBpedia entity the more valuable this entity will be. For this reason the ratio of the length of the matching stream of tokens by the original length indicates the quality of the discovered entity.

Since the value of semantic enrichment based on entities found at the individual term level might be very small, the *Entity Expansion* component is used to exploit the links and descriptive meta-data between such entities. This might be useful for queries like "*Ulysses by Joyce*", where Ulysses and Joyce will return a number of potential DBpedia entities, but each individual term alone does not yet allow drawing the conclusion to the actual concept. However, the missing link is contained in the connection between the two DBpedia entities "*Ulysses (book)*" and "*James Joyce*". For this reason the *Entity Expansion* component aims to resolve this relationship by exploring the links between DBpedia entities that were found for individual terms.

Another main task of this component is to extract content descriptions from known DBpedia entities. This is the basis for the final step of the SEM that takes places in the *Term Selection* component. It receives a list of terms, which had been extracted from DBpedia entities, and it creates a weighted list of term candidates for the reformulation of the query. In the first prototype implementation used for the experiments, this procedure was treated as an automatic query expansion process. This allowed the use of standard query expansion algorithms, but the corresponding figures like local and global term counts or document frequency had to be obtained from the indices. Due to the flexibility of Xtrieval this could be implemented in a straightforward way. CSCorrect[13] from the Terrier retrieval toolkit [8] was used as query expansion algorithm.

4 Experimental Set-up and Results

As the CHiC evaluation lab in 2012 was a pilot retrieval task, the first step was to analyse the structure of the document collection in order to create efficient and meaningful index structures. For mid-sized test corpora like the one at hand, previous experiments have shown that a manual selection of document content can help to reduce the noise in index structures, which results in better retrieval performance [4].

In its latest version the Xtrieval framework has a very flexible and fast *Data Collection Processor* (see Figure 1) implementation that is based on the Jaxen library[14]. It exclusively relies on XPath for selecting the content from documents and determining

[13] The documentation of the Terrier platform provides the following description: "CSCorrect implements the un-simplified Chi-square divergence for query expansion."

[14] Jaxen: Java XPath engine, http://jaxen.codehaus.org (retrieved on August 16th, 2012)

the fields in the index. Each index that was created based on the mappings listed in Table 1. Note that the base path to the root of each individual document has been omitted for better clarity. Table 1 also shows that most of the original content from the document collection was used for indexing.

The language-specific sub-collections for English, German, and French were indexed once for each language in order to conduct the experiments for the Mono- and Bilingual subtasks. A specific problem that was found in the entire document collection was that four types of language tags are used to indicate the language of the document descriptions. What made matters worse was the fact that for some documents these tags may indicate different languages for an individual document. In some other cases the indicated language for the document language was in fact wrong. This was problematic for our approach to apply language-specific content analysers for each individual document in the multi-lingual collection. For this reason an additional filter algorithm was implemented. It evaluates six available sources of evidence based on the following priority:

1. `ims:language`
2. `europeana:country`
3. `europeana:language`
4. `dc:language`
5. `europeana:isShownAt`
6. `europeana:isShownBy`

Although all of these tags may not be present in each document, the algorithm compares the content of the existing tags. In case of a mismatch the language is compared with the language obtained by treating the country code top level domain in the URI of 5. and 6. as an ISO 3166-2 language code. If this language does not match any of the previous languages the document content is fed to a language detection library (see Section 2) in order to obtain the actual language of the document.

Table 1. Mapping of the document structure for indexing

Field name	XPath construct for document to index mapping
content	dc:publisher|dcterms:isPartOf|dcterms:spatial|dcterms:alternative|dcterms:created|dcterms:temporal|dc:creator|dc:date|dc:description|dc:title|dc:subject
enrichment	*[ends-with(name(),'_label')]
enrichment_url	*[ends-with(name(),'_term')]
provider	europeana:dataProvider|europeana:provider
type	europeana:type
type_desc	dc:type|dc:format|dcterms:medium

A total number of four indices has been created for the experimental evaluation. Starting with the *StandardTokenizer* of Lucene that splits a text stream into tokens and recognizes some entities like URLs or e-mail addresses. Additional filters (marked with *) that are implemented in Xtrieval and further filters from Lucene packages applied were subsequently. Stemming filters for each language were applied if available. This was the case for the following languages: German, Swedish, French, Norwegian, Italian, Spanish, English, Dutch, Finnish, Estonian, Hungarian, Russian, Portuguese, Turkish, Romanian, Polish, Greek, Bulgarian, Czech, Slovak, Slovenian, and Danish. The token stream processing was implemented as follows:

1. *LowerCaseFilter* – converts the token to lower case.
2. *RemoveShortWordsFilter** – removes words shorter than 3 characters.
3. *StopFilter* – removes stop words depending on the language.
4. *SnowballFilter* – stems the token according to the document language.

4.1 Ad-hoc Retrieval Task

The aim of our experiments that were submitted to the Ad-hoc Retrieval Task was to compare the implemented approach to semantic enrichment with an automatic pseudo-relevance feedback (*qe_kl*) and a baseline (*base*) run. The restriction that only four experiments could be submitted for each of the sub-tasks allowed two different configurations for SEM (*qe_dbp_abs* and *qe_dbp_sub*). For this reason only the source of the expansion terms of the DBpedia entities was alternated:

```
http://dbpedia.org/ontology/abstract
http://purl.org/dc/terms/subject
```

Each of these two resources corresponds to specific content of the original Wikipedia page for a given DBpedia entity. The *abstract* contains a natural language description of the DBpedia entity and is available in different languages. This allows automatic translations based on the DBpedia abstracts. The *subject* description refers to a number of Wikipedia category pages that are identified by their corresponding concept label.

Monolingual Experiments

Four experiments were submitted to each of the monolingual sub-tasks on the English, German, and French collections. All experiments were based on a very simple retrieval algorithm that submits the created queries to the “content” field only (see Table 1). The obtained experimental results are listed in Table 2.

The results for the monolingual experiments demonstrate that the experiments based on the SEM have been clearly outperformed by the baseline runs on the English and German sub-collections. Only for the French sub-collection there is very little variance across the tested system configurations. Another observation is that automatic feedback also decreased retrieval performance when querying in the English and German languages. Regarding the two sources of expansion terms from DBpedia entities no clear conclusion can be drawn from the experiments in German and French. On the English test collection, however, extracting expansion terms from subject descriptions clearly outperformed the ones extracted from abstract descriptions of the DBpedia entities.

Table 2. Results for the monolingual sub-tasks

run id	lang	configuration summary	MAP
base	EN	Lucene for core retrieval, no feedback	**0.4860**
qe_kl	EN	Lucene retrieval, KLCorrect[15] exp., 3 docs, 10 terms	0.4072
qe_dbp_abs	EN	Lucene retrieval, DBpedia exp., 20 terms (abstract)	0.3036
qe_dbp_sub	EN	Lucene retrieval, DBpedia exp., 20 terms (subject)	0.4179
base	DE	see corresponding runs above	**0.6039**
qe_kl	DE		0.5854
qe_dbp_abs	DE		0.4240
qe_dbp_sub	DE		0.4141
base	FR	see corresponding runs above	0.3300
qe_kl	FR		**0.3590**
qe_dbp_abs	FR		0.3227
qe_dbp_sub	FR		0.3205

Bilingual Experiments

Eight runs were conducted for each of the three sub-collections in English, German, and French. A subset of experiments needed to be held back in order to comply with the restriction to four experiments per sub-task. To account for the translation problem there was a slight modification to the monolingual experiment set-up.

The baseline experiment (*base*) did not contain any kind of translation mechanism. Microsoft's translation service (see Section 2) was used to translate the queries into the collection language for a second experiment (*ms*). The third experiment (*qe_dbp_abs*) was modified in a way that the DBpedia expansion returned the abstracts in the required collection language. The final experiment (*qe_dbp_sub_ms*) took the returned subject contents from DBpedia as input and translated these with Microsoft's translation service. Table 3 lists the results for all experiments (including those that could not be submitted). Note that the official experiments are marked with a star (*).

Our experiments demonstrate that submitting queries in languages other than the actual language of the collection results in poor retrieval performance. In contrast to that using a translation service to translate the queries to the target language yielded substantial improvements over this baseline. For the German and French collections this experimental set-up performed better than any other configuration discussed in this contribution. Similar to the findings from the monolingual sub-task the semantic enrichment did not help to improve performance in general. For the English collection using terms extracted from subject descriptions of DBpedia entities did result in the best performance, but on the German and French collections this could not be confirmed. Another aspect that influences the retrieval performance in the bilingual retrieval scenario might be the source language of the topics. For the English collection French topics should be preferred over German ones and for the German collection

[15] The documentation of the Terrier platform provides the following description: "This class implements the correct Kullback-Leibler divergence for query expansion."

English topics should be preferred over French ones. This effect is not present on the French collection.

Table 3. Results for the bilingual sub-tasks (official runs are marked with *)

run id	lang	configuration summary	MAP
base	DE2EN	Lucene for core retrieval, no exp., no trans.	0.2784
ms	DE2EN	Lucene for core retrieval, no exp.	0.3240
qe_dbp_abs*	DE2EN	Lucene, DBpedia exp., 20 terms (abstract)	0.2805
qe_dbp_sub_ms*	DE2EN	Lucene, DBpedia exp., 20 terms (subject)	0.3399
base	FR2EN	see corresponding runs above	0.3031
ms	FR2EN		0.3513
qe_dbp_abs*	FR2EN		0.2780
qe_dbp_sub_ms*	FR2EN		**0.3549**
base	EN2DE	see corresponding runs above	0.3866
ms	EN2DE		**0.5092**
qe_dbp_abs*	EN2DE		0.3396
qe_dbp_sub_ms*	EN2DE		0.2898
base	FR2DE	see corresponding runs above	0.4000
ms	FR2DE		0.4670
qe_dbp_abs*	FR2DE		0.3724
qe_dbp_sub_ms*	FR2DE		0.3836
base	EN2FR	see corresponding runs above	0.2216
ms	EN2FR		**0.3238**
qe_dbp_abs*	EN2FR		0.1941
qe_dbp_sub_ms*	EN2FR		0.2646
base	DE2FR	see corresponding runs above	0.1882
ms	DE2FR		0.2424
qe_dbp_abs*	DE2FR		0.2294
qe_dbp_sub_ms*	DE2FR		0.3084

Multilingual Experiments

For the multilingual sub-task 18 experiments were conducted in total. Again, only four of these runs could be submitted for evaluation. In fact, three different system configurations were compared, but each of these were tested using the English, German, and French topics. Here our baseline experiment (*base_ms*) relied on Microsoft's translation service. A second experiment used the SEM based on abstracts for expansion and translation (*dbp_abs*). And the final experiment (*dbp_sub_ms*) also used the SEM, but with subjects from DBpedia entities that were translated using Mircrosoft's translation service.

This general set-up was then repeated for two different translation-based query formulation procedures. For the first group the topics were translated to English, German, and French only (see Table 4, "Xto2"). In the second group of experiments, the topics were translated to the nine most frequent languages of the CHiC collection: German, French, Swedish, Italian, Spanish, Norwegian, English, Dutch, and Finnish (see Table 5, “Xto8”). All officially submitted experiments are in the latter group and are marked with a star (*).

Table 4. Results for the multilingual subtask using 3 target languages

run id	lang	configuration summary	MAP
base_ms	ENto2	Lucene, no feedback, transl.	**0.3250**
dbp_abs	ENto2	Lucene, DBpedia exp., 20 terms (abstract)	0.1535
dbp_sub_ms	ENto2	Lucene, DBpedia exp., 20 terms (subject), transl.	0.1504
base_ms	DEto2	see corresponding runs above	**0.3308**
dbp_abs	DEto2		0.1746
dbp_sub_ms	DEto2		0.2292
base_ms	FRto2	see corresponding runs above	**0.2977**
dbp_abs	FRto2		0.1578
dbp_sub_ms	FRto2		0.1818

Table 5. Results for the multilingual subtask using 9 target languages (submitted runs: *)

run id	lang	configuration summary	MAP
base_ms	ENto8	Lucene, no feedback, transl.	**0.2377**
dbp_abs*	ENto8	Lucene, DBpedia exp., 20 terms (abstract)	0.0983
dbp_sub_ms*	ENto8	Lucene, DBpedia exp., 20 terms (subject), transl.	0.1085
base_ms	DEto8	see corresponding runs above	**0.2484**
dbp_abs	DEto8		0.1193
dbp_sub_ms	DEto8		0.1743
base_ms	FRto8	see corresponding runs above	**0.2197**
dbp_abs*	FRto8		0.1041
dbp_sub_ms*	FRto8		0.1333

The obtained results suggest an interesting conclusion regarding multilingual collections: translating queries to more languages can decrease retrieval performance (comparing each row in Table 4 with each corresponding row in Table 5), provided that the relevance assessments for the multilingual task covered documents in all languages. It can also be seen that using a translation service performs better than any other approach, regardless of the language of the topics. This observation is substantiated by the relationship between the two types of experiments with the SEM. Using the subject descriptions as source for translation outperformed the already translated abstract descriptions in all but on scenario (English as topic language, see Table 4).

4.2 Variability Task

For the sake of simplicity, all experiments from the Ad-hoc task were also used for the Variability task. This resulted in a total of 45 experiments, whereof 32 have been submitted for evaluation. The following modification was made to each of the experiment configurations to increase the diversity of the result sets.

The least recently used (LRU) algorithm [9] was adapted in order to restrict the original result list to different types of documents from different providers. First the collections were queried by type, i.e. each experiment was conducted by querying text, image, audio, and video documents separately (see Table 1). For each of these four result lists, only the first hit from each provider (up to the maximum of the twelve different providers) was stored in a list data structure with the addition of the type information and the document score (retrieval status value). According to the LRU algorithm the types were evaluated and if a type was returned too often, the corresponding document was refused, its score was discounted, and it was then put back into the list.

This process did not ensure that a total number of twelve hits were included for each topic. The reason for this was the fact that the original result sets did not necessarily contain documents originating from twelve different data providers.

At the time of writing, the evaluation of the experiments was exclusively based on MAP. Given the fact, that only twelve hits were returned for each of the experiments and that MAP does not reflect diversity at all, no results are reported here. The obtained results and suitable metrics are discussed in detail in [2].

4.3 Semantic Enrichment Task

This task required to return the ten most relevant concepts for each of the topics. Therefore, it was not necessary to rely on a retrieval system. Our approach used the semantic enrichment module based on DBpedia (see Section 3) with two significant modifications. First, a number of elements of DBpedia entities were used (instead of only two in Section 4.1). And second, to keep matters as simple as possible the *Term Selection* component (see Figure 2) was replaced with a straightforward weighted list data structure. This ensured that the experiments could be made without any dependency on the data collection.

So the problem was reduced to the extraction of terms from DBpedia entities that were found for each query and to weight these terms accordingly. The term extraction procedure followed several steps that used different resources of DBpedia entities:

```
1) entity selection by disambiguation:
    http://dbpedia.org/ontology/wikiPageDisambiguates
    http://www.w3.org/2000/01/rdf-schema#comment
    http://dbpedia.org/ontology/abstract
2) category extraction:
    http://purl.org/dc/terms/subject
3) category expansion (broader and narrower):
    http://www.w3.org/2004/02/skos/core#broader
4) finding related entities:
    http://purl.org/dc/terms/subject
5) exploring outgoing links:
    http://dbpedia.org/ontology/wikiPageWikiLink
```

A final step that represented a fallback mechanism for cases where no entities could be found for a given query required a list of default terms in three languages:

- English: "museum", "archive", "library", "text", "image", "audio", "video", "film", "memorial", "monument", "art", "photo", "architecture", "history", "painting", "picture",
- French: "musée", "archives", "bibliothèque", "texte", "image", "audio", "video", "film", "mémorial", "monument", "art", "photo", "architecture", "histoire", "peinture", "dessin",
- German: "museum", "archiv", "bibliothek", "text", "bild", "audio", "video", "film", "denkmal", "monument", "kunst", "foto", "architektur", "geschichte", "gemälde", "aufnahme".

The processing of the individual resources for the term extraction was straightforward for most of these steps (see [10] for more details). It has been decided to use different weighting schemes to select the terms for the four experiments that could be submitted to each of the sub-tasks (see Table 6).

Table 6. Weighting scheme for the term selection process

processing step	run1	run2	run3	run4
1a) entities	11	11	11	11
1b) disambiguation	10	10	10	10
2) category extraction	1.25	1.25	1.25	1.25
3a) cat. expansion (broader)	0.42	0.63	0.42	0.63
3b) cat. expansion (narrower)	0.14	0.21	0.21	0.31
4) related entities	0.13	0.19	0.19	0.28
5) exploring out-links	0.11	0.17	0.17	0.26
6) fallback list	0.10	0.16	0.16	0.16

Monolingual Experiments

For each of the three monolingual subtasks in English, French, and German, the four runs (see above) were submitted for evaluation. Table 7 illustrates the obtained results with respect to the three evaluation metrics Precision(weak), Precision(strong), and MAP. Note that for the calculation of the MAP metric the submitted concepts were tested with a retrieval system other than Xtrieval (see [2] for more details). For this reason, the MAP values are not directly comparable to the corresponding figures from Section 4.1.

The evaluation results show considerably weaker MAP values than in Section 4.1. This suggests that the approach used here may not be suitable for the task. Finding the reasons for this needs further analysis of the programming code. In general there seem to be only small differences in the result sets for the experiments, because all of the three evaluation metrics show only little variance across the four types of runs. The small variation of the weights for the term extraction process might be one reason for this. Another explanation could be the number of query terms, which was almost constant for all experiments and topics. The small amount of variance across the different experiments might also explain why the three metrics do not agree on the best experiment configuration.

Table 7. Results for the monolingual sub-tasks

run id	lang	Prec(weak)	Prec(strong)	MAP
cut_t3_run1	EN	**0.8000**	0.6160	**0.1092**
cut_t3_run2	EN	0.7640	0.6200	0.1069
cut_t3_run3	EN	0.7800	0.6160	0.1072
cut_t3_run4	EN	0.7880	**0.6520**	0.1056
cut_t3_run1	DE	**0.7720**	**0.6080**	0.2286
cut_t3_run2	DE	0.7640	0.6040	0.2383
cut_t3_run3	DE	0.7600	0.5840	**0.2600**
cut_t3_run4	DE	0.7640	0.5840	0.2403
cut_t3_run1	FR	0.5920	0.5480	**0.1467**
cut_t3_run2	FR	**0.6240**	**0.5720**	0.1450
cut_t3_run3	FR	0.6200	0.5680	0.1464
cut_t3_run4	FR	0.6120	0.5520	0.1458

Bilingual Experiments

For the preparation of the bilingual experiments another modification had to be made to the experiment in order to find terms in the corresponding collection language. Besides the strategy to use the multilingual abstract descriptions from DBpedia entities (see Section 4.1), another straightforward approach is to use the label of the entities, which is also available in different languages. Since the labels are short in general, they are very characteristic.

Table 8. Results for the bilingual sub-tasks

run id	lang	Prec(weak)	Prec(strong)	MAP
cut_t3_run1	DE2EN	0.7520	0.6400	**0.1312**
cut_t3_run2	DE2EN	**0.7680**	**0.6760**	0.1273
cut_t3_run3	FR2EN	0.6960	0.6360	-
cut_t3_run4	FR2EN	0.6880	0.6040	-
cut_t3_run1	EN2DE	**0.8400**	**0.7600**	-
cut_t3_run2	EN2DE	0.8160	0.7480	-
cut_t3_run3	FR2DE	0.6000	0.5240	-
cut_t3_run4	FR2DE	0.5320	0.4840	-
cut_t3_run1	EN2FR	0.7920	**0.6800**	**0.1913**
cut_t3_run2	EN2FR	**0.8000**	0.6680	0.1892
cut_t3_run3	DE2FR	0.6400	0.5680	0.1414
cut_t3_run4	DE2FR	0.5720	0.5040	0.1223

The results listed in Table 8 demonstrate that translation approach improved retrieval performance over all the corresponding monolingual experiments. This observation is also independent of the evaluation metric. Unfortunately, no MAP values could be obtained for some of the experiments, which may have helped to substantiate this conclusion. Obvious language-specific effects are a second key finding from the bilingual runs. Similar to the bilingual experiments discussed in Section 4.1, the topic language seems to have a considerable effect on the bilingual retrieval performance. French as the source language for the English sub-collection seems to be preferable over German, English topics seem to be superior to French ones for the German sub-collection, and for the French sub-collection English is the better source language than German. Note that the effect is the same as in Section 4.1, but the language preferences are different for the English and French sub-collections.

Multilingual Experiments

The multilingual experiments were based on the configuration of the bilingual runs, except for the translation. Here, the DBpedia entity labels were collected in all three target languages: English, French, and German. Four experiments have been submitted for evaluation. The corresponding results are presented in Table 9. As for all the previously discussed semantic enrichment experiments, the results are considerably weaker in terms of MAP than the corresponding baseline runs submitted to the Ad-hoc task. The evaluation of the expansion terms resulted in Precision values comparable to the mono- and bilingual runs. A possible impact of the source language of the query is not as obvious as for the bilingual sub-tasks.

Table 9. Results for the multilingual sub-task

run id	lang	Prec(weak)	Prec(strong)	MAP
cut_t3_run1	DE2X	0.7000	0.6040	0.0381
cut_t3_run2	FR2X	**0.7360**	**0.6440**	**0.0614**
cut_t3_run3	EN2X	0.6800	0.5800	0.0283
cut_t3_run4	DE2X	0.6680	0.5600	0.0246

5 Conclusion

The focus of our participation in the CHiC evaluation lab was on the development and evaluation of an extension for Xtrieval that exploits semantic resources like DBpedia. A maximum number of 32 experiments were submitted for each of the three tasks of the CHiC lab and yet not all prepared experiments could be included. The outcome of the conducted experiments is as follows:

- *Ad-hoc Task:*
  In the monolingual scenario, from the four submitted system configurations the baseline experiment without any specific modification outperformed the other three runs on the English and German sub-collections. The two experiments that

used the semantic enrichment module performed poorly compared to the straightforward baselines. For the bilingual scenario our confidence in the implemented concept enrichment process was proved wrong by the evaluation results. Again, the most straightforward configurations achieved the best MAP values. Using Microsoft's translation service for topic translation resulted in better bilingual retrieval performance than exploiting the DBpedia semantic web resource. In the multilingual retrieval scenario our confidence in the semantic enrichment process was proved wrong once more. In fact, we managed to choose the worst experiments for submission and hold back the best ones. Contrasting the latter evaluation results with the bilingual results indicates that multilingual retrieval performance is almost as good as bilingual performance.

- *Variability Task:*
  The experiments from the Ad-hoc task were re-used here. At the time of writing no detailed analysis can be provided for organisational reasons.
- *Semantic Enrichment Task:*
  Our experiments for the semantic enrichment task featured some additional challenges. Except for the DBpedia semantic web resource, no publicly available services or software projects were used. This may have contributed to the generally weak retrieval performance for the experiments based on our concept suggestions. The evaluation results demonstrated that using very short DBpedia entity labels as source for translation improved the retrieval results considerably in comparison to the other tested descriptors. This might be another cue that query expansion is a hard problem on this particular test collection.

Although, most of the presented experiments indicate that longer queries are not effective for retrieval in this particular collection, we think that there is still room for improvement. Further experiments are needed to study the observed effects in more detail. This could be achieved by incorporating tools for component level evaluation [11, 12]. Another opportunity for further analyses might be to submit the results of all contributions to online evaluation databases like EvalutIR.org [13].

Acknowledgements

This work was partly accomplished within the research project ValidAX (http://www.validax.de), funded by the German Federal Ministry of Education and Research under the grant reference number 16V0058. The authors take sole responsibility for the contents of this publication.

References

1. Gäde, M., Ferro, N., Paramita, M. L.: CHiC 2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. In *CLEF 2011 Labs and Workshop, Notebook Papers*, 19–22 September 2011, Amsterdam, The Netherlands. 2011.
2. Gäde, M., Petras, V.: Overview of the Cultural Heritage pilot evaluation lab, CLEF 2012 working notes, Rome, Italy, 2012.
3. Kürsten, J., Wilhelm, T., and Eibl, M.: Extensible Retrieval and Evaluation Framework: Xtrieval. In *Baumeister, J. and Atzmüller, M. (eds.), LWA 2008*. pp. 107–110, University of Würzburg. 2008.
4. Eibl, M., Kürsten, J.: The Importance of being Grid: Chemnitz University of Technology at Grid@CLEF. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September 30–October 2, 2009.
5. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A crystallization point for the Web of Data. In ***Web Semant.*** 7, 3 (September 2009), 154-165, 2009.
6. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R.: Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications* (ESWC 2009 Heraklion), Springer-Verlag, Berlin, Heidelberg, pp. 723-737, 2009.
7. Mendes, P. N., Jakob, M., and Bizer, C.: DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 21–27 May 2012, Istanbul, Turkey, 2012.
8. Ounis, I., Amati, G., Plachouras, V., He, H., Macdonald, C., and Lioma, C. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of OSIR Workshop 2006*, 2006.
9. Coffman, E. G., and Denning, P. J.: Operating Systems Theory. Prentice-Hall, 1973.
10. Richter, D.: Information Retrieval mit dem Xtrieval Framework für cultural heritage Datenbestände. Diploma Thesis, Chemnitz University of Technology, 2012.
11. Kürsten, J.: A Generic Approach to Component-Level Evaluation in Information Retrieval. Doctoral Thesis, Chemnitz University of Technology, 2012.
12. Wilhelm, T., Kürsten, J., and Eibl, M. A Tool for Comparative IR Evaluation on Component Level. In *Proceedings of the 34th International ACM SIGIR conference*, (New York, NY, USA, 2011), pp. 1291–1292.
13. Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. EvaluatIR: An Online Tool for Evaluating and Comparing IR Systems. In *Proceedings of the 32nd international ACM SIGIR conference*, (New York, NY, USA, 2009), ACM, pp. 833–833.