UniNE at CLEF - CHiC 2013

Mitra Akasereh, Nada Naji, Jacques Savoy

Computer Science Dept., University of Neuchâtel (Switzerland)

Mitra.Akasereh@unine.ch, Nada.Naji@unine.ch,
Jacques.Savoy@unine.ch

Abstract. This paper presents and analyzes the experiments done at the University of Neuchatel for both the multilingual and Polish CHiC tasks at CLEF 2013. Within these two tasks, our experiments explore the problem when facing with short text descriptions expressed in various languages having a richer morphology than English. For the multilingual task, each language and its corresponding CH object collection is managed separately. Thus for each query, the broker needs to merge 13 result lists to form a single ranked list of retrieved items. In this context, the best retrieval performance levels tend to be achieved when applying a stopword list for each language. The use of a language-dependent light stemmer may have either a positive or a negative but always slight impact. For the Polish task, we found that the use of a short stopword list and a light stemmer improves retrieval effectiveness. The use of words as indexing units is better than considering *n*-gram or trunc-*n* indexing schemes. Considering automatically generated enrichment descriptors does not improve the retrieval effectiveness neither does the use of pseudo-relevance feedback. Finally, the application of data fusion operator was not able to enhance the retrieval performance.

Keywords. Cultural Heritage; Multilingual IR; Polish IR; Evaluation.

1 Introduction

In this paper we describe our experiments done inside the CLEF – CHiC 2013 evaluation campaign [1] focusing both on the multilingual and Polish tasks. Searching for pertinent cultural heritage (CH) objects in response to a short user's query is a challenging task for various reasons. First, the available descriptions of the CH objects provided by the *Europeana* organization are rather short (e.g., in average 35 indexing terms per record for the English corpus). Those terms, manually selected for some of them, are rather broad and are produced by different content providers having different indexing policies. Therefore a direct comparison between these descriptors is not really possible. The CH domain is also characterized by a frequent use of names such as personal names (e.g., Picasso), works (e.g., Mona Lisa) as well as geographical entities (e.g., Paris) and temporal references (e.g., Baroque). Moreover, the described objects may originate from different media such as text, image, photo, video, music or

sound. Finally, the users do not form a homogenous group but are coming from different perspectives. We can find students, educators, tourists or "informed citizens". Of course, facing with short item descriptions and short query formulations is not frequent but can be found in other IR domains [2], [3].

This proposed *ad hoc* search task [4] was more complex due to the multilingual nature of the CH objects descriptors and topics. For each object, the given description is available in at least one language, and for many of them, a passage is available in a second language. However, no single language (e.g., English) covers all available records. The user's information needs are also given in various languages but only one must be selected to perform the search. This additional constraint can also be found in the commercial world as, for example, when users are searching for applications for their iPhone (or iPad). In this case, the users are coming from different linguistic backgrounds, express their needs with one or two terms to retrieve an item described by a few keywords or noun phrases.

Our experiments aim to explore the retrieval effectiveness of various IR models when facing with such multilingual short descriptions of CH objects and topic formulations. We also want to study the impact of automatic query translation and stemming approches to improve the mean average precision (MAP).

The rest of this paper is divided as follows. Section 2 describes the main characteristics of the *Europeana* collection. Section 3 presents an overview of our experiments and their evaluations in the multilingual task. Section 4 exposes our work done on the Polish task. Finally a conclusion draws the main findings of our study.

2 The Test Collection

The CLEF - CHiC 2013 multilingual collection is the same as last year and is composed of 20,310,425 CH object descriptions. We can find more than 1 M of records written in the German, French, Swedish, Italian, Spanish, Norwegian, Dutch, English, and Polish languages. With fewer objects, we can add the Finnish, Slovenian, Greek, and Hungarian languages, which sums up to 13 different languages.

This collection was made available by the *Europeana* organization. Each document corresponds to the relatively short description of a cultural object. Basically, these CH descriptions correspond to image objects but we can also find text objects as well as audio and video. Each CH object is mainly described by a set of metadata tags, in addition to some automatically enrichment appearing under the tag prefix *enrichment*:. However, the number of tags per record varies greatly. Some descriptors may contain many tags whereas fewer can be detected in other records.

The topic descriptions consist of a mixture of topical and named-entity queries. The 50 short topics in title-format only (e.g., "horse couriers", "Columbus ships") tend to reflect information needs as expressed by real *Europeana* users. Some topics descriptions contain personal names (e.g., "jean-jaques rousseau" with a spelling error in "Ja**c**ques"), but we also have topics with geographical names (e.g., "falkland islands", "rock of Gibraltar") or with historical names (e.g., "uprisings in 18th century"). It is noteworthy that some of these named-entities feature several spelling vari-

ants in different languages (e.g., *Geneva* (EN), *Genf* (DE), *Genève* (FR), *Ginevra* (IT), and *Genewa* (PL)).

Relevant document could not however be found for each topic in each language. For example, Topic #64 ("Crockery doll houses") does not have any relevant item among English documents and for Topic #91 ("Columbus ships") no French document can be found (more information is given in the Appendix). In German collection, objects can be found for all topics while for 35 topics no relevant item can be found in the Finnish corpus. The number of relevant documents per topic varies greatly. Topic #53 ("Postage stamp") has the largest number of relevant items (1,390) while Topic #91 ("Columbus ships") has the smallest number of relevant documents (19). In mean, we can find 56.7 relevant CH objects per topic (median: 302; stdev: 323).

For the Polish subtask, relevant items can be found for every topic, with a minimum of 5 relevant objects for Topic #17 ("Czeslaw Milosz" or "Czesław Miłosz") and a maximum of 562 pertinent items for Topic #20 ("PRL (People's Republic of Poland)"). In mean, we can find 170.6 relevant objects per topic (median: 125; stdev: 139.6).

3 Europeana Multilingual CHiC Experiment

For this *ad hoc* IR task, we face with more than 23 million of CH objects described in 13 different languages with their corresponding topics. In our experiments, we have used the 50 topics written in each language. This corpus forms a real multilingual test collection and various MLIR strategies can be evaluated [5]. We used two different approaches to perform our search. As a first approach, we built a single huge collection with all CH object descriptions. We then searched into this single corpus using the 50 multilingual topics. This first approach must be viewed more as a baseline than a realistic implementation. As the second strategy, we built 13 distinct corpora according to the language in use and associated a dedicated server per language. We then searched separately each corpus against its corresponding topics. In a final step, the broker needs to merge the 13 different result lists to generate a single ranked list of retrieved items (see Section 3.2).

3.1 Experimental Setup

To index the collection (or language), we extracted only the tags containing textual information from each corpus and based our indexing scheme on isolated words. However, we did not use all the available information. In fact, we removed the tags containing general information on the objects such as the publisher and the provider name. To generate a surrogate for each CH object, we only use the following six tags: <dc:contributor>, <dc:creator>, <dc:description>, <dc:subject>, <dc:date>, <dc:title>.

As mentioned before the documents are relatively short. Once the collection is parsed, considering all the languages, the minimum of distinct terms per record is 12

(Slovenian or Greek), with a maximum of 50 (Polish), with a median of 19[1]. As for topics, we used only the title section of each topic formulation. Nevertheless, we provided two different sets of topics. First, we used the original topics provided in each language. In a second experiment, we used only the English topics and then we automatically translated them into the other 12 languages. We conducted some of our experiments with these two sets of topics to be able to measure the impact of the automatic query translation process.

For all languages, we can apply a stopword removal [6]. These lists differ in size for each language (from the longest composed of 747 Finnish words to 138 Polish terms, see the Appendix). For each language, such a list contains terms having a relatively high frequency and are composed mainly by determiners, prepositions, conjunctions, pronouns, and some verbal forms (these lists are freely available at `members.unine.ch/jacques.savoy/clef/`). However, we were unable to generate such a pertinent list for the Slovenian and Greek languages.

Considering the frequent use of names as one of the characteristics of the CH domain, we suggest applying a light suffix-stripping stemmer for each language. In this perspective, each algorithmic stemmer is designed according to the grammar rules of the corresponding language. More precisely, these light stemmers try to remove only the inflectional suffixes attached to nouns or adjectives to denote the gender, number, and the different grammatical cases. For example, the English light stemmer removes only the plural suffix "-s" [7]. The French light stemmer removes the inflectional suffixes denoting the plural and feminine forms while for languages like Polish or Dutch, more rules were needed. Finally, each suffix removal step is controlled by quantitative and qualitative restrictions to guarantee some consistencies of the resulting terms [8]. We can mention that the performance difference between a light and a more aggressive stemmer is not significant for the English language [9]. As a variant when high precision is the main objective, we have also indexed the CH object descriptions without considering the stemming stage.

3.2 IR Models and Data Fusion

As an effective IR model, we chose the Okapi (BM25) [10] as our weighting scheme. As we deal with relatively short documents, we thought that this IR model would provide a high retrieval effectiveness level [11]. To define the parameter values, we applied the default setting of the Okapi BM25 with $b = 0.75$ and $k_1 = 1.2$. The same set of values was used for all languages. After this step, we have 13 servers, each corresponding to one language. As soon as they received the query in their corresponding language, each server produces a ranked list of retrieved CH objects.

In order to merge these result lists produced separately, the broker may apply different merging strategies. As a baseline approach, we merged these lists in a round-robin manner (denoted as "RR"). In this case, we took one document in turn from all

[1] In the Table A.1 in the Appendix and for each language, we can find the number of CH objects, the average number of distinct indexing terms per record, as well as the number of terms in the stopword list.

individual lists and repeat this process [12]. As an alternative, we also used a biased round-robin approach [13]. In this case, we assume that each server does not contain the same number of pertinent items for each query. In our implementation, we decide to favor languages having a larger number of items, expecting they will also contain more relevant items. To simplify the process, we took, per round, three documents from German and French result lists, two from the Swedish, Italian, Spanish, and one from the rest of the languages. We will denote this biased round-robin approach as "bRR".

As other merging schemes, we can take into account the document score (or retrieval status value, RSV) computed for each retrieved item. Accordingly, as third merging strategy, we normalize the document scores within each language (or server). To achieve this, we divide each document score by the maximum score (or the score achieved by the first document in each ranked list) [13]. We name this strategy "NormMax". For the *i*th collection, the new RSV' for the *k*th document is $RSV'_k = RSV_k / Max^i$, where Max^i denotes the document score having the maximal value in the *i*th result list. As fourth merging approach, we applied a variant of the previous one, called "MinMax". In this case, we normalize the document score by taking into account not only the maximum score but also the minimum one [13]. More formally, the new RSV' score is computed as: $RSV'_k = ((RSV_k – Min^i) / (Max^i - Min^i))$

As the final merging strategy, we apply the Z-score operator to merge the different ranked lists (denoted as "Z-score"). In this case, the document score is normalized by considering the average and the standard deviation of the document scores distribution in each result list [13]. Thus, the new $RSV'_k = ((RSV_k - Mean^i) / Stdev^i) + \delta^i$, with $\delta^i = ((Mean^i - Min^i) / Stdev^i)$ used to obtain always positive value.

3.3 Official Results

As a first baseline (run called *UnineMultiRun1* in Table 3), we formed a single collection with all the CH object descriptors. As a search query, we concatenated all the 13 original topic titles (forming a multilingual query). We do not apply any stopword list and we ignore the stemming stage. The resulting MAP was rather low with a value of 0.0476.

To improve this result, we have indexed the CH descriptors according to their given language and we have applied a language-dependent stopword list. When using this indexing strategy with a single inverted file, we can achieve a MAP of 0.1158. In order to verify the impact of an automatic query translation, we conducted the same experiment but using the translated queries instead the original ones. To achieve this, from the English topic title, we use the Google translate service to automatically translate the submitted English query into the 12 different languages. Finally, we concatenate all query translations with the original English topic. This approach achieves a MAP of 0.1200.

To have an overview about the retrieval performance according to each language, using either the translated queries or the original ones, we have reported in Table A.2 (in the Appendix) the MAP over 50 topics. Using the original topic formulations, the achieved MAP is higher, but the retrieval performance differences are usually small.

In a related vein, we have also compared, for each language, the retrieval effectiveness when applying or not a light stemmer. In some cases, the light stemmer improves the mean performance (e.g., with the English or French language). For other languages, the resulting effect is small and negative (e.g., with the Swedish, Norwegian, or Spanish languages).

Overall, this first strategy owns the advantage to be rather simple to implement and demonstrates the usefulness of a stopword list. As a second indexing and search strategy, we have divided the *Europeana* corpus according to the language and formed 13 servers. The topic title of the original formulation was then sent to each server. Separately, each server produces a ranked list of retrieved items. Finally, we need to merge the 13 result lists to generate the final answer presented to the end-user.

To evaluate the various steps in this multilingual search process, we first evaluate the quality of the various merging strategies and the usefulness of applying a light stemming strategy. When the submitted topics tend to contain many names, a light stemming may hurt the overall retrieval effectiveness. For example, the name "Baring" becomes "Bare" when using the Porter stemmer.

Table 1. Evaluation of different stemming and merging strategies (multilingual task)

Official Name	Parameter setting	Stem	Stop word	Language	MAP
	Separate indexes, RR	No	Yes	All	0.1388
	Separate indexes, bRR	No	Yes	All	0.1402
UnineMultiRun3	Separate indexes, NormMax	No	Yes	All	0.1444
UnineMultiRun2	Separate indexes, MinMax	No	Yes	All	0.1516
UnineMultiRun5	Separate indexes, Z-score	No	Yes	All	0.1545
	Separate indexes, RR	Yes	Yes	All	0.1065
	Separate indexes, bRR	Yes	Yes	All	0.1386
	Separate indexes, NormMax	Yes	Yes	All	0.1515
	Separate indexes, MinMax	Yes	Yes	All	**0.1592**
	Separate indexes, Z-score	Yes	Yes	All	0.1396

Table 2. Evaluation of different server selection approaches (multilingual task)

Official Name	Parameter setting	Stem	Stop-words	Language	MAP
	Separate indexes, bRR	No	Yes	All	0.1402
UnineMultiRun3	Separate indexes, NormMax	No	Yes	All	0.1444
UnineMultiRun2	Separate indexes, MinMax	No	Yes	All	0.1516
UnineMultiRun5	Separate indexes, Z-score	No	Yes	All	0.1545
UnineMultiRun4	Separate, −{SL, EL, HU}, bRR	No	Yes	−{SL, EL, HU}	0.1389
	Separate, −{SL, EL, HU}, NormMax	No	Yes	−{SL, EL, HU}	0.1604
	Separate, −{SL, EL, HU}, MinMax	No	Yes	−{SL, EL, HU}	**0.1735**
	Separate, −{SL, EL, HU}, Z-score	No	Yes	−{SL, EL, HU}	0.1622

As depicted in Table 1, we have considered all the 13 languages, with a stopword list adapted for each language and five distinct merging strategies. In this set of runs, we can find three official runs, namely *UnineMultiRun2*, *UnineMultiRun3,* and *Unin-*

eMultiRun5. When we ignored the stemming stage, the best result (MAP: 0.1545) is achieved by our official run *UnineMultiRun5* based on the Z-score merging operator. When we apply a light stemming strategy (bottom part of Table 1), the best overall performance is obtained with the MinMax merging operator (MAP: 0.1592).

In Table 2, we assume that some languages, owning clearly less records than others, can be ignored during the selection of the most useful servers. More precisely, we have conducted a set of experiments where the Slovenian (SL), Greek (EL), and Hungarian (HU) languages were not searched (bottom part of Table 2). As we can see, this arbitrary and prior selection seems to work by allowing better overall retrieval performance than searching into all 13 collections. The best result is achieved by a run based on the MinMax merging operator. Finally, Table 3 depicts the name and specifications of our five official runs.

Table 3. Evaluation of our official runs for the multilingual task

Official Name	Parameter setting	Stem	Stop word	Language	MAP
UnineMultiRun1	One index for all languages	No	No	All	0.0476
UnineMultiRun2	Separate indexes, MinMax	No	Yes	All	0.1516
UnineMultiRun3	Separate indexes, NormMax	No	Yes	All	0.1444
UnineMultiRun4	Separate indexes, bRR	No	Yes	−{SL,EL,HU}	0.1389
UnineMultiRun5	Separate indexes, Z-score	No	Yes	All	**0.1545**
UnineMultiRun6	Separate indexes, RR	No	Yes	−{EL}	0.1387

4 Europeana Polish CHiC Experiment

In the Polish task, we focus only on the Polish corpus which is composed of 1,093,705 documents. This task is a classical *ad hoc* search based on a morphologically rich language. As the Polish language was not present in previous CLEF campaigns, we need to investigate different indexing and search strategies for this language. Moreover, we also need to build a stopword list and a light stemmer [8] (freely available at `http://members.unine.ch/jacques.savoy/clef/`). The suggested stopword list is composed of 304 words (mainly determiners, prepositions, conjunctions, pronouns and auxiliary verbal forms). Certainly a longer list can be created to achieve a broader coverage of functional words in this language [6].

4.1 Indexing Strategies

As for the multilingual task, each cultural object is described by a rather short list of keywords, usually extracted from a predefined authoritative list. Each CH object descriptor is in average composed of around 35 indexing terms. During the indexing process, we had to consider the following tags as useful to extract pertinent indexing terms: <dc:contributor>, <dc:creator>, <dc:date>, <dc:language>, <dc:title>, <dc:type>, <dc:subject>, <dc:description>, <dcterms:alternative>, <dcterms:created>, <europeana:country>, <europeana:language>, <europeana:type>, <europeana:year>.

This set of tags will be denoted *partial*. We can form a *full* set of tags by adding to the partial set of tags originating from *Europeana* automatic enrichment process namely: <enrichment:concept_broader_label>, <enrichment:concept_label>, <enrichment: period_label>, <enrichment:place_broader_label>. When inspecting the difference in average document length between the two versions, we found no real variation. Moreover, the retrieval effectiveness differences were also unobservable between the two sets of tags, indicating that the automatic enrichment was not useful, for the Polish collection at least.

As a first indexing strategy, we have investigated different text representations based on *n*-gram [14], as well as trunc-*n* (with $n = 4$, 5, and 6). Such representations usually tend to form good overall baselines when facing with a new language (for which no good stemmer is available or known). From the word "computer", a trunc-5 approach will form the single indexing term "compu", while the 5-gram approach will generate "compu", "omput", … "puter". The benefit sought from implementing *n*-gram or truncation is to assign low indexing weights to frequent suffixes usually added to indicate grammatical cases, gender modifications, or derivational suffixes. In fact, the Polish language has seven grammatical cases, three genders, and two numbers, and the corresponding suffixes are attached to both nouns (four possible declensions) and adjectives.

Table 4. MAP of runs based on *n*-gram or trunc-*n* approaches (Polish task)

	DFR-I(n_e)B2		Okapi	
Parameter	*n*-gram	Trunc-*n*	*n*-gram	Trunc-*n*
$n = 4$	0.2350	0.2268	0.2466	0.2532
$n = 5$	0.2610	0.2968	0.2577	0.3038
$n = 6$	0.2611	0.3078	0.2640	**0.3211**

In Table 4, we have evaluated different sub-word indexing strategies, showing that the trunc-*n* tends to produce better retrieval effectiveness. Moreover, the value of the parameter *n* must be larger than with the French of English languages, with the best value being equal to 6.

As a second indexing strategy, we will opt for the whole words, with or without applying a light stemmer [8]. This word normalization is based on a set of grammatical rules trying to remove only inflectional suffixes from nouns and adjectives. For the Czech language, applying a stemming stage improves the retrieval effectiveness of around 40% [15]. For other languages having a complex morphology, a simple algorithmic stemmer does not provide the expected improvement [16]; this is mainly due to numerous exceptions or spelling irregularities.

As IR models, we first consider the classical *tf idf* (with cosine normalization) [4]. This approach was selected only to provide a baseline. As more effective IR models, we have used the Okapi (or BM25) [10], and the DFR-I(n_e)B2, one implementation of the DFR probabilistic paradigm [17].

In Table 5, we have evaluated some variations with the classical *tf idf* and the two probabilistic models. The performance measure indicated that the Okapi probabilistic model proposes the best performance. Moreover, the use of both a stopword list and a

light stemmer clearly tends to improve the overall effectiveness. When comparing the MAP values depicted in the second (no stopword, no stemming) and third column (stopword, no stemming), we can see an improvement after removing functional words with the two probabilistic models (e.g., from 0.3060 to 0.3140 (+2.6%) for the Okapi model). Applying a light stemmer clearly improves the retrieval effectiveness of both probabilistic models (from 0.3140 to 0.3433 (+9.3%) for the Okapi model).

Table 5. MAP of runs based on word-based indexing (Polish task)

IR Model	No stopword no stemming	Stopword no stemming	No stopword with stemming	Stopword with stemming
tf idf	0.2558	0.2566	0.2541	0.2579
Okapi	0.3060	0.3140	0.3258	**0.3433**
DFR-I(n_e)B2	0.2883	0.3028	0.3085	0.3308

4.2 Pseudo-Relevance Feedback & Data Fusion

As an additional strategy to improve the retrieval effectiveness, we can generate a new enlarged query based on pseudo-relevance feedback information. In this case, we assume that the top *k* retrieved items are pertinent without inspecting them. Then from these we can extract *m* additional terms to automatically enrich the original query. Based on previous CLEF evaluation campaigns based on newspaper articles, this strategy may improve the MAP of around 5% to 20%. In our implementation, we can adopt either the Rocchio scheme [18] or and *idf*-based approach [19]. Of course, the best values for the parameters *k* and *m* are unknown. In Table 6, we have evaluated the DFR-I(n_e)B2 search model with different parameter values. As we can see, this search technique tends to hurt the MAP achieved by the original query, using the word-based or 5-gram indexing scheme, Rocchio or *idf*-based selection schemes. Adding automatically terms in the query is clearly a hard task in our context.

Table 6. MAP of runs based on Rocchio pseudo-relevance feedback (Polish task)

IR Model Parameter	DFR-I(n_e)B2 5-gram Rocchio	DFR-I(n_e)B2 word-based, no stem Rocchio	DFR-I(n_e)B2 word-based, no stem *idf*-based
Without PRF	0.2610	0.3028	0.3028
k=5 docs, *m*=5 terms	0.1572	0.2189	0.2784
k=5 docs, *m*=10 terms	0.1590	0.2119	0.2780
k=5 docs, *m*=20 terms	0.1552	0.2013	0.2777

As a final search technique to improve the MAP, we can apply a data fusion operator [13]. In this case, we combine the result lists computed by different search techniques to hopefully generate a better ranking than that proposed by the different individual runs.

Table 7. Evaluation of our official runs for the Polish task

Name	Parameter setting	MAP
UniNEBaseline	*tf idf* (cosinus), no stemming	0.2566
UniNEDFR	DFR-I(n_e)B2, light stemming	0.3308
UniNEFusion	Data fusion (Okapi: no stem, light stem, trunc-5)	**0.3433**
UniNEPRF	Data fusion, DFR-I(n_e)B2, PRF (Rocchio, 5 docs, 10 terms)	0.2578
UniNEGramPRF	Data fusion, DFR-I(n_e)B2, 5-gram, PRF	0.2203

Table 8. Evaluation of the individual runs belonging to the different fused runs

In Run	Parameter setting	MAP
UniNEFusion	Okapi, no stemming	0.3433
UniNEFusion	Okapi, light stemming	0.3140
UniNEFusion	Okapi, trunc-5	0.3038
UniNEPRF	DFR-I(n_e)B2, light stemming, Rocchio (5 docs, 10 terms), full	0.2616
UniNEPRF	DFR-I(n_e)B2, no stemming, Rocchio (5 docs, 10 terms), partial	0.2119
UniNEGramPRF	DFR-I(n_e)B2, 5-gram, Rocchio (10 docs, 20 terms), full	0.1853
UniNEGramPRF	DFR-I(n_e)B2, 5-gram, *idf*-based (10 docs, 20 terms), full	0.2342

4.3 Official Runs

In applying this strategy, we have selected the Z-score operator [13]. In the list of our official runs (see Table 7), the run labeled *UniNEFusion* indicates the retrieval effectiveness that can be reached when combining two word-based Okapi models (with and without a light stemming procedure) with an Okapi model based on the trunc-5 indexing scheme. This strategy produces the best retrieval effectiveness (even if the fusion operator does not improve the final result (Table 8 depicted the MAP of each individual run)).

The result obtained by the run *UniNEPRF* is a fusion approach obtained after a blind query expansion based on adding 10 terms extracted from the top five best-ranked documents. In this case, we thought that combining a run based on all tags (as indicated by the term *full*) and a second run using only a subset (denoted by the term *partial*) may produce an effective solution. This was not the case because the performance difference between the *partial* and the *full* set of tags is null.

5 Conclusion

In this paper we have described and analyzed our participation to the CLEF – CHIC evaluation campaign, both in the multilingual and Polish tasks. These tasks were classical *ad hoc* search within a corpus composed of short descriptions of CH objects. The IR task is rather complex due to very short descriptions, written with broad terms and the difficulty of having a precise meaning of the real user's information needs. The complexity of the morphology of the various languages used to describe these CH objects clearly increased the difficulty of this IR task.

In the multilingual task, we have selected the probabilistic Okapi model, a search strategy well adapted when facing with short textual descriptions [11]. We have opted for a query-based translation approach meaning that we have one server per language. In evaluating simple merging strategies, our experiments indicate that the Z-score scheme [13] tends to offer the best performance levels. Another interesting finding is to note the importance of a good stopword list when working with the Okapi search model [20]. Applying such lists has a clear and positive impact on the overall retrieval effectiveness. This is not the case with the light stemming strategy that can improve or degrade the mean average precision, depending on the language.

For the Polish task and unlike our experiments achieved using newspaper corpora, we found that applying a data fusion approach does not always improve the overall retrieval effectiveness. Moreover, automatically enlarging the original query using either the Rocchio or an *idf*-based weighting scheme does not improve the MAP.

On the other hand, our set of experiments confirms that the classical *tf idf* vector-space model is not the most effective IR model. Clearly the Okapi or the DFR-I(n_e)B2 models produce better retrieval effectiveness. With the Polish language, we also demonstrate that a stemming stage will enhance the final result. For both the multilingual and Polish tasks, we cannot however specify whether a more aggressive stemmer may further enhance the retrieval effectiveness, nor may statistical stemmers [21], [22], [23]. Moreover, the effectiveness of a Polish lemmatizer must also be investigated.

Acknowledgments

This research was supported in part by the SNF under Grant 200020_129535.

6 References

1. Petras V., Ferro N., Gäde M., Isaac A., Kleineberg M., Masiero I., Nicchio M., Stiller J. Cultural Heritage in CLEF (CHiC) Overview 2012. In *Proceedings CLEF-2012, Working Paper* (2012). See also URL: http://www.promise-noe.eu/chic-2013/collections
2. Metzler D., Dumais S., Meek C. Similarity Measures for Short Segments of Texts. In *Proceedings ECIR'07*, (2007) 16-27.
3. Sahami M., Heilman T.D. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In *Proceedings WWW'06,* (2006) 377-386.
4. Manning C.D., Raghavan P., Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008).
5. Peters C., Braschler M., Clough P. *Multilingual Information Retrieval: From Research to Practice*. Springer-Verlag, Berlin (2012).
6. Fox C. A Stop List for General Text. *ACM-SIGIR Forum*, 24 (1990) 19-35.
7. Harman D.K. How Effective is Suffixing? *Journal of the American Society for Information Science*, 42 (1991) 7-15.
8. Savoy J. Light Stemming Approaches for the French, Portuguese, German, and Hungarian Languages. In *Proceedings ACM-SAC*, (2006) 1031-1035.

9. Fautsch C., Savoy J. Algorithmic Stemmers or Morphological Analysis: An Evaluation. *Journal of the American Society for Information Science & Technology*, 60 (2009) 1616-1624.
10. Robertson S.E., Walker S., Beaulieu M. Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management*, 36 (2000) 95-108.
11. Yuanhua L., Zhai C. When Documents are Very Long, BM25 Fails! In *Proceeding ACM SIGIR*, (2011) 1103-1104.
12. Fox E.A. & Shaw J.A. Combination of Multiple Searches. In *Proceedings TREC-2*, Gaithersburg: NIST Publication #500-215 (1994) 243-249.
13. Savoy J. Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. *IR Journal*, 7 (2004) 121-148.
14. McNamee P., Mayfield J. Character *n*-gram Tokenization for European Language Text Retrieval. *IR Journal*, 7 (2004) 73-97.
15. Dolamic L., Savoy J. Indexing and Stemming Approaches for the Czech Language. *Information Processing & Management*, 45 (2009) 714-720.
16. Korenius T., Laurikkala J., Järvelin K., Juhola M. Stemming and Lemmatization in the Clustering of Finnish Text Documents. In *Proceedings of the ACM-CIKM*, (2004) 625-633.
17. Amati G., van Rijsbergen C.J. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems*, 20 (2002) 357-389.
18. Rocchio J.J.Jr. Relevance Feedback in Information Retrieval. In G. Salon (Ed.), *The SMART Retrieval System*. Prentice-Hall Inc., Englewood Cliffs (NJ), 1971, 313-323.
19. Rasolofo Y. Savoy J. Term Proximity Scoring for Keyword-based Retrieval Systems. In *Proceedings ECIR*, Springer-Verlag, LNCS #2633, (2003) 207-218.
20. Dolamic L., Savoy J. When Stopword Lists Make the Difference. *Journal of the American Society for Information Sciences and Technology*, 61 (2010) 200-203.
21. Majumder P., Mitra M., Parui S.K., Kole G. YASS: Yet Another Suffix Stripper. *ACM-Transactions on Information Systems*, 25 (2007) Article #18.
22. Paik J.H., Mitra M., Parui S.K., Jarvelin K. GRAS: An Effective and Efficient Stemming Algorithm for Information Retrieval. *ACM-Transactions on Information Systems*, 29 (2011) Article #19
23. Paik J.H., Parui S.K., Pal D., Robertson S.E. Effective and Robust Query Biased Stemming. *ACM-Transactions on Information Systems*, 31 (2013).

7 Appendix

Table A.1 depicts the number of CH objects per language, and the mean number of distinct indexing terms per language. In this table, we can also find the number of topics having at least one relevant record and, in the last column, the number of words of each stopword list per language.

Table A.1. Various statistics about each language

Language	Corpus size	Mean distinct indexing terms	Number of topics	Size of the stopword list
German	3,865,680	19	50	578
French	3,635,388	18	49	464
Swedish	2,360,050	30	31	386
Italian	2,120,059	21	45	430
Spanish	1,953,124	23	46	307
Norwegian	1,557,820	15	40	176
Dutch	1,251,027	12	42	315
English	1,107,176	27	49	571
Polish	1,093,705	50	36	138
Finnish	800,302	14	15	747
Slovenian	246,952	12	29	
Greek	197,371	12	30	
Hungarian	121,771	35	33	737
Total	20,310,425		50	

In Table A.2 we have computed the MAP over 50 topics separately for each language. In the second column, we can find the retrieval effectiveness of the original topics when using a light stemmer. In the third, we ignore this word normalization procedure. In the fourth column, we use the translated topic titles from the English formulation and perform the search without considering the stemming stage.

Table A.2. MAP computed separately for each language, with original or automatically translated queries, with or without a light stemmer

Language	Light original queries	No stem original queries	No stem translated queries	Number of topics
German	0.2863	0.2963	0.2846	50
French	0.2596	0.2359	0.2176	49
Swedish	0.2054	0.2216	0.1664	31
Italian	0.2402	0.2584	0.2575	45
Spanish	0.2558	0.3056	0.3057	46
Norwegian	0.3511	0.3859	0.2830	40
Dutch	0.3299	0.3223	0.2599	42
English	0.3022	0.2490	0.2490	49
Polish	0.3042	0.3035	0.2120	36