Using the Divergence Framework for Randomness: CHiC 2013 Lab Report

Diana Tanase

University of Westminster, London, UK
diana.tanase@my.westminster.ac.uk

Abstract. For this first participation to the CHiC Lab, we focused on understanding the challenges of working with a collection of cultural heritage objects with short textual descriptions and on how to fine-tune a set of weighting models from the probability models based on Divergence From Randomness to perform uniformly in monolingual and multilingual scenarios. The official runs submitted used PL2 as the retrieval model and query expansion for four monolingual runs for English and Italian, and two multilingual runs against an English-Italian collection. Our best results were obtained in the unofficial runs using DLH13 with stemming and stopwords removal.

1 Introduction

The following lab notes describe our experiments for the multilingual ad-hoc retrieval task organized by PROMISE (Participative Research Laboratory for Multimedia and Multilingual Information Systems Evaluation). The task involved retrieving relevant documents from the CHiC multilingual Europeana collection for the 50 topics provided in 13 languages. For this first participation to the CHiC Lab, we focused on understanding the challenges of working with a collection of cultural heritage objects with short textual descriptions and on how to fine-tune a set of weighting models from the probability models based on Divergence From Randomness (DFR) [2] to perform uniformly in monolingual and multilingual scenarios. The official runs submitted used PL2 as the retrieval model and query expansion for four monolingual runs for English and Italian, and two multilingual runs against an English-Italian collection. Our best results were obtained in the unofficial runs using DLH13 with stemming and stopwords removal.

In the next sections we present a summary of retrieval results and the combination of experimental settings we worked with. The results obtained in the official runs are modest, with substantial improvements in the unofficial runs that use DLH13.

2 Experimental Setup

The retrieval models we chose for these experiments are PL2 and DLH13. They are DFR models obtained by instantiating the three components of the framework: selecting a basic randomness model, applying the first normalization and

than normalizing the term frequencies. The mathematical formulas [4] describe that terms with informative value abide by the distributional rule *the more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d* [1]. Our decision to consider DFR models was also based on the results reported by [1], where similar retrieval performances are obtained across languages with DFR models.

PL2 weighting model – a Poisson model with Laplace after-effect and second normalization for resizing the term frequency by document length.

$$score(d,Q) = \sum_{t \in Q} qtw \cdot \frac{1}{tfn+1}\big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn)\cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\big) \tag{1}$$

DLH13 weighting model – a generalization of the hypergeometric model in a binomial case (parameter free):

$$score(d,Q) = \sum_{t \in Q} qtw \cdot \frac{1}{tf+0.5} \cdot \left(\log_2(\frac{tf \cdot avg\_l}{l} \cdot \frac{N}{F}) + (l - tf)\log_2(1-f) + 0.5\log_2\left(2\pi tf(1-f)\right)\right) \tag{2}$$

where the normalized term frequency is:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}) \tag{3}$$

Notations:
tf is the within-document frequency of t in d
avg_l is the average document length in the collection
l is the document length of d, which is the number of tokens in d
N is the number of document in the whole collection
F is the term frequency of t in the whole collection
nt is the document frequency of t
tfn is the normalized term frequency given by relation 3,
where c is a free parameter
λ is the variance and mean of a Poisson distribution. It is given by F/N and F is much smaller than N
qtw is the query term weight given by qtf/qtf_{max}
qtf is the query term frequency and qtf_{max} is the maximum query term frequency among the query terms

We used only two of the 13 collections made available: the English collection with 1107176 documents and the Italian Collection with 2120059 documents. Prior experiments at CHiC were performed using Lucene, Solr, Indri, or Cheshire [6], while in this setup we used Terrier Retrieval Platform [5]. After indexing,

[1] http://terrier.org/docs/v3.5/dfr_description.html

using the English tokeniser, respectively the UTF tokeniser we obtained two indexes. The English index had 338248, while the Italian had 274009, with a much larger number of tokens for Italian.

3 Official Runs

Our results presented in Table 1 are also described in finer detail in [3]. The MAP was computed for the multilingual scenario, where a topic is in one source language and the relevant documents can be in any of the different language collections. We noticed that the query expansion did not always have a positive impact on performance. This is a known issue with query expansion only working well for queries which have a good top-ranked document set returned by the first-pass retrieval. Also, based on query average precision 10 topics from the name topic category had precision zero in the Italian runs (e.g *isola di madeira*, *isole falkland*,*sesame street*).

Model	Query Expansion	Stemming	Stopwords	Run	MAP
PL2	-	x	x	EN-EN	4.82
PL2+Bo1	x	x	x	EN-EN	4.75
PL2	-	x	x	IT-IT	2.55
PL2+Bo1	x	x	x	IT-IT	2.89
PL2	-	x	x	EN - Mixed EN/IT	6.30
PL2+Bo1	x	x	x	IT - Mixed EN/IT	5.97

Table 1: CHiC Ad-Hoc Multilingual Official Runs

Overall, our submission is slightly worse than the 5th best result obtained in the multilingual ad-hoc evaluation (MAP 6.43%) and the results submitted only used the English and Italian document collections. We merged the result lists from monolingual retrievals and ordered them based on the $score(d, Q)$ values. This was possible in this instance because the collections had a comparable number of terms.

4 Monolingual Explorations

The PL2 is a parametric model, so the parameter we set a-priori could not be tuned without relevance assessments, and for a second set of experiments we opted for the DLH13 weighting model a parameter-free weighting model, with all its variables being set automatically from the collection statistics.

In the unofficial runs, we varied the conditions for each of them by using light NLP processing (stemming, stopwords removal), query expansion, and query enrichment by adding new terms for each query based on Google's auto-complete feature.

Model	Query Expansion	Stemming	Stopwords	Query Enrichment	MAP_{EN}	MAP_{IT}
DLH13	-	-	-	-	**36.25**	8.42
DLH13	x	-	-	-	34.97	7.45
DLH13	-	-	-	x	25.76	6.08
DLH13	x	-	-	x	25.44	6.49
DLH13	-	x	x	-	35.19	32.44
DLH13	x	x	x	-	33.75	29.34
DLH13	-	x	x	x	25.87	24.09
DLH13	x	x	x	x	25.70	21.43

Table 2: Summary Results of the Monolingual EN & IT Unofficial Runs

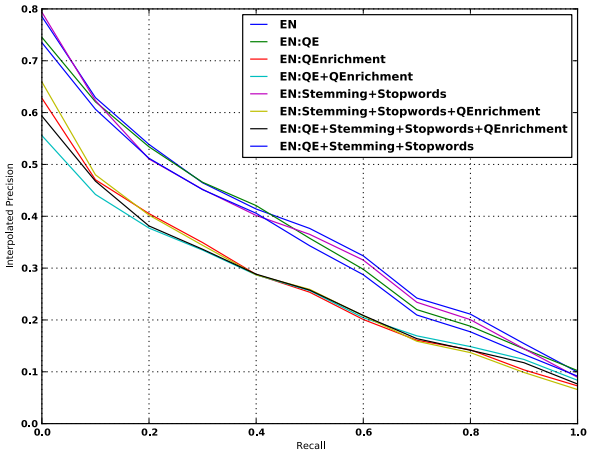

Fig. 1: CHiC Ad-Hoc EN Monolingual

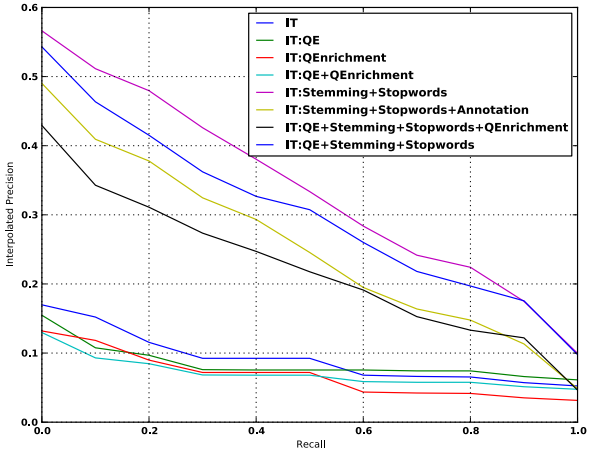

Fig. 2: CHiC Ad-Hoc IT Monolingual

Across the different setups (see Figure 1 and Figure 2), we noticed that the stemming and stopwords removal with DLH13 produces the most consistent results. We repeated the multilingual retrieval obtaining an improved MAP of 8.73% with only topic CHIC–91 (*navi di colombo*) having precision zero, an elusive query-topic with a 1.86 mean statistics for the number of relevant retrieved documents.

Precision at 1 : 0.6400	Precision at 0%: 1.4081
Precision at 2 : 0.6600	Precision at 10%: 0.6428
Precision at 3 : 0.6467	Precision at 20%: 0.2661
Precision at 4 : 0.6250	Precision at 30%: 0.1178
Precision at 5 : 0.5920	Precision at 40%: 0.0436
Precision at 10 : 0.5380	Precision at 50%: 0.0082
Precision at 15 : 0.5013	Precision at 60%: 0.0000
Precision at 20 : 0.4740	Precision at 70%: 0.0000
Precision at 30 : 0.4400	Precision at 80%: 0.0000
Precision at 50 : 0.3848	Precision at 90%: 0.0000
Precision at 100 : 0.3066	Precision at 100%: 0.0000
Precision at 200 : 0.2148	R-Precision: 14.30
Precision at 500 : 0.1146	
Precision at 1000 : 0.0665	
Average Precision: 8.73	

Fig. 3: CHiC Ad-Hoc Multilingual using DLH13, stemming, stopwords removal from EN, IT collections

5 Conclusions

The CHiC Lab 2013 Ad-Hoc Multilingual Task allowed us to experiment with two probabilistic models from the DFR family. The DLH13 outperformed PL2 in this instance, but with further tuning of the parameters for PL2 this could be reversed. We will continue to further our work using the topics and the Europeana collection having acquired the necessary baseline experience to expand to more languages from the collection.

References

1. Mitra Akasereh, Nada Naji, and Jacques Savoy. Unine at clef 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
2. G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.

3. Nicola Ferro and Ivano Masiero. Appendix CHiC 2013 Evaluation Lab. `http://www.promise-noe.eu/documents/10156/8f6af376-8095-48c1-badf-e317c4efdd46`, 2013.
4. Craig Macdonald, Ben He, Vassilis Plachouras, and Iadh Ounis. University of at 2005: Experiments in terabyte and enterprise tracks with terrier. In *In Proceedings of TREC-05*, 2005.
5. I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005.
6. Vivien Petras, Nicola Ferro, Maria Gde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, and Juliane Stiller. Cultural Heritage in CLEF (CHiC) Overview 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.