The CLEF Monolingual Grid of Points

Nicola Ferro and Gianmaria Silvello(✉)

Department of Information Engineering, University of Padua, Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. In this paper we run a systematic series of experiments for creating a *grid of points* where many combinations of retrieval methods and components adopted by *MultiLingual Information Access (MLIA)* systems are represented. This grid of points has the goal to provide insights about the effectiveness of the different components and their interaction and to identify suitable baselines with respect to which all the comparisons can be made.

We publicly release a large grid of points comprising more than 4 K runs obtained by testing 160 IR systems combining different stop lists, stemmers, n-grams components and retrieval models on CLEF monolingual tasks for nine European languages. Furthermore, we evaluate such grid of points by employing four different effectiveness measures and provide some insights about the quality of the created grid of points and the behaviour of the different systems.

1 Introduction

Component-based evaluation, i.e. the ability of assessing the impact of the different components in the pipeline of an *Information Retrieval (IR)* system and understanding their interaction, is a long-standing challenge, as pointed out by [24]: "if we want to decide between alternative indexing strategies for example, we must use these strategies as part of a complete information retrieval system, and examine its overall performance (with each of the alternatives) directly".

This issue is even more exacerbated in the case of *MultiLingual Information Access (MLIA)*, where the combinations of components and languages grow exponentially, and even the more systematic experiments explore just a small fraction of them, basically hampering a more profound understanding of MLIA.

In Grid@CLEF [15], we proposed the idea of running a systematic series of experiments and creating a *grid of points*, where (ideally) all the combinations of retrieval methods and components were represented. This would have had two positive effects: first, to provide more insights about the effectiveness of the different components and their interaction; second, to identify *suitable baselines* with respect to which all the comparisons have to be made.

However, even if Grid@CLEF succeeded in establishing the technical framework to make it possible to create such grid of points, it did not delivered a grid big enough, due to the high technical barriers to implement it.

N. Fuhr et al. (Eds.): CLEF 2016, LNCS 9822, pp. 16–27, 2016.
DOI: 10.1007/978-3-319-44564-9_2

More recently, the wider availability of open source IR systems [26] made it possible to run systematic experiments more easily and we see a renewed interest in creating grid of points, which also allow for *reproducible baselines* [14,21]. Indeed, in the context the "Open-Source Information Retrieval Reproducibility Challenge"[1] [1], we provided several of these baselines for many of the CLEF Adhoc collections as well as a methodology for systematically creating and describing them [11].

In this paper, we move a step forward and we release as an open resource the first fine-grained grids of points for many of the CLEF monolingual Adhoc tasks over a range of several years. The goal of these grids is to facilitate research in the MLIA field, to provide a set of standard baseline on standard collections, and to offer the possibility of conducting deeper analyses on the interaction among components in multiple languages.

The paper is organized as follows: Sect. 2 provides an overview of the used CLEF collections; Sect. 3 describes how we created the different grids of points; Sect. 4 presents some analyses to assess the quality of the created grids of points and get an outlook of the behaviour of the different systems; finally, Sect. 5 wraps up the discussion and provides an outlook of future work.

2 Overview of CLEF Monolingual Tasks

We considered the CLEF Adhoc monolingual tasks from 2000 to 2007 [2–6,12,13] in nine languages: Bulgarian, German, Spanish, Finnish, French, Hungarian, Italian, Portuguese and Swedish. The main information about the corpora, topics and relevance judgments of considered tasks are reported in Table 1.

The CLEF corpora are formed by document sets in different European languages but with common features: the same genre and time period, comparable content. Indeed, the large majority of the corpora are composed by newspaper articles from 1994–1995 with the exception of the Bulgarian and Hungarian corpora composed of newspaper articles from 2002.

The French, German and Italian news agency dispatches – i.e. ATS, SDA and AGZ – are all gathered from the Swiss news agency and are the same corpus translated in different languages. The Spanish corpus is composed of news agencies (i.e. EFE) from the same time period as the Swiss news agency corpus and thus it is very similar in terms of structure and content.

CLEF topics follow the typical TREC structure composed of three fields: title, description and narrative. The topic creation process in CLEF has had to deal with specific issues related to the multilingualism as described in [19].

As far as relevance assessments are concerned, CLEF adopted they standard approach based on the pooling method and the assessment based on the longest, most elaborate formulation of the topic, i.e. the narrative [25]. Typical pool depths are between 60 and 100 documents.

[1] https://github.com/lintool/IR-Reproducibility.

Table 1. Employed CLEF monolingual tasks: used corpora; number of documents; number of topics; size of the pool; number of submitted runs. Languages are expressed as ISO 639:1 two letters code.

Task	Year	Corpora	Docs	Topics	Pool	Runs
AH Mono BG	2005	SEGA 2002 STANDART 2002	69,195	49	20,130	20
	2006			50	17,308	11
	2007			50	19,441	16
AH Mono DE	2000	FRANKFURTER 1994	139,715	49	11,335	22
	2001	FRANKFURTER 1994	225,371	49	16,726	22
	2002	SDA 1994		50	19,394	28
	2003	SPIEGEL 1994 & 1995		57	21,534	38
AH Mono ES	2001	EFE 1994	215,738	49	14,268	22
	2002			50	19,668	28
	2003	EFE 1994 & 1995	454,045	57	23,822	38
AH Mono FI	2002	AMULEHTI 1994 & 1995	55,344	30	9,825	11
	2003			45	10,803	13
	2004			45	20,124	30
AH Mono FR	2000	LEMONDE 1994	44,013	34	7,003	10
	2001	LEMONDE 1994	87,191	49	12,263	15
	2002	ATS 1994		50	17,465	16
	2003	LEMONDE 1994 ATS 1994 & 1995	129,806	52	16,785	35
	2004	LEMONDE 1995 ATS 1995	90,261	49	23,541	38
	2005	LEMONDE 1994 & 1995	177,452	50	23,999	38
	2006	ATS 1994 & 1995		49	17,882	27
AH Mono HU	2005	MAGYAR 2002	49,530	50	20,561	30
	2006			48	20,435	17
	2007			50	18,704	19
AH Mono IT	2000	AGZ 1994 LASTAMPA 1994	108,578	34	6,760	10
	2001			47	10,697	14
	2002			49	17,822	25
	2003	AGZ 1994 & 1995 LASTAMPA 1994	157,558	51	20,902	27
AH Mono PT	2004	PUBLICO 1994 & 1995	106,821	46	20,103	22
	2005	FOLHA 1994 & 1995	210,734	50	20,539	32
	2006	PUBLICO 1994 & 1995		50	20,154	34
AH Mono SV	2002	TT 1994 & 1995	142,819	49	12,580	7
	2003			54	15,975	18

Figure 1 reports the box plots of the selected CLEF monolingual tasks grouped by language. We can see that in most cases the data are evenly distributed within the quantiles and they are not particularly skewed. For the monolingual tasks there is only one system with MAP equal to zero (i.e., an outlier for the AH-MONO-ES task) and for 78 % of the monolingual tasks the first quantile is above 10 % of MAP. Note that even amongst the tasks on the same

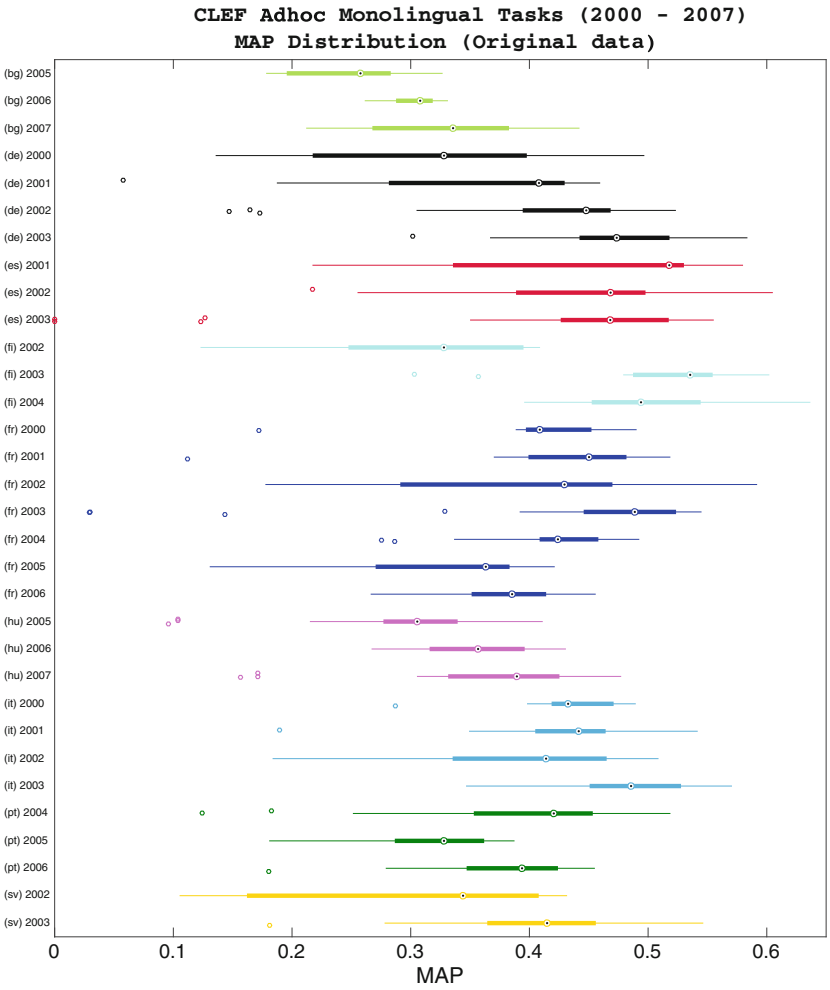

Fig. 1. MAP distribution of original runs submitted to the considered CLEF monolingual tasks.

language, the experimental collections differ from task to task and thus a direct comparison of performances across years is not possible; in [16] an across years comparison between CLEF monolingual, bilingual and multilingual tasks has been conducted by employing the standardization methodology defined in [28].

3 Grid of Points

We considered four main components of an IR system: stop list, stemmer, n-grams and IR model. We selected a set of alternative implementations of each component and by using the Terrier open source system [22] we created a run

for each system defined by combining the available components in all possible ways. Note that stemmers and n-grams are mutually exclusive alternatives since either you can employ a stemmer or a n-grams component.

stop list: nostop, stop;
stemmer: nostem, weak stemmer, aggressive stemmer;
n-grams: nograms, 4grams, 5grams;
model: BB2, BM25, DFRBM25, DFRee, DLH, DLH13, DPH, HiemstraLM, IFB2, InL2, InexpB2, InexpC2, LGD, LemurTFIDF, PL2, TFIDF.

The specific language resources employed such as the stoplist and the stemmers depend by the language of the task at hand. All the stoplists have been provided by the University of Neuchâtel (UNINE)[2]; in the Table 2 we report the number of words composing each stoplist. The stemmers have been provided by University of Neuchâtel (UNINE in the table) and by the Snowball Stemming language and algorithms project[3] (snowball in the table). We chose to use these stop lists and stemmers due to their availability as open source linguistic resources.

Table 2. The linguistic resources employed for each monolingual task.

Language	Stoplist	Weak stemmer	Aggressive stemmer
Bulgarian (bg)	UNINE 258 words	UNINE light stemmer	UNINE stemmer
German (de)	UNINE 603 words	UNINE light stemmer	Snowball stemmer
Spanish (es)	UNINE 307 words	UNINE light stemmer	Snowball stemmer
Finnish (fi)	UNINE 747 words	UNINE light stemmer	Snowball stemmer
French (fr)	UNINE 463 words	UNINE light stemmer	Snowball stemmer
Hungarian (hu)	UNINE 737 words	UNINE light stemmer	Snowball stemmer
Italian (it)	UNINE 399 words	UNINE light stemmer	Snowball stemmer
Portuguese (pt)	UNINE 356 words	UNINE light stemmer	Snowball stemmer
Swedish (sv)	UNINE 386 words	UNINE light stemmer	Snowball stemmer

To obtain the desired grid of points, we employed Terrier ver. 4.1 which we extended to work with UNINE stemmers and n-grams. For each task we obtained 160 runs and we calculated four measures: AP, RBP, nDCG20 and ERR20 which capture different performance angles by employing different user models; we chose these measures due to their large use in IR evaluation. The measures have been calculated by employing the *MATlab Toolkit for Evaluation of information Retrieval Systems (MATTERS)* library[4].

[2] http://members.unine.ch/jacques.savoy/clef/index.html.
[3] https://github.com/snowballstem.
[4] http://matters.dei.unipd.it/.

Average Precision (AP) [8] represents the "gold standard" measure in IR, known to be stable and informative, with a natural top-heavy bias and an underlying theoretical basis as approximation of the area under the precision/recall curve. AP is the reference measure in this study for all CLEF tasks and it is the measure originally adopted by CLEF for evaluating the systems participating in the campaigns.

Rank-Biased Precision (RBP) [23] is built around a user model based on the utility a user can achieve by using a system: the higher, the better. The model it implements is that a user always starts from the first document in the list and then s/he progresses from one document to the next with a probability p. We calculated RBP by setting $p = 0.8$ which represent a good trade-off between a very persistent and a remitting user.

nDCG [18] is the normalized version of the widely-known *Discounted Cumulated Gain (DCG)* which is defined for graded relevance judgments. We calculated nDCG in a binary relevance setting by giving gain 0 to non-relevant documents and gain 1 to the relevant ones; furthermore, we used a log_{10} discounting function.

Expected Reciprocal Rank (ERR) [10] is a measure defined for graded relevance judgments and for evaluating navigational intent and it is particularly top-heavy since it highly penalizes systems placing not-relevant documents in high positions. We calculated ERR in a binary relevance setting as we have done for nDCG.

The calculated measures, the scripts used to run Terrier on the CLEF collections along with the property files required to correctly setup the system and the modified version of Terrier comprising UNINE stemmers and n-grams components are publicly available at the URL: http://gridofpoints.dei.unipd.it/.

4 Analysis of the Grid of Points

In Fig. 2 we can see the MAP distributions for the runs composing the grid of points for each considered monolingual task. Given that these runs have been produced by adopting comparable systems, we can conduct an across years comparison between the different editions of the same task. Furthermore, given a task, we can compare the performances obtained by the runs in the grid of points with the performances achieved by the original systems reported in Fig. 1.

By analysing the performances reported in Fig. 2 we can identify two main groups of tasks, the first one comprising languages achieving the highest median and best MAP values which are Spanish, Finnish, French, German and Italian; and, a second group with the Bulgarian, Hungarian, Portuguese and Swedish languages. This difference in performances between different languages can be in part explained by the quality of the linguistic resources employed; indeed, the systems in the grid points obtained better performances for languages introduced in the early years of CLEF – e.g., French and Spanish – and lower performances for the languages introduced in the latter years – e.g., Bulgarian and Hungarian.

By comparing the box plots in Figs. 1 and 2 we can see the distribution of runs in the two sets and we can see where the grid of points runs are a good

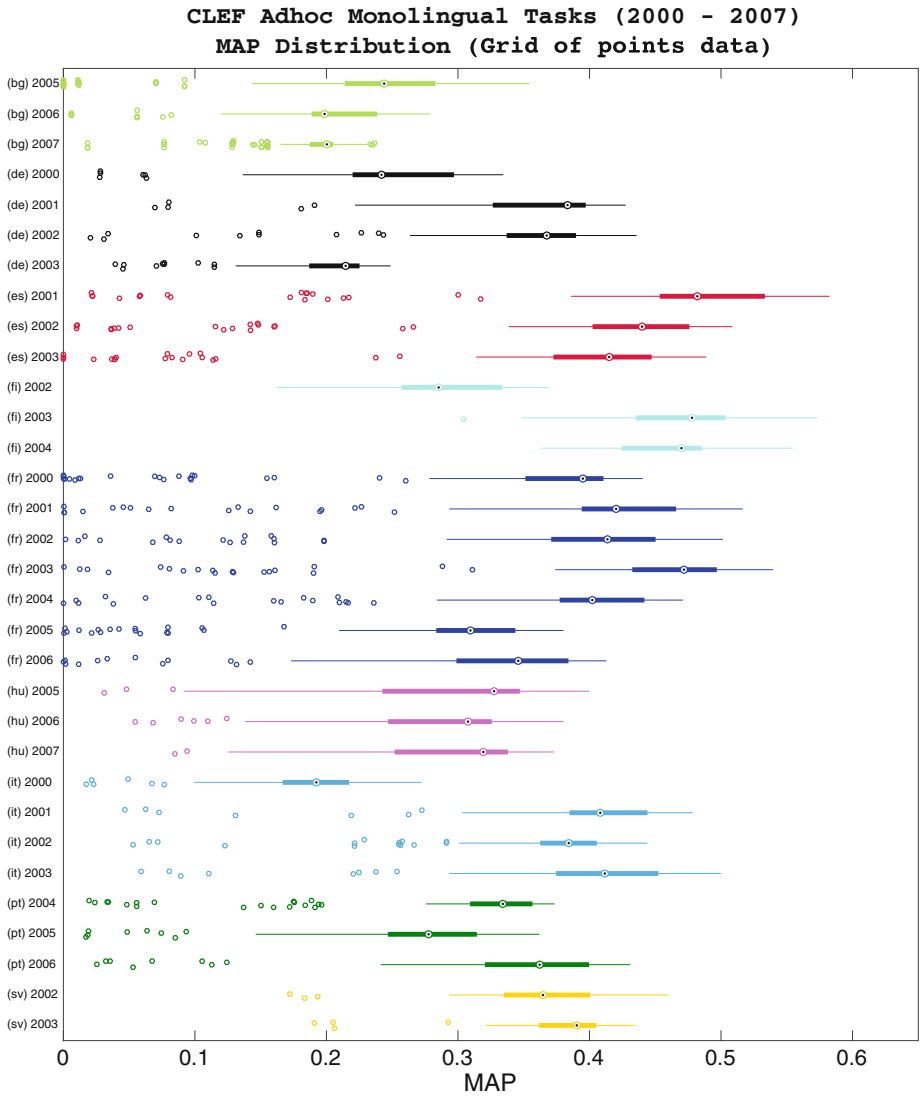

Fig. 2. Grid of points MAP distribution for the considered CLEF monolingual tasks.

representation of the original runs and where they differ one from the other. In the grid of points we have many more runs than in the original CLEF setting and this could explain the higher number of outliers we see in Fig. 2. If we focus on the median MAP values we can see several close correspondences between the original runs and the grid of points ones as for example for the Bulgarian 2005 task, the German 2001 task, the Spanish 2002 task, all Finnish tasks, French tasks from 2000 to 2004, the Italian 2001 task, the Portuguese 2006 task and

all Swedish tasks. On the other hand, there are tasks that do not find a close correspondence between the two run sets as for example the Bulgarian 2006 and 2007 tasks and the Hungarian tasks. Generally, when there is no correspondence, the performances of the grid of points runs are lower than those of the original runs. It must be underlined that some languages, as German and Swedish, get benefit from the use of a word decompounder component [7] which has not been included in the current version of the gird of points; this could lead to worse results in the grid of points with respect to the original CLEF languages.

We employ the *Kernel Density Estimation (KDE)* [27] to estimate the *Probability Density Function (PDF)* of both the original runs submitted at CLEF and the various grids of points. Then, we compute the *Kullback–Leibler Divergence (KLD)* [20] between these PDFs in order to get an appreciation of how different are the grids of points from the original runs. Indeed, KLD $\in [0, +\infty)$ denotes the information lost when a grid of points is used to approximate an original set of runs [9]; therefore, 0 means that there is no loss of information and, in our settings, that the original runs and the grid of points are considered the same; $+\infty$ means that there is full loss of information and, in our settings, that the grid of points and the original runs are considered completely different.

The values of the KLD for all the considered tasks are reported in Fig. 3. In our setting, we assume the "true"/reference probability distribution to be the one associated to the original runs and the "reference" probability distribution to be the one associated to the grid of points runs.

In Fig. 3 we can see that most of the KLD values are fairly low showing the proximity between the original AP values distributions and the grid of points ones. The bigger differences between the distributions are found for the Bulgarian 2006 and 2007 tasks, the German 2000 and 2003 tasks and the Italian 2000 task; for Bulgarian and German, this fact can be checked also by looking at the box plots in Figs. 1 and 2.

In Fig. 4 we can see a comparison between the KDEs of the PDF of AP calculated from the original runs and the grid of points ones; for space reasons

Task	KL-Divergence							
Bulgarian (bg)						6.3642	18.8435	131.6713
Finnish (fi)			5.4461	4.3093	1.9777			
French (fr)	7.8072	5.1439	4.3669	2.1618	3.7242	6.1699	8.5966	
German (de)	14.9074	2.9264	5.5079	61.2449				
Hungarian (hu)						1.4365	8.5492	7.7116
Italian (it)	1633.5	2.3715	3.1710	8.7488				
Portuguese (pt)					7.8868	6.8498	4.3388	
Spanish (es)		4.1603	3.8043	3.4854				
Swedish (sv)			9.0646	3.9338				
	2000	2001	2002	2003	2004	2005	2006	2007

Fig. 3. KLD for all the considered tasks.

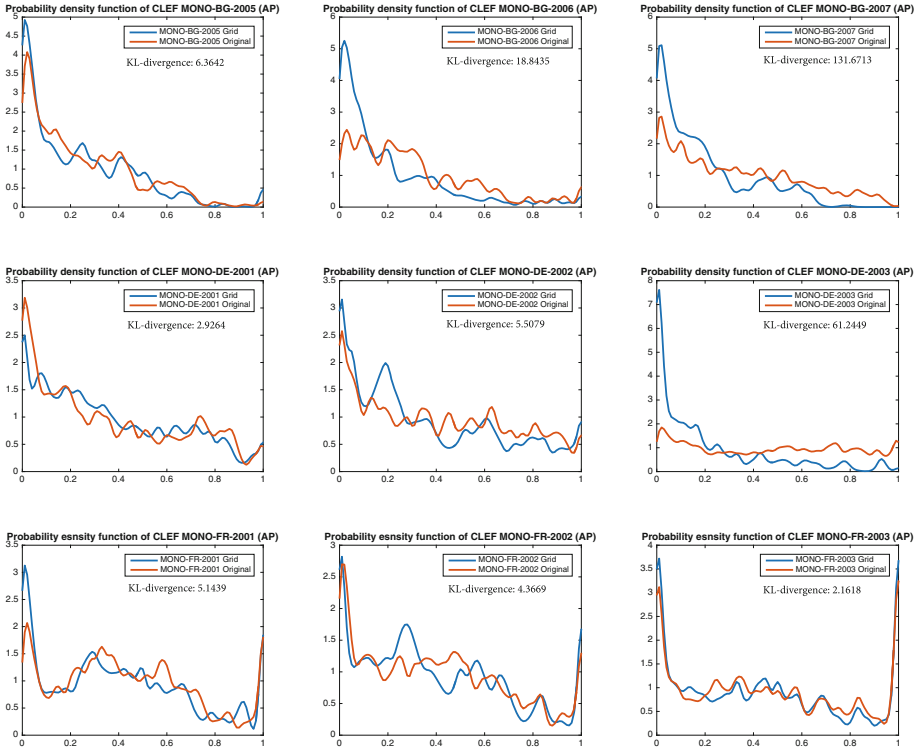

Fig. 4. The KDE of the PDF of AP calculated from the original runs and the grid of points ones.

we report the plots only for nine selected tasks – i.e. the 2005–2007 Bulgarian tasks, the 2001–2003 German tasks and the 2001–2003 French tasks. It is quite straightforward to see the correlation between the shape of the PDF curves and the KLD values reported in Fig. 4.

In Fig. 5 we present a multivariate plot for the CLEF 2003 Monolingual French task which reports the performances of the grid of points runs grouped by stop list, stemmer/n-grams and model. This figure shows a possible performance analysis allowed by the grid of points; indeed, we can see how the different components of the IR systems at hand contribute to the overall performances even though we cannot quantify the exact contribution of each component. For instance, by observing at Fig. 5 we can see that the effect of the stop list is quite evident for all the combinations of system components; indeed, the performances of the systems using a stop list are higher than those not using a stop list. The effect of the stemmer and n-grams components is also noticeable given that the lowest performing systems are consistently those employing neither a stemmer nor a n-grams component; we can also see that the employment of a n-grams component has a positive sizable impact on performances for the French language

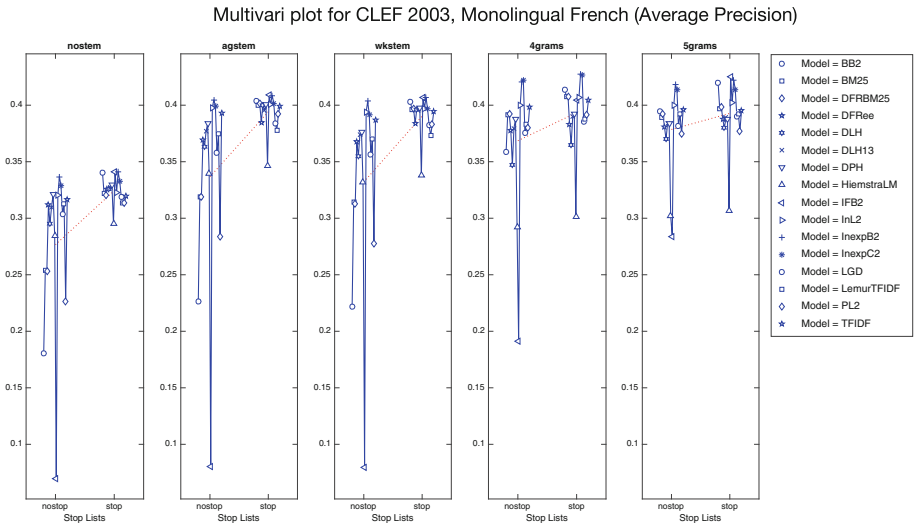

Fig. 5. Multivari plot grouped by stop list, stemmer/n-grams and model for the CLEF 2003 Monolingual French task.

and that it reduces the performance spread amongst the systems. Finally, we can also analyse the impact of different models and their interaction with the other components. For instance, we can see that the IFB2 model is always achieving the lowest performances of the group when the stop list is not employed, whereas it is among the best performing models when a stop list is employed. On the other hand, this model is not highly influenced by the use of stemmers and n-grams components.

5 Final Remarks

In this paper we presented a new valuable resource for MLIA research built over the CLEF Adhoc collections: a big and systematic grid of points combining various IR components – stop lists, stemmers, n-grams, IR models – for several European languages and for different evaluation measures – AP, nDCG, ERR, and RBP.

We assessed whether the produced grids of points are actually representative enough to allow for subsequent analyses and we have found that they have performance distributions similar to those of the runs originally submitted to the CLEF Adhoc tasks over the years.

Moreover, we have shown some of the analyses that are enabled by the grid of point and how they allow us to start understanding how components interact together.

These analyses are intended to show the potentialities of the grid of points that can be exploited to carry out deeper analyses and considerations.

For instance, the grid of points can be the starting point for determining the contribution of a specific component within the full pipeline of an IR system and to estimate the interaction of one component with the other. As a consequence, as far as future work is concerned, we will decompose system performance into components' ones according to the methodology we proposed [17] and we will try to generalize this decomposition across languages.

References

1. Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). SIGIR Forum **49**(2), 107–116 (2015)
2. Braschler, M.: CLEF 2000 - overview of results. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, p. 89. Springer, Heidelberg (2001)
3. Braschler, M.: CLEF 2001 - overview of results. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 9–26. Springer, Heidelberg (2002)
4. Braschler, M.: CLEF 2002 – overview of results. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 9–27. Springer, Heidelberg (2003)
5. Braschler, M.: CLEF 2003 – overview of results. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44–63. Springer, Heidelberg (2004)
6. Braschler, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF 2004: ad hoc track overview and results analysis. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 10–26. Springer, Heidelberg (2005)
7. Braschler, M., Ripplinger, B.: How effective is stemming and decompounding for german text retrieval? Inf. Retr. **7**(3–4), 291–316 (2004)
8. Buckley, C., Voorhees, E.M.: Retrieval system evaluation. In: TREC: Experiment and Evaluation in Information Retrieval, pp. 53–78. MIT Press (2005)
9. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, p. 488. Springer, Heidelberg (2002)
10. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of 18th International Conference on Information and Knowledge Management (CIKM), pp. 621–630. ACM Press (2009)
11. Di Buccio, E., Di Nunzio, G.M., Ferro, N., Harman, D.K., Maistro, M., Silvello, G.: Unfolding off-the-shelf IR systems for reproducibility. In: Proceedings of SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) (2015)
12. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 11–36. Springer, Heidelberg (2006)
13. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 21–34. Springer, Heidelberg (2007)
14. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing reproducibility in IR: findings from the Dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". SIGIR Forum **50**(1), 68–82 (2016)

15. Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF pilot track overview. In: Roda, G., Peters, C., Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 552–565. Springer, Heidelberg (2010)
16. Ferro, N., Silvello, G.: CLEF 15th birthday: what can we learn from ad hoc retrieval? In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 31–43. Springer, Heidelberg (2014)
17. Ferro, N., Silvello, G.: A general linear mixed models approach to study system component effects. In: Proceedings of 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). ACM Press (2016)
18. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) **20**(4), 422–446 (2002)
19. Kluck, M., Womser-Hacker, C.: Inside the evaluation process of the cross-language evaluation forum (CLEF): issues of multilingual topic creation and multilingual relevance assessment. In: Proceedings of 3rd International Language Resources and Evaluation Conference (LREC 2002) (2002)
20. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
21. Lin, J., et al.: Toward reproducible baselines: the open-source IR reproducibility challenge. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 408–420. Springer, Heidelberg (2016). doi:10.1007/978-3-319-30671-1_30
22. Macdonald, C., McCreadie, R., Santos, R.L.T., Ounis, I.: From puppy to maturity: experiences in developing terrier. In: Proceedings of OSIR at SIGIR, pp. 60–63 (2012)
23. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 2:1–2:27 (2008)
24. Robertson, S.E.: The methodology of information retrieval experiment. In: Jones, K.S. (ed.) Information Retrieval Experiment, pp. 9–31. Butterworths, London (1981)
25. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retr. **4**(4), 247–375 (2010)
26. Trotman, A., Clarke, C.L.A., Ounis, I., Culpepper, J.S., Cartright, M.A., Geva, S.: Open source information retrieval: a report on the SIGIR 2012 workshop. ACM SIGIR Forum **46**(2), 95–101 (2012)
27. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall/CRC, Boca Raton (1995)
28. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proceedings of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 51–58. ACM Press (2008)