A General Linear Mixed Models Approach to Study System Component Effects

Nicola Ferro
Department of Information Engineering
University of Padua
Padua, Italy
nicola.ferro@unipd.it

Gianmaria Silvello
Department of Information Engineering
University of Padua
Padua, Italy
gianmaria.silvello@unipd.it

ABSTRACT

Topic variance has a greater effect on performances than system variance but it cannot be controlled by system developers who can only try to cope with it. On the other hand, system variance is important on its own, since it is what system developers may affect directly by changing system components and it determines the differences among systems. In this paper, we face the problem of studying system variance in order to better understand how much system components contribute to overall performances. To this end, we propose a methodology based on *General Linear Mixed Model (GLMM)* to develop statistical models able to isolate system variance, component effects as well as their interaction by relying on a *Grid of Points (GoP)* containing all the combinations of analysed components. We apply the proposed methodology to the analysis of TREC Ad-hoc data in order to show how it works and discuss some interesting outcomes of this new kind of analysis. Finally, we extend the analysis to different evaluation measures, showing how they impact on the sources of variance.

1. INTRODUCTION

The experimental results analysis is a core activity in *Information Retrieval (IR)* aimed at, firstly, understanding and improving system performances and, secondly, assessing our own experimental methods, such as robustness of experimental collection or properties of the evaluation measures. When it comes to explaining system performances and differences between algorithms, it is commonly understood [10, 17, 23] that system performances can be broken down to a reasonable approximation as

$$\text{system performances} = \text{topic effect} + \text{system effect} + \text{topic/system interaction effect}$$

even though it is not always possible to estimate these effects separately, especially the interaction one.

It is well-known that topic variability is greater than system variability [23, 26]. Therefore, a lot of effort has been put in better understanding this source of variance [17] as well as in making IR systems more robust to it, e.g. [25, 28], basically trying to improve on the interaction effect. Nevertheless, with respect to an IR system, topic variance is a kind of "external source" of variation, which cannot be controlled by developers, but can only be taken into account to better deal with it.

On the other hand, system variance is a kind of "internal source" of variation, since it is originated by the choice of system components, may be directly affected by developers by working on them, and represents the intrinsic differences between algorithms. Its importance is witnessed by the wealth of research on how to compare systems performances in a reliable and robust way [1, 2, 4, 9, 20–23, 27].

However, a limitation of the current experimental methodology is that it allows us to evaluate IR systems only as a kind of "black-boxes", without an understanding of how their different components interact with each other and contribute to the overall performances. In other terms, we consider system variance as a single monolithic contribution and we cannot break it down into the smaller pieces (the components) constituting an IR system.

In order to estimate the effects of the different components of an IR system, we develop a methodology, based on *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* [13, 18], which makes us of a *Grid of Points (GoP)* containing all the possible combinations of inspected components. The proposed methodology allows us to break down the system effect into the contributions of stops lists, stemmers or n-grams and IR models, as well as to study their interaction.

We experimented on standard *Text REtrieval Conference (TREC)* Ad-hoc collections and produced a GoP by using the Terrier[1] open source IR system [12]. This gave us a very controlled experimental setting, which allowed us to systematically fit our *General Linear Model (GLM)* and break down the system variance. Note that such a controlled experimental setting is typically not available in evaluation campaigns, such as TREC, where participating systems do not constitute a systematic sampling of all the possible combinations of components and often are not even described in such a detail to know exactly what components have been used.

We applied the proposed methodology to TREC 5, 6, 7, and 8 Ad-hoc collections and we employed different measures – AP, Precision at 10, RBP, nDCG@20, and ERR@20. This setup allows us not only to highlight how components contribute to the overall system variance but also to gain

[1] http://www.terrier.org/

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2911530

insights on how different evaluation measures impact on system and component variances.

The paper is organized as follows: Section 2 presents related work; Section 3 introduces our methodology; Section 4 experiments the proposed methodology; and, Section 5 draws conclusions and discusses future work.

2. RELATED WORKS

The impossibility of testing a single component by setting it aside from the complete IR system is a long-standing and well-known problem in IR experimentation, as early remarked by [16]. Component-based evaluation methodologies [6–8] have tried to tackle this issue by providing technical solutions for mixing different components without the need of building a whole IR system. However, even if these approaches allowed researchers to focus on the components of their own interest, they have not delivered yet estimates of the performance figures of each component.

The decomposition of performances into system and topic effects has been exploited by [1, 23] to analyze TREC data; [4] proposed model-based inference, using linear models and ANOVA, as an approach to multiple comparisons; [10] used multivariate linear models to compare non-deterministic IR systems among them and with deterministic ones. In all these cases, the goal is a more accurate comparison among systems rather than an analysis and breakdown of system variance per se. [17] applied GLMM to the study of per-topic variance by using simulated data to generate more replicates for each (topic, system) pair in order to estimate also the topic/system interaction effect; however, they did not use real data nor did focus on breaking down the system effect.

The idea of creating all the possible combinations of components has been proposed by [7], who noted that a systematic series of experiments on standard collections would have created a GoP, where (ideally) all the combinations of retrieval methods and components are represented, allowing us to gain more insights about the effectiveness of the different components and their interaction; this would have called also for the identification of *suitable baselines* with respect to which all the comparisons have to be made. Even though [7] introduced the idea of a GoP and how it could have been central to the decomposition of system component performances, they did not come up with an full-fledged methodology for analyzing such data and breaking down component performances, which is the contribution of the present work instead.

More recently, the proliferation of open source IR systems [24] has greatly ameliorated the situation, allowing researchers to run systematic experiments more easily. This led the community to further investigate what *reproducible baselines* are [5, 11] and the "Open-Source Information Retrieval Reproducibility Challenge" provided several of these baselines, putting some points in the ideal GoP mentioned above. We move a step forward with respect to [11] since we propose an actual methodology for exploiting such GoPs to decompose system performances and we rely on a much finer-grained grid, in terms of number of components and IR models experimented.

3. METHODOLOGY

The goal of the proposed methodology is to decompose the effects of different components on the overall system performances. In particular, we are interested in investigating the effects of the following components: stop lists; *Lexical Unit Generator (LUG)*, namely stemmers or n-grams; IR models, such as the vector space or the probabilistic model.

We create a *Grid of Points (GoP)* on a standard experimental collection by running all the IR systems resulting from all the possible combinations of the considered components (stop list, LUG, IR model); we consider stemmers and n-grams as alternative LUG components, thus we do not consider IR systems using both stemmer and n-grams.

Given a performance measure, such as *Average Precision (AP)*, we produce a matrix Y, as the one shown in Figure 1, where each cell Y_{ij} represents a measurement on topic t_i of the system s_j. Note that the column average – i.e., $\mu_{\cdot j}$ – is the performance mean over all topics for a given system, e.g. *Mean Average Precision (MAP)*; the row average – i.e., $\mu_{i\cdot}$ – is the performance mean over all systems for a given topic.

A GLMM explains the variation of a dependent variable Y ("Data") in terms of a controlled variation of independent variables ("Model") in addition to a residual uncontrolled variation ("Error").

$$\text{Data} = \text{Model} + \text{Error}$$

The term "General" refers to the ability to accommodate distinctions on quantitative variables representing continuous measures (as in regression analysis) and categorical distinctions representing groups or experimental conditions (as in ANOVA). In our case, we deal with categorical independent variables, as for example different types of stemmers, which constitute the levels of such categorical variable. The term "Linear" indicates that the "Model" is expressed as a linear combination of factors, where the factors can be single independent variables or their combinations. In our case, we are interested both in single independent variables, i.e. the *main effects* of the different components alone, and their combinations, i.e. the *interaction effects* between components. The term "Mixed" refers to the fact that some independent variables are considered *fixed effects* – i.e. they have precisely defined levels, and inferences about its effect apply only to those levels – and some others are considered *random effects* – i.e. they describe a randomly and independently drawn set of levels that represent variation in a clearly defined wider population; a random factor is indicated by adding a single quote as superscript to the variable name. In our case, the different kinds of systems and components are fixed effects while topics are random effects.

The experimental design determines how you compute the model and how you estimate its parameters. In particular, it is possible to have *independent measures* designs where different subjects participate to different experimental conditions (factors) or *repeated measures* designs, where each subject participates to all the experimental conditions (factors). In our case systems and their components are the experimental conditions (factors) while topics are the subjects and, since each topic is processed by each system, we have a repeated measure design.

One advantage of repeated measures designs is a reduction in error variance due to the greater similarity of the scores provided by the same subjects; in this way, variability in individual differences between subjects is removed from the error. Basically, a repeated measure design increases the statistical power for a fixed number of subjects or, in other

terms, it allows us to reach a desired level of power with less subjects than those required in the independent measures design.

A final distinction is between *crossed/factorial* designs, where every level of one factor is measured in combination with every level of the other factors, and *nested* designs, where levels of a factor are grouped within each level of another nesting factor. In our case, we have a crossed/factorial design because in the generated GoP we experiment each possible combination of components.

3.1 Single Factor Repeated Measures Design

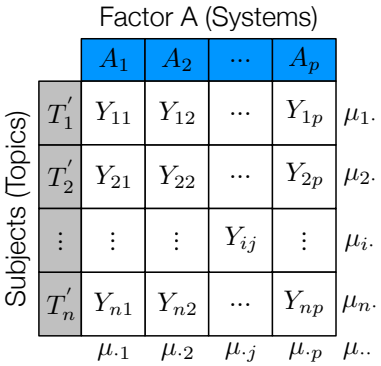

Figure 1: Single factor repeated measures design.

This design is the one typically used when ANOVA is applied to the analysis of the system performances in a track of an evaluation campaign, as in [1, 23], where the subjects are the topics and the factors are the system runs. Basically, in this context ANOVA is used to determine which experimental condition dependent variable score means differ, i.e. which systems are significantly different from others.

In our case, we are interested also in a second aspect, i.e. to determine what proportion of variation in the dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variables. This turns into determining which proportion of variation is due to the topics and which one to the systems.

3.1.1 Model

The full GLMM model for the one-way ANOVA with repeated measures is:

$$Y_{ij} = \underbrace{\mu_{\cdot\cdot} + \tau_i + \alpha_j}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}} \tag{3.1}$$

where: Y_{ij} is the score of the i-th subject (topic) in the j-th factor (system); $\mu_{\cdot\cdot}$ is the grand mean; τ_i is the effect of the i-th subject $\tau_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$ where $\mu_{i\cdot}$ is the mean of the i-th subject; α_j is the effect of the j-th factor $\alpha_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$ where $\mu_{\cdot j}$ is the mean of the j-th factor; ε_{ij} is the error committed by the model in predicting the score of the i-th subject in the j-th factor. It consists of a term $(\tau\alpha)_{ij}$ which is the interaction between the i-th subject and the j-th factor[2]; and, a term ϵ_{ij} which is any additional error due to uncontrolled sources of variance.

[2] In order to calculate interaction effects, you need to have several scores (*replicates*) for each cell. The mean of the cell scores is taken as the best estimate of the cell score and is used to calculate interaction effects, with the discrepancy between the mean and the actual score providing the estimates of experimental error. If there is only one score per subject per factor, then a mean and its error cannot be calculated per subject per factor and without these estimates, the factor ϵ_{ij} cannot be separated from the interaction effect $(\tau\alpha)_{ij}$.

3.1.2 Estimators

We have the following estimators for the parameters of the model above:

- grand mean: $\hat{\mu}_{\cdot\cdot} = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ij}$
- mean of the i-th subject $\hat{\mu}_{i\cdot} = \frac{1}{p}\sum_{j=1}^{p} Y_{ij}$ and its effect $\hat{\tau}_i = \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$
- mean of the j-th factor $\hat{\mu}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$ and its effect $\hat{\alpha}_j = \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$
- score of the i-th subject in the j-th factor $\hat{Y}_{ij} = \hat{\mu}_{\cdot\cdot} + \hat{\tau}_i + \hat{\alpha}_j = \hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$
- prediction error of the i-th subject in the j-th experimental condition $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_{\cdot\cdot} - \hat{\tau}_i - \hat{\alpha}_j = Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})$

3.1.3 Assessment

We can write the model of equation (3.1) introducing the estimated parameters as

$$\begin{aligned} Y_{ij} &= \hat{\mu}_{\cdot\cdot} + \hat{\tau}_i + \hat{\alpha}_j + \hat{\varepsilon}_{ij} \\ &= \hat{\mu}_{\cdot\cdot} + (\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}) + (\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}) + (Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})) \end{aligned}$$

which leads to the following decomposition of the effects

$$\underbrace{Y_{ij} - \hat{\mu}_{\cdot\cdot}}_{\text{Total Effects}} = \underbrace{\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}}_{\text{Subject Effects}} + \underbrace{\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}}_{\text{Factor Effects}} + \underbrace{Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})}_{\text{Error Effects}} \tag{3.2}$$

From equation (3.2), we can compute the sum of squares (SS), degrees of freedom (DF), and mean squares (MS) as follows:

- total effects $SS_{tot} = \sum_{j=1}^{p}\sum_{i=1}^{n}(Y_{ij} - \hat{\mu}_{\cdot\cdot})^2$ with mean squares $MS_{tot} = \frac{SS_{tot}}{df_{tot}}$ where $df_{tot} = pn - 1$ where df_{tot} comes from the fact that we are summing up pn scores and one degree of freedom is lost because of the grand mean $\hat{\mu}_{\cdot\cdot}$;
- subject effects

  $$SS_{subj} = \sum_{j=1}^{p}\sum_{i=1}^{n}(\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot})^2 = \sum_{i=1}^{n} p\,(\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot})^2$$

  with mean squares $MS_{subj} = \frac{SS_{subj}}{df_{subj}}$ where $df_{subj} = n-1$ where SS_{subj} considers that the quantity $\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$ is the same for all the p factors which the i-th subject experiences; df_{subj} is calculated by summing up n times the subject mean $\hat{\mu}_{i\cdot}$ where one degree of freedom is lost because of the grand mean $\hat{\mu}_{\cdot\cdot}$;
- factor effects

  $$SS_{fact} = \sum_{j=1}^{p}\sum_{i=1}^{n}(\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})^2 = \sum_{j=1}^{p} n\,(\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})^2$$

  with mean squares $MS_{fact} = \frac{SS_{fact}}{df_{fact}}$ where $df_{fact} = p-1$ where SS_{fact} considers that the quantity $\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$

is the same for all the n subjects which experience the j-th factor; df_{fact} is calculated by summing up p times the factor mean $\hat{\mu}_{\cdot j}$ where one degree of freedom is lost because of the grand mean $\hat{\mu}_{\cdot\cdot}$;

- error effects

$$SS_{err} = \sum_{j=1}^{p}\sum_{i=1}^{n}(Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}))^2$$

with mean squares $MS_{err} = \frac{SS_{err}}{df_{err}}$ where $df_{err} = (p-1)(n-1)$ where df_{err} is calculated by summing up n times the scores where one degree of freedom is lost in the subject scores because of the subject mean $\hat{\mu}_{i\cdot}$ and one degree of freedom is lost in the factor scores because of the factor mean $\hat{\mu}_{\cdot j}$.

Note that $SS_{tot} = SS_{subj} + SS_{fact} + SS_{err}$.

In order to determine if the factor effect is statistically significant, we compute the F statistics defined as:

$$F_{fact} = \frac{MS_{fact}}{MS_{err}} \tag{3.3}$$

and compare it with the distribution $F_{(df_{fact}, df_{err})}$ under the null hypothesis H_0 that there are not significant differences in order to estimate the probability (p-value) that F_{fact} has been observed by chance. We can set a significance level α (typically $\alpha = 0.05$) and, if p-value $< \alpha$, the factor effect is considered statistically significant.

As introduced above, we are not only interested in determining whether the factor effect is significant but also which proportion of the variance is due to it, that is we need to estimate its *effect-size measure* or *Strength of Association (SOA)*. The SOA is a "standardized index and estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables" [15]. We use the $\hat{\omega}^2_{\langle fact\rangle}$ SOA:

$$\hat{\omega}^2_{\langle fact\rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + pn} \tag{3.4}$$

which is an unbiased estimator of the variance components associated with the sources of variation in the design.

The common rule of thumb [14] when classifying $\hat{\omega}^2_{\langle fact\rangle}$ effect size is: 0.14 and above is a large effect, 0.06–0.14 is a medium effect, and 0.01–0.06 is a small effect. $\hat{\omega}^2_{\langle fact\rangle}$ values could happen to be negative and in such cases they are considered as zero.

When you conduct experiments, two types of error may happen. A *Type 1* error occurs when a true null hypothesis is rejected and the significance level α is the probability of committing a Type 1 error. A *Type 2* error occurs when a false null hypothesis is accepted and it is concerned with the capability of the conducted experiment to actually detect the effect under examination. Type 2 errors are often overlooked because if they occur, although a real effect is missed, no misdirection occurs and further experimentation is very likely to reveal the effect.

The *power* is the probability of correctly rejecting a false null hypothesis when an experimental hypothesis is true

$$\text{Power} = 1 - \beta$$

where β (typically $\beta = 0.2$) is the Type 2 error rate.

To determine the power of an experiment, we compute the effect size parameter:

$$\phi = \sqrt{n \cdot \frac{\hat{\omega}^2_{\langle fact\rangle}}{1 - \hat{\omega}^2_{\langle fact\rangle}}} \tag{3.5}$$

and we compare it with its tabulated values for a given Type 1 error rate α to determine β.

3.2 Factorial Repeated Measures Design

While single factor designs manipulate a single variable, factorial designs take into account two or more factors as well as their interaction. As an example a two factors repeated measure design can be defined extending the design described above, where we manipulated one factor (A), by adding an additional factor (B) and the interaction between them (AB).

We can therefore define a three factors design where we manipulate factors A, B and C which correspond to the stop lists, the LUG and the IR models respectively; with this design we can also study the interaction between component pairs as well as the third order interaction between them.

In Figure 2 we can see a table which extends to three factors the design presented in Figure 1 for a single factor. We can see that the systems are now decomposed into three main constituents: (i) factor A (stop lists) with p levels where, for instance, A_1 corresponds to the absence of a stop list, A_2 to the indri stop list, A_3 to the terrier stop list and so on; (ii) factor B (LUG) with q levels where B_1 corresponds to the absence of a LUG, B_2 to the Porter stemmer, B_3 to the Krovetz stemmer and so on; (iii) factor C (IR models) with r levels where C_1 corresponds to BM25, C_2 to TF*IDF and so on. We call this design a $p \times q \times r$ factorial design. Each cell of the table in Figure 2, say Y_{npqr}, reports the measurement (e.g., AP) on topic $T_n^{'}$, for the system composed by the stop list A_p, the LUG B_q and IR model C_r.

The full GLMM model for the described factorial ANOVA for repeated measures with three fixed factors (A, B, C) and a random factor ($T^{'}$) is:

$$\begin{aligned} Y_{ijkl} = & \underbrace{\mu_{....} + \tau_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \\ & \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl} + \alpha\beta\gamma_{jkl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \end{aligned} \tag{3.6}$$

where: Y_{ijkl} is the score of the i-th subject in the j-th, k-th, and l-th factors; $\mu_{....}$ is the grand mean; τ_i is the effect of the i-th subject $\tau_i = \mu_{i...} - \mu_{....}$ where $\mu_{i...}$ is the mean of the i-th subject; $\alpha_j = \mu_{.j..} - \mu_{....}$ is the effect of the j-th factor, where $\mu_{.j..}$ is the mean of the j-th factor; $\beta_k = \mu_{..k.} - \mu_{....}$ is the effect of the k-th factor, where $\mu_{..k.}$ is the mean of the k-th factor; and, $\gamma_l = \mu_{...l} - \mu_{....}$ is the effect of the l-th factor where $\mu_{...l}$ is the mean of the l-th factor; ε_{ijkl} is the error committed by the model in predicting the score of the i-th subject in the three factors j, k, l. It consists of all the interaction terms between the random subjects and the fixed factors, such as $(\tau\alpha)_{ij}$, $(\tau\beta)_{ik}$ and so on, plus the error ϵ_{ijkl} which is any additional error due to uncontrolled sources of variance. As in the single factor design to calculate interaction effects with the subjects, you need to have replicates; when there is only one score per subject per factor the factor ϵ_{ijkl} cannot be separated from the interaction effects with the random subjects.

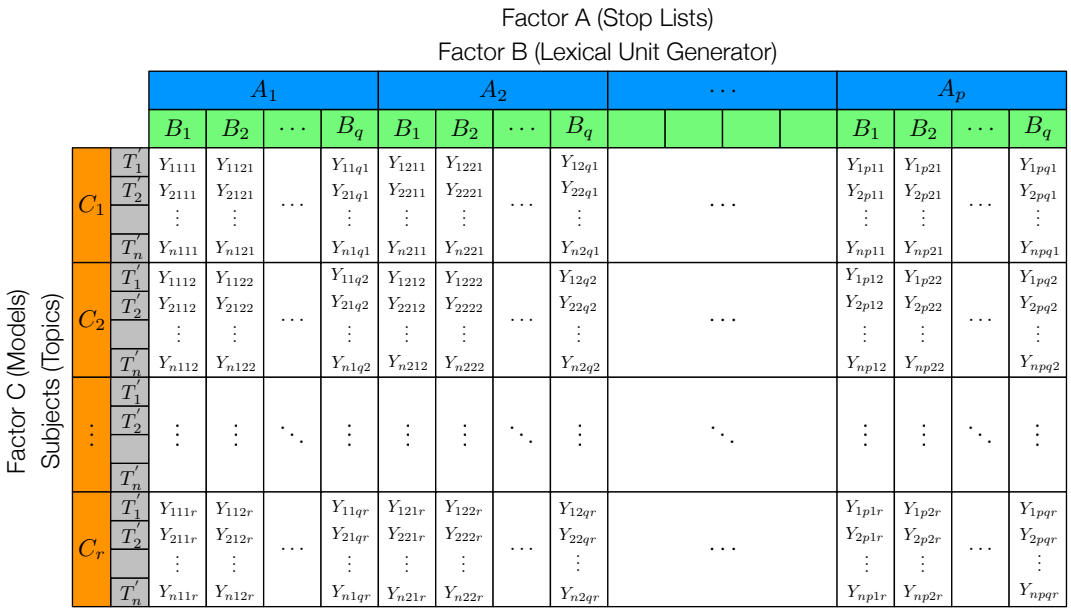

Figure 2: Three factors repeated measures design.

The estimators of the main effects can be derived by extension from those of the single factor design; for instance, the grand mean is $\hat{\mu}_{....} = \frac{1}{rqpn}\sum_{l=1}^{r}\sum_{k=1}^{q}\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ijkl}$, the mean of the k-th effect is $\hat{\mu}_{..k.} = \frac{1}{rpn}\sum_{l=1}^{r}\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ijkl}$ and its estimator is $\hat{\beta}_k = \hat{\mu}_{..k.} - \hat{\mu}_{....}$.

The estimators of the interaction factors are calculated as follows, let us consider $(\alpha\beta)_{jk}$:

$$\widehat{\alpha\beta}_{jk} = \hat{\mu}_{.jk.} - (\hat{\mu}_{....} + \hat{\alpha}_j + \hat{\beta}_k) \tag{3.7}$$

where $\hat{\mu}_{.jk.} = \frac{1}{nr}\sum_{i=1}^{n}\sum_{l=1}^{r} Y_{ijkl}$; $\hat{\alpha}_j = \hat{\mu}_{.j..} - \hat{\mu}_{....}$; and, $\hat{\beta}_k = \hat{\mu}_{..k.} - \hat{\mu}_{....}$.

Similarly, we calculate the estimators for all the other interaction factors – i.e. $\widehat{\alpha\gamma}_{jl}$ and $\widehat{\beta\gamma}_{kl}$; $\widehat{\alpha\beta\gamma}_{jkl}$ is calculated by extending equation (3.7):

$$\widehat{\alpha\beta\gamma}_{jkl} = \hat{\mu}_{.jkl} - (\hat{\mu}_{....} + \hat{\alpha}_j + \hat{\beta}_k + \hat{\gamma}_l) \tag{3.8}$$

where $\hat{\mu}_{.jkl} = \frac{1}{n}\sum_{i=1}^{n} Y_{ijkl}$ and $\hat{\gamma}_l = \hat{\mu}_{...l} - \hat{\mu}_{....}$.

In this design the error $\varepsilon_{ijkl} = Y_{ijkl} - \hat{Y}_{ijkl}$ contains the variance not explained by the main and interaction effects discussed above and it is composed by all the interactions of the subjects τ_j with the other factors in the model in addition to the uncontrolled sources of variance.

The sum of squares, mean squares and degrees of freedom of the main effects can be derived by extension form those of the one factor design. As an example, the degrees of freedom of factor A are $p-1$ and its sum of squares is:

$$SS_A = \sum_{l=1}^{r}\sum_{k=1}^{q}\sum_{j=1}^{p}\sum_{i=1}^{n}\hat{\alpha}_j^2 = rqn\sum_{j=1}^{p}(\hat{\mu}_{.j..} - \hat{\mu}_{....})^2$$

As an example of the computations for the interaction terms, we consider the term $A\times B$ whose degrees of freedom are $(p-1)(q-1)$ and whose sum of squares is:

$$SS_{A\times B} = \sum_{l=1}^{r}\sum_{k=1}^{q}\sum_{j=1}^{p}\sum_{i=1}^{n}\widehat{\alpha\beta}_{jk}^2 = rn\sum_{k=1}^{q}\sum_{j=1}^{p}(\hat{\mu}_{.jk.} - \hat{\mu}_{.j..} - \hat{\mu}_{..k.} + \hat{\mu}_{....})^2$$

As in the single factor design case, the mean squares of a factor (both main and interaction) are calculated by dividing its sum of squares by its degrees of freedom, the F-test is calculated with equation (3.3), the SOA measure with equation (3.4), and the power with equation (3.5).

4. EXPERIMENTATION AND DISCUSSION

We considered three main components of an IR system: stop list, LUG and IR model. We selected a set of alternative implementations of each component and by using the Terrier open source system we created a run for each system defined by combining the available components in all possible ways. The components we selected are:

stop list: nostop, indri, lucene, smart, terrier;

LUG: nolug, weak Porter, Porter, Krovetz, Lovins, 4grams, 5grams;

model: BB2, BM25, DFRBM25, DFRee, DLH, DLH13, DPH, HiemstraLM, IFB2, InL2, InexpB2, InexpC2, LGD, LemurTFIDF, PL2, TFIDF.

Note that the stemmers and n-grams of the LUG component are used as alternatives, this means that we end up with two distinct groups of runs, one using the stemmers and one using the n-grams; the `nolug` component is common to both these groups. The group using the stemmers defines a $5\times5\times16$ factorial design with a grid of points consisting of 400 runs; the group using the n-grams defines a $5\times3\times16$ factorial design with a grid of points consisting of 240 runs.

Table 1: Single factor, ANOVA table for TREC 08 (stemmer group) using AP.

Source	SS	DF	MS	F	p-value
Topics′	820.99	49	16.75	694.7235	0
Systems	36.44	399	0.09	7.4464	0
Error	88.20	19551	0.0045		
Total	945.63	19999			

We conducted single factor and three-factors ANOVA tests for both the groups on TREC 05, 06, and 08 collections, and by employing the following five measures: AP, P@10, nDCG@20, RBP and ERR@20. All the test collections are composed by 50 different topics and have binary relevance judgments; the corpus of TREC 05 is the TIPSTER disk 2 and 4 counting 525K documents, the corpus of TREC 06 is TIPSTER disk 4 and 5 counting 556K documents and the corpus of TREC 07 and 08 is the TIPSTER disk 4 and disk 5 (minus Congressional Record) counting 528K documents.

To ease reproducibility, the code for running the experiments is available at: http://gridofpoints.dei.unipd.it/.

4.1 Single Factor Repeated Measures Effects

We conducted 40 single factor ANOVA tests (4 collections $\times$ 5 measures $\times$ 2 run groups), so for space reasons we cannot report all the result; as an example, Table 1 reports the synthesis data of the ANOVA test for TREC 08 using the stemmer group of runs measured with AP.

From the sum of squares (SS) and the mean squares (MS), we can see that topics explain a large portion of the total variance. Nonetheless, the effect of the IR systems is statistically significant (p-value 0). We can also see that the sum of squares of the error is not negligible since it contains both the variance of the unexplained topics/systems interaction effect and the the other uncontrolled sources of variance. From this table we can calculate the statistical power of the experiment, which is 1 with a Type 1 error probability $\alpha = 0.05$, indicating that we are observing effects in a reliable way.

Table 2 reports the $\hat{\omega}^2_{\langle system \rangle}$ SOA measure and the p-value of the ANOVA test for the single factor models on all the test collections for all the considered evaluation measures. The "LUG" column indicates the runs group we are considering (stemmers or n-grams). This table shows that despite the high variance of the topics, the system effect sizes are generally large and this is consistent across all the collections and measures. Moreover, system effect sizes of stemmer runs group systems are large (> 0.14) for all the collections and measures with the solely exception of AP for TREC 05. Whereas, for the n-grams runs group we can see that the system effect sizes are consistently smaller than those of the stemmer group; this, supports the observation that "for English, n-grams indexing has no strong impact" [3].

Table 2 shows that measures impact on the amount of variance explained by the system effect. Generally, system effect sizes are higher when nDCG@20 is used, followed by RBP, P@10, AP and ERR@20. This could be related to two characteristics of the measures: their discriminative power and their user model. Indeed, if a measure is less discriminative than another one, it could be able to grasp less variance in the system effect; on the other hand, different user models mean looking at (very) different angles of system performances and this can change the explained variance.

To explore a bit this hypothesis, in Table 3 we report the discriminative power of the five considered measures over the test collections calculated by employing the paired bootstrap test defined in [19]. We can see that there is some agreement between the system effect sizes for a measure and its discriminative power; for instance, ERR@20 explains less system variance than the other measures and this can be explained by its discriminative power which is the lowest amongst all measures; similarly, RBP and nDCG@20 have both comparable discriminative power and close system effect sizes. The main exception is AP which typically has the highest discriminative power but the smallest system effect size; this could be due to the user model behind AP, which is quite different from the one of the other measures and may counterbalance the higher discriminative power leading to a final lower system effect size.

4.2 Three Factors Repeated Measures Effects

In Table 4 we report the ANOVA table of a three factors test for the stemmer group of runs on TREC 08 measured with AP.

We can see that the sum of squares of the topics is the same as the one determined with the single factor design, as well as the error and the total sum of squares. The main difference with the one factor design is that the variance of the systems is now decomposed into three main effects (stop list, stemmer and IR model) and four interaction effects. In this case all the main effects are statistically significant meaning that they have a role in explaining systems variance; in particular, the stop list explains more variance than the model and the stemmer is the component with the lowest impact in this design. Amongst the interaction effects, only the stoplist*model effect is significant explaining a tangible portion of the systems variance. The statistical power for the main effects is 0.97 for the stop list, 0.66 for the stemmer and 0.99 for the model with a Type 1 error probability $\alpha = 0.05$.

Table 5 reports the estimated ω^2 SoA for all the main and interaction effects and the p-values for all the ANOVA three-way tests we conducted; from this this table we can see that main and interaction effect sizes are consistent across the different collections.

Analyzing the main effect sizes reported in Table 5 we can see that for the stemmer group of runs the stop list has always a higher $\hat{\omega}^2$ than the IR model and the stemmer and, with the solely exceptions of TREC 05 for AP and ERR@20, the stop list has a medium effect size. Whereas, n-grams tend to reduce the stop list effect and to increase the IR model one; this can be also seen from the *n-grams*model* interaction effect which is small but statistically significant, differently from the *stemmer*model* effect which is never significant.

These observations cast a light on the importance of linguistic pre-processing and linguistic resources, given that the role of the stop list is significant in an IR system as well as choosing between stemmers or n-grams. We can further analyze these aspects by looking at Figure 3; the plot on the left reports the main effects for the TREC 08 stemmer group case and we show the marginal means (response means) described in Section 3.2 for the effect under investigation on the y-axis and the various components on the x-axis.

From the first plot we see that the presence or absence of a stop list affects the system performances because the line connecting "no stop" and "indri" is not horizontal, whereas

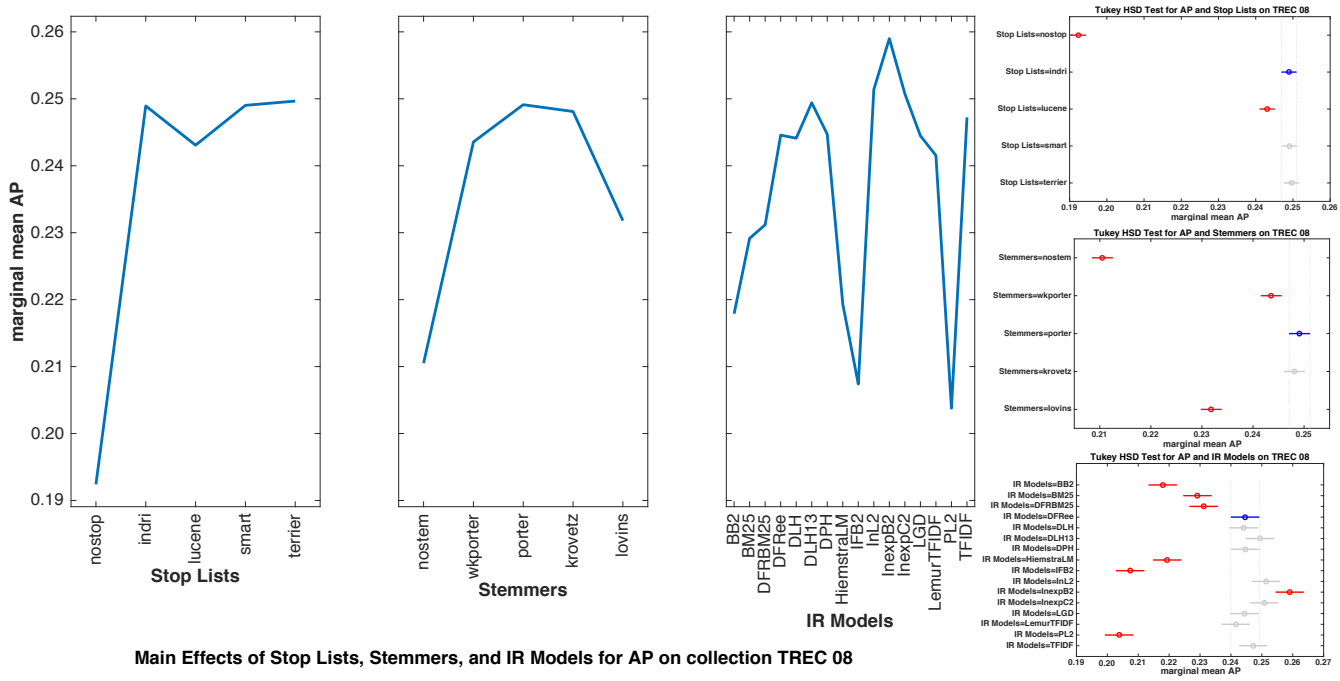

Figure 3: Main effects plots and Tukey HSD test plots for the stemmer group of runs on TREC 08 with AP.

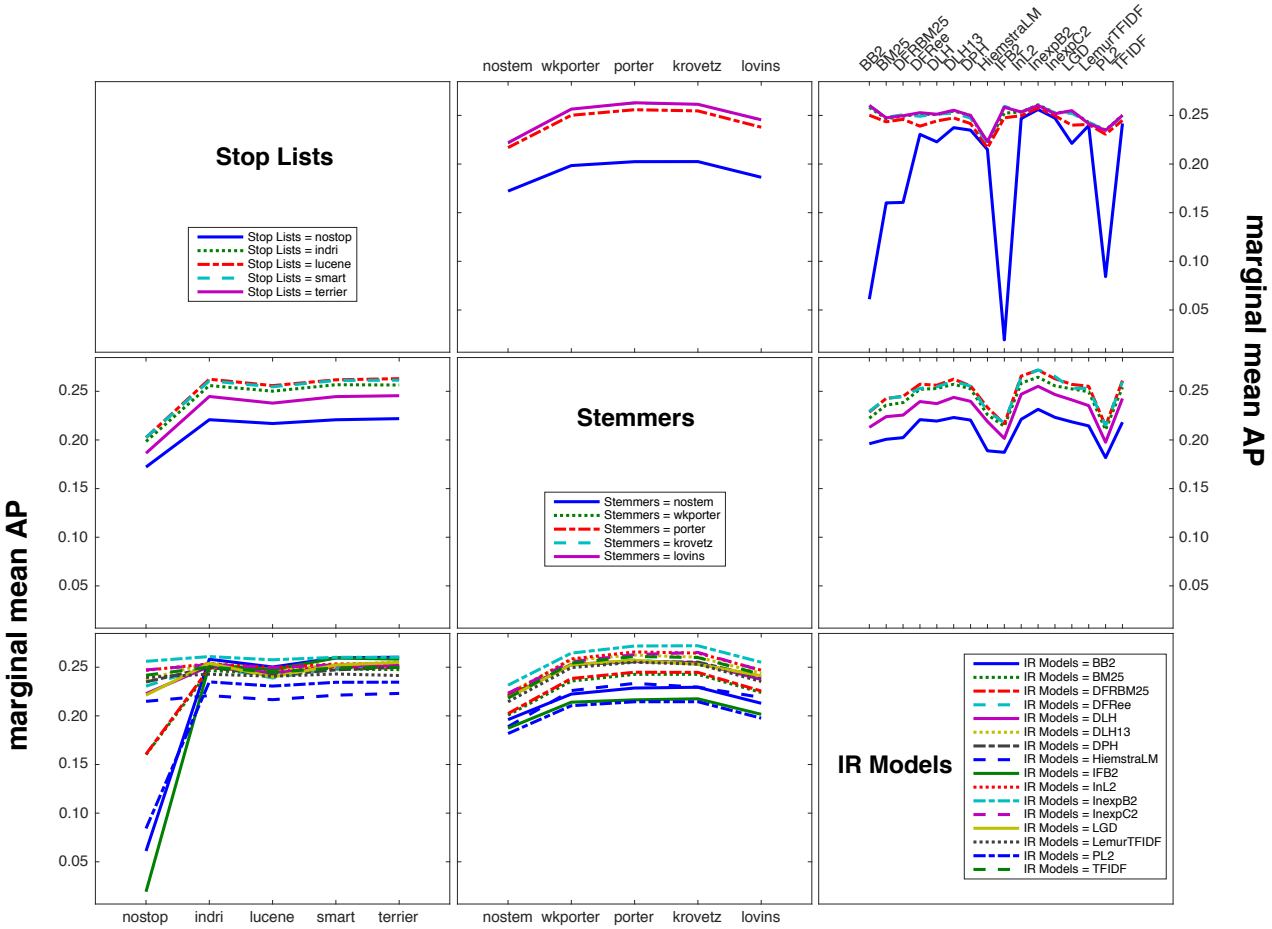

Figure 4: Interaction plots for the stemmer group of runs on TREC 08 with AP.

Table 2: Summary of single factor models on TREC collections. Each cell reports the $\hat{\omega}^2$ for the System effects and, within parentheses, the p-value for those effects. Large effect sizes ($\hat{\omega}^2_{\langle Systems \rangle} > 0.14$) are in bold.

Collection	LUG	Effects	AP	P@10	RBP	nDCG@20	ERR@20
TREC 05	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.1223 (0.00)	**0.2023** (0.00)	**0.1970** (0.00)	**0.1879** (0.00)	**0.1406** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.0794 (0.00)	0.1178 (0.00)	0.1349 (0.00)	0.1200 (0.00)	0.1063 (0.00)
TREC 06	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2108** (0.00)	**0.2458** (0.00)	**0.2716** (0.00)	**0.2742** (0.00)	**0.2377** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.1350 (0.00)	**0.1496** (0.00)	**0.1597** (0.00)	**0.1725** (0.00)	**0.1469** (0.00)
TREC 07	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2155** (0.00)	**0.2568** (0.00)	**0.2894** (0.00)	**0.2977** (0.00)	**0.2445** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.1502** (0.00)	**0.1658** (0.00)	**0.1920** (0.00)	**0.1898** (0.00)	**0.1480** (0.00)
TREC 08	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2774** (0.00)	**0.2780** (0.00)	**0.3025** (0.00)	**0.3118** (0.00)	**0.2484** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.1758** (0.00)	**0.1907** (0.00)	**0.2006** (0.00)	**0.2135** (0.00)	**0.1530** (0.00)

Table 3: Discriminative power of the evaluation measures on TREC 05, TREC 06, TREC 07 and TREC 08 for the stemmers and n-grams groups.

Group		TREC 05	TREC 06	TREC 07	TREC 08
stemmer	AP	3011	.2748	.3591	.4743
	P@10	.3774	.2687	.3222	.3171
	RBP	.3152	.2589	.3302	.3422
	nDCG@20	.3448	.2698	.3169	.3834
	ERR@20	.2014	.2235	.2096	.2388
n-grams	AP	.3180	.3553	.5184	.3498
	P@10	.3025	.2656	.3660	.2977
	RBP	.3852	.2539	.4193	.2797
	nDCG@20	.3260	.3130	.4292	.2938
	ERR@20	.2832	.1978	.2549	.2416

Table 4: Three factor, ANOVA table for TREC 08 (stemmer group) using AP.

Source	SS	DF	MS	F	p
Topics'	820.99	49	16.75	3713.90	0.00
Stop list	9.89	4	2.47	548.06	0.00
Stemmer	4.16	4	1.04	230.76	0.00
Model	5.16	15	0.3443	76.32	0.00
Stop list*Stemmer	0.05	16	0.03	0.67	0.83
Stop list*Model	17.01	60	0.28	62.84	0.00
Stemmer*Model	0.07	60	0.001	0.26	1.00
Stop list*Stemmer*Model	0.09	240	0.00	0.08	1.00
Error	88.20	19551	0.005		
Total	945.63	19999			

the lines connecting the different stop lists have much lower slope. In particular, we see that the choice of the stop list does not make a big difference with respect to use or not use a stop list; this can be further explored looking at the Tukey HSD test plot on the upper-right corner of the figure (in blue the selected component; in grey the components in the same group, i.e. not significantly different; in red the components in a different group, i.e. significantly different), where we can see that there are no significant differences between the "indri", "smart" and "terrier" stop lists, whereas the "lucene" stop list (which is composed by 15 words) is significantly different from the other three.

The main effect of the stemmer is always significant even though its size is quite small; nevertheless, the central plot of Figure 3 shows that there is a tangible difference between systems using or not using a stemmer. This can be seen also from the Tukey HSD test plot on the right; in particular, we can observe that there is no significant difference between the Porter and the Krovetz stemmer which are the stemmers with the highest impact on variance followed by the weak Porter and the Lovins ones.

Lastly, the plot on the right of Figure 3 reports the main effects of the IR models: they behave differently, as shown by several lines with high slopes, but the corresponding Tukey HSD shows that a many models are not significantly different one from the other. This can explain why the IR models effects are statistically significant but their effect sizes are not large.

For all the collections, consistently across the measures and both for the stemmer and the n-grams group, the higher effect size is reported by the *stop list*model* interaction effect which is always of medium or large size. This effect shows us that the variance of the systems is explained for the bigger part by the stop list and the model components. For the stemmer group of TREC 08, this can be seen in the plots on the upper-right and lower-left corners of Figure 4 where the lines of the interaction between the stop lists and the models intersect quite often. Indeed, the interaction plots show how the relationship between one factor and a response depends on the value of the second factor. These plots display means for the levels of one factor on the x-axis and a separate line for each level of another factor; if two lines (or segments) are parallel then no interaction occurs, if the lines are not parallel then an interaction occurs and the more nonparallel the lines are, the greater the strength of the interaction.

The *stop list*stemmer* interaction effects are always not significant as we can see from the p-values of Table 5 and the interaction plots in the upper-left part of Figure 4 where the line segments are parallel. A very similar trend can be observed for the *stemmer*model* interaction effect.

It is interesting to note that the second order interactions for the n-grams group are all statistically significant and that, in particular, we can see that n-grams, differently than the stemmers, have a bigger effect on the stop list than on the IR model.

We observe that different measures see the stop lists in a comparable way in terms of effect size and this is consistent with what we have seen in the one factor analysis. This is valid also for the stemmer, with the exception of ERR@20 for which it has an almost negligible effect size even though it is statistically significant. In Table 5 we can see that AP and ERR@20 weight the effects in a similar way as it happened in the single factor analysis reported in Table 2. For the n-grams group all the measures are comparable and ERR@20 is not as low as it happens for the stemmers.

Lastly, we can see that the third order interaction are never significant.

5. CONCLUSIONS AND FUTURE WORK

In this paper we faced the issue of how system variance contributes to the overall performances and how to break it down into some of the main components constituting an IR system. To this end, we developed an analysis methodology consisting of two elements: a *Grid of Points (GoP)* created on standard experimental collections, where all the combinations of system components under examination are considered; and, a GLMM model to decompose the contribution of these components to the overall system variance, paired with some graphical tools for easily assessing the main and interaction effects.

Collection	LUG	Effects	AP	P@10	RBP	nDCG@20	ERR@20
TREC 05	Stemmers	$\omega^2_{\langle \text{Stop Lists} \rangle}$	0.0432 (0.00)	**0.0632** (0.00)	**0.0638** (0.00)	**0.0605** (0.00)	0.0476 (0.00)
		$\omega^2_{\langle \text{Stemmers} \rangle}$	0.0178 (0.00)	0.0217 (0.00)	0.0116 (0.00)	0.0188 (0.00)	0.0000 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0219 (0.00)	0.0458 (0.00)	0.0452 (0.00)	0.0409 (0.00)	0.0311 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0005 (0.98)	-0.00 (0.46)	-0.0004 (0.97)	-0.0004 (0.94)	-0.0005 (0.99)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.0632** (0.00)	**0.1118** (0.00)	**0.1145** (0.00)	**0.1047** (0.00)	0.0826 (0.00)
		$\omega^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0019 (1.00)	-0.00 (0.49)	-0.00 (0.48)	-0.0008 (0.95)	0.0009 (0.05)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0115 (1.00)	-0.0099 (1.00)	-0.0109 (1.00)	-0.0107 (1.00)	-0.0102 (1.00)
	n-grams	$\omega^2_{\langle \text{Stop Lists} \rangle}$	0.0165 (0.00)	0.0272 (0.00)	0.0288 (0.00)	0.0256 (0.00)	0.0225 (0.00)
		$\omega^2_{\langle n\text{-grams} \rangle}$	0.0170 (0.00)	0.0105 (0.00)	0.0211 (0.00)	0.0288 (0.00)	0.0188 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0208 (0.00)	0.0341 (0.00)	0.0391 (0.00)	0.0275 (0.00)	0.0308 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.0016 (0.00)	0.0015 (0.00)	0.0020 (0.00)	0.0019 (0.00)	0.0015 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0296 (0.00)	0.0544 (0.00)	0.0571 (0.00)	0.0483 (0.00)	0.0424 (0.00)
		$\omega^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0050 (0.00)	0.0047 (0.00)	0.0049 (0.00)	0.0050 (0.00)	0.0040 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0063 (1.00)	-0.0040 (0.99)	-0.0034 (0.99)	-0.0056 (1.00)	-0.0048 (1.00)
TREC 06	Stemmers	$\omega^2_{\langle \text{Stop Lists} \rangle}$	**0.0750** (0.00)	**0.0852** (0.00)	**0.0904** (0.00)	**0.0932** (0.00)	**0.0673** (0.00)
		$\omega^2_{\langle \text{Stemmers} \rangle}$	0.0112 (0.00)	0.0082 (0.00)	0.0068 (0.00)	0.0126 (0.00)	0.0015 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0557 (0.00)	0.0596 (0.00)	**0.0692** (0.00)	**0.0696** (0.00)	**0.0638** (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0007 (1.00)	-0.0007 (0.99)	-0.0007 (0.99)	-0.0004 (0.94)	-0.0001 (0.64)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.1153** (0.00)	**0.1483** (0.00)	**0.1709** (0.00)	**0.1671** (0.00)	**0.1539** (0.00)
		$\omega^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0020 (1.00)	-0.0016 (0.99)	-0.0017 (1.00)	-0.0017 (1.00)	-0.0013 (0.99)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0119 (1.00)	-0.0109 (1.00)	-0.0116 (1.00)	-0.0112 (1.00)	-0.0107 (1.00)
	n-grams	$\omega^2_{\langle \text{Stop Lists} \rangle}$	0.0241 (0.00)	0.0282 (0.00)	0.0305 (0.00)	0.0306 (0.00)	0.0296 (0.00)
		$\omega^2_{\langle n\text{-grams} \rangle}$	0.0340 (0.00)	0.0144 (0.00)	0.0126 (0.00)	0.0249 (0.00)	0.0104 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0404 (0.00)	0.0516 (0.00)	0.0563 (0.00)	0.0545 (0.00)	0.0494 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.0026 (0.00)	0.0034 (0.00)	0.0036 (0.00)	0.0033 (0.00)	0.0032 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0465 (0.00)	**0.0628** (0.00)	**0.0673** (0.00)	**0.0746** (0.00)	**0.0646** (0.00)
		$\omega^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0058 (0.00)	0.0091 (0.00)	0.0111 (0.00)	0.0093 (0.00)	0.0080 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0033 (0.99)	-0.0019 (0.94)	-0.0008 (0.72)	0.0004 (0.36)	-0.0010 (0.78)
TREC 07	Stemmers	$\omega^2_{\langle \text{Stop Lists} \rangle}$	**0.0747** (0.00)	**0.0830** (0.00)	**0.0997** (0.00)	**0.1023** (0.00)	**0.0802** (0.00)
		$\omega^2_{\langle \text{Stemmers} \rangle}$	0.0227 (0.00)	0.0157 (0.00)	0.0163 (0.00)	0.0146 (0.00)	0.0056 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0441 (0.00)	0.0525 (0.00)	**0.0601** (0.00)	**0.0653** (0.00)	0.0513 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	0.0001 (0.36)	0.0009 (0.00)	0.0004 (0.09)	0.0004 (0.08)	0.0002 (0.21)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.1209** (0.00)	**0.1624** (0.00)	**0.1856** (0.00)	**0.1919** (0.00)	**0.1571** (0.00)
		$\omega^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0018 (1.00)	-0.0009 (0.95)	-0.0014 (0.99)	-0.0018 (1.00)	0.0007 (0.12)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0113 (1.00)	-0.0103 (1.00)	-0.0111 (1.00)	-0.0110 (1.00)	-0.0107 (1.00)
	n-grams	$\omega^2_{\langle \text{Stop Lists} \rangle}$	0.0237 (0.00)	0.0344 (0.00)	0.0395 (0.00)	0.0362 (0.00)	0.0290 (0.00)
		$\omega^2_{\langle n\text{-grams} \rangle}$	0.0208 (0.00)	0.0059 (0.00)	0.0132 (0.00)	0.0154 (0.00)	0.0112 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0563 (0.00)	0.0552 (0.00)	**0.0623** (0.00)	**0.0663** (0.00)	0.0382 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.00 (0.0001)	0.0014 (0.00)	0.0023 (0.00)	0.0025 (0.00)	0.0017 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0517 (0.00)	**0.0818** (0.00)	**0.0958** (0.00)	**0.0874** (0.00)	**0.0793** (0.00)
		$\omega^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0200 (0.00)	0.0126 (0.00)	0.0116 (0.00)	0.0145 (0.00)	0.0082 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0055 (1.00)	-0.0044 (1.00)	-0.0031 (0.99)	-0.0030 (0.99)	-0.0034 (0.99)
TREC 08	Stemmers	$\omega^2_{\langle \text{Stop Lists} \rangle}$	**0.0986** (0.00)	**0.0913** (0.00)	**0.1000** (0.00)	**0.1006** (0.00)	**0.0799** (0.00)
		$\omega^2_{\langle \text{Stemmers} \rangle}$	0.0439 (0.00)	0.0165 (0.00)	0.0190 (0.00)	0.0268 (0.00)	0.0071 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0535 (0.00)	**0.0615** (0.00)	**0.0666** (0.00)	**0.0707** (0.00)	0.0521 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0003 (0.83)	-0.0005 (0.98)	-0.0005 (0.98)	-0.0006 (0.99)	-0.0004 (0.95)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.1565** (0.00)	**0.1765** (0.00)	**0.1969** (0.00)	**0.2006** (0.00)	**0.1622** (0.00)
		$\omega^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0022 (1.00)	-0.0014 (0.99)	-0.0020 (1.00)	-0.0018 (1.00)	-0.0016 (0.99)
		$\omega^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0111 (1.00)	-0.0105 (1.00)	-0.0110 (1.00)	-0.0110 (1.00)	-0.0102 (1.00)
	n-grams	$\omega^2_{\langle \text{Stop Lists} \rangle}$	0.0396 (0.00)	0.0423 (0.00)	0.0445 (0.00)	0.0479 (0.00)	0.0304 (0.00)
		$\omega^2_{\langle n\text{-grams} \rangle}$	0.0037 (0.00)	0.0031 (0.00)	0.0008 (0.00)	0.0023 (0.00)	0.0093 (0.00)
		$\omega^2_{\langle \text{IR Models} \rangle}$	0.0550 (0.00)	0.0545 (0.00)	0.0548 (0.00)	**0.0637** (0.00)	0.0307 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.0035 (0.00)	0.0023 (0.00)	0.0024 (0.00)	0.0029 (0.00)	0.0032 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.0928** (0.00)	**0.1129** (0.00)	**0.1231** (0.00)	**0.1277** (0.00)	**0.0940** (0.00)
		$\omega^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0080 (0.00)	0.0050 (0.00)	0.0059 (0.00)	0.0050 (0.00)	0.0040 (0.00)
		$\omega^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0038 (0.99)	-0.0040 (0.99)	-0.0032 (0.99)	-0.0034 (0.99)	-0.0028 (0.99)

Table 5: Summary of three factor models on the TREC Ad-hoc collections. Each cell reports the estimated ω^2 SoA for the specified effects and, within parentheses, the p-value for those effects. Medium and large effect sizes are in bold; not significant effects are highlighted.

We conducted a thorough experimentation on TREC collections and used different evaluation measures to show how the proposed approach works and to gain insights on the considered components, i.e. stop lists, stemmers and n-grams, and IR models.

We found that the most prominent effects are those of stop lists and IR models, as well as their interactions, while stemmers and n-grams play a smaller role. Moreover, we have seen that stemmers produce more variation on system performances than n-grams. Overall, this highlights importance of linguistic resources.

Finally, measures explain system and component effects differently one from the other and not all the measures seem to be suitable for all the cases as it happens for ERR@20 which almost does not detect the stemmer effect. These insights can be useful to understand where to invest effort and resources for improving components, since they give us an idea of the actual impact of a family of components on the overall performances.

As far as future work is concerned, we plan to extend the proposed methodology in order to be able to capture also interaction between topics/systems and topics/components. Indeed, to estimate interaction effects, more replicates would be needed for each (topic, system) pair, as [17] simulated, and they are not possible in the present settings, since running more than once the same system on the same topics produces exactly the same results.

Moreover, we plan to further investigate the impact of the measures on the determination of effect sizes. A possible approach could be to conduct a four-factor analysis, using measures as additional factor. However, even if the measure scores are normalized in the range $[0, 1]$, they do not mean the exactly the same thing, i.e. AP $= 0.20$ is not exactly the same as ERR $= 0.20$ because of their different user models. A possibility for smoothing these differences and make the scores more directly comparable could be to normalize them by the maximum value achieved on the dataset, thus reasoning in term of ratios.

Lastly, an open challenge is how to run this kind of analysis on the systems which participated to past TREC editions. A first obstacle is that often there is no precise description of all the components used in these systems and so their metadata should be enriched in the way we suggested in [5]. A second obstacle is that the GoP would be very sparse and many combinations would be missing; therefore, we would need to rely on unbalanced GLMM and, probably, to consider the components as random factors.

6. REFERENCES

[1] D. Banks, P. Over, and N.-F. Zhang. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1:7–34, May 1999.

[2] L. Boytsov, A. Belova, and P. Westfall. Deciding on an Adjustment for Multiplicity in IR Experiments. In *SIGIR 2013*, pp. 403–412, 2013.

[3] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, USA, 2010.

[4] B. A. Carterette. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM TOIS*, 30(1):4:1–4:34, 2012.

[5] E. Di Buccio, G. M. Di Nunzio, N. Ferro, D. K. Harman, M. Maistro, and G. Silvello. Unfolding Off-the-shelf IR Systems for Reproducibility. In *SIGIR RIGOR 2015*, 2015.

[6] N. Ferro, R. Berendsen, A. Hanbury, M. Lupu, V. Petras, M. de Rijke, and G. Silvello. PROMISE Retreat Report – Prospects and Opportunities for Information Access Evaluation. *SIGIR Forum*, 46(2):60–84, 2012.

[7] N. Ferro and D. Harman. CLEF 2009: Grid@CLEF Pilot Track Overview. In *CLEF 2009*, pp. 552–565. LNCS 6241, 2010.

[8] A. Hanbury and H. Müller. Automated Component-Level Evaluation: Present and Future. In *CLEF 2010*, pp. 124–135. LNCS 6360, 2010.

[9] D. A. Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR 1993*, pp. 329–338, 1993.

[10] G. K. Jayasinghe, W. Webber, M. Sanderson, L. S. Dharmasena, and J. S. Culpepper. Statistical comparisons of non-deterministic IR systems using two dimensional variance. *IPM*, 51(5):677–694, 2015.

[11] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In *ECIR 2016*, pp. 357–368. LNCS 9626, 2016.

[12] C. Macdonald, R. McCreadie, R. L. T. Santos, and I. Ounis. From Puppy to Maturity: Experiences in Developing Terrier. *OSIR at SIGIR*, pp. 60–63, 2012.

[13] S. Maxwell and H. D. Delaney. *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Lawrence Erlbaum Associates, 2nd ed, 2004.

[14] K. R. Murphy and B. Myors. *Statistical power analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests (2nd ed.)*. Lawrence Erlbaum, 2004.

[15] S. Olejnik and J. Algina. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8(4):434–447, 2003.

[16] S. E. Robertson. The methodology of information retrieval experiment. In *Information Retrieval Experiment*, pp. 9–31. Butterworths, 1981.

[17] S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In *SIGIR 2012*, pp. 891–900, 2012.

[18] A. Rutherford. *ANOVA and ANCOVA. A GLM Approach*. John Wiley & Sons, 2nd ed, 2011.

[19] T. Sakai. Evaluating Evaluation Metrics based on the Bootstrap. In *SIGIR 2006*, pp. 525–532, 2006.

[20] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

[21] J. Savoy. Statistical Inference in Retrieval Effectiveness Evaluation. *IPM*, 33(44):495–512, 1997.

[22] M. D. Smucker, J. Allan, and B. A. Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *CIKM 2007*, pp. 623–632, 2007.

[23] J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In *Overview of TREC-3*, pp. 385–398. NIST, SP 500-225, 1994.

[24] A. Trotman, C. L. A. Clarke, I. Ounis, J. S. Culpepper, M.-A. Cartright, and S. Geva. Open Source Information Retrieval: a Report on the SIGIR 2012 Workshop. *ACM SIGIR Forum*, 46(2):95–101, 2012.

[25] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust Ranking Models via Risk-Sensitive Optimization. In *SIGIR 2012*, pp. 761–770, 2012.

[26] W. Webber, A. Moffat, and J. Zobel. Score Standardization for Inter-Collection Comparison of Retrieval Systems. In *SIGIR 2008*, pp. 51–58, 2008.

[27] W. J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *J. Inf. Science*, 20(4):270–284, 1994.

[28] P. Zhang, S. Dawei, J. Wang, and Y. Hou. Bias–variance analysis in estimating true query model for information retrieval. *IPM*, 50(1):199–217, 2014.