Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review

Giuseppe Rizzo[1⋆], Federico Tomassetti[2], Antonio Vetrò[3], Luca Ardito[2], Marco Torchiano[2], Maurizio Morisio[2], Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France
[giuseppe.rizzo, raphael.troncy]@eurecom.fr,
[2] Politecnico di Torino, Turin, Italy
[federico.tomassetti, luca.ardito, marco.torchiano, maurizio.morisio]@polito.it
[3] Technische Universität München, Germany
vetro@in.tum.de

Abstract. A Systematic Literature Review (SLR) identifies, evaluates and synthesizes the literature available for a given topic. This generally requires a significant human workload and has subjectivity bias that could affect the results of such a review. Automated document classification can be a valuable tool for recommending the selection of studies. In this paper, we propose an automated pre-selection approach based on text mining and semantic enrichment techniques. Each document is firstly processed by a named entity extractor. The DBpedia URIs coming from the entity linking process are used as external sources of information. Our system collects the bag of words of those sources and it adds them to the initial document. A Multinomial Naive Bayes classifier discriminates whether the enriched document belongs to the positive example set or not. We used an existing manually performed SLR as benchmark dataset. We trained our system with different configurations of relevant documents and we tested the goodness of our approach with an empirical assessment. Results show a reduction of the manual workload of 18% that a human researcher has to spend, while holding a remarkable 95% of recall, important condition for the nature itself of SLRs. We measure the effect of the enrichment process to the precision of the classifier and we observed a gain up to 5%.

1 Introduction

A Systematic Literature Review (SLR) is a research methodology used *to identify, analyze and interpret all available evidences related to a specific research question in a way that is unbiased and (to a degree) repeatable* (Kitchenham

⋆ Corresponding author.

2007). A SLR has to be performed according to a pre-defined protocol describing how primary studies[4] are selected and categorized, reducing as much as possible subjectivity bias. Depending on the research field where it is applied, the protocol changes. In this paper, we focus on a SLR applied to the field of Software Engineering, where the protocol can be summarized by the following steps (Kitchenham 2004): (i) identification of research, (ii) selection of primary studies, (iii) study quality assessment, (iv) data extraction and monitoring progress, (v) data synthesis. The first step defines the search space, i.e. the set of documents in which researchers select papers. A small sample set of relevant documents is used to define the search space. The second step identifies and analyses all possible useful studies among the papers which are contained in the search space that can help to answer some research questions. In the third step, an assessment about the quality of the studies collected is performed, while in the fourth step, the data extraction forms are delivered according to the review under evaluation. The last step delivers the data synthesis methods. Although these steps seem to be sequential, it is worth considering them as iterative steps and, therefore, the outputs may evolve according to the evolving topics.

The entire process is supervised and guided by researchers who summarize all existing information about some phenomena in a thorough and, potentially, unbiased manner. The final goal is to draw more general conclusions about some phenomena derived from individual studies, or as a prelude to further research activities. A SLR has a crucial importance in all research fields but it is extremely time-consuming, requiring an important human workload which is costly and error prone. Even though full automation of SLR is not possible due to the need of human reasoning for the aggregation and interpretation of scientific results, we believe that a tool support in the selection of the primary studies can reduce the human workload necessary in that phase, without loosing knowledge (which is a particularly important condition for the nature itself of SLRs).

Therefore, the objective of this paper is to reduce the human workload in a SLR, semi-automating the selection of primary studies (i.e. the second step of the SLR process). This depends on the dimensions of the search space. The larger the search space is the more effective our proposed approach will be. Our method focuses on a filter strategy resorting to semantic enrichment and text mining techniques to reduce the number of papers that researchers, who perform a SLR, should read. We use a text classifier to filter potentially interesting documents within the search space. The classifier produces a reduced set which contains a higher percentage of interesting document than the initial set. Afterwards, this reduced set is manually examined by researchers. In this way, we reduce the workload required to all researchers, limiting the human error rate. This phenomenon usually occurs when a set is sparse and searching through it requires more efforts than in a clean set, where the noise is smaller.

[4] A primary study is *(in the context of evidence) an empirical study investigating a specific research question (Kitchenham 2007).*

RQ1 Does the automatic selection process based on the Multinomial Naive Bayes classifier and semantic enrichment (enriched process) reduce the amount of manual work of a SLR with respect to the original process?

RQ2 Does the automatic selection process based on Multinomial Naive Bayes classifier and semantic enrichment (enriched process) reduce the amount of manual work of the alternative version of the process with only Multinomial Naive Bayes classifier (non-enriched process)? In other words, we aim to validate the idea behind the use of enriched papers as test samples instead of using original papers as test samples.

The approach presented in this paper is based on a previous work (Tomassetti et al. 2011). The following improvements are proposed: while previously the automatic classification was planned to fully automate the entire selection process step, in this paper, we propose a semi-supervised approach. This is because papers selected by the automatic classifiers could be immediately discarded by a human researcher just looking at the title and the abstract and do not need necessarily to be fully read. In addition, we perform an evaluation on a much larger dataset, extending the benchmark dataset size from the previous 111 papers to the current 2215 papers (almost 20 times larger). Finally, we present an exhaustive task-based evaluation.

The remainder of this paper is organized as follows. Section 2 compares our approach with the state of the art in the SLR domain. Section 3 details the steps of selecting primary studies and Section 4 presents our approach to improve this step. Section 5 describes the use case we use to validate our approach. In Section 6, we report and discuss the results we obtained. Finally, we give our conclusions and outline future work in Section 7.

2 Related Work

The automatic text classification applied to a systematic review is more challenging than the typical classification task. This is basically due to the dynamic nature of a SLR which is a supervised and iterative process where the initial scope of the SLR often evolves during the review process. Numerous research efforts have been spent to reduce the human workload when a SLR is performed. We focus on two different types of studies: i) machine learning based, and ii) ontology based.

Cohen *et al.* proposed a first attempt to reduce the human workload in the SLR field (Cohen et al. 2006). They used automatic classification to discard non-interesting papers from a set of them in fifteen different medical systematic literature reviews, each one considering the validity of a particular drug. Their classification model uses a reduced set of the features gathered from the paper such as author name, journal name, journal references, abstract, introduction, and conclusion. The classification model is built using negative examples as well as positive examples, where negative examples are selected from the pool of papers which do not adhere to the chosen SLR. Finally, this model is used to

create a perceptron modified vector for each feature in the feature set. Negative examples bias the model. In order to limit this phenomenon, they introduced a perceptron learning adjustment just evaluating the false negatives and false positives, monitoring them according to the False Negative Linear Rate (FNLR). A test article is classified by taking the scalar product of the document feature vector with the perceptron vector and comparing the output values. Considering a recall of 95%, the reduction of workload ranges from 0% to 68% according to the SLR they took under evaluation. Similarly to Cohen *et al.*'s work, in our approach we evaluate the reduction of human workload, while holding a 95% of recall for the classifier. The experiment we conduct is inspired to this, but we differentiate in terms of feature selection and the classifier used. For the former, we use a bag of words model enriched with further descriptions available in an external knowledge base, and we used a Multinomial Naive Bayes classifier. The human workload and the precision we achieve are in order of magnitude comparable with the ones observed by Cohen *et al.* (above the average) on fifteen medical literature reviews. However, due to the difference of the SLR domains (medical for Cohen *et al.*, Software Engineering in this paper), we cannot exhaustively compare the two approaches. Among the findings, Cohen *et al.* suggested that the automatic classification may be useful to regularly monitor new relevant journal issues in order to identify interesting primary studies, easing the task to keep a SLR constantly updated. According to this result, it is crucial to consider the classification problem in the SLR field as a semi-supervised approach in which a human being supervises the inclusion or exclusion of possible relevant studies selected by the classifier.

Another attempt to reduce the human workload in selecting relevant primary studies was performed by (Matwin et al. 2010). They proposed an approach mainly based on the Naive Bayes classifier with some optimizations which are based on the Complement Naive Bayes (CNB) (Rennie et al. 2003). The results they achieved outperform what detailed in (Cohen et al. 2006), but using a different configuration parameters (they consider only title and abstract for each document instead of the large set of features considered by Cohen). Leveraging on Natural Language Processing techniques (NLP), Cohen *et al.* tackle the problem of paper handling once the review starts (Cohen 2008). This is practically done to allow the reviewer to first analyze the documents which are labelled as potentially relevant documents, leaving at the end the evaluation for the remaining ones. They combined the approach of unigram and Medical Subject Headings (MeSH) to create the histogram of documents which potentially fits the scope of the review.

In (Ruttenberg et al. 2009), the authors proposed a hybrid approach for automating scientific literature search by means of data aggregation and text mining algorithms to make easy the search process. The key point of their work was to find a way to represent and share knowledge learned by human beings reading relevant papers, by means of an ontology. Through it, it was possible to combine outcomes of each single document and to represent it into a graph, which is mapped to the ontology. The first step of this process consists of identifying

the key phrases of the document (outcomes). Then, key phrases are used to link different concepts in the graph. Following this process, concepts are linked together, obtaining a chain of relationships. This work is usually made by human beings, who are experts of the domain. Ideally, they shoud be objective but the authors assessed that the graph mapping is strongly affected by the expert subjectivity. Then, they proposed a mechanism based on text mining algorithms to be able to navigate and cluster inferences. This work represents the first attempt to introduce the concept of knowledge representation in a SLR and, among the findings, they stated that a pre-clustering and linking of documents limit the human subjectivity improving the overall result.

3 Selection of primary studies

In this section, we detail the selection step of the SLR process analyzing its strengths and weaknesses according to the guidelines described in (Kitchenham 2004). This step takes as input the set of primary studies W gathered from a collection assumed to be the universe of all scientific papers in the domain of interest of the review. W results from the first step of the process and it is obtained as the output of the search process performed by human beings using keywords on dedicated sources. For instance, W could be composed by all papers published by a given set of journals or by all papers that a digital library provided as result of the search with keywords. The selection of primary studies is divided in two sub-steps: the former operates a selection based on reading titles and abstracts (*first selection*), the latter is the decision based on the full text human analysis (*second selection*). Both steps are basically affected by the following choice criteria: does it fit the research field? We define C (*candidate studies*) the set of studies that successfully passed the first selection and are eligible to be processed by researchers in the second selection step. It has the goal to split C in I (*included studies*) and E (*excluded studies*) where those sets are:

- I is the set of studies $\in C$ which successfully passed the second manual selection and will contribute to the systematic review. The following relation holds: $I \subseteq C$.
- E is the set of studies $\in C$ which did not pass the second manual selection and will *not* contribute to the systematic review and synthesis. Hence, $E \subseteq C$ and $E \cap I = \oslash$.

Figure 1 illustrates the selection of primary studies step. As introduced in the previous section, the selection of primary studies is performed by human beings who usually apply selection criteria . However, the application of those criteria could rarely be completely objective, and it is frequently instead affected by the subjective opinions of the involved researchers. A semi-supervised approach aims to reduce this potential bias.

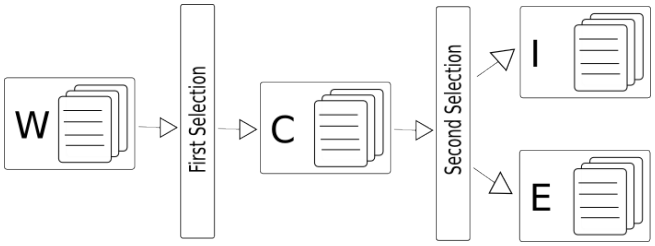

Fig. 1. Selection of primary studies in a Systematic Literature Review

4 Approach

The proposed approach relies on text mining techniques and semantic enrichment to reduce the set of interesting papers a researcher has to evaluate. The approach consists of a semi-supervised iterative process built on top of the following assumption: $W \neq \oslash$ (as a result of the applied search strategy) and $I \neq \oslash$ at the beginning (the set of relevant documents already known addedis not emply when the systematic review starts. The output of this approach is the set of most interesting papers W' gathered from a larger set of unread papers W.

4.1 I_0 construction

The initial set of sources contained in I is named I_0 and it is composed of primary studies already classified as relevant for the review: this is the first step of our process and it is needed to start the iterative part of the algorithm. I_0 can be built in two different ways. The first way is to ask researchers to use their previous knowledge indicating the most well known and fundamental papers in the field of interest. This strategy considers that, often, systematic reviews are undertaken by experts in the field. The second way is to explore a portion of the search space using the basic process, e.g. searching on digital libraries or selecting the issues of (a) given journal(s). This portion is marked as I_0 and the enriched process is used to explore the remaining search space.

4.2 Model building

The second step of our approach consists in computing automatically a model M from I_0. The idea is to build a bag of words (BoW) model starting from the primary studies in I_0. For each study, we considered the words from the abstract and introduction. According to (Cohen et al. 2006) words which appear at the beginning and at the end of a document (such as title, abstract, introduction and conclusion) are more significant. We empirically assessed that using a reduced set of words, coming only from abstract and introduction, provides the same results of considering the extended set of words (i.e. set of words coming from the title,

abstract, introduction and conclusion). The explanation is that the semantic enrichment stage (cfr. Section 4.3) compensates a reduced cardinality of the BoW through linking external sources and gathering from them textual data. Finally, we perform stop words elimination and stemming process, using the Porter algorithm (Porter 1980). The model built is used to train a Multinomial Naive Bayes classifier which computes the weight for each word according to the TF-IDF normalized approach (Kibriya et al. 2005).

4.3 Semantic enrichment

We define w_i a document composed by the BoW collected from the abstract and the introduction of one paper $w_i \in W$. Each w_i is processed to get a bag of named entities N which features w_i. A named entity is a name of a person or an organization, a location, a brand, a product, a numeric expression including time, date, money and percent found in a sentence (Grishman & Sundheim 1996). Basically, it is an information unit described by a set of classes (e.g. person, location, organization) which may be further disambiguated by an entry in a knowledge base such as DBpedia or Freebase. In this work we disambiguate entities to DBpedia (Bizer et al. 2009), with the rationale of linking them to external knowledge base entries. We then will fetch the abstract description of those entries and we join the existing textual content with the retrieved textual data. The encyclopedic nature of this dataset is appropriate to enrich the content of each w_i. Once we have extracted the bag of named entities N, we link each $n_i \in N$ to the corresponding DBpedia resource (when it is available). The extraction of named entities is performed using OpenCalais[5]. OpenCalais provides a classification for each named entity and suggests a URI of an external source where the information is disambiguated. Relying on it, we point to a DBpedia resource defined by the *owl:sameAs* property. Since not all the instances in the OpenCalais knowledge base have the owl:sameAs property, to minimize the loss, we used a logic that looks up entries in DBpedia that match the labels of the extracted entities (e.g. an occurrence of Systematic Literature Review is mapped to `http://dbpedia.org/resource/Systematic_review`). Once the resource is found, then we collect all words contained in the description field (`dbpedia-owl:abstract` property). The abstract property is one of the descriptive property , whose usage is consistent across the entire DBpedia dataset. After collecting these descriptions, we add them to the bag of words natively taken by the document w_i. We call it the enrichment process and the resulting document is defined as $w+_i$, and with BoW+ we refer to the bag of words extracted from $w+_i$. Finally, it is compared with the trained model M using a Naive Bayes classifier which is described below.

4.4 Classification

We used a Multinomial Naive Bayes (MNB) classifier and we implement the TF-IDF weight normalization. The choice of the Multinomial Naive Bayes clas-

[5] `http://www.opencalais.com`

sifier was based on two criteria: **(1)** the characteristics of the specific data and classification problem, and **(2)** the focus of the approach:

1. A first characteristic in this use case is the small training set, which is a peculiarity of the problem under the study (i.e. the common situation is that the initial set of available papers is not large at the beginning of a literature search).
   Usually, specific configuration of the classification algorithm parameters can improve the performances of a classifier (Forman & Cohen 2004). However, this is not a task that we expect from a normal user, given that we address a very transversely and general problem. Instead Naive Bayes models are more robust towards shift in training distribution (Elkan 2001). Another characteristic is the data heterogeneity because every word is interpreted as feature, thus leading to the well known problems of sparsity (which produces the so-called curse of dimensionality). Common text classifiers such a Support Vector Machines (SVMs), which are more often used for text classification purposes (Murphy 2012), particularly suffer leading to consequent overfitting issues (Cawley & Talbot 2010). In such fuzzy contexts, Naive Bayes (NB) approaches corrected with TF-IDF are competitive (Rennie et al. 2003). We then opt for the MNB setting since it is proven to lead the best results compared with other NB variants for such a context (Kibriya et al. 2005). Finally, SLRs produce highly imbalanced datasets.
   As a matter of fact, in our case study only 50 articles over 2215 are interesting (cfr. Section 5.1). Typical solutions to this type of problem are resampling techniques or hybrid algorithms (Chawla et al. 2004, Chawla 2005). While the first type of solutions is not applicable to the case of systematic literature reviews, the second one has the risk of a too specific implementation, which is not in the focus of our study.
2. The classification task in our case is subordinate to the enrichment process. For this reason our focus is to show that even with a very simple classifier, such as the MNB, the enrichment process is worthy: in fact, we show that using the BoW+ produces better results than using the original BoW in terms of saved manual work (from 15% to 18% reduction), preserving the recall beyond 95%, which is a very high value for all type of classifications.

We use the classifier to compare $w+_i$ with the model M and we determine whether the conditional probability that $w+_i$ belongs to I is significant or not. This allows to still preserve the context of the initial documents where the entities are extracted, hence favoring the classifier to decide also according to the entire bag of words instead of the extracted named entities. We assume that all papers which do not belong to I, belong to E adopting the Boolean algebra. The comparison is done for each $w+_i \in W$: papers with $P[w+_i \in I] \geq threshold$ are moved to W' and they are manually analyzed by researchers. Finally, all the papers whose $P[w+_i \in I] < threshold$ remain in W.

4.5 Iteration

The papers with a $P[w+_i \in I] \geq threshold$ are moved to W' to be manually processed, whilst the remaining ones still remain in W. It is likely that some of the papers moved in W' will pass the manual selection and will go to I, while the others will go to E. When I is modified, M becomes obsolete and it is necessary to re-build the model and repeat the classification step for all papers $w+_i \in W$. Again, if $P[w+_i \in I] \geq threshold$, $w+_i$ is moved to W' to be manually analyzed. If any $w+_i$ goes to W', i.e. $W' = \oslash$ after a classification, the iteration stops. Papers that remain in W after the last iteration are finally discarded and not considered by researchers. The exclusion of these papers represents the reduction in workload for the human researchers. At each iteration, the model will be progressively tailored to the domain of interest, allowing to refine the selection of primary studies.

Algorithm 1 Enriched selection process algorithm

```
Define I_0
Init I with I_0
repeat
   /* automatic recommendation of primary studies */
   Train classifier with I
   Extract model M
   for all w_i in W do
      Enrich w_i obtaining w+_i
      Compare w+_i with model M:
      if P[w+_i in I] ≥ threshold then
         move w_i to W'
      end if
   end for
   /* first selection */
   for all w'_i ∈ W' do
      Manually read title and abstract (w'_i ∈ I ) ? move w'_i to C : discard w'_i
   end for
   /* second selection */
   for all c_i ∈ C do
      Manually read full paper (c_i ∈ I ) ? move c_i to I : move c_i to E
   end for
until C ≠ ⊘
Discard ∀ w_i ∈ W
```

We provide in Algorithm 1 the synopsis of the whole study selection process proposed in this paper and in Figure 2 its complementary graphical representation. Comparing this picture with Figure 1 which represents the selection process provided by the guidelines (Kitchenham 2004), we observe that the original process is not changed, but we have added a selection of primary studies that recommends papers similar to the model at each iteration. We also reported in Figure 2 the steps of the new process described in subsections 4.1 to 4.4: the use of a model of bag of words (b) derived from I_0 or I (a), the enrichment of papers through semantic enrichment (c) and the comparison of the model M with the studies through a Multinomial Naive Bayes classifier (d).

5 Experimental Settings

The proposed approach has been implemented in the Semantic Systematic Review tool which is publicly available at `https://github.com/ftomassetti/semreview`.[6] The tool allows the loading of an already performed SLR from which are already known both the set of interesting papers and the set of non-interesting ones. This enables experiments to be run to assess the effectiveness of our approach. The tool creates the initially set of relevant papers I_0 (papers which belong to the I set) randomly selecting a sub-set of the interesting papers defined by the SLR. Doing that, the tool simulates the operation performed by human researchers at the beginning of the SLR. The other interesting papers, together with the non-interesting ones, end in the W. This set is used for assessing the performance of the approach. From I_0, the tool extracts the corresponding BoW and initializes the model M. Then, for all the papers in W, the tool automatically performs the recommendation of the primary studies (the second step in the SLR process) implementing the approach described in Section 4. Finally, the tool reports the performance of the approach using as ground truth the SLR taken as reference. The performance is measured as the amount of the saved manual work. The baseline in the experiment is given by the semi-supervised automatic approach without the semantic enrichment mechanism.

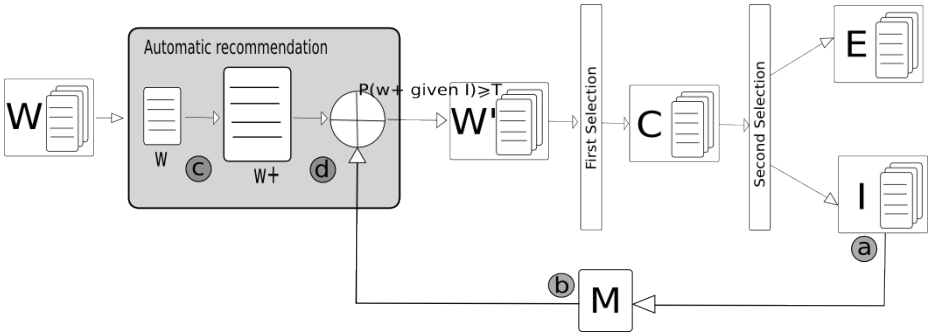

Fig. 2. The enriched study selection process and its principal steps: model extraction (b) after I is built (a), enrichment of papers through semantic enrichment (c) and comparison with the model through a Multinomial Naive Bayes classifier (d).

5.1 Benchmark dataset

As a case study we selected a SLR on Software Cost Estimation done by (Jorgensen & Shepperd 2007) and we limit the ground truth to all the papers mentioned

[6] The version released is a research prototype. It does not include some of the additional scripts used to run the experiments.

in the SLR coming from the IEEE Transactions on Software Engineering (IEEE TSE) journal. They cover a timeframe ranging from 1977 to April 2004. We had to exclude the first volume of IEEE TSE because it is not accessible from the IEEEXplore portal[7]. The resulting set contains 2215 candidates, all of them evaluated from the SRL taken as reference. The original SLR contains 51 interesting papers. However, only 50 of them are actually present in the set of the candidates available from the IEEEXplore, the missing one having been published in the first volume of IEEE TSE. Our benchmark dataset is therefore composed of 2215 papers, 50 of which belong to the I set. The others are considered as non-interesting papers, i.e. they do not pass the selection criteria defined at the beginning of the performed study and they belong to the E set.

5.2 Variable selection

The main outcome under measurement is the manual work, consisting of reading primary studies either entirely or only title and abstract, to select the interesting ones for the subject of the SLR. We measure the manual work as the number of papers that are read assuming the number as a proxy for the actual time that would be spent reading the articles. The minimum manual work ideally required is the total number of interesting papers. However, this minimum could reasonably never be reached in SLR. Indeed, the relation $I \subset W$ holds, where I is the set of relevant papers and W is the set of containing papers defined by the search criterion. This choice is motivated by the fact that the SLR, selected as subject of the case study, does not report neither the time spent for papers selection nor which papers were read entirely and which partially (only title and abstract). As a consequence, we define the following two metrics:

mw is the manual work. More specifically mw_O is the manual work performed in the original SLR, i.e. manually selecting and reading all papers, mw_{NE} is the manual work obtained applying the selection based on the Multinomial Naive Bayes classifier using original papers (non-enriched process), mw_E is the manual work obtained applying the selection based on the Multinomial Naive Bayes classifier using enriched papers (enriched process).

t is the applied task. Three levels are possible: manual, non-enriched, enriched.

5.3 Hypothesis formulation

The last step of the design is the hypothesis formulation. We formulate a pair of null and alternative hypothesis for each of the two research questions. Goal of the experiment is to reject the null hypothesis H_0 monitoring the `p-value` (Hubbard & Lindsay 2008). In other words, we discard the null hypothesis and we validate the alternative one H_A if the probability to reject the H_0 is lower than the 0.001. Moreover, it tells that when choosing the alternative hypothesis H_A, the probability to commit an error is lower than 0.001.

[7] `http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=32`

1. $H1_0 : mw_O \leq mw_E$, recall= 0.95
   $H1_A : mw_O > mw_E$, recall= 0.95

2. $H2_0 : mw_{NE} \leq mw_E$, recall=0.95
   $H2_A : mw_{NE} > mw_E$, recall=0.95

5.4 Parameter configuration

We decided to assess the validity of our process with different sizes of I_0 ranging between 1 and 5. In order to limit the bias introduced by a particular configuration of selected papers, we built 30 different I_0 sets per each dimension choosing them randomly among 50 relevant papers. We used each generated I_0 to kick-off the two variants of the process: enriched and non-enriched. Moreover, we replicated the experiment varying the classification threshold between 0 and 1 with steps of 0.01. The classifier threshold represents the posterior probability for a sample to belong to I (interesting set). Overall, we executed the complete algorithm 30,300 times = 5 (number of I_0 sizes) x 30 (number of I_0 sets for each size) x 2 (variants of the algorithm) x 101 (thresholds).

A preliminary step consisted to define the best classifier threshold T which maximizes the recall for the two variants. According to (Cohen et al. 2006), we decided to aim at a recall of 95%. Although this recall value is a strong constraint, we adopted it for limiting as much as possible the elimination of interesting papers. In Table 1, we report the distribution of the maximum classifier threshold which permits to obtain the target recall using the different I_0 sets. We chose the maximum threshold because is the one which minimizes the workload while it still satisfies the requirement of a recall equal to or greater than 95%. We select the median values to set the classifier, that means 0.22 for the enriched process and 0.17 for the non-enriched one.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
non-enriched	0.11700	0.1700	0.1700	0.1729	0.1775	0.1900
enriched	0.2100	0.2100	0.2200	0.2201	0.2200	0.2600

Table 1. Analysis of the best classifier threshold for both enriched and non-enriched process across different I_0 sets. The first and last column show the minimum and maximum values, second and fifth columns respectively the first and third quartile of the distribution, then mid columns show median and the mean of it.

5.5 Analysis methodology

The goal of data analysis is to apply proper statistical tests to reject the null hypotheses we formulated. Since the values are not normally distributed (according

to the Shapiro test), we adopt a non parametric test. In particular, we select the Mann-Whitney test (Hollander & Wolfe 1973) that compares the medians of the vectors of mw. To do that, we considered all papers extracted from the dataset except those papers used to build the I_0.

6 Results and Discussion

Figure 3 shows the comparison distributions for different settings of I_0 according to the two different types of recommendation approaches proposed: enriched process or non-enriched process. On the y-axis, the workload needed for a human being after both processes (enriched E and non-enriched NE) is reported. On the x-axis, we indicate the number of papers used for training the I_0 set and the process used (e.g. 1.E means an I_0 composed of 1 paper and the process has been performed using the enrichment mechanism). We observe a reduction of the workload in both approaches. Comparing the semantic enrichment with the baseline, we observe a greater reduction of the workload. This increment ranges from 2.5% to 5% for all I_0 settings, except for the I_0 composed of 1 paper (1.E in Figure 3) where the increment is lower then 1% with respect to the not-enriched (e.g. 1.NE in Figure 3).

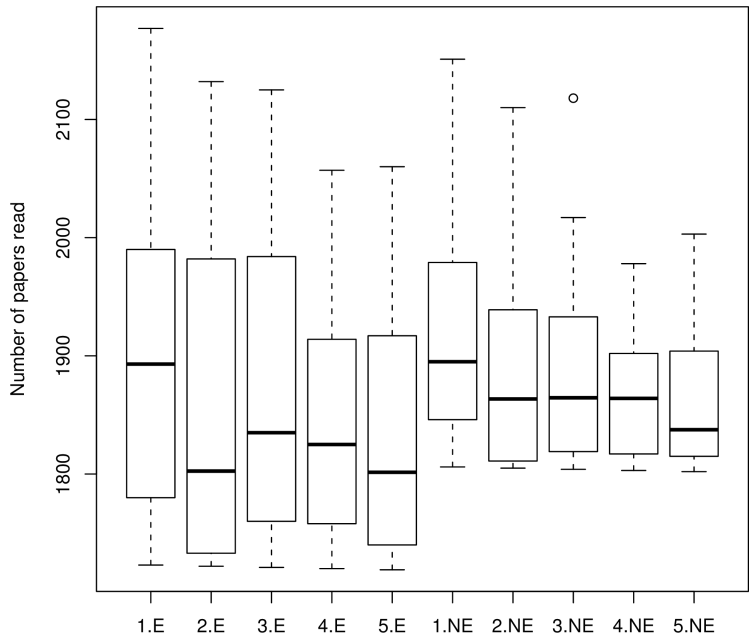

Fig. 3. Number of papers to read for different I_0 sizes and tasks applied: E (with enrichment) and NE (without).

We present below the results according to the two research questions addressed in this paper (see Section 1): evaluating whether the semantic automatic process classification reduce the amount of work of a SLR or not (RQ1) and evaluating if the semantic enrichment increases the performance of the simple classification process (RQ2).

6.1 RQ1: Reduction of the Human Workload

The results from the Mann-Whitney test are shown in Table 2. The table reports the I_0 size (column 1), the manual work in the original SLR process (column 2), the manual work obtained with our enriched process (column 3), the estimated percentage of manual work to be performed with our enriched approach with respect to the total work required using the common approach (column 4) and the `p-value` obtained from the Mann-Whitney test. The `p-value` for all the configurations indicates that the null hypothesis can be rejected and we assume the alternative which motivates the choice to use the semantic enrichment approach. In addition, we notice that the workload reduction increases as the size of I_0.

	Workload		Manual workload vs enriched workload	
$\|I_0\|$	mw_O	mw_E	median	$p - value$
1	2214	1897.567	**85%**	< 0.001
2	2213	1864.367	**84%**	< 0.001
3	2212	1863.833	**84%**	< 0.001
4	2211	1843.133	**83%**	< 0.001
5	2210	1829.1	**82%**	< 0.001

Table 2. For each I_0 configuration, we first compare the workload required to a human being in the original SLR and the workload mean if our process is performed. To verify the goodness of our process, we compute the Mann-Whitney test and we reject the hypothesis $mw_O \leq mw_E$ with a recall = 0.95.

6.2 RQ2: Assessing the Performance of the Enrichment Process

We used the Mann-Whitney test to reject the null hypothesis by which we state that $mw_{NE} \leq mw_E$. Table 3 reports the I_0 size (column 1), the estimated difference of manual workload between the two processes (column 2), and the `p-value` of Mann-Whitney test (column 3). While we can observe that the enriched process requires less workload for every size of I_0, we can affirm it with $p < 0.001$ just when the size of I_0 is 5.

$\|I_0\|$	workload median pairwise difference	$p - value$
1	26.67	0.0192
2	66.00	0.0073
3	40.83	0.0090
4	33.00	0.0083
5	**49.99**	0.0009

Table 3. For each I_0 configuration, we performed the Mann-Whitney test, evaluating median pairwise difference and `p-value` to estimate the minimum workload using both process: enriched and not-enriched. As for RQ1, the minimum recall is 0.95.

6.3 Discussion

The results show that our approach actually reduces the human workload to perform a SLR, while aiming to maintain a high level of completeness. Indeed, by limiting the recall to 95%, we adhere to the state of the art in the automation of SLR field maintaining its high quality. However, relying only on positive papers, this approach introduces one more configuration step for defining the threshold. The threshold can change according to the field of the SLR. In our test, we empirically observed that the probability threshold is almost consistent in different test scenarios. For this reason, we consider it as a baseline value for further investigations. In addition, we observed that the enriched process performs better than the variant without enrichment up to 5%. There are still two shortcomings: *i)* the extracted entities from OpenCalais sometimes point to resources in the OpenCalais knowledge base which do not contain sameAs links to DBpedia resources. We observe that the enrichment process fails in around 20% of the cases. The fallback strategy, to rely on another interlinking step using the named entity labels and lookup in DBpedia, partially fills the gap, since we observe that 19.9% of resources can be located, holding a loss of 0.1% of matched resources. However, this does not entirely fulfill the semantic gap since the interlinking step empowered as fallback does not consider the context from which the named entity has been extracted (raising an ambiguity issue which should be further analyzed with domain adaptive techniques). *ii)* a massive use of encyclopedic sources can bias the content of the enriched paper, penalizing words which do not appear often in the linked source but that are frequent in the initial document.

Differently from what we expected, the I_0 configuration does not affect the recall. Indeed, our results suggest that the number of papers in I_0 is not relevant. Its composition in terms of which papers are used to create it may play a more important role. For instance, let us consider an initialization of I_0 with papers that are not strictly related or if they represent just a niche of the research field, or if we select papers which are completely out of argument and they represent different meaning. While in the latter case, a wrong initialization affects all process and requires the initial set, in the former case the enrichment process enlarges I evading from the niche. Experiments show that the subjective bias in

the composition of I_0 is reduced when we use the semantic enrichment approach. While we do not have statistical evidence for that, I_0 size seems to play a role on workload reduction.

An important positive consequence of the use of automatic classification is the possibility to operate on larger search spaces because the effort of exploring W is reduced by means of partial automation. As consequence the search strategies can also explore potential interesting sources. For example, using the standard approach, search on a high number of journals and conferences is commonly quite expensive. Instead resorting on partially automatic classification, this search is more affordable. Moreover, using an external knowledge base we are able to capture not just papers we recognize being similar to the ones already selected, but we are able to capture papers that have conceptual relations (named entities) to the content expressed in the already selected papers. This strategy allows to deal with an incomplete description of the field of interest, which can not be completely described by the set of already selected papers. Therefore the proposed approach allows, as reported by the results, to use also a I set which is relative small and not representative of the whole field and to obtain results which outperform the classification process using only original sources. In addition, the experimental results show that these improvements are obtained with a still high recall (above 95%), which means loosing a negligible amount of relevant information, which is an essential condition for the nature itself of SLRs.

7 Conclusion and Future Work

In this paper, we presented a semantic enrichment recommendation of primary studies in a SLR. Resorting on text mining techniques and semantic enrichment, we improved the second step of the SLR process in order to filter the set of possible studies a researcher should read, automatically discarding the not relevant papers. Our approach has two main advantages: i) reduction of workload requested to classify sources and ii) reduction of subjectivity in the overall process. We tested our approach using a real SLR (Jorgensen & Shepperd 2007) which is used as benchmark dataset. Keeping a recall of 95% (i.e. we expected to discard papers only when the system is at least 95% sure that the paper is out the scope) we gained a percentage of workload saved of 18% when I_0 is composed of 5 papers. In addition, we demonstrated that the enrichment process outperforms up to 5% the automatic recommendation process without enrichment which is used as baseline.

As future work, we plan to improve the classification step, using besides positive examples also negative examples. We believe that using also negative examples the process may have a more accurate value of the plausible probability if a sample belongs to the interesting set. The first idea is to use some of the papers not included in the SLR for training negative examples. Although this may be intuitive, we may address the problem of a short distance from positives and negatives, due to the cross topics which these papers may report. A further

evaluation of the distance among papers from different journal issues may give a better idea about the use of negative examples. Therefore a deep analysis of which studies may be considered as negative is needed. In addition, we have planned to extract one paper i at a time from the set of relevant papers I, and to use the remaining papers $\in I$ to train the classifier and, then, to evaluate if it recognizes i as similar to the others. In this way, the classifier is used to give a "second opinion" on the selection process, potentially reducing the number of researchers necessary to undertake this step.

In the presented approach, we rely on the MNB classifier. It is considered as the baseline for text classification, but its results are often comparable to the state of the art in text classification, such as SVM and Markov chain (Rennie et al. 2003) and as shown in Section 4.4. We plan to validate the use of the semantic enrichment with other classifiers to investigate the changes in performance. The experiments addressed an important weakness in the named entity extraction task. The disambiguation mechanism provided by OpenCalais often links, via the sameAs link, to DBpedia resources. The loss of this process is recovered by an in-house interlinking logic which disambiguates the entity to DBpedia only considering the name of the entity.

Currently we are investigating the effect of NERD (Rizzo et al. 2014) which disambiguates to DBpedia considering the surroundings of the text where the entity has been spotted, hence preserving the semantics. Finally, the semantic enrichment mechanism has been validated using one SLRs. We plan to validate it also using other SLRs especially coming from other field of research. We believe that our approach could be adopted by scientific content providers such as journal portals, to index sources and to automatically classify and cluster the papers they publish. This approach may be used to propose a faceted view of sources queried by a user. The challenge will be to compute this operation in real-time to limit human efforts.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911).

References

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R & Hellmann S 2009 DBpedia - A crystallization point for the Web of Data Web Semantics: Science, Services and Agents on the World Wide Web **7**(3), 154–165.

Cawley G C & Talbot N L 2010 On over-fitting in model selection and subsequent selection bias in performance evaluation The Journal of Machine Learning Research **11**, 2079–2107.

Chawla N V 2005 Data Mining for Imbalanced Datasets: An Overview Data Mining and Knowledge Discovery Handbook pp. 853–867.

Chawla N V, Japkowicz N & Kotcz A 2004 Editorial: Special Issue on Learning from Imbalanced Data Sets ACM SIGKDD Explorations Newsletter **6**(1), 1–6.

Cohen A M 2008 Optimizing feature representation for automated systematic review work prioritization in 'Annual Symposium of the American Medical Informatics Association (AMIA)' pp. 121–125.

Cohen A M, Hersh W R, Peterson K & Yen P Y 2006 Reducing Workload in Systematic Review Preparation Using Automated Citation Classification Journal of the American Medical Informatics Association (JAMIA) **13**(2), 206–219.

Elkan C 2001 The Foundations of Cost-sensitive Learning in '17^{th} International Joint Conference on Artificial Intelligence' IJCAI'01.

Forman G & Cohen I 2004 Learning from Little: Comparison of Classifiers Given Little Training Knowledge Discovery in Databases: PKDD 2004 .

Grishman R & Sundheim B 1996 Message Understanding Conference-6: a brief history in '16^{th} International Conference on Computational linguistics (COLING'96)' pp. 466–471.

Hollander M & Wolfe D A 1973 Nonparametric Statistical Methods John Wiley and Sons New York.

Hubbard R & Lindsay R M 2008 Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing Theory & Psychology **18**(1), 69–88.

Jorgensen M & Shepperd M 2007 A Systematic Review of Software Development Cost Estimation Studies IEEE Transactions on Software Engineering **33**(1), 33–53.

Kibriya A, Frank E, Pfahringer B & Holmes G 2005 Multinomial Naive Bayes for Text Categorization Revisited in '17^{th} Australian joint conference on Advances in Artificial Intelligence (AI'05)'.

Kitchenham B 2004 Procedures for performing systematic reviews Technical Report TR/SE-0401 Software Engineering Group, Department of Computer Science, Keele University.

Kitchenham B 2007 Guidelines for performing systematic literature reviews in software engineering Technical Report EBSE-2007-01.

Matwin S, Kouznetsov A, Inkpen D, Frunza O & O'Blenis P 2010 A new algorithm for reducing the workload of experts in performing systematic reviews Journal of the American Medical Informatics Association (JAMIA) **17**(4), 446–453.

Murphy K P 2012 Machine Learning: a Probabilistic Perspective The MIT Press.

Porter M 1980 An algorithm for suffix stripping Program **14**(3), 130–137.
URL: *http://www.emeraldinsight.com/doi/abs/10.1108/eb046814*

Rennie J D M, Shih L, Teevan J & Karger D R 2003 Tackling the Poor Assumptions of Naive Bayes Text Classifiers in '20^{th} International Conference on Machine Learning (ICML'03)'.

Rizzo G, van Erp M & Troncy R 2014 Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web in '9^{th} edition of the Language Resources and Evaluation Conference (LREC'14)'.

Ruttenberg A, Rees J A, Samwald M & Marshall M S 2009 Life sciences on the Semantic Web: the Neurocommons and beyond Briefings in Bioinformatics **10**(2), 193–204.

Tomassetti F, Rizzo G, Vetro A, Ardito L, Torchiano M & Morisio M 2011 Linked Data Approach for Selection Process Automation in Systematic Reviews in 'Evaluation and Assessment in Software Engineering (EASE'11)'.