Semantic Search-by-Examples for Scientific Topic Corpus Expansion in Digital Libraries

Hussein T Al-Natsheh, Lucie Martinet, Fabrice Muhlenbach, Fabien Rico, Djamel Zighed

Semantic Search-by-Examples for Scientific Topic Corpus Expansion in Digital Libraries

Hussein T. Al-Natsheh*¶ Lucie Martinet†* Fabrice Muhlenbach‡ Fabien Rico§
Djamel A. Zighed*
* Université de Lyon, Lyon 2, ERIC EA 3083, 5 Avenue Pierre Mendès France – F69676 Bron Cedex – France
†CESI EXIA/LINEACT, 19 Avenue Guy de Collongue – F69130 Écully – France
‡Université de Lyon, UJM-Saint-Étienne, CNRS, Laboratoire Hubert Curien UMR 5516 – F42023 Saint Étienne – France
§Université de Lyon, Lyon 1, ERIC EA 3083, 5 Avenue Pierre Mendès France – F69676 Bron Cedex – France
¶CNRS, Institut des Sciences de l'Homme FRE 3768, 14 avenue Berthelot – F69363 Lyon Cedex 07 – France

***Abstract*—In this article we address the problem of expanding the set of papers that researchers encounter when conducting bibliographic research on their scientific work. Using classical search engines or recommender systems in digital libraries, some interesting and relevant articles could be missed if they do not contain the same search key-phrases that the researcher is aware of. We propose a novel model that is based on a supervised active learning over a semantic features transformation of all articles of a given digital library. Our model, named *Semantic Search-by-Examples* (SSbE), shows better evaluation results over a similar purpose existing method, *More-Like-This* query, based on the feedback annotation of two domain experts in our experimented use-case. We also introduce a new semantic relatedness evaluation measure to avoid the need of human feedback annotation after the active learning process. The results also show higher diversity and overlapping with related scientific topics which we think can better foster transdisciplinary research.**

I. Introduction

Scientists who work on a multi-disciplinary research topic need to explore related work that goes beyond keyword search matching. Depending on the discipline, special research topics are usually expressed using different terminologies. Relying only on term matching mostly does not work in such cases. The problem is that the scientist does not necessary know all variations of the terminologies used to express the research topic. Therefore, exploring relevant research articles from different disciplines than the scientist's discipline is considered as a challenging task. We denote such a task by "scientific topic corpus expansion." This corpus expansion is also needed to perform some statistical studies on that research topics, e.g., the estimation of the number of publications per year or the discovery of emerging terminologies.

In this work, we study the use of semantic representation of the article abstract for exploring semantically similar articles in a multi-disciplinary digital library. Given a set of articles annotated by the scientist as examples from the research topic of interest, the task is to generate the scientific topic corpus by expanding these examples to other articles that are relevant to the given examples but does not necessary contain the same topic naming terminologies. Instead of using special queries, i.e., *More-Like-This* query [1], [2], we transform all the article abstracts of the digital library into its semantic representation space. Then we look for articles having an abstract representation that is close to the provided examples in the semantic space. We call this approach "Semantic Search-by-Examples" (SSbE). With this approach, we wish to promote the fortunate discoveries obtained by accident. Digital libraries are a source of extraordinary knowledge about the world. It would then be a pity not to be able to exploit such a treasure due to disciplinary compartmentalization and scientific jargon. The objective of our proposal is to bring to the researcher the same serendipitous results as those obtained in a non-virtual library: sometimes, when we are looking for a book, we can accidentally find another one with an appealing title [3].

In order to experiment and verify the model, we worked on a use-case with two senior scientists of a given research topic. They provided us with some articles as examples of that topic and asked us to explore semantically related articles from different disciplines that might use different terminologies. In this use-case, we utilized an open access meta-data scientific digital library *ISTEX*[1] with millions of articles from different publishers and several disciplines.

We can summarize our contribution in the following:

- the introduction of a novel pipeline of a supervised active learning ranking model on semantic features
- the study of the importance of *active learning* in the model pipeline
- the use of sentence semantic similarity as a measure of document semantic similarity in order to aid the evaluation of the active learning process without the need for further domain expert annotation
- the analysis of topic diversity on the expanded scientific topic corpus.

Note that since this problem is rarely studied, possibly due to the lack of experimental dataset, we published our dataset and the code for repeatability and further comparative studies by other interested researchers.

In section II, we present a brief of techniques that are related to the study. We then describe *SSbE* model in section III. Our conducted experiment is presented afterwards in section IV followed by the evaluation results and its discussion in section V.

[1]Excellence Initiative of Scientific and Technical Information www.istex.fr

II. Related Work

According to the position expressed in [4], we consider our work to be in the general domain of text mining, in a view that allows us to unify the concepts "natural language processing" (for the use of linguists properties and their adaptation by computer specialists), machine learning and information retrieval (for the application as search engine technology) and, more generally, data mining (for the possibility of processing large volumes of data). In this section, a literature review of the general approach of text mining and its application in the recommendation of scientific articles extracted from digital libraries is given. Due to ever-easier access to large libraries of scientific articles, the work carried out in the joint field of data mining, information retrieval and natural language processing has given rise to numerous advances in recent years in the specific field of research-paper recommender system [5], which is the specific application domain of our contribution. We will also discuss in this section the related work on text representation, and especially text embedding, that allows to represent the documents in a form on which automated techniques of machine learning and data mining can be applied.

To facilitate the exploration of the articles in the scientific digital libraries, numerous works were carried out following several tracks. Most often they rely on topic modeling realized with latent Dirichlet allocation [6], this modeling is used to establish a similarity between the documents, this similarity is then used to link the documents together in different ways, such as a graph, and then allow a graphical exploration of this graph [7], [8], [9]. Some approaches focus on the human aspects of the document exploration interface [10], others tend to detect the evolution of scientific topics in the time [11], or try to promote serendipity [12].

Semantic space text representation is in the core of many natural language understanding research and application [13]. Recently, using new implementations of word embedding techniques trained on large text corpus showed a breakthrough in many computational linguistics benchmarks [14], [15], [16]. The main idea of the word embedding is based on a famous quotation of J. R. Firth in 1957: "You shall know a word by the company it keeps." The concept of word embedding has been extended to a sequence of words, i.e., sentences and paragraphs [13], [17], [18], [19], [20], [21]. However, such embedding techniques could be simplified and compared with classical matrix factorization techniques of text co-occurrence matrix, i.e., Latent Semantics Analysis (LSA) [22] and Singular Value Decomposition (SVD) [23] . At the end, what mainly makes difference between all such techniques is the hyper-parameter tuning and the performance of the implementation in addition to few tweaks to handle common issues like rare words.

The retrieval of semantically related documents in digital libraries is not a new problem and there are many approaches to overcome it. For example, in many digital libraries the meta-data of documents is enriched by set of tags like keywords or subject categories. However, these meta-data may lack a standardized taxonomy and also suffer from the coverage as it is usually a human annotated process [24]. More advanced information retrieval systems accept structured sentences as a query, e.g., question-answering systems [25], [26]. While using high number of documents (e.g., few hundreds) as a query is considered a new application, using one or very few examples as a query is not a new problem. JSTOR digital library recently introduced a system[2] that accepts one –and only one– document, applies topic modeling to extract search keywords and then uses them to retrieve related documents with topics faceted navigation. Other document-input query types are able to accept more than one document: *More-Like-This* query in *ElasicSearch* [1]. Starting from very few number of documents, this special query type retrieves other similar ones. Active learning is also used in an interactive information retrieval approach to enhance the relevancy of the recommended documents [27]. Our approach, SSbE, that also utilized active learning, has the ability to take much bigger number of documents as *initial corpus* and construct a model able to recognize semantically similar ones. Such initial corpus usually generated by using the topic key-phrases as search query.

III. SSbE Model

"Semantic Search-by-Examples", or *SSbE*, is the name we give to the method we propose to solve the problem of scientific domain expansion. In the following sections we will present each element in the pipeline of the model. We will show the model in two stages. The first stage would be denoted by $SSbE_p$, which is the **p**artial pipeline (without the active learning process). The second stage, denoted by *SSbE*, would be the completed pipeline (with the active learning process).

A. Model Overview

The purpose of this model is to expand a bibliography of a certain *focus scientific topic*. Such a topic is defined by a set of articles and possibly a topic label that we denote as a *topic key-phrase*, e.g., "human machine interface," "breast cancer," or "biological water treatment."

As it is presented on Figure 1, we define the input of the model from two main sources: the *scientific corpus* and the *seed articles* belonging to the topic. The scientific corpus should have big amount of articles from many disciplines. Those articles may be extracted from a scientific digital library. To be usable they must have a minimum set of meta-data like the title, the abstract, and a unique index to be retrievable. It is possible to benefit from other meta-data fields like the set of keywords, authors, references but they are not required for the model. In order to be able to evaluate the model, the content of the articles would be necessary, so that the expert annotator could provide their feedback. The set of *seed articles* consists in a few number of examples, preferably between 100 and 300, with the same requirements of the scientific corpus, i.e.,

[2]http://www.jstor.org/analyze/

a title, an abstract and a unique index. These seed articles are provided as a kind of query-of-examples in which the user aims to find semantically similar articles possibly from other disciplines. Practically, seed articles are articles belonging to the *scientific corpus*, or which can be added to it, and that are annotated as *focused topic*. This set of articles is retrieved by matching the *topic key-phrase* with the meta-data of the articles of the *scientific corpus*. We will denote this set of articles as *extended positive articles*.

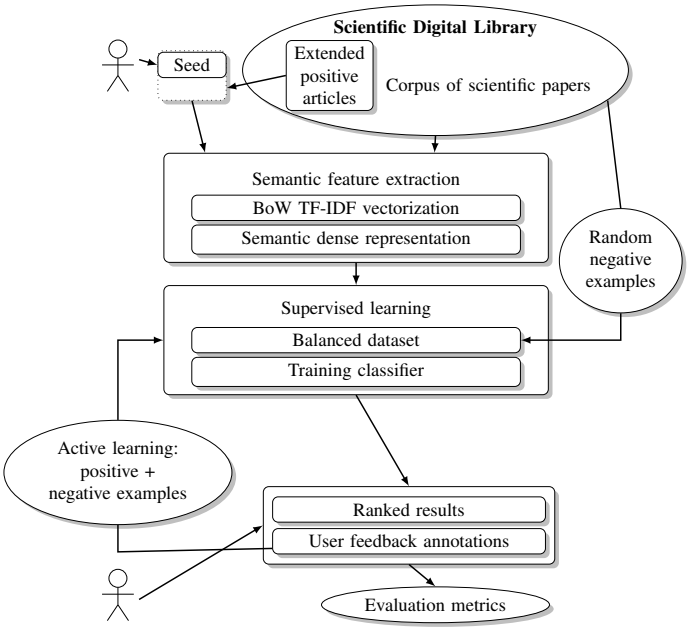

Fig. 1. The *SSbE* Model Pipeline. The input of the system is an initial corpus that consists in the seed articles and the extended positive examples which are search key-phrase matches to the focus scientific topic. After transforming all the articles into their semantic feature representations, a supervised learning classifier is trained on a balanced set of positive (initial corpus) and negative (randomly selected) article examples. The results are then ranked by the probability value that the trained binary classifier predicted each article in the digital library as the positive class. Finally the user provide his annotation on the top results which are used to regenerate a new training set with negative examples with the active learning process to enhance the results in which the top ranked results would be the output scientific topic expanded corpus

The SSbE model consists in a few high-level phases illustrated in Figure 1. The output is the ranked list of recommended articles that may extend the knowledge about the focus scientific domain by including semantically relevant articles from other disciplines. The first process in the model is to vectorize the whole corpus in addition to the seed articles using the bag-of-words (BoW) method. The second step is to transform the BoW into the vector semantics dense representation. Next, a balanced dataset will be generated from both positive examples, that are the *seed* and the *extended positive articles*, and negative examples, that are randomly selected from the *scientific corpus* other than the *matched key-phrase* articles. This dataset is then used to train a supervised binary classifier. The trained classifier is finally used to rank all the articles of the *scientific corpus* with the probability of belonging to the *focused topic*. A complementary enhancing step is the active learning process where the user feedback is used to regenerate the balanced training dataset.

B. Vectorization

Our system uses the common TF–IDF weighted BoW method to initially obtain a vectorized representation of the documents. The main drawback with BoW vectorization is that the information of the order of the words in the text is lost. Even there are few techniques to overcome this issue, i.e., n-gram, using BoW alone still lacks of encoding the semantic and syntactic information [28].

The system then extract the dense semantic representation from the weighted BoW. This could be done either by Latent Semantics Analysis (LSA) based decomposition or by a technique based on learning semantic features [28]. In order to find a good semantic representation space, we computed the average inner cosine similarity (*AICS*) as in Equation 1 for two lists of documents: the positive ones, and negatives ones. The negative list is constructed by randomly selecting the same number of documents from the corpus. The function we maximize searching for a good semantics space transformer is given by Equation 2.

$$AICS = \frac{\sum_{i=1}^{n-1} cosine\_similarity(list[i], list[i+1:n])}{number\_of\_comparisons} \tag{1}$$

$$argmax_{transformer}(AICS_{positive\_list} - AICS_{random\_list}) \tag{2}$$

The vector semantics transformation is constitutionally a long and expensive process, however it is luckily needed to be run only once. This is true not only for a certain *focused topic* use case but for any other *focused topic* that the users would like to apply later on if the *seed articles* are found in the same corpus.

C. Learning Process

1) Balanced training set generation: After transforming all the articles corpus (i.e., title + abstract) into its semantic vectorized representation, our method relies on building a classifier that would be able to predict if a given example is of the *focused topic* or not. To build such classifier, we built a balanced training set of both positive and negative examples. At the beginning, the negative examples are randomly sampled from the corpus excluding positive examples. This is of course based on the assumption that a uniformly randomly picked samples from such corpus would less likely be positive examples. In case the number of positive examples are small even after adding the *extended positive articles*, the system randomly duplicated some positive examples to match the experimented size of the dataset. At the active learning stage, this balanced dataset would be regenerated with better negative examples provided by the user feedback.

2) Supervised Learning: In order to generate the aimed results, our method uses a binary classifier trained on the generated balanced dataset in order to compute the prediction probability of each article in the *scientific corpus* to be of the *focused topic* class. A ranked list of all the corpus articles, sorted by that probability value as a score, is finally considered as the system output. This result excludes all the positive examples used in the learning process, as the aim is to find any unexpected relevant article with our semantic-based recommendation approach.

Choosing the type of the classifier is a design parameter of the model and could be decided experimentally. We recommend ensemble learning methods like gradient boosting or random forest because of the ability of such methods of providing the predicted probability value of a document to belong to the class. Otherwise, regression could be also used.

Unlike the vector semantics transformation phase, the supervised learning is a very fast and repeatable process which is practically very useful for the *active learning* process.

3) Active Learning: In this complementary but important process, the user feedback is used to regenerate the balanced training dataset. This aims to extend the negative examples with the related but marked-irrelevant results by the user. The positive examples will also be enriched by providing marked-relevant articles but from different disciplines. Accordingly, the classifier will continuously better learn how to semantically separate the articles than only using randomly sampled negative examples as in the first generated dataset.

In case of many users providing relevancy annotation to the results, the model compute the average score for each annotation. The numeric value used to indicate relevance is 1, 0 in the case of an irrelevant document, and 0.5 when experts can not decide.

4) Using sentence semantic relatedness for evaluation: In order to avoid asking the user to provide his feedback annotation, which is not an easy task, we introduced an automated comparative evaluation criterion of the results after applying the active learning process. This criterion is based on sentence semantic relatedness [29] between the titles of the seed articles and the titles of the results. We first use the Cartesian product composed of the titles of the seed and the results articles set to generate the set of titles pairs. Then, we use a pre-trained model that takes a set of sentences, i.e., title pairs as input and estimates the semantic relatedness score for each pair as an output. Our proposed evaluation method is then to count the title pairs that exceed a semantic relatedness score threshold.

IV. Experimentation

A. Use Case from Sports Science: Mental Rotation

For our experiments, we chose to focus on a field of research far from our own field (i.e., computer science) for which there were possibilities of transdisciplinary inputs because this field is already in close connection with related disciplines. This research discipline is *sports science*. This field interconnected with other scientific domains, e.g,, physiology, psychology, anatomy, biomechanics, biochemistry and biokinetics.

B. Data description

In this experiment we are interested in the domain of "mental rotation" which is a good case of study because it rises interest in different disciplines such as education, social sciences, psychology or medical science. The data we use in our SSbE model is based on the *meta-data* describing the articles, which composed of the DOI, the title, the authors, the key-words when they exist and the abstract of the documents.

The documents composing the *scientific corpus* set comes from ISTEX scientific digital library (SDL) whose aims is first to gather the publications of different publishers of the last decades, second to offer an interface to access this large amount of research documents, and third to develop some useful statistical and research functions in order to exploit the available documents.

Out of many document types, e.g., slides, posters and conference articles, we only considered English research papers that were published after 1990 with sufficient abstract size (35 to 500 words). The extracted meta-data dataset contains more than 4.17 millions articles.

1) Construction of the seed article: The number of the seed articles of our use case experiments was 182 articles. They are all annotated by the focus domain experts as the focused topic: mental rotation. In this seed article set, 29 tagged articles do not even contain the *topic key-phrase* in their meta-data. Only 25 documents tagged by the specialists are also part of *the scientific corpus*. For each article, we extracted the same meta-data than in the SDL: DOI, authors, title, abstract, keywords when exist, and source.

2) Expansion of the seed articles set: We increase the number of the positive examples by extracting from the SDL database the research articles containing the expression "mental rotation" in the meta-data. Thanks to this strategy, we extracted 199 additional documents out of the SDL and consider them as positive examples. We will denote this 199 additional articles as *extended positive articles*.

C. Model Experimental Design

Truncated Randomized SVD [30] and *Paragraph Vector* [28] are two examples of vector semantics transformation techniques we considered in our experiments. The choice of these two methods among others was based on the availability of a scalable implementation in addition to the recent claimed efficiency. We first run comparative experiments of the two transformers based on Equation 2. Unexpectedly, the Paragraph Vector transformer did not result in any good vector representation using our experimented corpus. This could be due to the size and the specialty of such text corpus. However, the SVD transformer showed good results. Accordingly, we focused on finding a good design parameters of the SVD transformer. The parameter values we found the best among several experiments are listed in Table I:

TABLE I
BEST PARAMETER SETTINGS FOUND BY OUR EXPERIMENTAL DESIGN

Parameter	Best value
Minimum term frequency	20
Maximum term frequency percentage to keep	0.95
N-gram range (constrained by the memory size)	range (1,2)
Filtering out stop-words or not	yes
Using lemmatization or not	no
Dense semantic space dimension	150
Using TF–IDF transformer or not	yes

The results of the cosine difference of Equation 2 on these parameters was 0.31 detailed as follows:

- Average cosine similarity within seed articles: 0.4;
- Average cosine similarity within articles randomly selected from corpus articles other than the positive ones (same set size of mental rotation ones): 0.09.

As a binary classifier, we used the ensemble learning method that is random-forest classifier. This choice was based on the performance of such type of machine learning in many applications reported recently in many publications. Another important feature of this method is the ability to provide the probability score of class prediction which is needed for our method. The design parameters were set as the defaults of the classifier implementation of scikit-learn machine learning library (python2 version 0.18.1). In order to decide on the best number of estimators to use and to validate the accuracy of the classifier, we used cross validation and 30:70 test-training dataset splitting. Using 500 estimators for that classifier, the prediction accuracy was higher than 0.95. This accuracy value was the average of several runs with different randomly sampled negative examples. The number of runs were 100 so that we can somehow neutralize our assumption of that the randomly selected samples from the scientific corpus are more likely to be negative examples. This assumption will be also handled in the active learning process as we will discuss later in this paper.

After obtaining our trained classifier, we apply it to all the documents, more than 4 millions, predicting the probability for each document to be classified as a mental-rotation article. We used this probability value as a score value in which we ranked all the documents in a descending order. The top few thousands documents can then be evaluated and thus considered as a potential expanded scientific corpus of the topic "mental rotation."

D. Active Learning

For the active learning process of the SSbE model, we generate a balanced dataset as follows:

- Negative examples that are composed of 2 sets:
  - the annotated results by the domain experts in which at least one of them marked it as irrelevant
  - In case the number of positive examples are higher than the annotated irrelevant articles, we randomly extract articles from the digital library corpus other than the positive ones in order to have a balanced dataset
- Positive examples that are composed of 3 sets:
  - seed articles (182)
  - extended positive articles (199)
  - the annotated results by the domain experts in which at least one of them marked it as relevant while the other could not decide

This new balanced dataset is then used to re-train the classifier we used in SSbE. We then use this newly trained classier to predict the probability of each article in the digital library corpus. Finally, we sort all the articles by the score value with descending order to form the new results of the model. The new results should not have any of the irrelevant-annotated articles in the top results. This would be verified in section V-B

E. Sentence Semantic Relatedness Measure

We extracted the titles the top 200 results for each of:

- More-Like-This method: MLT
- Partial SSbE model (without active learning): $SSbE_p$
- SSbE model with active learning: $SSbE$

We then generated 3 set of pairs from the seed titles and the titles of each method. The size of each set was $200 \times 182 = 36,400$ pairs. In order to estimate the semantic relatedness score of each pair for each of the 3 sets, we used a pre-trained model[3] [31] which provides an estimation score between 0.0 to 5.0. This model was trained on an open access datasets[4]. We finally counted the pairs with the semantic relatedness score above a threshold $t = 3.0$ in order to compare the results of the 3 methods.

F. Diversity Analysis

Exploring relevant articles from different disciplines, by definition, should lead to a higher diversity and related topic overlapping in the expanded scientific topic corpus. Accordingly, we need to identify and define measures that quantify the rate of diversity and relevancy in order to compare the results obtained by different methods. This task is not that simple due to the large possible number of parameters even with simple aggregation like counting or averaging. In our case, we proposed to base the statistics on the words appearing in the title of the article, the author affiliations, the journal names, the keywords, or even a compilation of the keywords appearing in one or more of these fields. We should keep in mind that any derived diversity measure must maintain the results relevancy, e.g., such diversity indicators should still be relevant to the studied scientific topic expansion. We can assume that we achieve this purpose if we extract such keywords from the relevant articles in the results.

Our proposed diversity measure compares the distribution of the vocabulary extracted from the titles, author affiliation, and other interesting elements of the articles. We can base these

[3]https://github.com/natsheh/sensim

[4]http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

analysis on all the top ranked articles of the two compared methods. Considering only relevant articles, according to the feedback annotation of the domain experts, could be risky in case the amount of relevant articles is not balanced between the two methods. To overcome this risk, we extract diversity indicator keywords from equivalent number of relevant articles of both methods. This means applying a kind of random sub-sampling from the method that has bigger set of results. So, we run the experiment a certain number of times and then we apply the Wilcoxon test [32] that counts the number of times that a method has a higher results than the other.

G. Repeatability

The developed code for all conducted experiments of this article is available as open-access in a github repository[5]. We think that this shared code would be useful for the repeatability and further comparative research. The dataset generation script is also included in the repository.

V. Results and Discussion

In this section we will show and discuss the evaluation results of our proposed model SSbE with and without the active learning process in comparison with another method that is the More-Like-This query method (MLT).

A. Model Result Evaluation without Active Learning

In order to generate comparative results to partial $SSbE_p$ model (without active learning), we passed to the MLT query the seed articles and the 199 articles that we found "mental rotation" in its meta-data, i.e., extended positive examples. Using the default parameters of this query in ElasticSearch resulted into low number of results. So, we looked into these parameters and tuned them according to the design parameters of our method in order to return a sufficient number of results for our evaluation experiment. The number of results we achieved was 391 articles.

The tuned MLT query parameters were:

- *max query terms* was set to 150 to reflect the vector size we have in our method;
- *min term freq.* was set to 20;
- *max doc freq.* was set to 0.95 $\times$ number of articles;
- and not providing a list of stop-words.

For a comparative evaluation, we took 100 articles from the top results of our SSbE method and another 100 articles from the top results of MLT method. The resulted 200 articles were then shuffled and blindly handed to two experts, same who provided the initial corpus, to manually annotate each article if it is relevant or not to the focused topic, i.e., "mental rotation." Keeping in mind that none of these articles have "mental rotation" in their meta-data, the experts needed to look carefully through the whole article content to give their annotation. Inexpert annotators would be inadequate as the task requires deep understanding of the research topic to decide whether an article from different discipline is relevant.

[5]https://github.com/ERICUdL/ISTEX_MentalRotation

TABLE II
Confusion matrix of the two domain expert judgment of both of $SSbE_p$ (SSbE without active learning) and MLT method on 100 results randomly picked from the top 200. S corresponds to $SSbE_p$ and M corresponds to MLT. CND indicates that the expert Can Not Decide

	relevant		CND		irrelevant		**Total**	
Method	S	M	S	M	S	M	S	M
relevant	8	2	3	3	0	0	11	5
cannot decide	10	1	10	4	17	4	37	9
irrelevant	2	0	13	5	37	81	52	86
Total	20	3	26	12	54	85	100	100

TABLE III
Cohen's kappa scores for annotation of the two domain experts. The table shows results for different combination of annotation labels. The scores are rounded to 4 decimals

Labels	**Cohen's kappa score**
[relevant, irrelevant]	0.9008
[relevant, cannot decide]	0.1751
[cannot decide, irrelevant]	0.2764
[relevant, irrelevant, cannot decide]	0.3797

Accessing to more 2 experts in such rare domain was not easy but we think it would be sufficient for a fair comparison. Given that the annotation efforts were big, we could barely reach our minimal target of 200 annotated articles. In addition to [relevant, irrelevant], we found a third case in which the domain expert find the recommended article related and useful being partially relevant. So, they cannot label it as relevant nor as irrelevant. After discussion with the experts we decide to include such a case that would be denoted as *cannot decide*.

A confusion matrix for each method were then computed in order to check the agreement level between both experts. These confusion matrices are shown in Table II.

+We also computed the Cohen's kappa coefficient [33] that measures the agreement between the two domain experts with their annotation labels. As we can see in Table III, the kappa score is very high for the labels [*relevant, irrelevant*]. It means that they were mostly agreed on the extreme judgment on the resulted articles. However, The two domain experts have less agreement when one of them use the label *cannot decide*.

To come up with a relevancy score for each article in the list of ranked results, we assign a numeric value for each expert annotation (i.e., 1.0 for relevant, 0.5 when experts can not decide, and 0.0 for irrelevant). The final score of each item is then the average of two expert scores. Thus, the possible score values we have for each result item are $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. The corresponding score results are listed in Table IV.

Afterwards, we computed the accuracy of both methods at the top *n*, such that *n* is the number of results, based on Equation 3. Iterating over the ranked results from 1 to the number of annotated articles, we could see in Figure 2 the accuracy curves of our SSbE method and MLT method.

TABLE IV

FREQUENCIES OF THE EVALUATION SCORES VALUES FOR BOTH THE $SSbE_p$ METHOD AND THE MLT METHOD. THE BLUE SCORE LABELS ARE GOOD WHILE THE RED SCORE LABELS ARE BAD

Score	1	0.75	0.5	0.25	0
$SSbE_p$	**8**	13	12	30	37
MLT	2	4	4	9	**81**

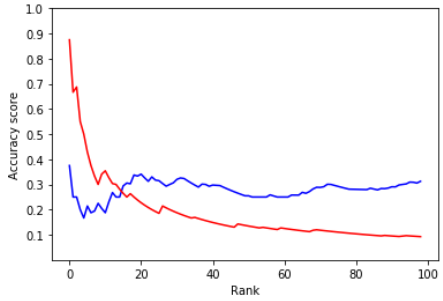

Fig. 2. Accuracy curves of $SSbE_p$ method in blue and MLT method in red. Considering the top 100 $SSbE_p$ scored 0.3125 while MLT scored 0.09. At the very top results, MLT has better score but with very few total number of relevant results.

$$\frac{\sum_{rank=1}^{top_n} score_{rank}}{top_n} \tag{3}$$

Another point of view for comparing the quality of the results obtained with the two methods using very simple measures: count of really irrelevant articles, i.e., scored 0, in the set of documents proposed to the reader by the classification method. We notice that our method obtain better results on a long ranked list of articles than the MLT method which is a little bit more efficient on the first results as we can see on figure 3.

B. Evaluation of the Model with Active Learning

The first results verification step we did was to make sure that the new results after applying the active learning process

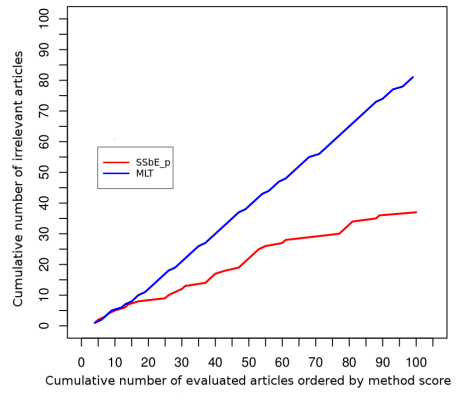

Fig. 3. Number of irrelevant documents proposed to be in the field asked by the user (here, "mental rotation"), that have a lower rank than the value in abscissa using MLT or $SSbE_p$ method.

TABLE V

COMPARATIVE RESULTS OF THE 3 METHODS USING SENTENCE SEMANTIC RELATEDNESS MEASURE BASED ON COUNT OF PAIRS WITH SCORE HIGHER THAN 3.0 OUT OF 5.0

Method	MLT	$SSbE_p$	$SSbE$
Count of pairs	124	217	382

TABLE VI

COMPARATIVE RESULTS OF THE 3 METHODS ON THE TOP 959 RESULTS OF EACH METHOD USING A TEST SET EXTRACTED FROM THE DIGITAL LIBRARY META-DATA THAT WAS HIDDEN FROM OUR EXPERIMENT. THE NUMBER OF 959 RESULTS WERE SELECTED AS A RESULT OF EXCLUDING EXTENDED POSITIVE ARTICLES, WHICH HAVE BEEN USED IN THE TRAINING PHASE, FROM THE TOP 1000 RESULTS OF $SSbE_p$. *:THE TOTAL NUMBER OF THE MLT RESULTS IS 391 ARTICLES

Method	MLT^*	$SSbE_p$	$SSbE$
matches count	1	1	6
rank of them	1	851	$24, 82, 227, 567, 699, 929$

do not contain any of the irrelevant-user-annotated results. We verified that the top results of the new ranked list of articles does not contain any. The second step is then to find away to evaluate the new results. For that, we will show two evaluation criteria: first, by using the sentence semantic relatedness measure on the article titles, and second, by using a test set generated from the meta-data annotation of the digital library corpus that was hidden from our experiment.

1) Evaluation using sentence semantic relatedness: As described in section IV-E, we want to evaluate 3 models: *MLT*, *SSbE$_p$* and *SSbE* by pairing the titles of top 200 results of each method with the titles of the seed articles. Using the introduced evaluation measure in section IV-E and a threshold semantic relatedness score value of 3.0, we obtained the results in Table V.

We can see from Table V that the results of the introduced measure correlate with the evaluation results of the two domain experts showing that the *SSbE$_p$* method is better than the *MLT* method. Using the same measure, we can observe that we achieved a higher evaluation value of *SSbE* than *SSbE$_p$* thanks to the active learning process.

2) Evaluation using a test-set: In this measure, we checked how many matches and at which ranks we can find a test set of articles. These test set articles were hidden from the experiment and were extracted from the meta-data of the digital library corpus. The query criteria we used to extract this test set was finding the phrase *"mental rotation"* in the list of *subjects* or *keywords* but not mentioned in the *abstract* nor the *title*. The results of this test set evaluation are shown in Table VI.

We can notice in Table VI that we have 6 matches for $SSbE$ comparing to only 1 match for the other two methods. Looking to the rank of these matches, we observe that MLT method was better than $SSbE$ method using this type of evaluation. However, by using the domain experts annotation as shown in section V-A, we could find much more relevant articles using $SSbE$. We can also see that 5 out of 6 ranks of matches for

TABLE VII
AMOUNT OF DISTINCT VOCABULARY OVER THE FIRST 200 ARTICLES RANKED BY THE THREE SYSTEMS, BASED ON CATEGORIES : MLT, $SSbE_p$, $SSbE$

System	MLT	$SSbE_p$	SSbE
	76	57	57

the $SSbE$ method were higher than the rank of the $SSbE_p$ method.

C. Results of Diversity Analysis

As introduced in section IV-F, we propose to observe the diversity of the documents, using indicators like journal names, departments of the authors, assigned topics or keywords. We may have a clue of the coherence of the results looking at the number of articles concerned by categories of subjects.

The initial idea was to simply observe the amount of vocabulary we can extract from titles, affiliations or scientific categories of the articles that is ranked in the top 200 results of each method, without taking into account the relevancy of this documents. We first considered single-word tokens extraction from the titles of the articles, author affiliations and journals. This single-word tokens strategy produced a set of vocabulary that are noised with a lot of irrelevant vocabulary. We also face the problem of very generic words that can be applied in a lot of domains, especially with the MLT system. On the contrary, working on key-phrases such as domain categories given in the meta-data of the articles seems to give very good and relevant results. Simply by counting the number of different phrases that we can find in the articles excluding the completely marked irrelevant ones produced the results summarized in Table VII.

The illustrated diversity analysis in Figure 4 presents the number of documents containing the vocabulary in abscissa for $SSbE_p$, compared to MLT system. We select only the vocabulary that appears enough time (20 at least) to be considered as relevant because well used in the domain. In addition of a slightly better diversity of the $SSbE_p$ method, shown in with more blue color than red, we also notice that the vocabulary of our method includes more related vocabulary of the target domain, i.e. mental rotation. Most of the MLT vocabulary we get is very generic and can be applied to various domains. For example, the three category names in the top of the figure that have more articles produced by MLT are actually irrelevant to the studied domain. On the other hand, the key-phrases extracted from the results of our systems are more precise such as “biological psychology” or “physiology” compared to “gerontology” or “psychiatry.” So, we could conclude that the amount of diversity is not enough to judge; We should additionally take into account the relevancy of this diversity.

D. Topic Modeling Analysis

To have a good overview of the organization of our data, and especially identify the sub-topics of the mental rotation domain we use two complementary strategies described in the following sub-sections. Both of these strategies were based on the topic modeling algorithm called Latent Dirichlet Allocation (LDA) ([6]) to retrieve the main topics. The input of the LDA model consists in a balanced set of positive (annotated as mental rotation) and negative (randomly selected) articles examples.

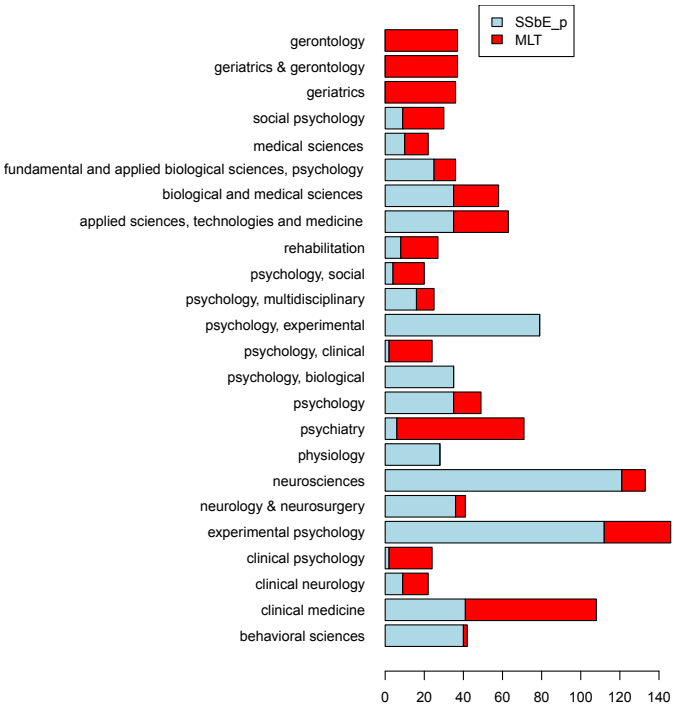

Fig. 4. Distribution of the vocabulary of each documents over the global vocabulary based on categories discovered in the 200 best ranked documents for each the MLT and $SSbE_p$ systems. We show here only the vocabulary appearing more than 20 times globally for the two methods.

1) Sub-Topics on Seed Articles: In order to extract a list of key-phrases that can describe the mental rotation topic as well as possible subtopics, we applied a topic modeling approach that is able to define and categorize a given set of documents into a number of topics. In our experiments we generated 400 positive examples out of the seed articles as well as the articles we found containing the key-phrase *mental rotation* in their meta-data. Another 400 negative examples uniformly randomly selected from the corpus, more than 4 million documents, compose the negative set. Since we do not know in advanced the number of topics that the topic modeling will best partition this 800 articles dataset, we iterated from 2 to 50 topics. We know that at least one topic must be a mental-rotation related topic as we know that half of the dataset consists of such documents. Indeed, the results showed that there is at least 1 topic in which its top featuring tokens are mental rotation related topics. We had this a-priori knowledge of such key-phrases from the experts who firstly provided the list of 182 seed documents. Using 8 topics as an input of LDA model, we found 2 topics with sufficient amount of articles related to mental rotation according to the features. One of these 2 topics is mainly described by the tokens list (mental

rotation, task, motor, stimuli, orientation...) while the other one is mainly described by the list of tokens (spatial ability, sex differences, patient...). These results are aligned with the initial analysis of the experts who provided the 182 documents about mental rotation. This analysis provides us with extracted key-phrases for the main topic "mental rotation," as well as the 2 sub-topics, i.e., mental rotation tasks and spatial ability studies on demographics differences.

2) Emerging Topics on Results: We address the same question of the topics representation in the results and top list of documents extracted by $SSbE_p$, first, to have a way to evaluate the quality of our results and to underline the diversity of the fields that are concerned by the scientific aspects risen by the inter-disciplinary "mental rotation" research domain. We track the emerging key-phrases related to mental rotation by analyzing the top 10K results of the $SSbE_p$ model with LDA topic modeling. We could find the following main additional cognitive science related concepts that seems to overlap with mental rotation (e.g., event related potentials, mismatch negativity, or attention deficit hyperactivity disorder).

E. Examples of some Surprising Articles

By identifying how accidental discovery processes occur, [34] resumed the words of Louis Pasteur who said "accidents favor the prepared mind," adding that "it is well known that attention is often attracted to phenomena that are familiar to the observer but that turn up in an unusual environment, or to new phenomena in a familiar environment, provided that the phenomena are relevant to the viewer's usual range of interests." We consider that our approach with *SSbE* model will favor such accidental discoveries by connecting scientific papers describing relevant similarities seen on a higher level than the topic targeted by a given discipline.

The articles found by $SSbE_p$ model and recommended to the researchers are not always considered to be relevant. However, since these proposed articles contain semantic similarities to those used as input (i.e., the initial corpus), the recommended papers share some topic connections with the input papers and open the research on new thematics. In our study, some recommended articles surprised the experts who evaluated these documents and gave them ideas for further research in new directions. For information purposes, the sports science experts came across an article which, without mentioning the mental rotation task, evoked a near theme concerning the studies on abilities to read a map in different orientations [35]. This discovery has led the sports scientists working on mental rotation to see extensions of their work to the field of *orienteering*, a sport that requires navigational skills using map and compass to run in an unfamiliar terrain.

Another example of transdisciplinary discovery made by the mental rotation experts is the following: through a similarity of activation of brain areas, they find that there are some connections between the mental rotation and the sign language [36]. Indeed, sign language and lip-reading used by deaf signers are actions that require some mental rotation abilities for reading the manual communication. Scientific bridges had not been made between such fields of study until now.

VI. Conclusion

We proposed a novel model to expand a given set of scientific article examples into a corpus of semantically relevant articles of the scientific topic. Beyond keyword matching, these explored articles might belong to variant disciplines that tend to use different terminologies. We call this model Semantic Search-by-Examples *SSbE*. We conducted an experiment of our proposed model over *ISTEX*, a big digital library corpus, on a use-case of a multi-disciplinary scientific domain, i.e., *Mental Rotation*. The experiment showed the superiority of our model against an existing method, i.e., More-Like-This query which exists in a widely used open-source search engine for digital libraries. The comparative evaluation was possible by having a feedback annotation of two domain experts. We also showed the applicability and the importance of active learning process in the model pipeline. Additionally in this paper, we introduced a new semantic relatedness evaluation measure to avoid the need of human annotators for result evaluation. The measure we introduced is based on a pre-trained sentence semantic relatedness estimator. We finally presented a further result analysis of the topic extraction and topic diversity. Our proposed approach produced more diversity on a set of related topic categories rather than the compared method. The code used in our experiment in addition to the script for downloading the dataset is made available for other researchers for repeatability and further comparative studies in this open research problem. This model could be applied not only for scientific corpus expansion but also for enriching the metadata of the digital library in off-line fashion. Once the articles are annotated with this semantically related scientific topic, it would be much easier for researchers to query such articles using any semantic variation of the topic key-phrase.

As a future work, we would like to study the usability of the sentence semantic relatedness measure inside the model pipeline to boost-up articles with high semantically related sentences. We also want to examine a topic modeling approach on the top ranked results trying to identify the clusters that are mostly relevant to the initial corpus. Finally, based on an enhanced semantic sentence relatedness model, we can also introduce a semantic sentence highlighter that will identify interesting part in the text of the recommended articles. This will make it easier for the user to provide her/his annotation to the system and thus to feed in the active learning process.

Acknowledgment

This work was kindly funded by ISTEX project. The use-case study was funded by the "Centre de Recherche et d'Innovation sur le Sport" who also technically contributed though the expert annotation process by Dr. Patrick Fargier and Prof. Raphaël Massarelli. We would like also to thank ARC6 of the French Region Auvergne-Rhône-Alpes that funds the current PhD studies of the first author.

REFERENCES

[1] B. Dixit, "Chapter 2. the improved Query DSL," in *Mastering Elasticsearch 5.x*. Birmingham, UK: Packt Publishing, Limited, 2017, pp. 74–141.

[2] M. Hagen and C. Glimm, "Supporting more-like-this information needs: Finding similar web content in different scenarios," in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, ser. Lecture Notes in Computer Science, E. Kanoulas, M. Lupu, P. D. Clough, M. Sanderson, M. M. Hall, A. Hanbury, and E. G. Toms, Eds., vol. 8685. Springer, 2014, pp. 50–61.

[3] N. Ramakrishnan and A. Grama, "Data mining: From serendipity to science - guest editors' introduction," *IEEE Computer*, vol. 32, no. 8, pp. 34–37, 1999.

[4] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining, Second Edition*, ser. Texts in Computer Science. Springer, 2015.

[5] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, November 2016.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[7] L. F. Klein, J. Eisenstein, and I. Sun, "Exploratory thematic analysis for digitized archival collections," *Digital Scholarship in the Humanities*, vol. 30, no. Suppl-1, pp. i130–i141, 2015.

[8] J. He, Y. Huang, C. Liu, J. Shen, Y. Jia, and X. Wang, "Text network exploration via heterogeneous web of topics," in *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain.*, C. Domeniconi, F. Gullo, F. Bonchi, J. Domingo-Ferrer, R. A. Baeza-Yates, Z. Zhou, and X. Wu, Eds. IEEE, 2016, pp. 99–106.

[9] T. M. V. Le and H. W. Lauw, "Semantic visualization with neighborhood graph regularization," *J. Artif. Intell. Res. (JAIR)*, vol. 55, pp. 1091–1133, 2016.

[10] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. U. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM TIST*, vol. 3, no. 2, pp. 23:1–23:26, 2012.

[11] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles, "Detecting topic evolution in scientific literature: how can citations help?" in *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, D. W. Cheung, I. Song, W. W. Chu, X. Hu, and J. J. Lin, Eds. ACM, 2009, pp. 957–966.

[12] E. C. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher, "Serendip: Topic model-driven visual exploration of text corpora," in *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014, Paris, France, October 25-31, 2014*, M. Chen, D. S. Ebert, and C. North, Eds. IEEE Computer Society, 2014, pp. 173–182.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013). Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[17] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, C. Chung, A. Z. Broder, K. Shim, and T. Suel, Eds. ACM, 2014, pp. 373–374.

[18] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 3294–3302.

[19] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *CoRR*, vol. abs/1703.02507, 2017.

[20] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1188–1196.

[21] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. P. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi, Eds. ACM, 2013, pp. 2333–2338.

[22] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[23] C. F. Van Loan, "Generalizing the singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 13, no. 1, pp. 76–83, 1976.

[24] A. Abrizah, A. N. Zainab, K. Kiran, and R. G. Raj, "Lis journals scientific impact and subject categorization: a comparison between web of science and scopus," *Scientometrics*, vol. 94, no. 2, pp. 721–740, Feb 2013.

[25] A. Severyn and A. Moschitti, "Modeling relational information in question-answer pairs with convolutional neural networks," *arXiv preprint arXiv:1604.01178*, 2016.

[26] K. Tymoshenko, D. Bonadiman, and A. Moschitti, "Convolutional neural networks vs. convolution kernels: Feature engineering for answer sentence reranking." in *HLT-NAACL*, 2016, pp. 1268–1278.

[27] L. Chen, R. Bao, Y. Li, K. Zhang, Y. An, and N. N. Van, "An interactive information-retrieval method based on active learning." *Journal of Engineering Science & Technology Review*, vol. 10, no. 3, 2017.

[28] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1188–1196.

[29] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe, "Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, S. Bethard, D. Cer, M. Carpuat, D. Jurgens, P. Nakov, and T. Zesch, Eds. San Diego, California: Association for Computational Linguistics, June 2016, pp. 497–511.

[30] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[31] H. T. Al-Natsheh, L. Martinet, F. Muhlenbach, and D. A. Zighed, "Udl at semeval-2017 task 1: Semantic textual similarity estimation of english sentence pairs using regression model over pairwise features," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, August 2017, pp. 115–119. [Online]. Available: http://www.aclweb.org/anthology/S17-2013

[32] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] P. W. Langley, H. A. Simon, G. Bradshaw, and J. M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Process*. Cambridge, MA: The MIT Press, 1987.

[35] M. Tlauka, "Orientation dependent mental representations following real-world navigation," *Scandinavian Journal of Psychology*, vol. 47, pp. 171–176, 2006.

[36] N. Sadato, T. Okada, M. Honda, K.-I. Matsuki, M. Yoshida, K.-I. Kashikura, W. Takei, T. Sato, T. Kochiyama, and Y. Yonekura, "Cross-modal integration and plastic changes revealed by lip movement, random-dot motion and sign languages in the hearing and deaf," *Cerebral Cortex*, vol. 15, no. 8, pp. 1113–1122, August 2005.