Textual Concept Similarity

Navid Rekabsaz**,Ralf Bierig**, Alexandru L. Ginsca*, Mihai Lupu**, Adrian Popescu*, Phong Vo*

*CEA, LIST, LVIC France

** Vienna University of Technology, ISIS, IMP

Contacts: lupu@ifs.tuwien.ac.at, adrian.popescu@cea.fr

MUCKE Project, Deliverable 4.1

01/12/2014

Contents

1 Introduction 4

2 From word to document concept similarity 4

2.1 Similarity Methods 5

2.2 Experiments 5

2.3 Results and Discussion 6

2.3.1 SemEval 2014 Task 10 Results 6

2.3.2 MediaEval Results 6

3 Knowledge base matching 8

3.1 Introduction 8

3.2 Explicit Semantic Analysis (ESA) 9

3.2.1 Classical Formulation of ESA 10

3.2.2 ESA Adaptation for ad-hoc multilingual IR (ESA-C) 11

3.3 Europeana Collection Processing 13

3.4 Enrichment and retrieval framework 13

3.4.1 Semantic Topic Enrichment Framework 13

3.4.2 Ad-hoc Retrieval Framework 14

3.5 Experiments 16

3.5.1 Semantic Enrichment 16

3.5.2 Ad-hoc retrieval 18

4 Semantic Indexing with Payloads 21

5 Conclusion 22

Abstract

Two approaches to concept similarity for text data are explored in this report, both in the context of multimodal image retrieval in the context of the project and related evaluation exercises. First, we show that random indexing and deep learning can provide added benefit to text-based retrieval of images from Flickr. Second, we show that Explicit Semantic Analysis, i.e. the mapping of terms against a set of predefined concepts, improves multimodal and multilingual retrieval for wikipedia images.

1 INTRODUCTION

The similarity of textual concepts has been a matter of research not only in computer science, but also in linguistics and philosophy. In this deliverable we present work done in the context of MUCKE to establish relationships between terms based on their meaning.

Wittgenstein [15] stated that the meaning of words must depends on the use and the context of these words. Methods that employ statistical surface modelling, or corpus-based methods, model the co-occurence of words to identify the probability of their meaning. If two words co-occur often in the same context (of other words), then they are more likely to share a similar meaning.

The following two sections explore this idea based on computer science approaches.

2 FROM WORD TO DOCUMENT CONCEPT SIMILARITY

Text similarity measures are central to information retrieval when ranking results. Despite their successes, they fail when two texts use a disjunct vocabulary to describe the same fact or situation. Measuring the degree of how well two texts relate semantically can be seen as the natural succession of text similarity and has been investigated in a number of related fields, including information retrieval, natural language processing and artificial intelligence [14]. Despite a generally strong interest on word-to-word or sentence-to-sentence similarity [3], research on any text-to-text and document retrieval level remains limited.

In this report we focus on text-to-text similarity and our approach is based on statistical corpus features of text. Avoiding NLP and knowledge-bases allows it to be applied to other, less supported languages and to apply it to more specific domains without pre-existing knowledge. We tested two semantic text-to-text similarity methods and two word representations. We provide extension experiments incorporating two test collections, to evaluate the performance and limitations of the methods and word representations. We first review related work in the next Subsection. Our methods are described in Section 2.1 before describing the experiments in Section 2.2 and presenting and discussing results in Section 2.3.

Latent Semantic Analysis/Indexing (LSA/LSI) is the pioneer approach that initiated a new trend in surface text analysis. Latent Dirichlet Allocation (LDA) models texts as a finite mixture over an underlying set of topics. Random Indexing (RI) is an alternative to LSA/LSI that creates context vectors based on the occurrence of words contexts [12] with the benefit of being incremental and operating with significantly less resources while producing similar inductive results as LSA/LSI and LDA. Word2Vec further expands this approach while being highly incremental and scalable [6]. When trained on large datasets, it is also possible to capture many linguistic subtleties (e.g. relations between cities to their counties) that allow basic arithmetic operations within the model. This, in principle, allows exploiting the implicit knowledge within corpora. All of these methods represent the words in a vector spaces.

This sets these methods apart from approaches that use external knowledge, such as WordNet or OpenCyc, for determining word meanings by explicitly expressed, human-entered knowledge.

2.1 SIMILARITY METHODS

In order to measure the semantic similarity between two documents, we consider two methods that use the word representation in vector space.

The first method, called $SimAgg$, is shown in Equation 1. The method creates a representation vector for each document by aggregating the vector representations of the words in the document. We define the aggregation method as the weighted sum of the elements of the word vectors. Having the document vectors, we calculate the similarity with the traditional cosine function.

$$V_{A,k} = \sum_{i=1}^{n} idf_i * A_{i,k} \qquad SimAgg(A,B) = Cos(V_A, V_B) \tag{1}$$

where V_A represents the vector representation of the document A, A_i is the vector representation the ith word, n is the number of words in the document, idf_i is the Inverse Document Frequency of the ith word in the corpus and k stands for the value of the kth element in each vector.

The second method, called $SimGreedy$ [5] is based on $SimGreedy(A,B)$ [5] defined in Equation 2. Each word in the source document is aligned to the word in the target document to which it has the highest semantic similarity. Then, the results are aggregated based on the weight of each word to achieve the document-to-document similarity. As shown in Equation 3, $SimGreedy$ is the average of SimGreedy(A,B) and SimGreedy(B,A).

$$SimGreedy(A,B) = \frac{\sum_{i=1}^{n} idf_i * maxSim(A_i, B)}{\sum_{i=1}^{n} idf_i} \tag{2}$$

$$SimGreedy = \frac{SimGreedy(A,B) + SimGreedy(B,A)}{2} \tag{3}$$

Rus et al. [11] expand the method by introducing a penalizing factor which factors out very low similarities as noise. Adding this penalizing factor was not efficient in our experiment e.g. instead of filtering the noise, it reduces all values evenly without any re-ranking benefit.

Taking a closer look at the two methods, we can see significant differences between the time complexities. If the number of words in the document A and B is indicated by n and m, the complexity of SimAgg method is $O(n+m)$ while SimGreedy is $O(n*m)$. Experiments show that the difference in performance of SimGreedy to SimAgg is aggravated as the dimension of the vectors increases.

2.2 EXPERIMENTS

We performed two sets of experiments, one using the SemEval 2014 Task 10[1] for Semantic Textual Similarity and one using the MediaEval Retrieving Diverse Social Images Task 2013/2014[2].

[1] http://alt.qcri.org/semeval2014/task10

[2] http://www.multimediaeval.org/mediaeval2014/diverseimages2014

The *SemEval 2014 Task 10 (Semantic Textual Similarity)* consists of two separate subtasks, one for English and one for Spanish. We selected the English subtask (STS-En). The goal of this task is to measure the semantic similarity of two sentences and express it as a similarity score. Participating systems are compared by their mean Pearson correlation between the system output and the human-annotated gold standard. The original SemEval Task 10 corpus was not suitable for training due to its limited in size. We therefore used the English Wikipedia text corpus to train all our word representations in this experiment: Word2Vec with 600 dimensions and Random Indexing with 200 and 600 dimensions. We found that training Random Indexing was about 100 times faster than Word2Vec. We then applied the SimGreedy and SimAgg methods (Section2.1) using these three word representations.

The *MediaEval Retrieving Diverse Social Images Task* addresses result relevance and diversification in social image retrieval. The dataset of both the 2013 and 2014 editions consists of about 110k photos of 600 famous world locations (e.g. the Eiffel tower). Each location is provided with a ranked list of photos, a representative text, Flickr's metadata, a Wikipedia article of the location and user tagging credibility estimation. For semantic text similarity, we focused on the relevance of the representative text of the photos containing title, description and tags. After preprocessing (HTML tag removal and decompounding terms) we expanded the topic names with the first sentence of its Wikipedia page to gain more descriptive queries. We trained both Word2Vec and Random Indexing on the text corpus of MediaEval with 200 dimensions (due to the much smaller MediaEval corpus rather than Wikipedia). We then applied SimAgg and SimGreedy. Additionally, since this MediaEval task addresses an information retrieval problem, we also applied SimGreedy(Q,D) and SimGreedy(D,Q) to consider the similarities between query and document.

2.3 RESULTS AND DISCUSSION

2.3.1 SEMEVAL 2014 TASK 10 RESULTS

Table 1 shows the mean Pearson correlations between the similarity methods and the gold standard. The most impressive result is that SimGreedy with Word2Vec achieved an average correlation of 0.71 as the best overall performance. This relates to rank 11 from a total of 38 evaluated runs. All 10 runs above it use a knowledge base and/or NLP. Across similarity methods, SimGreedy shows better performance than SimAgg. It also appears that the similarity method has more effect on the results rather than the number of dimensions or word representation.

2.3.2 MEDIAEVAL RESULTS

The MediaEval runs were all evaluated with precision at a cutoff of 20 documents (Table 2). A standard Solr index was used as the baseline that produced a P@20 of 0.7908. On the 2014 test set, using SimGreedy and Word2Vec trained on the MediaEval corpus, we achieved the state-of-the-art result of 0.8524 for P@20 between 14 participating teams. Based on the results in Table 2,

Repres.	Dim	SimAgg	SimGreedy
W2V	600	0.685	**0.715**
RI	600	0.691	0.706
RI	200	0.678	0.702

TABLE 1: MEAN PEARSON CORRELATION OF SEMEVAL 2014 TASK 10. W2V AND RI STAND FOR WORD2VEC AND RANDOM INEDIXING WORD REPRESENTATION

Corpus	Repres.	Dim	SimAgg	SimGreedy	SimGreedy(Q,D)	SimGreedy(D,Q)
Wiki	W2V	600	0.756 (6)	0.786 (1)	0.706	0.782
Wiki	RI	600	0.766 (7)	0.794 (2)	0.719	0.785
Wiki	RI	200	0.768 (8)	0.794 (3)	0.714	0.786
MediaEval	W2V	200	0.78 (9)	0.792 (4)	0.706	0.777
MediaEval	RI	200	0.795 (10)	0.788 (5)	0.693	0.776

TABLE 2: PERFORMANCE MEASURED WITH P@20 (NON-STAT. SIG. DIFFERENCES ARE HIGHLIGHTED). THE DIGITS IN THE PARENTHESES INDICATE THE ID OF EACH RUN.

SimGreedy(D,Q) shows better results than SimGreedy(Q,D) since documents are generally longer and more descriptive than queries. However SimGreedy outperformed both SimGreedy(Q,D) and SimGreedy(D,Q). We suspect that the (Q,D) version, which performs very poorly on its own, acts as a length normalization factor for the (D,Q) version, therefore contributing to the improved result.

In order to compare the result of SimGreedy and SimAgg, we tested for significance using Fisher's two-sided paired randomization test. The highlighted areas in Table 2 show all results with no significant difference. This shows that SimGreedy outperforms SimAgg regardless of the training method when using an external corpus for learning word representations. However, using the same corpus for learning representations, the two methods show very similar results.

For more insight in the differences between the runs, we additionally compared all our combinations by calculating their pairwise Pearson rank evaluation (see Figure 1). The average correlation between runs using SimGreedy is larger than those that using SimAgg. This means that regardless of the training method or corpus for word representation, using SimGreedy produces more similar results. SimGreedy methods correlate highest when using Wikipedia as training corpus demonstrating the small effect of the selected training method when using SimGreedy with an external training resource. We also observe very high correlations between the models trained on Wikipedia using Random Indexing with 200 and with 600 dimensions which demonstrates that increasing the dimensionality does not aid performance.

Traditionally, the processing time is key for the semantic analysis of large datasets. Based on the time complexity discussion in Section 2.1, we measured the execution time and found out that

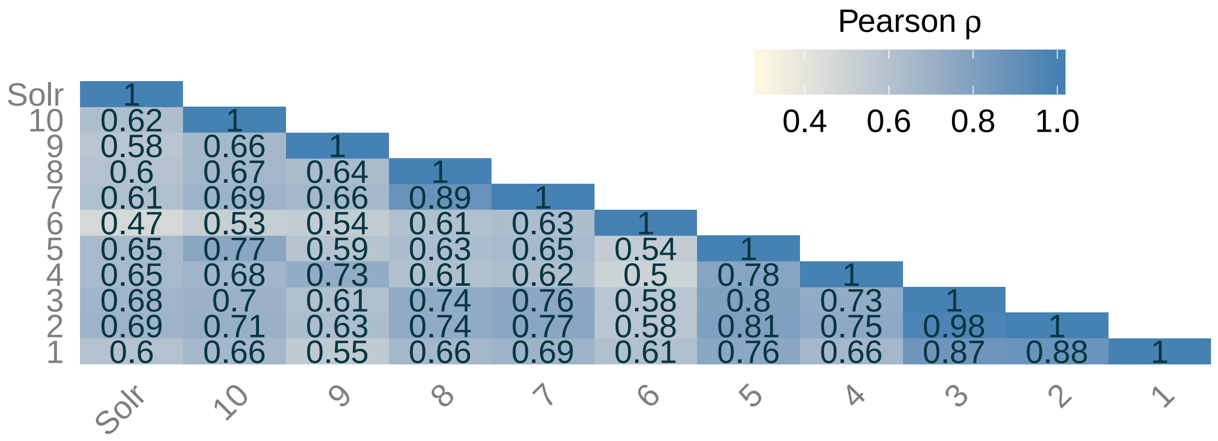

FIGURE 1: PEARSON CORRELATIONS BETWEEN ALL 10 COMBINATIONS OF APPROACHES AND THE SOLR BASELINE THE NUMBERS REFER TO THE ID OF THE RUNS IN TABLE 2

SimGreedy is approximately 40 times slower than SimAgg for 200 dimensions and 45 times slower for 600 dimensions. We therefore turned the procedure into a two-phase process [1]. In the first phase, we applied the SimAgg method to obtain a first ranking of the results. As the second phase, we used n percent of the top documents ranked by the first phase and re-ranked them using SimGreedy. We calculated all combinations with all the values of n from the 1 to 100. We found that all combinations show an extremely similar behaviour over the different levels and summarized this in Figure 2, which shows their average performance. In order to find the best value for n as the cutting point, we spotted the highest precision value that is not significantly different from the best one (i.e. when n is 100 percent). Tracing the results, we found the percentage value of 49 as a good approximation for the cutting point. Assuming the second phase (SimGreedy) is about 40 times slower than the first (SimAgg), using this approach reduces about 48 percent of the execution time while the performance remains the same.

3 KNOWLEDGE BASE MATCHING

3.1 INTRODUCTION

The CHiC evaluation lab[3] aims to explore different aspects related to the retrieval of cultural heritage content stored in digital libraries such as Europeana[4]. There are two subtasks in the multilingual task, dealing with multilingual ad-hoc retrieval and multilingual semantic enrichment. For the ad-

[3] http://www.promise-noe.eu/chic-2013/home

[4] http://www.europeana.eu/

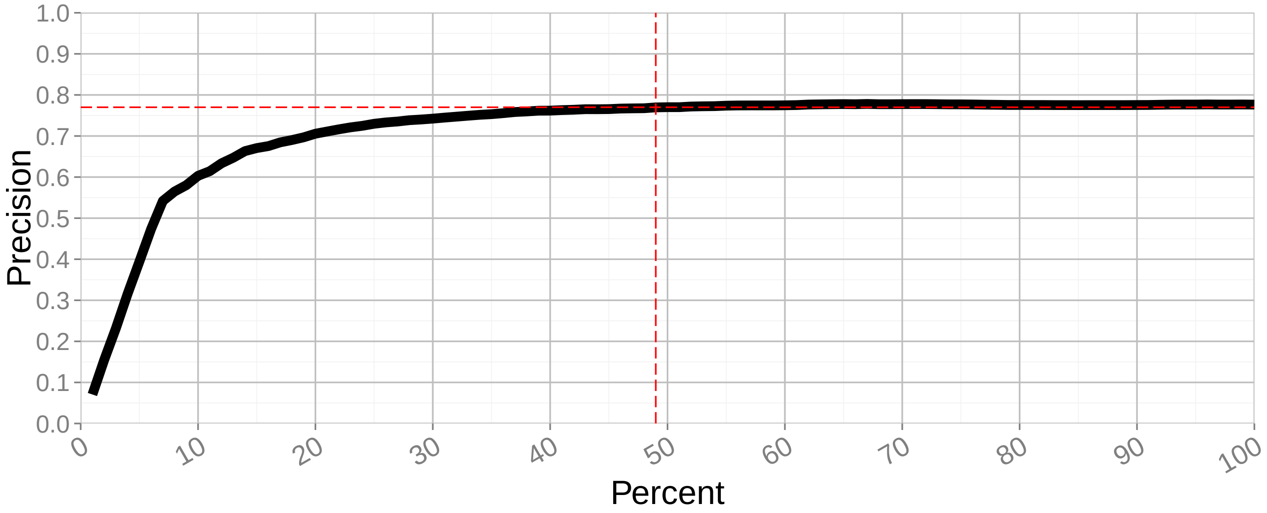

FIGURE 2: AVERAGE PERFORMANCE OF THE TWO-PHASE APPROACH WITH BEST VALUE AT AROUND 49%.

.

hoc task, participants were provided with metadata in 13 different languages and were free to use any automatic method in order to return ranked lists of results for a set of 50 diversified topics. For the semantic enrichment task, a subset of 25 topics from the initial pool was provided and the objective was to return a ranked list of 10 related concepts that could be used to enrich the initial topic and might help to precise the user's information need.

3.2 EXPLICIT SEMANTIC ANALYSIS (ESA)

Explicit Semantic Analysis [2] is a method that maps textual documents onto a structured semantic space. Since its introduction in 2007, ESA was successfully exploited in different natural language processing and information retrieval tasks. The success of this simple method lies in the richness and the quality of the underlying conceptual space. In the original evaluation, ESA outperformed state of the art methods in a word relatedness estimation task and different developments were subsequently proposed.

Radinsky and al. [10] added a temporal temporal dimension to ESA vectors and showed that this addition improves the results for word relatedness.

Hassan and Mihalcea [4] introduced Salient Semantic Analysis, a variant of ESA that relies on the detection of salient concepts prior to linking words and concepts. The merits of their method are difficult to estimate since the comparison is often made with an in-house ESA implementation whose results are significantly poorer than those presented in [2].

We proposed an ESA adaptation to information retrieval tasks that gives priority to categorical information [9]. The comparison with a classical ESA implementation showed that a significant improvement was obtained in an image retrieval setting. Moreover, the method compared favorably with other state of the art indexing and retrieval schemes. Here, we extend the work in [9] and propose to use ESA for query expansion and consolidation, two operations that are explained in more details in Subsection 3.4.

ESA has only weak language dependence and was already deployed in several languages. Sorg and Cimiano [13] proposed an extension of the method to different languages and showed that the method is useful in cross-lingual and multilingual retrieval settings. Here we create ESA vectors in the 10 most represented languages out of the 13 present in the Europeana collection. The following languages are supported: English, German, French, Spanish, Italian, Dutch, Swedish, Norwegian, Polish and Finnish. Adaptations to different languages include detection and removal of Wikipedia disambiguation and list pages and detection of category section.

3.2.1 CLASSICAL FORMULATION OF ESA

Put simply, ESA exploits classical text weighting schemes, such a tf-idf, to model concepts from a structured resource, such a Wikipedia. A relation between words and the concepts that structure the space is established by inverting the concepts' vectorial representations. Thus, each word of the vocabulary has an associated high-dimension projection onto the concept space of the underlying resource. Finally, in order to compare two words or two documents, the representations of individual words are summed and the resulting vectors compared. In information retrieval, the most useful component of ESA is the mapping of words onto concepts that can be used for topic expansion or consolidation.

Classical ESA representations are well adapted for single words, since there is nothing to be done, and for long documents, since the summing operation smooths individual contributions and an accurate semantic representation of the document is obtained. However, the method has some drawbacks for documents such as retrieval topics that contain only few words (typically 2 to 4 words). Here, the smoothing of individual contributions is not sufficient because the contribution of a single word can be higher than that of the others and the obtained related concepts could be related to a part of the topic only. An illustration of this type of problem is provided in table 3 which presents the top 10 ESA concepts associated to topics *Freshwater Fish* and *Jean-Jaques Rousseau*[5]. The results from table 3 indicate that most ESA top ranked concepts are not related to the entire query. When examining results for topic CHiC-051, *Freshwater bivalve* and *Freshwater, Isle of Wight* are related to *freshwater* while *Bait fish* and *Bank fishing* are related to *fish*. Similarly, when examining results for topic CHiC-058, we notice that several ESA top concepts are brought up by the family name *Rousseau* and have little semantic relatedness with the original topic. Concepts found for topic CHiC-064 (*crockery doll house*) are related to doll but not the other terms from the query.

We use an in-house implementation of ESA that includes only the optimization cues publicly available until recently [6]. To validate our implementation, we performed the word similarity task

[5] The wrong spelling of Rousseau's name was extracted from Europeana logs and provided as such for the task.

[6] A full list was recently made public at https://github.com/faraday/wikiprep-esa/wiki/roadmap but the remaining cues were not yet integrated in our implementation.

described in [2], with the same version of Wikipedia, and the method achieved a 0.72 correlation with human judgments (to be compared with 0.75 reported by [2]).

3.2.2 ESA ADAPTATION FOR AD-HOC MULTILINGUAL IR (ESA-C)

We already proposed a version of ESA that gives a privileged role to categorical information in [9]. There, we used two scores to rank Wikipedia concepts:

- a boolean score that captures the number of common words between the initial topic and the words found in the categories associated to Wikipedia concepts.
- the score used in the classical ESA in order to rank concepts, based on the sum of the contributions of the individual words.

Since topics are often short, ties are often obtained with the boolean score and their are broke using the second, finer grained score.

The introduction of the boolean score has two main objectives. First, categorical information should be favored in order to obtain concepts that are hierarchically related (i.e. *isA* relation) to the initial topic or to parts of it. Second, it is possible to identify which parts of the initial query an ESA related concept is related to. For instance, the categories of *Tropical Fish* are *Fish stubs* and *Aquaria* and the topic would have a boolean score of 1 (out of a maximum of 2). Similarly, *Freshwater bivalve*, the top ranked concept with classical ESA, only loosely related to the initial topic, has a boolean score of 0 since its only category is *Bivalves*. The categorical ranking rightly gives a better position to *Tropical Fish* compared to *Freshwater bivalve* since the first concepts is more closely related to *Freshwater fish*.

Here we modified our ESA adaptation for IR in two directions. First, given that categorical information is often sparse, we added the words contained in the first 150 characters after the concepts name in the first paragraph of the category words. This enrichment of the categorical space is motivated by the fact that the first paragraph of Wikipedia article is often a definition that contains salient concepts related to the target one. The limitation to words contained in string of 150 character is useful since the first paragraph has varying length and contains information that is only loosely related to the concepts when it is long. The second modification is a concept detection that is used to produce a third score which favors articles that contain longer concepts from the initial query over other articles. At equal categorical scores, the inclusion of concept detection allows us to favor a Wikipedia concept that includes *Jean-Jacques Rousseau* in its text when compared to another concept that includes *Jean-Jacques* and *Rousseau* separately. The top related concepts obtained with ESA-C for *Freshwater fish* and *Jean-Jacques Rousseau* are presented in the third column of table 3. In both cases, the top 10 concepts are much more closely related to the initial topics compared to the use of classical ESA. The list for topic CHiC-051 contains only fish and the list for CHiC-058 includes different works of *Jean-Jacques Rousseau*. *Crockery doll house* is a specialized one, which is not well represented in Wikipedia and the retrieved concepts are still

TABLE 3: TOP 10 ESA RELATED CONCEPTS FOR TOPICS *FRESHWATER FISH*, *JEAN-JAQUES ROUSSEAU* AND *CROCKERY DOLL HOUSE*. THE SECOND COLUMN CONTAINS RESULTS FOR CLASSICAL ESA, WHILE THE THIRD RESULTS FOR THE ADAPTED VERSION OF ESA (ESA-C).

Topic CHiC-051 *Freshwater Fish*		
Rank	ESA	ESA-C
1	Freshwater bivalve	Eastern freshwater cod
2	Freshwater mollusc	Ide (fish)
3	Tropical fish	New Zealand longfin eel
4	Freshwater, Humboldt County, California	Common galaxias
5	Fish fillet processor	European perch
6	Bait fish	Green swordtail
7	Fish marketing	Rainbowfish
8	Bottom fishing	Common rudd
9	Freshwater, Isle of Wight	Spotted bass
10	Bank fishing	Common bream
Topic CHiC-058 *Jean-Jaques Rousseau*		
Rank	ESA	ESA-C
1	Confessions (Rousseau)	Confessions (Rousseau)
2	Saint-Jean	Considerations on the Government of Poland
3	Considerations on the Government of Poland	Discourse on the Arts and Sciences
4	Eugène Rousseau (chess player)	Emile, or On Education
5	John Jacques, Baron Jacques	Essay on the Origin of Languages
6	Eugene Rousseau (saxophonist)	Discourse on Inequality
7	Jean-Jacques Henner	Letter to M. D'Alembert on Spectacles
8	Victor Rousseau	Pygmalion (Rousseau)
9	Bobby Rousseau	Julie, or the New Heloise
10	Discourse on the Arts and Sciences	Le devin du village
Topic CHiC-064 *Crockery doll house*		
Rank	ESA	ESA-C
1	Peg wooden doll	Mabel Lucie Attwell
2	Composition doll	Bringing Up Father
3	Anatomically correct doll	The Tale of Mrs. Tiggy-Winkle
4	Bisque doll	China doll
5	Black doll	Japanese traditional dolls
6	Paper doll	Queen Mary's Dolls' House
7	Madame Alexander	Bild Lilli doll
8	Fashion doll	Vivien Greene
9	Doll	Paper Dolls (band)
10	China doll	Wall House (Elkins Park, Pennsylvania)

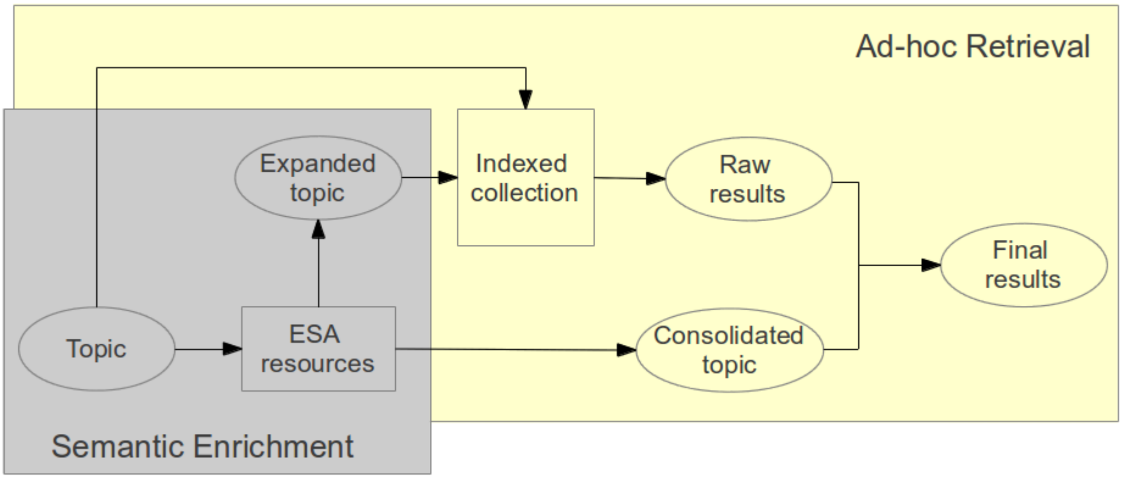

FIGURE 3: OVERVIEW OF THE SEMANTIC ENRICHMENT AND RETRIEVAL FRAMEWORK.

unrelated to the entire topic in a large majority of cases. This last topic illustrates one limitation of all ESA implementation, namely the poor mapping between the initial document and the knowledge included in the underlying conceptual space.

The use of the lists of related ESA concepts for both semantic enrichment and for ad-hoc retrieval is detailed in Section 3.4.

3.3 EUROPEANA COLLECTION PROCESSING

We kept documents of the Europeana Collection that belong to the 10 languages processed with ESA. Separate indexes were created for each of the modeled languages. Then selected metadata associated to the following fields: "dc:title", "dc:description", "dc:subject", "dc:type", "dcterms:medium", "dc:date". The retained fields were merged into a bag-of-words representation and then modeled using a tf-idf scheme. Due to the use of probabilistic models of the topics (see Subsection 3.4.2), the tf-idf representation of documents was subsequently transformed into a probabilistic form by dividing the weight of each word by the sum of the scores of all words in the document. Monolingual collections are stemmed using the corresponding Perl Snowball stemmer implementation.

3.4 ENRICHMENT AND RETRIEVAL FRAMEWORK

The framework devised here was used for both semantic enrichment and ad-hoc retrieval and is summarized in figure 3. The semantic enrichment process exploits only topic expansion with the ESA versions (ESA and ESA-C) and returns ranked lists of results using different ranking schemes detailed in Subsection 3.4.2.

3.4.1 SEMANTIC TOPIC ENRICHMENT FRAMEWORK

The purpose of the semantic enrichment process is to return a ranked list of concepts that are semantically related to the initial topic and could be use for query expansion. To test multilingual rankings, we introduced fusion methods that exploit the explicit interlingual links available in Wikipedia using either different fusion schemes based on the scores in individual languages. In all cases, the proposed enrichments are collection independent. Only Wikipedia concepts formed of at most 4 words were retained in the final rankings. Lists of related concepts obtained with the original version of ESA and with the adapted ESA-C version are presented in table 3.

3.4.2 AD-HOC RETRIEVAL FRAMEWORK

Within CHiC, the objective of the ad-hoc retrieval process is to return the best results possible using whatever automatic method at hand. In our approach, the target topic is first processed using ESA resources to expand and consolidate it. The initial words and the expanded concepts are then compared to the index of the collection in order to retrieve a raw list of results. The elements of this list are then compared to the consolidated version of the topic in order to obtain the final list of results. Similarly to the ranking of ESA related concepts, two similarity measures are used:

- a boolean score to measure a coarse similarity between the initial topic or its related concepts and the documents in the collection.
- the cosine similarity is used to measure the degree of similarity between a topic and corresponding documents.

The boolean score has a higher priority than the cosine similarity, which is used only to break ties. For multilingual runs, the process is performed for each of the languages processed and then results are combined by ranking results by decreasing scores.

Topic expansion is performed in a way similar to description provided in Subsection 3.4.1. Consolidation is a by-product of the expansion process and aims to obtain an expanded version of the initial topic that contains, in addition to the original words, other words that are semantically related to the topic but are not part of it. The words are ranked by summing their individual scores associated to the top 100 ESA related concepts and then by multiplying this sum with the log of the number of different articles in which they appear. This last score is used in order to favor words that are associated to a large number of Wikipedia concepts related to the initial articles. Up to 1000 related words are retained for each topic and a probabilistic model of the consolidated versions is obtained by dividing individual word scores by the sum of all scores. In table 4, we present top 10 words related to each topic obtained with ESA and ESA-C. A majority of the obtained words are semantically related to the initial topic, although some outliers appear. For *freshwater fish*, all top 10 words are related to the initial topic and thus useful for ranking results. In the case of *Jean-Jaques Rousseau*, there are French stop words that were not removed. For *Crockery doll house*, *nrhp* (abbreviation of National Register of Historic Places) appears since this acronym is strongly related to *house*. Similarly to the collection processing, the consolidated versions of the topics are stemmed.

TABLE 4: TOP 10 WORDS FORM THE CONSOLIDATED VERSION OF THE TOPICS FOR *FRESHWATER FISH*, *JEAN-JAQUES ROUSSEAU* AND *CROCKERY DOLL HOUSE*. THE SECOND COLUMN CONTAINS RESULTS FOR CLASSICAL ESA, WHILE THE THIRD COLUMN PRESENTS RESULTS FOR THE ADAPTED VERSION OF ESA (ESA-C).

Topic CHiC-051 *Freshwater Fish*		
Rank	ESA	ESA-C
1	fish	fish
2	freshwater	freshwater
3	acquarium	galaxia
4	fillet	aquarium
5	fishery	species
6	bait	fin
7	water	water
8	species	river
9	lake	trout
10	fin	carp
Topic CHiC-058 *Jean-Jaques Rousseau*		
Rank	ESA	ESA-C
1	jacques	rousseau
2	jean	jean
3	rousseau	de
4	de	jacques
5	french	french
6	le	le
7	saint	paris
8	paris	philosopher
9	la	pygmalion
10	baptiste	pierre
Topic CHiC-064 *Crockery doll house*		
Rank	ESA	ESA-C
1	doll	doll
2	house	house
3	toy	toy
4	barbie	barbie
5	goo	mattel
6	album	album
7	mattel	goo
8	nrhp	film
9	licca	licca
10	song	song

3.5 EXPERIMENTS

As we mentioned, we have evaluated runs for both the semantic enrichment and the ad-hoc retrieval subtasks and we analyse them here. Unfortunately, this analysis is altered by the fact that an important bug in the scoring of related was discovered after the release of official results. This bug had a strong negative impact on the quality of results for all runs that exploited fusion techniques for semantic enrichment and automatic topic expansion for ad-hoc retrieval. The bug biases individual boolean scores but the order of concepts is not affected and a comparison of ESA versions remains possible. Boolean scores of expanded concepts were overrated compared to the boolean scores of documents found using terms from the original topic. All affected runs are indicated by a "*" sign in the following subsections.

3.5.1 SEMANTIC ENRICHMENT

Evaluated runs The following eight runs were submitted to the semantic enrichment subtask:

- *ceaListEnglishMonolingual* - Related concepts are obtained with ESA-C. The experiment is monolingual since it only exploits the English Wikipedia version. Proposed expansions are collection independent.

- *ceaListEnglishMonolingualOriginal* - Related concepts are obtained with classical ESA. The experiment is monolingual since it only exploits the English Wikipedia version. Proposed expansions are collection independent.

- *ceaListEnglishRankEnglish* - Rank fusion for monolingual results obtained with ceaListEnglish-Monolingual. The rank of the concept is obtain by averaging its ranks in different languages. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only concepts that have an English version are considered.

- *ceaListEnglishRankMultilingual* - Rank fusion for monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score in an individual language is presented.

- *ceaListEnglishBooleanEnglish ** - Fusion of boolean scores for monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only English versions of Wikipedia concepts are presented. Proposed expansions are collection independent. Only concepts that have an English version are considered.

- *ceaListEnglishBooleanMultilingual* * - Fusion of boolean scores monolingual results obtained with ceaListEnglishMonolingual. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score in an individual language is presented. Only concepts that have an English version are considered.
- *ceaListEnglishCosineEnglish* * - Fusion of cosine similarity scores for monolingual results obtained with ceaListEnglishMonolingualOriginal. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Only English versions of Wikipedia concepts are presented. Only concepts that have an English version are considered.
- *ceaListEnglishCosineMultilingual* * - Fusion of cosine similarity scores monolingual results obtained with ceaListEnglishMonolingualOriginal. For a Wikipedia concept to be considered, it has to appear in at least three languages. The experiment is multilingual since different Wikipedia versions (9 languages: en, fr, de, es, it, nl, no, sv, pl) are used. Given different translations of a concept, the one that has the highest score is presented. Only concepts that have an English version are considered.

Results Even though the results for 6 out of 8 runs are biased here, there are some interesting conclusions that we can draw from table 5. The comparison between ceaListEnglishMonolingualOriginal and ceaListEnglishMonolingual is favorable to the latter method. The original ESA implementation has significantly poorer performances compared to the adapted method introduced (P@10 0.365 vs. 0.66). The privileged role given to categories and to the first words in the concept text, coupled with concept detection in the queries have a positive impact on semantic enrichments.

None of the fusion methods proposed improves results compared to the best submitted run but this is at least in part due to the bug that affected the values of boolean concept scores. When comparing the fusion schemes, there are no significant differences between monolingual and multilingual fusions. Since the same concepts were proposed but languages differed, this results show that the ground truth is of high quality. The cosine-based fusion strongly degrades results, while the fusion based on ranks is closer to the original results.

Important differences occur at the topic level. For ceaListEnglishMonolingual, when examining CHiC-51 (*freshwater fish*) and CHiC-58 (*Jean-Jacques Rousseau*), all top 10 related concepts are at least partly related to the initial topic. Inversely, results are very poor (9 out of 10 irrelevant enrichments) for topics CHiC-64 (*crockery doll houses*) and CHiC-65 (*sea sunset*). These failures are probably due to a poor mapping of the topic in the Wikipedia corpus for CHiC-64 and to the very small number of Wikipedia concepts that cover both *sea* and *sunset*.

TABLE 5: SEMANTIC ENRICHMENT ACCURACY MEASURED USING P@10 OF RELEVANT AND OF RELEVANT + PARTLY RELEVANT RESULTS.

Run name	P@10	P@10 (rel + part.rel)
ceaListEnglishMonolingual	**0.468**	**0.66**
ceaListEnglishMonolingualOriginal	0.212	0.364
ceaListEnglishRankEnglish	0.34	0.56
ceaListEnglishRankMultilingual	0.3382	0.5556
ceaListEnglishBooleanEnglish *	0.228	0.436
ceaListEnglishBooleanMultilingual *	0.22	0.428
ceaListEnglishCosineEnglish *	0.076	0.164
ceaListEnglishCosineMultilingual *	0.076	0.164

3.5.2 AD-HOC RETRIEVAL

Evaluated runs For "noExpansion" runs, results are ranked first by the number of terms from the initial topic that appear in the document and then by the the cosine similarity between the consolidated version of the topic and document representations. For the other runs, the boolean score of related ESA concepts biases the results. As we mentioned, multilingual fusion was performed by The following 16 runs were submitted to the semantic enrichment subtask:

- **ceaListMultilingualNoExpansion** - Multilingual run that retrieves only documents which contain at least one word from the initial topic.
- **ceaListFrenchNoExpansion** - Monolingual French run that retrieves documents which contain at least one word from the initial topic.
- **ceaListGermanNoExpansion** - Monolingual German run that retrieves documents which contain at least one word from the initial topic.
- **ceaListMultilingualOriginal *** - Multilingual run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Wikipedia concepts obtained with classical ESA.
- **ceaListMultilingualFiltered *** - Multilingual run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Wikipedia concepts obtained with ESA-C.
- **ceaListDutchFiltered *** - Monolingual Dutch run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Dutch Wikipedia concepts obtained with a version of Explicit Semantic Analysis adapted to short documents (topics).

- **ceaListEnglishFiltered *** - Monolingual English run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related English Wikipedia concepts obtained with ESA-C.

- **ceaListFrenchFiltered *** - Monolingual French run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related French Wikipedia concepts obtained with ESA-C.

- **ceaListGermanFiltered *** - Monolingual German run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related German Wikipedia concepts obtained with ESA-C.

- **ceaListItalianFiltered *** - Monolingual Italian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Italian Wikipedia concepts obtained with ESA-C.

- **ceaListNorwegianFiltered *** - Monolingual Norwegian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Norwegian Wikipedia concepts obtained with ESA-C.

- **ceaListPolishFiltered *** - Monolingual Polish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Polish Wikipedia concepts obtained with ESA-C.

- **ceaListSpanishFiltered *** - Monolingual Spanish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Spanish Wikipedia concepts obtained with ESA-C.

- **ceaListSwedishFiltered *** - Monolingual Swedish run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Swedish Wikipedia concepts obtained with ESA-C.

- **ceaListEnglishOriginal *** - Monolingual English run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related English Wikipedia concepts obtained with ESA-C.

- **ceaListItalianOriginal *** - Monolingual Italian run that retrieves documents which contain at least one word from the initial topic and/or the full name of one of query's 1000 most related Italian Wikipedia concepts obtained with classical ESA.

Results Due to the bug that affected all the runs that involved ESA based topic expansion, it is difficult to compare runs that did not involved expansion and the others. However, it is worthwhile noticing the the best submitted run, i.e. ceaListMultilingualNoExpansion a simple fusion

TABLE 6: MAP PERFORMANCES FOR AD-HOC RETRIEVAL RUNS.

Run name	MAP
ceaListMultilingualNoExpansion	**0.1878**
ceaListFrenchNoExpansion	0.0478
ceaListFrenchFiltered *	0.0290
ceaListGermanNoExpansion	0.0631
ceaListGermanFiltered *	0.0505
ceaListMultilingualOriginal *	0.0805
ceaListMultilingualFiltered *	0.0977
ceaListDutchFiltered *	0.0377
ceaListEnglishOriginal *	0.0304
ceaListEnglishFiltered *	0.0321
ceaListItalianOriginal *	0.0165
ceaListItalianFiltered *	0.0222
ceaListNorwegianFiltered *	0.0251
ceaListPolishFiltered *	0.0109
ceaListSpanishFiltered *	0.0204
ceaListSwedishFiltered *	0.0123

of results obtained for individual languages, gave interesting results compared to monolingual runs that involved no ESA expansion.

When comparing ceaListMultilingualOriginal and ceaListMultilingualFiltered, the two multilingual runs that exploit ESA and ESA-C, obtained results are better for the second run (MAP 0.0805 vs. O.0977). This result confirms the one obtained for semantic enrichment, where ESA-C was also superior to classical ESA. It is also in line with our findings from [9], which showed that giving a privileged role to categorical information is beneficial in an image retrieval scenario. The favorable comparison of ESA-C with ESA is also confirmed for English (MAP 0.321 vs. 0.304) and Italian (MAP 0.165 vs. 0.0222).

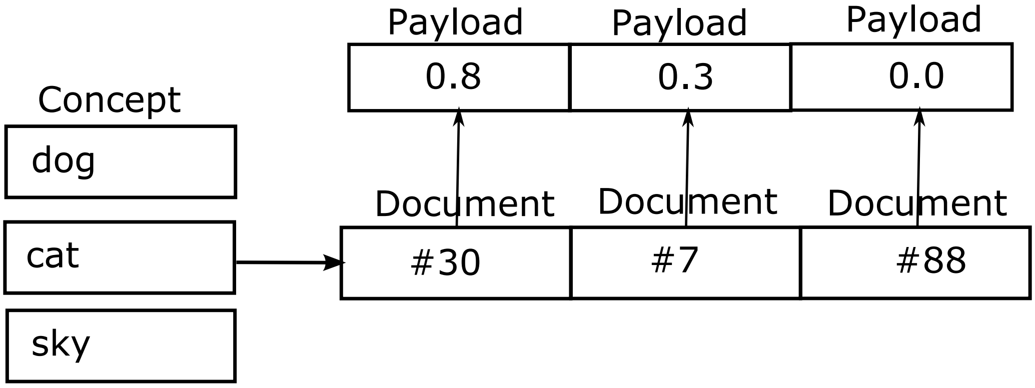

FIGURE 4: USE OF LUCENE PAYLOAD FEATURE FOR EMBEDDING SEMANTIC PROBABILITIES

.

4 SEMANTIC INDEXING WITH PAYLOADS

The Payload feature gives control of score-boosting on term level and is one of the latest additions to Lucene[7]. Specifically, payloads were introduced to allow boosting individual terms based on additional meta information about these terms to further diversify the significance of these terms in the scoring process of a search engine. This could, for example, be the visual appearance of the term as displayed on a website (e.g. boosting terms that are displayed in red color over those that are presented in standard color) or it could be its linguistic role in a sentence (e.g. boosting nouns over prepositions as identified by a part-of-speech tagger). We apply the payload feature semantically as part of the concept index. The concept index stores unique concept labels and associates them with documents and facets (e.g. text facets). The payload represents the probability of a concept being similar to another concept, such as a concept extracted from a query or a concept that has been learned from an image. Figure 4 depicts how the payload is integrated in an a standard inverted index.

The index is shown in figure 4. During the indexing process concepts are linked to documents that have a semantic link to this concept. This is done by adding an array of concept fields to each document covering the entire semantic spectrum of the document. Each document, in relation the the concept is is related to is now extended with a payload that can generally contain any kind of information. We use this payload extension for storing an additional semantic weight (e.g. document #30 is with a probability of 0.7 related to the concept "dog"). This allows the scoring algorithm (e.g. in vector space ranking) to promote document #30 over document #7 and document #88) in addition to pure term matching. This enables the application of semantic probabilities of concept relatedness for indexing and searching.

The semantic probabilities can be determined by an outside source (e.g. a machine learning classifier that uses multiple source of evidence to determine the relatedness of a document with a concept to determine its weight before this is added to the index as a payload).

[7] http://lucene.apache.org/core

One possibility for determining the conceptual relatedness between a document with all its underlying concepts (i.e. the probability of a document to be related to a particular concept) is to apply a neighbourhood search in its conceptual space. We are currently investigating the FLANN library as an optimized high-speed approximate K-Nearest-Neighbour (KNN) search in high-dimensional hyperspaces [8] and apply it to the high-dimensional word vector spaces based on the content of documents. FLANN provides a range of state-of-the-art algorithms that work best for K-Nearest-Neighbour (KNN) searches and additionally offers a automatic system that selects the most appropriate algorithm including an optimized parametric setting based on the dataset at hand[7]. FLANN is used to store, access and extract semantic word vectors based on their high-dimensional neighbourhood with a selection of possible distance metrics. Currently, the main challenge is that is requires to search every concept of all documents to find local neightbourhoods which is clearly sub-optimal and requires improvement.

5 CONCLUSION

We explored a variety of text semantic approaches to multimodal and multilingual image retrieval: random indexing, explicit semantic analysis, deep learning, in the context of two evaluation campaigns and showed the benefits brought by these methods in comparison to the state of the art. The obtained results are especially encouraging for the semantic enrichment task. The ESA-C adaptation of Explicit Semantic Analysis clearly outperforms the original version of the method.

References

[1] Van Dang, Michael Bendersky, and W.Bruce Croft. Two-stage learning to rank for information retrieval. In *Proc. of ECIR*, 2013.

[2] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[3] Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proc. of Joint Conference on Lexical and Computational Semantics*, 2013.

[4] Samer Hassan and Rada Mihalcea. Semantic relatedness using salient semantic analysis. In *AAAI*, 2011.

[5] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of AAAI*, 2006.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, pages 331–340. INSTICC Press, 2009.

[8] Marius Muja and David G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36, 2014.

[9] Adrian Popescu and Gregory Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval*, pages 33:1–33:8, 2011.

[10] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.

[11] V. Rus, M. Lintean, R. Banjade, N. Niraula, and D. Stefanescu. SEMILAR: The Semantic Similarity Toolkit. In *Proc. of ACL*, 2013.

[12] Magnus Sahlgren. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop int the Proc. of TKE*, 2005.

[13] P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, 2012.

[14] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1), 2010.

[15] L. Wittgenstein. *Philosophical Investigations – 3rd ed.* Basil Blackwell, 1963.