Analyzing Semantic Concept Patterns to Detect Academic Plagiarism

Norman Meuschke, Nicolas Siebeck, Moritz Schubotz, Bela Gipp
Department of Computer and Information Science
University of Konstanz, Germany
{first.last}@uni-konstanz.de

ABSTRACT

Detecting academic plagiarism is a pressing problem, e.g., for educational and research institutions, funding agencies, and academic publishers. Existing plagiarism detection systems reliably identify (nearly) copied text, but often fail to detect disguised forms of academic plagiarism, such as paraphrases, translations, and idea plagiarism. We present Semantic Concept Pattern Analysis - an approach that performs an integrated analysis of semantic text relatedness and structural text similarity. Using 25 officially retracted cases of academic plagiarism, we demonstrate that our approach can detect cases that established text matching approaches would not identify. We see the approach as a promising addition to improve the detection capabilities for strong paraphrases. We plan to further improve Semantic Concept Pattern Analysis and include the approach as part of an integrated detection process that analyzes heterogeneous similarity features to better identify the many possible forms of plagiarism in academic documents.

1 INTRODUCTION

Academic plagiarism is "the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected" [10]. Detecting academic plagiarism is a pressing problem, e.g., for educational and research institutions, funding agencies, and academic publishers. Research on information retrieval (IR) approaches for plagiarism detection (PD) has yielded mature systems that employ text retrieval to find suspiciously similar documents. These systems reliably retrieve documents containing (nearly) copied text, but often fail to identify disguised forms of academic plagiarism, such as paraphrases, translations, and idea plagiarism [38].

Researchers pursue several approaches to improve the detection capabilities for disguised forms of academic plagiarism. Methods that analyze the semantic information in academic documents are promising to complement text matching methods for identifying obfuscated instances of plagiarism, such as strong paraphrases.

In this paper, we propose a new approach that combines the analysis of semantic text relatedness with an analysis of structural text similarity. We demonstrate that the approach can complement established text matching approaches in identifying real-world cases of academic plagiarism. We structure the presentation of our contributions as follows. Section 2 briefly reviews technologies for determining semantic relatedness as well as existing semantic detection approaches and their drawbacks. Section 3 presents a new PD approach that addresses these weaknesses by adapting Explicit Semantic Analysis (ESA), a successful approach to determine the semantic relatedness of texts, to the PD use case. We use Wikipedia as our semantic background, which enables the approach to be applied to academic documents from a wide range of disciplines. Section 4 demonstrates the capability of the new approach to detect real-world cases of academic plagiarism that established text matching approaches would not identify. Section 5 concludes the paper and presents our plans for future research.

2 RELATED WORK

This section summarizes approaches to quantify the semantic relatedness of words or texts. In particular, we present Explicit Semantic Analysis as a well-established approach for this task. By reviewing approaches that use semantic features for PD, we motivate that adapting ESA for this task and combining it with an assessment of structural similarity holds promise to overcome some of the weaknesses of current PD approaches.

2.1 Semantic Relatedness

Quantifying the semantic relation between a pair of words or texts is essential for many Natural Language Processing (NLP) and IR tasks [24]. Budanitsky and Hirst categorize semantic relations into semantic similarity and semantic relatedness [3].

Semantic similarity covers linguistic relations between words, such as synonymy (e.g., "forest" and "wood"), abbreviations (e.g., "bicycle" and "bike"), and hypernymy (e.g., "tree" and "plant").

Semantic relatedness covers any relation between words, including those of similarity. Semantic relatedness includes additional lexical associations, such as meronymy ("is-a-part-of" relations, e.g., "tree" and "leaf") and antonymy (e.g., "hot" and "cold"), but also more general relations, which Morris and Hirst characterize as "non-classical lexical semantic relations" [28]. While classical semantic relations are context-free, non-classical relations are context-dependent. For example, a non-classical relation exists between "referee" and "ball" in the context of soccer.

Approaches to determine semantic relatedness fall into two categories: knowledge-based and corpus-based [20]. Some methods combine both approaches [27]. *Knowledge-based approaches* use information derived from semantic networks, such as, dictionaries, thesauri, or other lexical resources. The methods use the connection between term nodes in the network to determine the relation between the terms. WordNet[1] is a well-known example of a semantic network. This dictionary and thesaurus for the English language groups words by their part of speech as well as into sets of synonyms (synsets). Additionally, WordNet contains many linguistic relations, making it especially suitable for the computation of semantic similarity. Researchers proposed numerous approaches to

[1] http://wordnet.princeton.edu/

quantify semantic relations with the help of WordNet [3]. Other lexical resources include PropBank[2], VerbNet[3], and FrameNet[4]. In theory, any ontology can function as a semantic network [35].

The major drawback of knowledge-based approaches is their domain-specificity [22]. Most resources focus on lexical information about individual words, but contain little information on the different word senses or "world knowledge". Creating and maintaining lexical resources requires expertise, time, effort, and money. Since the resources still only cover a small portion of the natural language lexicon, the applicability of such resources is limited [14].

Corpus-based approaches exploit the idea that semantically related words occur in similar contexts to extract semantic information from large corpora. Models like hyperspace analogue to language [4] and Latent Semantic Analysis (LSA) [8] learn semantic relations from patterns of word co-occurrence in the corpus. The drawback of LSA is its limitation to using the knowledge encoded in the text collection as is. In other words, the approach does not use human-organized knowledge. By relying on Singular Value Decomposition, LSA is essentially a dimensionality reduction technique that identifies the most significant dimensions in the data, which are assumed to represent "latent concepts" [14]. The next section describes ESA, a corpus-based approach to determine semantic relatedness from explicitly encoded human knowledge.

2.2 Explicit Semantic Analysis

Explicit Semantic Analysis [13] is an approach to model the semantics of a text by representing the text as a vector in a high-dimensional vector space of semantic concepts. Semantic concepts are topics that are explicitly encoded in a knowledge base corpus, i.e. a collection of individual texts attributable to specific concepts ("topics").

Encyclopedias are prime examples for knowledge base corpora. Each article in an encyclopedia covers one specific topic. Thus, each article can be considered as a concept, which can be labeled with the article's title. The text of an article is an explicit, man-made description of the semantic content of the concept. In theory, any collection of documents that can be mapped to a topic, can serve as a knowledge base. For example, initial implementations of ESA used the Open Directory Project[5], an open-content directory of Web links [11] as a knowledge base. Gabrilovich and Markovitch showed that Wikipedia is a suitable knowledge base corpus[13], which is why we use it for our approach.

Figure 1 illustrates how ESA derives the representation of input text in the semantic concept vector space. Each article, i.e. concept, in the knowledge base corpus (in our case Wikipedia) is parsed and represented as a *tf/idf*-weighted vector in the high-dimensional vector space of all terms in the collection. These *tf/idf*-weighted vectors for articles are then transformed into a weighted inverted index, called Semantic Interpreter. The Semantic Interpreter maps each term t_i in the knowledge base corpus to a vector $\vec{c_i}$ of concepts, i.e. Wikipedia articles. The value of each component $c_k \in \vec{c_i} \quad (k = 1 \dots N)$ corresponds to the *tf/idf* value of t_i for the article represented by c_k. In other terms, each $\vec{c_i}$ reflects how

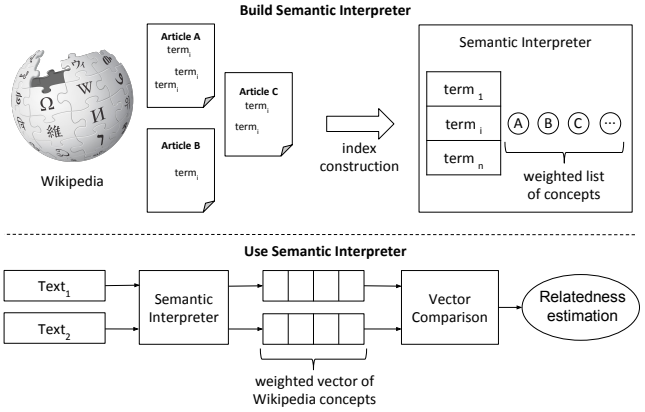

Figure 1: Concept of Explicit Semantic Analysis [13].

descriptive the term t_i is for each of the N concepts (articles), in the knowledge base corpus.

To determine the semantic relatedness of input texts, ESA represents each text x_j as a *tf/idf*-weighted term vector $\vec{v_j}$ with elements $v_k \quad (k = 1 \cdots N)$. Each term t_i occurring in x_j is then queried to the Semantic Interpreter to retrieve the weighted vector of concepts $\vec{c_i}$. To form the semantic concept vector $\vec{s_j}$ of length N that represents x_j, ESA computes the components $s_k \in \vec{s_j}$ as $s_k = \sum_{t_i \in x_j} v_k \cdot c_k$. Finally, the semantic relatedness of the texts is quantified as the cosine distances of their semantic concept vectors.

The major advantage of ESA over term-based vector space retrieval is that ESA does not require a high overlap in literally matching terms between the texts. For example, if key terms in a text are replaced with synonyms, the effectiveness of term-based vector space retrieval decreases rapidly. Since ESA maps multiple terms to the same concept, the approach is better suited to identify the high semantic overlap of texts in which key terms have been replaced with semantically equivalent terms [9].

ESA has been shown to perform well in modeling semantic relatedness for various use cases, such as text categorization [12], word sense disambiguation, and ontology matching [22] as well as mono-lingual [9] and cross-lingual Information Retrieval [32].

2.3 Semantic Plagiarism Detection

Plagiarism detection is a specialized IR task with the objective of comparing an input document to a large collection and retrieving all documents exhibiting similarities above a certain threshold. PD systems typically follow a two-stage process consisting of candidate retrieval and detailed comparison [36]. In the candidate retrieval stage, the systems employ computationally efficient retrieval methods, such as n-gram fingerprinting, vector space models, or citation analysis to limit the retrieval space [25, 37]. Traditionally, exhaustive string comparisons are applied in the detailed comparison stage. However, such approaches are limited to finding near copies of text. To detect disguised forms of plagiarism, researchers proposed a variety of mono-lingual approaches that employ semantic or syntactic feature analysis, as well as cross-lingual IR methods.

We focus our review on PD approaches that consider semantic features. Such approaches commonly use lexical resources, such

[2]https://verbs.colorado.edu/ mpalmer/projects/ace.html
[3]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[4]https://framenet.icsi.berkeley.edu/fndrupal/
[5]www.dmoz.org

as WordNet, and pairwise sentence comparisons to analyze the set of exactly matching and semantically related words [31, 34]. Other works go beyond comparing word-based semantic similarity by also considering similarity in the argument structure of the sentences [29, 30]. These approaches apply semantic role labeling using lexical resources such as PropBank, VerbNet, or FrameNet. Semantic role labeling is an automated process to identify the arguments of a sentence, i.e. the subject, object, events, and relations between these entities, using a pre-defined set of roles. The detection approaches typically combine the information on semantic arguments with the word-based semantic similarity. For instance, Osman et al. only consider exactly matching words and WordNet derived synonyms for the similarity assessment if they belong to the same argument in both sentences [29].

Few researchers investigated the use of corpus-based semantic analysis methods for PD. Ceska employed Singular Value Decomposition to improve the detection of slightly obfuscated instances of plagiarism [5]. His test collection consisted of 150 texts that students had "synthetically" plagiarized by cutting, pasting and slightly altering content from source articles.

In previous research, we analyzed patterns of in-text citations in academic documents as language-independent features to model both semantic relatedness and structural similarity [16, 17, 19]. We showed that analyzing citation patterns is a computationally modest approach to identify heavily disguised academic plagiarism in real-world, large-scale collections [15, 18].

The success of citation-based PD lies rooted in two factors. First, citations encode a large amount of semantic information that cannot easily be substituted. Second, analyzing in-text citation patterns, i.e. identical citations occurring in proximity and / or similar order within two documents, can indicate structural similarity of the texts in addition to similar semantic content.

We see the combined analysis of semantic text relatedness and structural text similarity as most promising to overcome the limitations of current PD approaches [26]. The next section presents an approach that uses semantic concepts obtained by performing ESA for an integrated analysis of semantic relatedness and structural similarity of texts.

3 PROPOSED APPROACH

The idea of our approach, which we name *Semantic Concept Pattern Analysis*, is to model the semantic relatedness and structural similarity of texts in terms of shared semantic concepts and the order in which such concepts occur in the texts. Documents whose similarity according to our model exceeds defined thresholds are retrieved as potential instances of plagiarism.

Topically related academic documents, e.g., papers in the same research area, naturally share semantic content. Therefore, we expect that exclusively analyzing the amount of shared semantic concepts is an insufficient indicator for potential plagiarism. The purpose of academic writing is to present a logical sequence of arguments to arrive at a conclusion. We hypothesize that sharing semantic content in similar order is therefore a stronger indicator for potentially suspicious similarity in academic documents. Our past research on analyzing patterns of in-text citations in academic documents to model semantic relatedness and structural similarity provided evidence for the validity of this assumption (cf. Section 2.3).

Semantic Concept Pattern Analysis partitions documents into fragments and represents the fragments as semantic concept vectors. To derive the vectors, we employ ESA and use the English version of Wikipedia as the knowledge base corpus. As presented in Section 2.1, Wikipedia has been proven to be a highly qualitative knowledge-base for a broad spectrum of domains.

We developed two approaches, which emphasize different similarity characteristics, to identify and score semantic concept patterns. Both approaches seek to primarily detect paraphrased instances of plagiarism by identifying semantic relatedness and structural similarity. The following sections present the two approaches.

3.1 Semantic Sequence Scoring

Semantic Sequence Scoring (SSS) extends ESA with a heuristic procedure for identifying and scoring sequential concept patterns that indicate structural text similarity. SSS performs a pairwise document comparison for which it partitions the documents into text fragments, i.e. paragraphs or sentences. SSS then represents all text fragments in two documents A and B as semantic concept vectors and calculates the relatedness of all vector pairs $r(\vec{a_i}, \vec{b_j})$.

We developed two variants of SSS. The first variant, SSS_a, uses semantic concept vectors with full dimensionality, i.e. all components of the concept vector are considered (also such with low values). The second variant, SSS_t, only considers the k components of the semantic concept vector with highest value, i.e. the semantic concepts being most descriptive of a text fragment. Aside from the different approach to creating the semantic concept vectors, SSS_a and SSS_t also employ different procedures for scoring semantic concept patterns. The next two sections explain the variants.

3.1.1 Variant SSS_a. Figure 2 illustrates the SSS_a approach. After constructing the semantic concept vectors with full dimensionality, SSS_a uses the cosine metric to determine the relatedness score for each vector pair. The semantic relatedness scores for all concept vector pairs are inserted into a $n \times m$ matrix spanned over all fragments in document A and document B.

Identifying patterns in the occurrence of semantic concepts in texts requires setting a similarity threshold above which to consider two semantic concept vectors a match. The vector space for semantic concept vectors typically spans several thousand or tens of thousands of dimensions. Exclusively matching identical vectors is likely too restrictive of an approach to identify any similarities except for copied text.

To find a suitable similarity threshold for semantic concept vectors and to investigate our hypothesis that plagiarized documents exhibit patterns of similar semantic concepts, we employed a visual analytics approach. Visual analytics seeks to combine the reasoning skills of humans with the data processing capabilities of computers by providing interactive data visualizations. We computed the semantic relatedness scores for the 25 confirmed cases of plagiarism and their respective source documents in our test collection (cf. Section 4.1.2). We used ESA as proposed in [13] and partitioned the documents a) into sentences and b) into paragraphs. We plotted a heatmap of the semantic relatedness scores (see Figure 3). The axes of the heatmap represent all sentences (left plot) or paragraphs

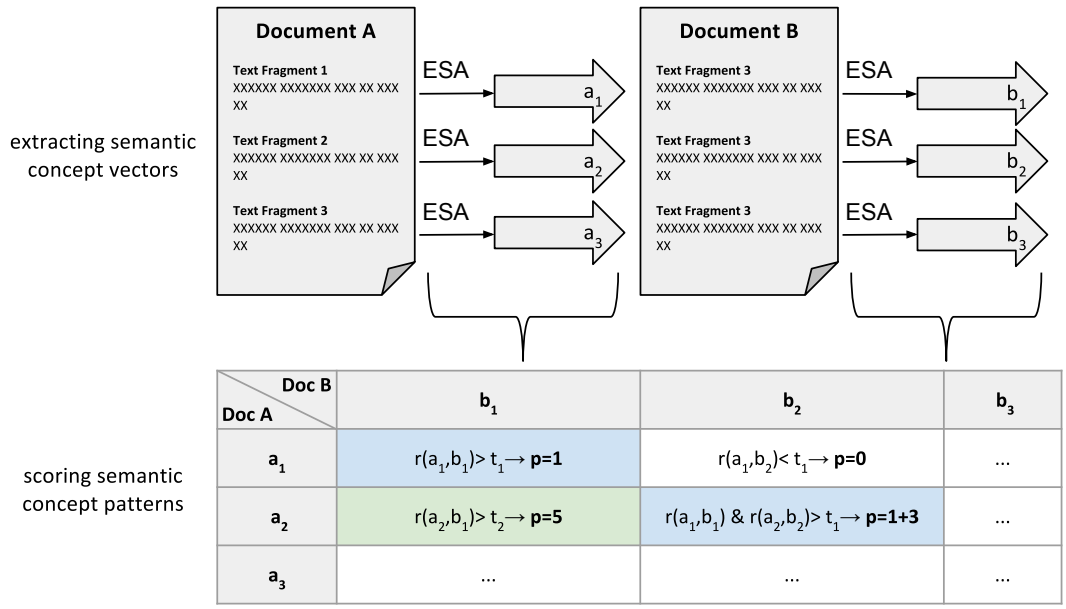

Figure 2: Semantic Sequence Scoring (variant SSS_a).

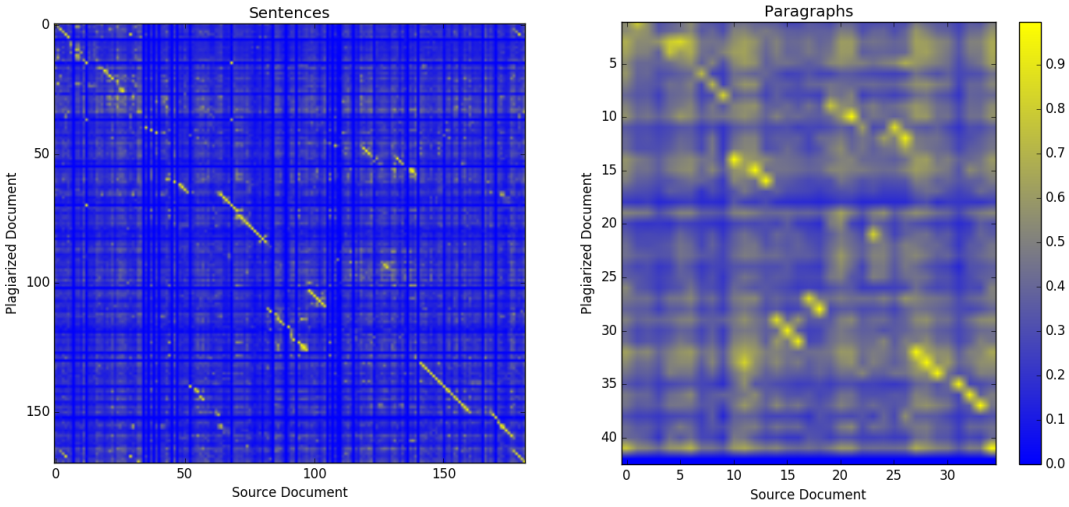

Figure 3: Heatmap of semantic relatedness scores in a plagiarized document and its source.

(right plot) in the source document (x-axis) and the plagiarized document (y-axis). The pixel color indicates the semantic relatedness score according to the scale depicted on the right side of Figure 3.

By investigating the heatmaps for several cases and selectively checking the corresponding text fragments, we derived two insights. First, patterns of similar semantic concepts are observable in many cases. For instance, in Figure 3 sequential patterns are observable, particularly in the heatmap for sentences, but also in the heatmap for paragraphs. Sequential patterns appear as accumulations of yellow pixels approximately following a negative linear function. Second, given our observations, we defined two similarity thresholds, $t_1 = .60$ and $t_2 = .75$, for semantic concept vectors. We consider vectors that exceed t_1 to be related and vectors that exceed t_2 to be highly related.

To identify and score patterns of semantic concepts, SSS_a employs the two similarity thresholds t_1 and t_2. The scoring procedure computes the pattern score $p(\vec{a}_i, \vec{b}_j)$ for each concept vector pair as follows: SSS_a assigns a score of $p = 1$ if the semantic relatedness score r of a vector pair exceeds $t_1 = .60$ and a score of $p = 5$ if r exceeds $t_2 = .75$. To account for sequences of semantically related text fragments, the score of a vector pair is increased by 3 if the semantic relatedness score $r(\vec{a}_{i-1}, \vec{b}_{j-1})$, i.e. the score in the diagonally adjacent cell in the matrix, also exceeds t_1.

3.1.2 Variant SSS_t. The semantic concept vectors used for the SSS_a variant include many components with low value. Considering such components can be useful to quantify weak semantic relations. However, for the PD use case, a focus on identifying strong semantic relatedness of text fragments seems most promising. Therefore, the SSS_t variant reduces the dimensionality of the semantic vectors formed with the help of ESA to the k most significant components, i.e. the concepts having the highest values. We experimented with different values for k and found that setting $k = 10$ yielded the best results.

SSS_t considers how many of the k (here $k = 10$) most significant concepts in the semantic concept vector for one text fragment are also among the 10 most significant concepts in the semantic concept vector of the comparison fragment. Analogously to SSS_a, SSS_t uses a matrix whose dimensions are the text fragments in the two documents under comparison. The entries of the matrix are the number of identical concepts among the top-k concepts in the vectors for each fragment pair.

SSS_t uses the score matrix and two additional heuristic thresholds for identifying and scoring consecutive sequences of semantically related text. The threshold m_{min} defines the smallest number of identical concepts in the vector representations of both text fragments to consider the text fragments related. In our experiments, we set $m_{min} = 2$. The threshold l_{min} defines the smallest number of consecutive related text fragment pairs that are considered as a sequence. Likewise, we set $l_{min} = 2$ in our experiments.

To identify sequences, SSS_t finds all scores in the matrix that exceed m_{min}. In the next step, the procedure identifies all occurrences of diagonally adjacent cells that exceed m_{min}. Occurrences that exceed $l_{min} = 2$ are considered a sequential pattern. The score p for each identified sequential pattern is calculated as the sum of all identical concepts in the vector representations that form the pattern times the length of the sequential pattern.

3.2 Concept Combination Frequency Indexing

Concept Combination Frequency Indexing (CCFI) searches for text fragments that contain rare combinations of semantic concepts. The intuition is that academic documents typically address highly specific topics. CCFI is inspired by the classical *tf-idf* weighting scheme in text retrieval and seeks to capture the semantic specificity of content. Instead of words, CCFI considers how often semantic concepts co-occur in text fragments within the collection to increase the weight assigned to rare concept combinations.

In a pre-processing step, CCFI partitions all documents in the collection into text fragments, employs ESA to determine the semantic concept vector for each fragment, and inserts the k most significant concepts (here $k = 10$) for each fragment into an inverted index. The analysis step partitions each analyzed document into text fragments and employs ESA to determine the k most significant concepts for each fragment. CCFI then forms all combinations of the top-k concepts of each fragment and queries the index for fragments that contain the specific combination of concepts. Every concept combination is assigned a score that reflects the combination's inverse collection frequency, i.e. the score is 1 if a concept combination occurs once in the collection, and 0 if the concept occurs in every fragment of the collection. The semantic relatedness score of a fragment pair is calculated as the sum of the scores for the concept combinations that occur in the fragments.

4 EVALUATION

Conclusively evaluating plagiarism detection approaches is difficult due to the covert nature of plagiarism and the lack of reliable methods for detecting disguised forms of plagiarism. Two evaluation options with inherent advantages and disadvantages exist. The first and widely accepted option is to use test collections with artificially created, "synthetic", plagiarism instances. Prominent collections of synthetic plagiarism instances are the collection used in the PAN-competitions for evaluating plagiarism detection systems [33], the collection by Clough [7], and the collection by Alzahrani [1]. Reasons for relying on evaluation frameworks that include synthetic plagiarism instead of real-world instances include:

- *The lack of ground truth data*: Academic plagiarists are highly motivated to avoid detection and meet the high quality standards of peer-reviewed journals. Plagiarism is therefore often disguised and hard to detect. The presence or absence of plagiarism in real-world collections can therefore only be approximated.
- *The bias towards less-obfuscated forms of plagiarism*: Due to the effort necessary to detect disguised forms of plagiarism and the lack of tools to support users with that task, identified cases of plagiarism typically exhibit a low level of disguise.
- *The limited reproducibility and comparability of results*: Academic documents are often subject to copyright, which prevents public sharing of test collections that include real-world cases of plagiarism. This restriction impedes comparing a new approach to the state of the art or reproducing the results of other researchers.

Despite these valid reasons for using artificially created test collections, such collections exhibit a critical disadvantage. Synthetic plagiarism instances are typically created by automated methods, e.g., using random text replacements or synonym substitutions, or non-experts, e.g., students or workers hired via crowd-sourcing platforms, such as Amazon Mechanical Turk. We argue that such plagiarism instances are typically not representative of the sophisticatedly disguised real-world plagiarism committed by experienced researchers with a strong incentive to hide their doing.

The second option for evaluating PD approaches are test-collections that includes real-world instances of plagiarism. Given that Semantic Concept Pattern Analysis is conceptually different to existing PD approaches, the goal of our initial evaluations was to gauge whether the approach holds promise to detect real-world cases of academic plagiarism. In our view, this is the crucial requirement for any new PD approach, since a variety of reliable methods for detecting less-obfuscated instances of plagiarism exist. Therefore, we chose to accept the limitations of using real-world instances of plagiarism for our initial evaluation. The following sections present the methodology and results of our evaluation.

4.1 Methodology

4.1.1 Evaluated Methods. We implemented the two variants of Semantic Sequence Scoring, SSS_a and SSS_t and the Concept Combination Frequency Indexing approach using paragraphs as the unit to partition documents ($CCFI_p$). Due to the positive results

of the CCFI approach (see Section 4.2), we additionally tested this approach using sentence partitioning ($CCFI_s$).

To compare Semantic Concept Pattern Analysis to a representative, text-based PD approach, we used the open-source tool *Sherlock*[6] (SHL). Sherlock uses word-based text chunking with probabilistic chunk selection. This detection approach – called text fingerprinting – is representative for most plagiarism detection systems available for productive use. Sherlock calculates the similarity s of two documents as: $s = \frac{100 l_s}{l_1 + l_2 - l_s}$ where l_s denotes the overall length of the passages identified as similar in both documents and l_1 and l_2 denote the lengths of the two documents.

To increase the comparability of semantic concept pattern scores to Sherlock's scores, we normalized the scores of our methods. Using each of our four methods, we compared the 25 plagiarized documents to themselves and used the resulting scores as normalization factors for the scores of the specific method.

4.1.2 Test Collection. To compile a collection of research papers that had been retracted for plagiarism, we relied on a study by Halevi and Bar-Ilan [23]. The two authors reviewed 998 retracted articles retrieved from Elsevier's full text database ScienceDirect. We restricted their collection to articles in Chemistry, Medicine, and other Life Sciences to enable acquisition of topically related full-text articles from the publicly available PubMed Central Open Access Subset[7]. Furthermore, we restricted the selection to articles, for which the text of the retraction notice contains the word "plagiarism". These restrictions retained 32 articles and their respective source documents. We excluded additional 7 cases, because we could not obtain the source document(s) or because the source documents were only available as scanned images. Thus, our test collection includes 25 retracted journal articles.

We embedded the 25 test cases in a collection of related articles retrieved by the recommender system of PubMed Central. For each of the 25 plagiarized articles, we obtained a list of 200 related articles, which we filtered for articles that are publicly available in NXML format as part of the Open Access Subset. These restrictions reduced the number of related articles per case. The fewest articles (70) were retained for case 10; the most articles (152) for case 17. The average number of related articles per case was 107. The final collection of related articles contains 2,688 documents.

4.1.3 Ground Truth. Our ground truth approximation for the 25 test cases consists of 27 documents, which expert reviewers of the respective journals have confirmed to be the source for content in the plagiarized articles. Establishing a ground truth approximation on the sub-document-level, i.e. to determine which particular content has been plagiarized, requires judgments by domain experts, which exceeds our resources. Therefore, we restrict our performance evaluation to the candidate retrieval task of the plagiarism detection process, i.e. to retrieving potential sources for content in the plagiarized documents (cf. Section 2.3).

4.1.4 Semantic Backgrounds. The main requirement for the effectiveness of ESA is a substantial overlap in the vocabularies of the knowledge base and the analyzed documents. Gottron et al. showed that using a domain-specific knowledge base corpora improves the performance of ESA for documents of that domain [21]. Anderka and Stein analyzed the influence of the corpus size on the performance of ESA and suggested that a corpus of 1,000 - 10,000 documents typically achieves a good trade-off between accuracy and computational effort.

To consider these findings of previous research on ESA for our use case, we tailored the semantic background to the domains of the articles in our test collection. We also experimented with two different sizes for the knowledge base corpus to explore if and to what degree detection effectiveness increases with increasing size of the corpus. We compiled the two semantic backgrounds by extracting Wikipedia articles, i.e. concepts, from the Wikipedia categories Biology, Chemistry and Medicine. Within these categories, we traversed and included articles up to a maximum depth of two levels below the main category for the smaller background and up to three levels below the main category for the larger background. This procedure yielded the following two semantic backgrounds:

- *Background 1*: 2,620 articles, 53,623 index words
- *Background 2*: 53,636 articles, 136,831 index words

4.1.5 Performance Metrics. For 23 of the 25 test cases, the ground truth approximation is limited to one known item of relevance; for the other two test cases to two relevant items. Thus, we essentially evaluate our approach in performing a *known item retrieval* task. For such tasks, precision-related performance metrics provide little information, since precision is essentially reduced to a binary figure. Therefore, the rank at which the relevant item is retrieved is most descriptive of the effectiveness of a retrieval approach [6].

To quantify the retrieval effectiveness of an approach, we report the *Mean Reciprocal Rank* $\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$. MRR is the average of the reciprocal ranks at which each query q in a set of queries Q retrieves the first relevant item. In our case, the 25 plagiarized documents are the queries. A detection method would achieve the best possible score of 1 if it retrieves a source document at rank 1 for each test case. To quantify overall retrieval success, we report recall at rank 5, i.e. the fraction of all source documents that a detection method identifies among its five top-ranked results.

4.2 Results

A first insight of our evaluation is that the larger semantic background achieved significantly better retrieval effectiveness than the smaller semantic background. Due to space limitations, we only report the results obtained using the larger background.

Table 1 shows the Mean Reciprocal Rank and recall at rank 5 for the evaluated detection methods. The table indicates that SSS_a, which extends native ESA with a heuristic scoring function, is clearly outperformed by all other methods. SSS_a achieves poor recall (.59) and the worst ranking performance (MRR = .50). We assume that using semantic concept vectors of full dimensionality entails too much noise to reliably distinguish potentially suspicious similarity in semantic content from legitimate semantic relatedness among articles in the same research area.

[6] http://www.cs.usyd.edu.au/ scilect/sherlock/
[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

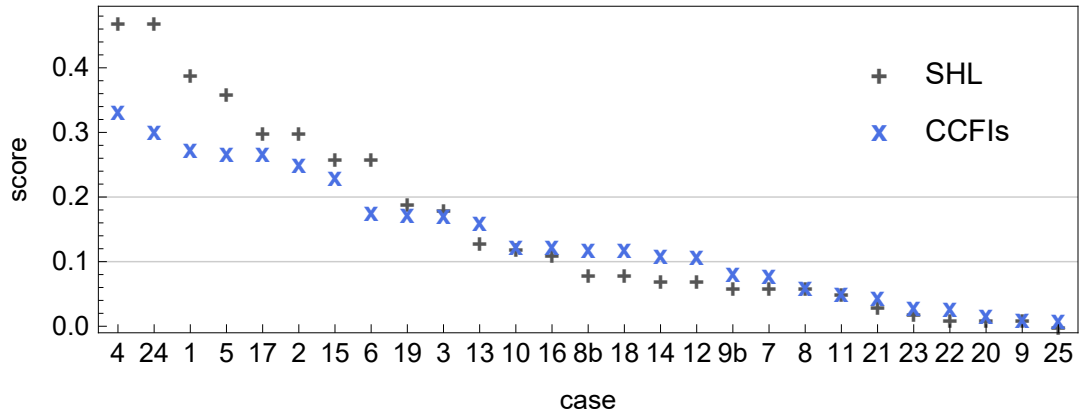

Figure 4: Similarity scores of Sherlock and CCFI_s for all source documents.

Table 1: Mean Reciprocal Rank and recall at rank 5.

Method	MRR	R@5
SSS_a	.50	.59
SSS_t	.72	.85
CCFI_p	.78	.81
CCFI_s	.79	.81
SHL	.85	.89

SSS_t, which considers only the 10 most significant components of the concept vectors and assigns higher weights to concept sequences than SSS_a, achieves the best recall (.85) of all semantic methods and a notably better MRR (.72) than SSS_a.

The results of CCFI, which achieved the best MRR performance of all semantic detection methods, also indicate that focusing on the most significant semantic concepts of a text fragment is most promising for PD. Given the good results of CCFI on paragraph level (CCFI_p), we tested whether the performance of the approach can be further increased when partitioning documents into sentences. However, the performance increase (MRR +.01) of applying CCFI on sentence level (CCFI_s) is negligible.

The good performance of Sherlock (SHL), which achieved the best MRR (.85) and recall (.89), is partially attributable to limitations of our exploratory evaluation. In this first evaluation of Semantic Concept Pattern Analysis we wanted to explore the behavior of the scoring heuristics we devised and how they reflect different forms and levels of similarity in academic documents. Therefore, we did not impose thresholds for the similarity scores of our methods.

To create equal conditions, we deactivated the similarity threshold in Sherlock. By default, the threshold is .20, i.e. Sherlock does not retrieve documents with lower scores. With this threshold deactivated, Sherlock retrieved 31, 145, and 12 documents with a score of 1 for the case 20, 22, and 9 (among them the correct source documents). Also in other cases Sherlock retrieved multiple documents at the same rank.

Figure 4 plots the similarity scores that Sherlock (SHL) and the best performing semantic detection method (CCFI_s) assigned to each of the 27 source documents. The cases are ordered according to Sherlock's similarity score. Only for 8 of the 27 documents Sherlock assigned a score that exceeds the tool's default threshold of .20. The remaining 19 source documents would have been disregarded.

Overall, a correlation between text-based and semantic-based similarity scores is observable. For the 8 documents with high textual similarity (Sherlock scores $s > .20$), Sherlock's text-based approach performs better in identifying these documents within the collection. However, for documents with Sherlock scores between $.20 \leq s \leq .10$ i.e. with low textual similarity, (see horizontal lines in Figure 4), CCFI_s assigns a higher similarity score.

Checking text fragments with high semantic concept pattern scores in documents with low textual similarity confirmed that semantic-based detection approaches reflect similarity in such cases better than text-based similarity measures. Visualizing paragraphs with high semantic relatedness provided a notable benefit over visualizing literal text matches to identify paraphrased text.

Clearly, this first evaluation of Semantic Concept Pattern Analysis only provides initial circumstantial evidence for the strengths of the approach and leaves much room for future improvement and more comprehensive evaluation. Nevertheless, we expect that Semantic Concept Pattern Analysis can help to increase the detection capabilities for instances of plagiarism with low textual similarity. We explain our plans for improving and more comprehensively evaluating Semantic Concept Pattern Analysis in the next Section.

5 CONCLUSION AND FUTURE WORK

We present Semantic Concept Pattern Analysis as a new approach to improve the detection capabilities for paraphrased instances of plagiarism. The approach combines Explicit Semantic Analysis with a heuristic assessment of structural document similarity. Our initial evaluation using 25 retracted cases of plagiarism demonstrated that Semantic Concept Pattern Analysis can help to identify documents whose textual similarity is too low to raise suspicion during an analysis with established text-based methods.

In future research, we plan to improve the scoring functions for concept patterns and evaluate the approach in more detail. Given our initial results, normalizing the pattern scores by the identity score of a document does not truthfully reflect the subjective similarity in semantic content we observed in the documents. Assigning additional weight to rarely co-occurring concepts and to concept

sequences as well as rethinking the normalization procedure seems promising to improve the ability of semantic pattern scores to more clearly indicate potentially suspicious semantic similarity. Additionally, we need to evaluate Semantic Concept Pattern Analysis more comprehensively to define suitable thresholds for the similarity scores computed by the approach. Deriving these thresholds requires: i) embedding test cases in a significantly larger collection to better understand the characteristics of legitimate and potentially suspicious semantic pattern similarity, ii) obtaining a balanced amount of test cases for specific forms of plagiarism, iii) obtaining a ground truth approximation on the sub-document level.

Requirement i) is easy to accomplish, e.g., by using more documents from the PubMed Central Open Access Subset. Requirement ii) can be achieved by reviewing more retractions, e.g., from the collection of Halevi and Bar-Ilan [23]. Requirement iii) is hard to accomplish, since reviewing and annotating cases requires substantial efforts by domain experts. The crowd-sourced project VroniPlag[8] offers real-world plagiarism cases that were manually annotated on the text passage level. However, since those cases originate from different domains, compiling a suitable collection to embed the cases and gathering a suitable semantic background requires effort. Although using a collection of real-world cases of plagiarism is desirable (cf. Section 4) resorting to collections with synthetic instances of plagiarism, such as the PAN-PC corpus [33], may help to improve Semantic Concept Pattern Analysis.

Our long-term goal, as described in [26], is to embed Semantic Concept Pattern Analysis as a component of an integrated detection process. Our research indicates that not a single, but combined PD approaches are most promising to reliably detect the many possible forms of plagiarism ranging from blatant copying to strongly disguised idea plagiarism [15]. The idea is to accumulate evidence on potentially suspicious similarity using heterogeneous similarity features. The integrated detection process will analyze literal text matches, academic citations, images, mathematical content as well as semantic and syntactic features. Including a wide range of similarity features increases the effort required for hiding plagiarism, hence increases the deterrent effect of PD systems and thus helps to prevent plagiarism in the first place.

REFERENCES

[1] Salha Alzahrani, Vasile Palade, Naomie Salim, and Ajith Abraham. 2011. Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications. *JASIST* 63(2) (2011), 286–312.

[2] Maik Anderka and Benno Stein. 2009. The ESA Retrieval Model Revisited. In *Proc. SIGIR*. 670–671.

[3] Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.* 32, 1 (2006), 13–47.

[4] Curt Burgess. 1998. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers* 30, 2 (1998), 188–198.

[5] Zdenek Ceska. 2008. Plagiarism Detection Based on Singular Value Decomposition. In *Advances in NLP*. LNCS, Vol. 5221. 108–119.

[6] Paul Clough and Mark Sanderson. 2013. Evaluating the Performance of Information Retrieval Systems using Test Collections. *Informat. Res.* 18, 2 (2013).

[7] Paul Clough and Mark Stevenson. 2009. Creating a Corpus of Plagiarised Academic Texts. In *Proc. Corpus Linguistics Conf.*

[8] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JASIST* 41, 6 (1990), 391.

[9] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based Information Retrieval using Explicit Semantic Analysis. *TOIS* 29, 2 (2011), 8.

[10] Teddi Fishman. 2009. "We know it when we see it"? is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright. In *Proc. Asia Pacific Conf. on Educational Integrity.*

[11] Evgeniy Gabrilovich and Shaul Markovitch. 2005. Feature Generation for Text Categorization using World Knowledge. In *Proc. IJCAI*, Vol. 5. 1048–1053.

[12] Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, Vol. 6. 1301–1306.

[13] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis.. In *IJcAI*, Vol. 7. 1606–1611.

[14] Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based Semantic Interpretation for Natural Language Processing. *J. Artif. Intell. Res.* 34 (2009), 443–498.

[15] Bela Gipp. 2014. *Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis.* Springer.

[16] Bela Gipp and Norman Meuschke. 2011. Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In *Proc. DocEng*. 249–258.

[17] Bela Gipp, Norman Meuschke, and Joeran Beel. 2011. Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. In *Proc. JCDL*. 255–258.

[18] Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. *JASIST* 65, 2 (2014), 1527–1540.

[19] Bela Gipp, Norman Meuschke, Corinna Breitinger, Mario Lipinski, and Andreas Nuernberger. 2013. Demonstration of Citation Pattern Analysis for Plagiarism Detection. In *Proc. SIGIR*.

[20] Wael H Gomaa and Aly A Fahmy. 2013. A Survey of Text Similarity Approaches. *Int. J. of Computer Applications* 68, 13 (2013).

[21] Thomas Gottron, Maik Anderka, and Benno Stein. 2011. Insights into Explicit Semantic Analysis. In *Proc. CIKM*. 1961–1964.

[22] Jorge Gracia and Eduardo Mena. 2008. Web-based Measure of Semantic Relatedness. In *Proc. Int. Conf. on Web Informat. Sys. Eng*. 136–150.

[23] Gali Halevi and Judit Bar-Ilan. 2016. Post Retraction Citations in Context. In *Proc. BIRNDL Workshop at JCDL*. 23–29.

[24] Michael D Lee, DAniel J Navarro, and Hannah Nikkerud. 2005. An Empirical Evaluation of Models of Text Document Similarity. In *Proc. of the Cognitive Science Society*, Vol. 27.

[25] Norman Meuschke and Bela Gipp. 2013. State of the Art in Detecting Academic Plagiarism. *Int. J. for Educational Integrity* 9, 1 (2013), 50–71.

[26] Norman Meuschke and Bela Gipp. 2014. Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space. In *Proc. JCDL*. 197–200.

[27] Rada Mihalcea, Courtney Corley, Carlo Strapparava, and others. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *AAAI*, Vol. 6. 775–780.

[28] Jane Morris and Graeme Hirst. 2004. Non-classical Lexical Semantic Relations. In *Proc. of the HLT-NAACL Ws. on Computational Lexical Semantics*. Assoc. for Computat. Linguist., 46–51.

[29] Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Ssennoga Twaha, Yogan Jaya Kumar, and Albaraa Abuobieda. 2012. Plagiarism detection scheme based on Semantic Role Labeling. In *Proc. Int. Conf. on Information Retrieval Knowledge Management*. 30–33.

[30] Merin Paul and Sangeetha Jamal. 2015. An improved SRL based plagiarism detection technique using sentence ranking. *Proc. Comput. Sc.* 46 (2015), 223–230.

[31] Maria Soledad Pera and Yiu-Kai Ng. 2011. SimPaD: a Word-Similarity Sentence-Based Plagiarism Detection Tool on Web Documents. *Web Intelligence and Agent Sys.* 9, 1 (2011), 24–41.

[32] Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. In *Proc. ECIR*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). 522–530.

[33] Martin Potthast, Benno Stein, Alberto Barrón Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proc. ACL*. 997–1005.

[34] Salha Alzahrani and Naomie Salim. 2010. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In *CLEF 2010 Notebook Papers.*

[35] David Sánchez, Montserrat Batet, David Isern, and Aida Valls. 2012. Ontology-based Semantic Similarity: A new Feature-based Approach. *Expert Systems with Applications* 39, 9 (2012), 7718–7728.

[36] Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. 2007. Strategies for Retrieving Plagiarized Documents. In *Proc. SIGIR*. 825–826.

[37] K. Vani and Deepa Gupta. 2016. Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *J. Engin. Sc. & Techn. Review* 9, 5 (2016).

[38] Debora Weber-Wulff. 2014. *False Feathers: A Perspective on Academic Plagiarism.* Springer.

[8] http://vroniplag.wikia.com

Citation for this Paper

Citation Example:

N. Meuschke, N. Siebeck, M. Schubotz, and B. Gipp. Analyzing semantic concept patterns to detect academic plagiarism. In *Proceedings International Workshop on Mining Scientific Publications (WOSP) held in conjunction with the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017.

Bibliographic Data:

RIS Format	**BibTeX Format**
Y - CONF AU - Meuschke, Norman AU - Siebeck, Nicolas AU - Schubotz, Moritz AU - Gipp, Bela T1 - Analyzing Semantic Concept Patterns to Detect Academic Plagiarism T2 - Proceedings International Workshop on Mining Scientific Publications (WOSP) held in conjunction with the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) AD - Toronto, Canada Y1 - 2017	@InProceedings{Meuschke2017, author = {Meuschke, Norman and Siebeck, Nicolas and Schubotz, Moritz and Gipp, Bela}, title = {Analyzing Semantic Concept Patterns to Detect Academic Plagiarism}, booktitle = {{P}roceedings {I}nternational {Workshop} on {M}ining {S}cientific {P}ublications ({WOSP}) held in conjunction with the {ACM/IEEE-CS J}oint {C}onference on {D}igital {L}ibraries ({JCDL})}, year = {2017}, location = {Toronto, Canada} }