Journal Name manuscript No.
(will be inserted by the editor)

Paraphrase Plagiarism Identification with Character-level Features

Fernando Sánchez-Vega · Esaú Villatoro-Tello · Manuel Montes-y-Gómez · Paolo Rosso · Efstathios Stamatatos · Luis Villaseñor-Pineda

Received: date / Accepted: date

Abstract Several methods have been proposed for determining plagiarism between pairs of sentences, passages or even full documents. However, the majority of these methods fail to reliably detect paraphrase plagiarism due to the high complexity of the task, even for human beings. Paraphrase plagiarism identification consists in automatically recognizing document fragments that contain re-used text, which is intentionally hidden by means of some rewording practices such as semantic equivalences, discursive changes, and morphological or lexical substitutions. Our main hypothesis establishes that the original author's writing style fingerprint prevails in the plagiarized text even when paraphrases occur. Thus, in this paper we propose a novel text representation scheme that gathers both *content* and *style* characteristics of texts, represented by means of character-level features. As an additional contribution, we describe the methodology followed for the construction of an appropriate corpus for the task of paraphrase plagiarism identification, which represents a new valuable resource to the NLP community for future research work in this field.

F. Sánchez-Vega (✉) · M. Montes-y-Gómez · L. Villaseñor-Pineda
INAOE, Luis Enrique Erro No.1, Tonantzintla,
Puebla 72840, México
E-mail: fer.callotl@ccc.inaoep.mx

E. Villatoro-Tello (✉)
Information Technologies Dept.,
UAM Cuajimalpa, México, Mexico City.
E-mail: evillatoro@correo.cua.uam.mx

P. Rosso
PRHLT Research Center,
Universitat Politècnica de València, Spain.

E. Stamatatos
Dept. of Information & Communication Systems Engineering,
University of the Aegean, Greece.

Keywords Plagiarism identification · Paraphrase Plagiarism · Text Reuse · Character n-grams · Stylistic representation

1 Introduction

Plagiarism is known as intellectual theft: it consists in using words (ideas) of others and presenting them as your own [34]. Nowadays, the phenomenon of plagiarism is growing very rapidly given the easiness for sharing and using information extracted from electronic media. In particular, it is very frequent in the journalism and academia fields, where journals' editors, teachers, etc., habitually employ commercial software aiming at detecting cases of plagiarism.

Although current technologies are very effective detecting the verbatim type of plagiarism (*i.e.,* copy-paste), most of them fail in detecting cases where the plagiarists attempt to hide the similarity with the original document by modifying the plagiarized text fragments. This type of plagiarism, generated by means of applying some rewording operations to the original text, is known as *paraphrase plagiarism*. More specifically, we will refer to paraphrase plagiarism as the act of taking text fragments containing (original) ideas from others and present them as your own after performing any type of paraphrase operations such as: semantic or lexical equivalences (*e.g.*, the substitution of words by synonyms or related words), syntactic or discursive changes (*e.g.*, active to passive voice conversions), and morphological substitutions (*e.g.*, affixes modification).

1.1 The plagiarism identification problem

Generally speaking, a complete automatic plagiarism detection scenario involves two main steps: *i)* the retrieval of candidate source texts for a given suspicious document, and, *ii)* the identification of all likely plagiarized text fragments in the suspicious document, and their corresponding passages in the source documents. Each one of these steps has its own particular goals and difficulties [33][1]. First, the retrieval sub-task has been mainly approached by using information retrieval models suited to the particular conditions of the problem [21, 23]. Then, once a set of candidate source documents have been retrieved, the second subtask focuses on measuring their *textual similarity* with the suspicious document. The documents are segmented to process the passage alignment based on an exhaustive pair comparison process [13]. At the end, collected similarities at this fine grained level will help in concluding if a suspicious document is in fact or not result of plagiarism and to what extent.

Given the importance of the second subtask, the plagiarism identification task, current research has focused on proposing different methods to measure the similarity between *two* text fragments. However, as detailed in Section 2, the main drawback of these approaches is that they carry out the plagiarism

[1] The PAN competition (`http://pan.webis.de`)

decision considering only information about the degree of *content overlap* between the suspicious and source texts. Hence, these strategies are affected by the thematic correspondence of the texts, which implies the existence of common domain-specific word sequences, and causes an overestimation of their overlap [9]. Additionally, if some paraphrases exist within the suspicious text, these techniques face serious difficulties when determining the existence, or not, of plagiarism.

In addition to the content comparison, according to [31] measuring structural similarities reveals important information when detecting plagiarism; such information exposes (to some extent) the author's *writing style.* Thus, our intuition is that when paraphrasing, a plagiarist can easily and successfully obfuscate re-used passages, since it is possible to mislead current technologies by means of modifying/replacing words or phrases by synonyms, which eventually leads to syntactic modifications. However, it is very hard to completely hide some of the original author's writing style, since its obfuscation would involve a major change in the entire text.

1.2 About our proposal

In this paper we propose using character-level features for identifying paraphrase plagiarism on passage level. Thus, we consider that documents have been segmented into passages (i.e. paragraphs) as is performed in [24,28,30].

Currently, character n-grams are the single most successful features in authorship attribution [14,17,18,26,32]. These works have shown that character n-grams provide an excellent trade-off between sparseness and information content, while at the same time they combine different types of information: punctuation, morphology, lexicon and even context [12]. Hence, we aim at exploring the pertinence of using all this information by means of defining different categories of short character n-grams [25]. Although character n-grams have been used in the past for plagiarism identification [27,29], these works are based on long character n-grams ($n > 9$) capturing aleatory text chunks, nevertheless they are not a real proposal where character-level features are evaluated. The small character n-grams (3 and 4-grams) were only used for Cross-language Plagiarism Detection (CLPD) [20]

For evaluating automatic paraphrase plagiarism identification methods we need an appropriate corpus containing both positive and negative cases, *i.e.,* examples of both *plagiarism* and *not-plagiarism.* For the construction of such corpus, we started from the corpus P4P[2] [2]. Our contribution here was the inclusion of *not-plagiarism* examples, *i.e.,* including pairs of texts samples with likely thematic or stylistic similarity. This new enriched corpus represents a valuable contribution to the NLP community given that it may be used for future research work in this field. An additional contribution of this paper is

[2] The P4P corpus and guidelines used for its annotation are available at `http://clic.ub.edu/corpus/en/paraphrases-en`

the evaluation of the effectiveness of several current state-of-the-art methods in the posed task.

Our goal in this paper is two-fold: on the one hand we want to provide evidence of the effectiveness of character-level features in the task of paraphrase plagiarism identification; on the other hand we want to investigate what type of information is captured by different categories of character n-grams and how useful is this information for solving the posed task. Particularly, the research questions we aim to answer are:

Q1: How effective are current state-of-the-art plagiarism detection methods for the problem of paraphrase plagiarism identification?

Q2: What type of features, word-level or character level, are more effective for the problem of paraphrase plagiarism identification?

Q3: How effective is the proposed representation on detecting specific types of paraphrases?

Q4: What type of character n-grams are the most informative for solving the problem of paraphrase plagiarism identification?

1.3 Structure of the paper

The rest of the paper is organized as follows. Section 2 describes recent works in plagiarism detection. Section 3 describes the construction of the evaluation corpus and gives some examples of both positive and negative cases of paraphrase plagiarism. Next, Section 4 explains the proposed method based on short character n-grams, and Section 5 presents the obtained results; it is in this section where our posed research questions are answered. Finally, Section 6 depicts our conclusions and some future work directions.

2 Related Work

The comparison process between the suspicious document and each candidate source text is usually done at fragment level in order to provide detailed evidence for the plagiarism case [13]. Over the years, many approaches have addressed this issue by measuring the lexical and structural similarity of texts by means of different kinds of features such as single words [10,37], fixed length substrings (*i.e.*, word $n-$grams) [1,10,16,29], variable length substrings [3,10], and dependency relations or a combination of them through comparing syntactic structures [6,8,15]. These techniques are able to accurately detect the verbatim case of plagiarism (*i.e.*, copy-paste) as well as some simple cases of paraphrasing; for example, reordering words practices, which is a type of syntactic paraphrase. However, they are unable to detect more complex cases of paraphrases including lexical or semantic substitutions.

As it is possible to infer, detecting more complex cases of paraphrases represents a more challenging task for automatic methods; they have to be able to measure the semantic overlap between texts. Accordingly, some works that have been proposed in the field of *semantic textual similarity* consider the use of external knowledge resources such as WordNet[3] for determining the semantic proximity of texts [4,11]. Particularly, [4] defines a method that assigns a degree of semantic similarity by measuring the path distance between every pair of words from the texts. A similar approach is the work proposed by [11], where the information extracted from WordNet in combination with several word occurrence statistics is employed to identify paraphrases practices. Although these two methods have been widely applied for measuring the degree of paraphrases between two given texts, just the work described in [22] evaluates its pertinence on the plagiarism detection problem.

From another perspective, the research works described in [29,31] employ different sets of features for plagiarism detection. Particularly, in the work proposed by [29] character n-grams are used to make a very raw and fast comparison of text passages, where, by means of using long character n-grams ($n = 10$), it is possible to accurately detect big verbatim plagiarism cases. In the work of [31], a method for determining structural similarities among documents by means of stopword n-grams is proposed, indicating (to some extent) that when plagiarism occurs, structural similarities are preserved among suspicious and source texts. Another study using structural information is the one described in [35], having as major weakness that is not an efficient method.

From the analysis of the related work, we identified three main problems: *i)* overestimation of the overlap due to thematic correspondences; *ii)* high dependency on external knowledge resources; and, *iii)* the assumption of similar syntactic forms between the suspicious and source documents. Thence, in order to face these problems, we propose using a representation based on different categories of short character n-grams, which aims at measuring distinct types of information that helps in detecting different types of paraphrase operations.

3 Construction of the Evaluation Corpus

In this section we describe the construction process and main characteristics of the compiled corpus. First, Section 3.1 briefly describes the original P4P corpus. Next, Section 3.2 explains the methodology followed for incorporating not-plagiarism examples in the P4P corpus. Finally, Section 3.3 provides some examples of the included examples as well as some statistics of the extended corpus.

3.1 The P4P corpus

The P4P corpus [2] already had positive paraphrase plagiarism cases. This corpus contains pairs of text fragments where one fragment represents the original source text, and the other represents a paraphrased version of the

[3] A large lexical database of English (`https://wordnet.princeton.edu/`).

original. The pairs of text fragments comes from PAN-PC-10 corpus[4]. The paraphrased versions were obtained from asking several volunteers to manually construct a paraphrased version of the original fragment by crowdsourcing via Amazon Mechanical Turk [5][5].

The P4P corpus represents a high quality corpus including cases of paraphrase plagiarism. It contains 847 manually constructed paraphrase cases, which have been reviewed and categorized by expert linguists. Such examples of paraphrases were tagged into their corresponding types of paraphrases phenomenon. Authors of the P4P corpus [2] employed a paraphrases typology, which includes four general classes, four sub-classes and nineteen types of paraphrases. For our purposes we focused on the second categorization level of paraphrases types[6], namely: *morphology, lexicon, syntax, discourse, semantic* and *miscellaneous* changes. Next, we provide a brief description of the different categories of paraphrases contained in the P4P corpus.

- *Morphology-based changes* include inflectional changes (*e.g.,* affixes modification), modal verb modification (*e.g., might* → *could*) and derivation changes.
- *Lexicon-based changes* comprise modifications such as synthetic and analytic reconstruction, spelling and format change, polarity substitutions and converse substitutions; in general these types of changes alter only one lexical unit within a sentence preserving the original meaning.
- *Syntax-based modifications* cause structural changes in a sentence, allowing to have the same meaning but redirecting the main focus to different elements within the sentence; paraphrase types included in this category are: diathesis alterations, negation switching, ellipsis, coordination changes and subordination with nesting changes.
- *Discourse-based modifications* alter the sentences' form and order; they include changes in punctuation marks, modifications in the syntactic structure, modality changes as well as some direct or indirect style alternations.
- *Semantic-based changes* consider modifications involving substitution of some elements within a sentence that results in lexical and syntactical modifications without interfering with the original meaning of the sentence. Semantic-based changes represent the highest level of modifications.
- *Miscellaneous-based changes* recollect all types of modifications that do not correspond to specific linguistic paraphrase phenomena, such as addition, deletion or changing the order of lexical units.

[4] `http://www.uni-weimar.de/en/media/chairs/webis/corpora/pan-pc-10/#webis-download`

[5] Workers from the Amazon Mechanical Turk (`https://www.mturk.com/mturk/welcome`) perform simple tasks in exchange for a monetary reward.

[6] The second categorization level consists of four sub-classes and two classes without subclasses

3.2 Obtaining not-plagiarism examples for the P4P corpus

We addressed the task of expanding the P4P corpus by including not-plagiarism cases, *i.e.,* pairs of unrelated texts samples with likely thematic or stylistic similarity. For this, we established three different conditions for the inclusion of not-plagiarism examples, namely: *form*, *content* and *author-controlled* conditions. By fulfilling these conditions it is guaranteed to obtain relevant and non-trivial examples of not-plagiarism cases.

Form condition. Similarly to plagiarism examples, not-plagiarism examples must represent complete sentences (ideas), *i.e.,* not-plagiarism examples do not represent random text chunks, extracted from aleatory text fragments. Additionally, the size (in characters) of the not-plagiarism examples must be very close to the average length of the source text fragment. This condition guarantees that both source and suspicious texts look alike in terms of their length and structure.

Content condition. It establishes that every not-plagiarism example must contain a considerable content overlap against the source text fragment. This condition emulates a real-life scenario, where a pair of texts on a similar topic (*i.e.*, using similar words) is not necessarily a case of paraphrase plagiarism.

In order to accomplish this condition we referred to the original documents from the PAN-PC-10 corpus. We identified the original document from where each suspicious text fragment was generated, and then we extracted the text fragments that maximize the *discursive* and *thematic* similarities.

a. *Discursive similarity*: These examples are extracted from the adjacent paragraphs to the source text fragment (see Figure 1). By following this strategy we guarantee that the selected not-plagiarism examples have enough thematic overlap as well as a similar writing style, since they belong to the same discursive formation.
b. *Thematic similarity*: For this, we search for the text fragment that has the highest thematic similarity against the source text fragment (*e.g.,* "Text fragment A" in Figure 1a). To compute this similarity we employed a traditional lexical overlap measure, namely the Dice's coefficient.

At the end, all of these types of text fragments (A,B and C in Figure 1b) are considered to form the set of not-plagiarism examples.

Author-controlled condition. This condition dictates that the not-plagiarism examples must belong to the same author of the original text fragment. To fulfill this condition we extracted the not-plagiarism examples from the corresponding source document (see Figure 1a).

As we mentioned in Section 3.1, positive paraphrase plagiarism examples were manually generated from source text fragments and, therefore, a special stylistic mark is preserved in the paraphrasing process [36]. By means of accomplishing the author-controlled condition, we imposed difficulties to

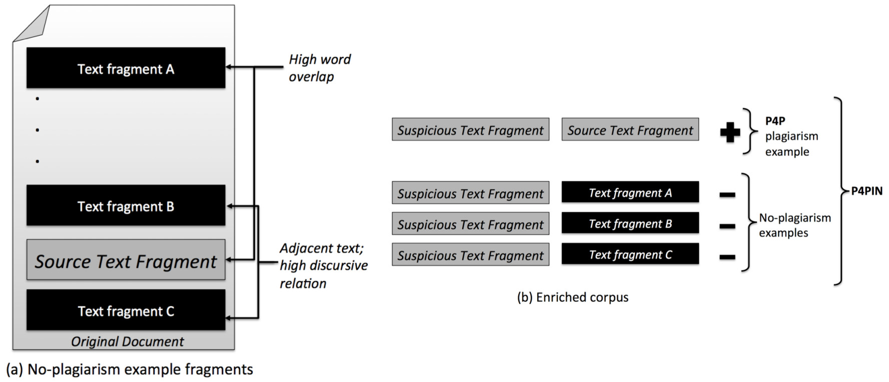

Fig. 1 Graphical representation of the text fragments used as not-plagiarism examples. Figure 1a shows text fragments that are good candidates for being not-plagiarism examples. Figure 1b depicts the final composition of the P4PIN corpus.

plagiarism identification techniques that are based on stylistic features since text fragments pairs belonging to the not-plagiarism class use a similar writing style to those belonging to the plagiarism class.

3.3 The P4PIN corpus

The final corpus contains 847 plagiarism examples from the P4P corpus and 2507 not-plagiarism cases, where 1660 are discursive related and 847 are thematic related cases. We henceforth refer to this extended version of the P4P corpus as **P4PIN**, where the suffix **IN** refers to ***I**ncluding **N**egative* examples[7].

Table 1 exhibits a pair of paraphrase plagiarism examples from the original P4P corpus. Each example (*i.e.,* each column) contains both the original source text fragments as well as its respective paraphrased version. The types and general categories of the applied paraphrases are indicated below of the text fragments. Finally, last row indicates the percentage of content overlap. Notice that example #1 suffered from more changes (*i.e.*, four types from three different categories of paraphrases) in comparison with example # 2 (only Lexicon-based changes); nevertheless the content overlap is greater in example # 1 than in example # 2. It is also important to remark that in both examples some style information is preserved (*e.g.,* verbal tenses and discursive style, first person in the first example and impersonal style in the second), indicating to some extent that in fact, when paraphrasing, plagiarists retain some characteristics of the original author's writing style.

On the other hand, Table 2 shows two not-plagiarism examples. These examples exhibit cases where the percentage of common words between the

[7] Available at: `http://ccc.inaoep.mx/~mmontesg/resources/corpusP4PIN.zip`

Table 1 Text fragments showing plagiarism examples from the P4P corpus. Underlined words represent common words between the original and the suspicious texts; below each column, it is indicated the sub-type and type of paraphrase phenomena as well as the percentage of common words between text fragments.

	Plagiarism Example # 1	**Plagiarism Example # 2**
Original	I pored through these pages, and as I perused the lyrics of The Unknown Eros that I had never read before, I appeared to have found out something wonderful: there before me was an entire shining and calming extract of verses that were like a new universe to me.	The sincere compassion of the country has been articulated in many forms, and with ever-lasting impact, and has decorated the resting place of those who have gone before us in the Northwest wilderness. Allow me to add one blossom to the wreath.
Suspicious	I dipped into these pages, and as I read for the first time some of the odes of The Unknown Eros, I seemed to have made a great discovery: here was a whole glittering and peaceful tract of poetry which was like a new world to me.	The heartfelt sympathy of the country has been ex-pressed in many forms, and ever with deep effect, and has twined a garland to drop upon the graves of those who sleep to-night away in the wilds of the North-West. Permit me to add one flower to that chaplet.
Type	Semantic changes Same Polarity substitution Synthetic/analytic substitution Addition/deletion	Spelling and format Same-polarity substitution Synthetic/analytic substitution
Category	Semantic Miscellaneous Lexical	Lexical
Common words	57.4%	48.3%

suspicious and the original text fragments is very high, around 50%, very alike to the plagiarism cases showed in Table 1.

By analyzing the plagiarism and not-plagiarism examples (Tables 1 and 2 respectively), it is possible to visualize the inherent difficulties involved in the problem of paraphrase plagiarism identification. Shown examples illustrate the main difficulties that either purely thematic or purely stylistic methods will face when dealing with a real paraphrase plagiarism evaluation corpus.

4 Proposed Method

The proposed method relies on representing a pair of text fragments by means six features. These features depict similarities of content and style, estimated at character-level, between the analyzed pair of text fragments. Particularly, we propose representing the relation between an original text (t_o) and a suspicious text (t_s) through their overlaps across different categories of short character n-grams.

We characterize each text fragment (*i.e.* both t_o and t_s) by means of a *Bag of Character n-grams*, henceforth referred as BoC. During the construction

Table 2 Text fragments showing not-plagiarism examples (*i.e.,* pairs of texts fragments with likely thematic or stylistic similarity) in the evaluation P4PIN corpus. Underlined words represent common words between the original and the suspicious document; below each column, it is indicated the percentage of common words between text fragments.

	Not-plagiarism Example # 1	**Not-plagiarism Example # 2**
Original	Since the man was wearing all of his clothes, Sir Horace was able to ascertain that the man was murdered before he went to bed and before Burchill broke into the house. All of this evidence shows that the murder was committed before dark.	The fact that an omnipresent God exists is the one universal factor that governs the laws of nature. God has set in place the laws of the universe for His own purposes.
Suspicious	Your only alternatives to that conclusion are that the murdered man went to bed with his clothes on, or that the murderer broke into the house before Sir Horace had gone to bed and after killing Sir Horace went coolly round the house turning out the lights instead of fleeing in terror at his deed without even waiting to collect any booty.	The laws of nature are the art of God. Without the presence of such an agent, one who is conscious of all upon which the laws of nature depend, producing all that the laws prescribe. The laws themselves could have no existence.
Common words	48.3%	54.8%

of such characterization approach we do not discard any symbol. In other words, all alphanumeric characters, blank spaces and punctuation marks are considered for the construction of the BoC. Given that we are working with short text fragments (about one paragraph), we did not considered character n-grams frequency values, instead we focus on the presence/absence of character n-grams, meaning that we employed a binary weighting scheme.

Once we have extracted all character n-grams for the pair of text fragments t_o and t_s, we build their joint representation using a total of six features, one for each n-gram category ($\text{cat}_1 \rightarrow \text{cat}_5$) and one for the general BoC which considers all n-grams of size n. The proposed representation is as follows:

$$(t_o, t_s) = \langle f^n_{\text{cat}_1}, f^n_{\text{cat}_2}, \ldots, f^n_{\text{cat}_5}, f^n_{\text{all}} \rangle \tag{1}$$

Each feature value ($f^n_{\text{cat}_i}$) represents the similarity of both texts calculated over the category cat_i of the BoC. This similarity is computed using the *Dice* coefficient:

$$\text{sim}_{\text{cat}_i}(t_s, t_o) = \frac{2|\text{BoC}_{\text{cat}_i}(t_s) \cap \text{BoC}_{\text{cat}_i}(t_o)|}{|\text{BoC}_{\text{cat}_i}(t_s)| + |\text{BoC}_{\text{cat}_i}(t_o)|} \tag{2}$$

where $\text{BoC}_{\text{cat}_i}$ depicts the set of BoC belonging to the category cat_i from the distinct character n-grams categories. By means of this representation we extract a multi-perspective on the similarities between two text passages, which serves as a richer and more detailed source of information for an automatic classifier in discriminating paraphrase plagiarism cases.

Accordingly, we proposed the following five short character n-grams categories aiming at highlighting distinct stylistic and content characteristics from texts. For illustration purposes, consider as a running example the following sentence: "*The n-grams type analysis is useful for understanding how the method is working*". Its corresponding short character n-grams categories are indicated in Table 3. For ease of understanding, we replace spaces in n-grams with underscores (_).

Punctuation. This set of character n-grams comprises all character sequences including at least one punctuation mark, such as periods, commas, colons, semi-colons, apostrophe, exclamation mark, etc. Punctuation marks help representing the writing style.

Inner-Words. This set is formed by all the character sequences appearing in the inner part of a word. Words of length greater than $n-1$ will be considered into this category, hence this type of character n-grams will filter out many stopwords. These n-grams do not capture any structural information but reflect thematic related aspects.

Between-Words. They comprise all the sequences of characters formed by the ending and the beginning parts of the words. Contrary to the other categories, it captures (to some extent) part of the structure within a sentence.

Prefixes. Set of character sequences formed by the beginning part of the word. This type of n-grams represents an important thematic feature given that its elements tend to capture word's lemmas.

Suffixes. This set is formed by the ending parts of the words. As it is possible to infer, this type of n-grams helps to capture several stylistic features such as the verbal form, singular/plural nouns, etc.

Finally, for determining the category of *plagiarism* or *not-plagiarism* between a pair of text fragments t_s and t_o, we adopted a supervised approach that takes advantage of the capabilities of machine learning techniques to handle multiple features representations. Accordingly, as a validation strategy we employed a 10-fold-cross-validation technique. Even though we evaluate the performance of our proposed approach using different machine learning algorithms, we only report results obtained with the Naïve Bayes[8] algorithm due to its high performance. Note that for every method considered in the next section, we did the construction of their respective classification model following the exact same approach.

5 Experimental Results

In this section, we present various results on the task of paraphrase plagiarism identification using the enriched P4PIN corpus (see Section 3.3). We performed

[8] We used the implementation provided by Weka (`http://www.cs.waikato.ac.nz/ml/weka/`)

Table 3 Character 3-grams belonging to each category for the example sentence shown above. We only show the corresponding character 3-grams for ease of understanding. All 3-grams appear separated by coma.

Category	**Character n-grams**	**Category**	**Character n-grams**
Punctuation	ng., _n-, n-g, -gr	*Between-Words*	he_, e_n, _n-, ms_, s_t, _ty, pe_, e_a, _an, is_, s_i, _is, is_, s_u, _us, ul_, l_f, _fo, or_, r_u, _un, ng_, g_h, _ho, ow_, w_t, _th, he_, e_m, _me, od_, d_i, _is, is_, s_w, _wo
Inner-Words	The, n-g, -gr, gra, ram, ams, typ, ype, ana, nal, aly, lys, ysi, sis, use, sef, efu, ful, for, und, nde, der, ers, rst, sta, tan, and, ndi, din, ing, how, the, met, eth, tho, hod, wor, ork, rki, kin, ing, ng.	*Prefixes*	_n-, _ty, _an, _is, _us, _fo, _un, _ho, _th, _me, _is, _wo
		Suffixes	he_, ms_, pe_, is_, is_, ul_, or_, ng_, ow_, he_, od_, is_

several experiments using the most representative and well adopted state-of-the-art methods, and compared the obtained results against the performance of our proposed method. For all the experiments we carried out a ten-fold cross-validation strategy and considered as main evaluation metric the F_1-measure given it considers both generality and accuracy.

This section is organized as follows: *i*) the first set of experiments focuses on evaluating the performance of previously proposed methods for plagiarism and paraphrase identification in the task of paraphrase plagiarism identification; *ii*) the second set of experiments aims at showing the pertinence of our method based on character-level features; various sizes of character n-grams were used with the purpose of highlighting the relevance of content and style information for solving the posed task; *iii*) the third set of experiments shows a detailed analysis on the relevance of the proposed representation for detecting different categories of paraphrase plagiarism; finally, *iv*) the last set of experiments studies the impact of the different categories of short character n-grams on the identification of paraphrase plagiarism.

5.1 Baselines performance

We compare our proposed representation against four different broad strategies for estimating the similarity between a pair of texts, namely: *(1)* content-based, *(2)* structural-based, *(3)* knowledge-based, and *(4)* ensembles (see Table 4). We refer to all these experiments as our *baseline* methods.

Among the *content-based* techniques we considered the traditional Bag-of-Words approach as well as some more elaborated techniques such as word n-grams and the Longest Common Subsequence (LCS) [7]. As it is known, these

Table 4 Performance of some state-of-the-art methods in the P4PIN corpus.

Method's configuration		F_1-measure
Content based	BOW	0.886
	word 2-grams	0.873
	word 3-grams	0.792
	word 4-grams	0.425
	word 5-grams	0.284
	LCS	**0.887**
Structural based	SW 1-grams	0.681
	SW 2-grams	0.684
	SW 3-grams	0.659
	SW 4-grams	0.629
	SW 5-grams	0.512
	POS 1-grams	0.565
	POS 2-grams	**0.734**
	POS 3-grams	0.718
	POS 4-grams	0.702
	POS 5-grams	0.677
Knowledge based	**WordNet+PD**	**0.882**
	WordNet+PD+C	0.756
Ensembles	**Content-based**	**0.884**
	Structural-based	0.742
	Knowledge-based	0.870
	All methods	0.865

methods are the typical configurations employed for performing content-based comparisons and, to some extent, word n-grams and LCS approaches are able to capture some structural information from texts.

Regarding the *structural-based* methods, we considered a representation based on stopword n-grams. This approach proposed by [31] is able to capture how a plagiarist uses very frequent stopwords, allowing to detect plagiarism by means of comparing structural features between texts. Another similar approach proposes using Part-of-Speech (POS) n-grams, which is also able to detect similar syntactic patterns in texts.

Although paraphrase and plagiarism identification share similar challenges, for the former very different and more complex approaches have been proposed. Most of these approaches use large external resources such as WordNet or controlled corpora. Therefore, we refer to these approaches as *Knowledge-based* methods. To evaluate this type of methods in the task of paraphrase plagiarism identification, we replicated the method proposed by [4] and [11]. The first method determines text similarity by means of matching words through their path distance in WordNet; we call this method WordNet+PD. Similarly, the method by Courtney and Mihalcea uses the path distance approach in conjunction with a reference corpus for evaluating the semantic similarity of two given texts. We refer to this method as WordNet+PD+C[9].

In order to determine if the combination of the above methods would be beneficial in the task of paraphrase plagiarism identification, we configured

[9] For this experiment we keep the best configuration obtained using the Brown corpus.

ensembles techniques to adequately combine all different approaches and evaluate whether or not they are complementary in the posed task. Our ensemble method refers to a configuration where the decision process employs as features all the distinct similarity values obtained by the considered approaches, *i.e.*, this configuration is a kind of meta-learning approach. Accordingly, we configure four types of ensembles using as features: *(a)* only *content-based* measures; *(b)* only *structural-based* measures; *(c)* only *knowledge-based* measures; and, *(d)* a combination of the three types of measures (*all methods*).

For the baseline experiments we did the following: *i)* texts characterization is computed according to the considered approach (*content, structural*, or *knowledge-based*), *ii)* for building the representation of each text pair, the similarity feature is computed using the DICE coefficient; and, *iii)* a Naïve Bayes classifier determine the appropriate similarity levels for assigning the *plagiarism* or *no-plagiarism* label. Table 4 shows the obtained results for all the proposed baselines. As it is possible to observe, the best global result was achieved by LCS, a *content-based* strategy, followed by the WordNet+PD configuration which is a *knowledge-based* strategy. As previously described, they represent very distinct approaches (lexical and semantic respectively). An additional aspect to remark is the fact that the method proposed by [11] (WordNet+PD+C) does not perform as well as the one proposed by [4], even when they both use WordNet. We consider that this is due to the noise introduced by the Brown corpus. Similarly, approaches using word n-grams (with large n values) perform very poorly compared with the LCS approach, which indicates that if using large contexts, the similarity must be computed with few elements instead of the whole vocabulary.

Although the semantic information shows to be useful for solving the posed task, employing external knowledge resources such as WordNet represents an expensive approach compared with the *content-based* techniques, which, at the end, allows achieving similar results.

From the results by *structural-based* methods, we observe that the best configuration was the POS n-grams (particularly the POS 2-grams). These results help to highlight the complexity of the posed task, since in traditional plagiarism identification the most successful method, among the *structural-based* approaches, was the use of stopwords n-grams.

Finally, ensemble results demonstrate that combining different set of features, extracted from different types of text characterization methods, does not leads to better results. The only configuration that it is able to outperform its corresponding single view configuration is the structural-based ensemble. Such result indicates that this kind of late fusion of different structural features allows better capturing the author's writing style.

5.2 Importance of style

An important parameter of the proposed method is the value of n during the definition of the BoC representation. Such parameter determines the kind of characteristics emphasized by the character n-grams characterization. Thus, when n is set to a large value (*i.e.,* 9+), the BoC is more suitable for detecting verbatim matches. For values between 5 and 8 (*i.e.,* more or less the average length of words), then the BoC representation is more appropriate for detecting thematic overlaps. By setting n to small values (*i.e.,* from 2 to 4 characters length), it is possible to capture important elements that reflect the author's writing style, for example *prefixes, suffixes* as well as the usage of *stopwords.* Furthermore, according to some previous studies [19], short n-grams represent (to some extent) the syllables' size, which reflects many phonetic properties inherent to texts related to the acoustic affinity, an additional stylistic feature that determines sonority, fluency and the rhythm followed by the author during the construction of a particular text.

Another important parameter considered by our method is the number of preprocessing operations. As it is known, many content related NLP tasks (*e.g.,* information retrieval, text classification and document clustering) consider as a common step within their pipeline applying preprocessing operations in order to allow focusing on content related terms. Traditionally, preprocessing includes lowercase conversion, punctuation marks elimination, stopwords removal, and stemming. As expected, preprocessing operations filter most of the style related characteristics, and highlight content related terms.

We performed a series of experiments aimed at demonstrating the importance of the style-based features when detecting paraphrase plagiarism cases. Figure 2 shows the performance of the proposed method when varying the value of the parameter n from 2 to 9, when no preprocessing is applied (solid line), and when preprocessing operations are applied (dotted line). Notice that the constant valued horizontal line represents the best baseline result (see Section 5.1), *i.e.* the LCS method[10].

As it is possible to observe, preprocessing operations are detrimental. Hence, these results support our initial hypothesis, which establishes that style-based features are important elements for identifying paraphrase plagiarism. Similarly, notice that small values of n (particularly $n = 3, 4$) allow the best performance. On the contrary, larger values of n generate lower results; notice that when n is around 8, the obtained performance is close to the one obtained by the LCS method.

Finally, it is worth noticing that our best configurations ($n = 3$ and $n = 4$, applying no preprocessing operations) obtain a 4% of relative improvement over the best baseline method. This result supports our intuition regarding the importance of including stylistic features when detecting paraphrase plagiarism. The results indicate that there is a statistically significant difference (according to the paired Student's t-test, $\alpha = 0.001$) between our proposed approach and the best baseline method. *i.e.*, F-measures from the best configuration. Hereafter, reported experiments use $n = 3$ and $n = 4$ together, and consider no preprocessing operations.

[10] The LCS performance is constant since the n parameter is not applicable for it.

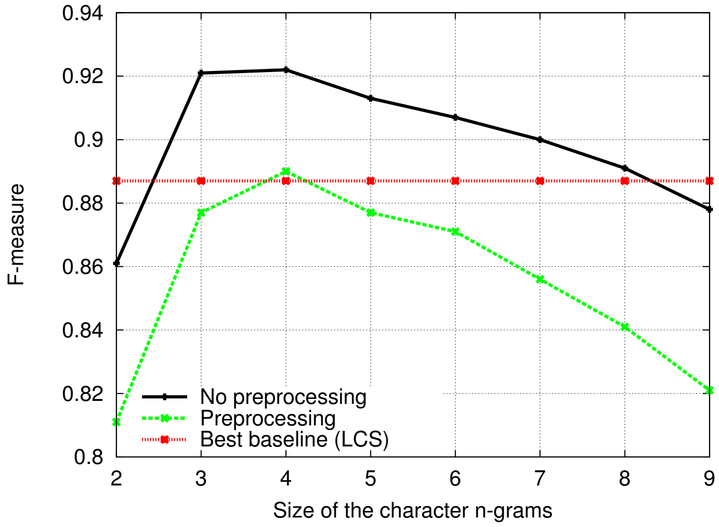

Fig. 2 Obtained performance of the proposed method varying the size of parameter n, and when preprocessing operations are considered vs. when they are not.

Table 5 The robustness of the proposed method across different paraphrases categories. For the baseline methods we selected the configurations achieving the best results.

Paraphrases types	**Best baseline method** F	*Configuration*	**Proposed Representation** F (% improvement)
Semantic	0.727	LCS	0.789 (+8.52%)
Morphological	0.865	BOW	0.906 (+4.73%)
Discourse	0.860	LCS	0.900 (+4.65%)
Syntactical	0.880	BOW	0.918 (+4.31%)
Miscellaneous	0.877	BOW	0.907 (+3.42%)
Lexical	0.882	WordNet+PD	0.928 (+5.21%)

5.3 Robustness on different paraphrase categories

As mentioned in Section 3, the P4PIN corpus has been labeled according to different categories of paraphrases phenomena. This section focuses on measuring the robustness of the proposed method against different paraphrase practices. For comparison purposes, we evaluated the performance of the best baseline methods under the same circumstances.

Results from these experiments are reported in Table 5. As it is possible to observe under the column "Best baseline method", there is no single baseline that performs equally good for all types of paraphrases. On the contrary, our proposed method outperforms the results from the baseline methods across all different paraphrases types. The relative improvement by the proposed representation appears between parentheses.

Besides the levels of relative improvement provided by the proposed representation, Table 5 shows some other interesting aspects. For example, in terms of F-measure, character-level features obtained the highest results detecting lexical-based changes in text fragments ($F = 0.928$). Although this type of paraphrases includes modifications such as the insertion of synonyms, it does not necessarily imply a modification of the writing style. Similarly, our representation was able to reach an F value above 0.9 for morphological and syntactical changes. Our explanation for having achieved such good results on these paraphrase categories is that the proposed representation is able to capture some style characteristics through detecting some stopwords usage patterns.

It is also important to notice that the lowest value of F was obtained for the semantic-based changes ($F = 0.789$); though, semantic changes represent the most difficult and elaborated type of paraphrase. Notice that for this type of paraphrases, the best baseline (*i.e.,* LCS) also obtains a poor performance. Using character-level features improved this result by a 8.52%, showing the best relative improvement among all paraphrase types.

An important observation regarding the results shown in Tables 4 and 5 is that LCS and WordNet+PD were the best methods in the overall evaluation, but they did not perform as well in the fine grained evaluation. The LCS method obtained the best baseline results only for the discourse and semantic-based changes, whilst the WordNet+PD method was the best option only for the lexicon-based changes. On the one hand, the fact that LCS performed well for discourse-based changes indicates that this method is able to capture core messages instead of discursive changes. On the other hand, the WordNet+PD method did not contribute in detecting semantic-based changes given the high complexity of such type of paraphrases. However, it was very accurate in detecting lexical-based changes because of the easiness of extracting synonyms from WordNet. Generally speaking, our proposed method is robust enough for accurately identifying several types of paraphrase plagiarism.

5.4 Are all short character n-grams equally important?

The main goal of this section is to help understanding the contribution of the information being captured by each type of character n-gram on the paraphrase plagiarism identification process. In order to achieve this goal, we performed a detailed analysis on the performance of the different categories of short character n-grams.

Figure 3 shows the obtained results when the paraphrase plagiarism task is approached using only one particular category of short character n-grams, *i.e.,* a single feature for representing the similarity of a pair of text passages. From these results we can observe that each category of short character n-grams conveys its own important information. In particular, the *prefixes* category achieved the best performance by its own ($F = 0.91$), indicating the relevance of thematic information for identifying paraphrase plagiarism. However, *prefixes* n-grams do not show an statistically significant difference with respect to the LCS method (best baseline method, $F = 0.88$), but the number of computed text only represents 14.22% with respect to the amount text employed by the LCS method.

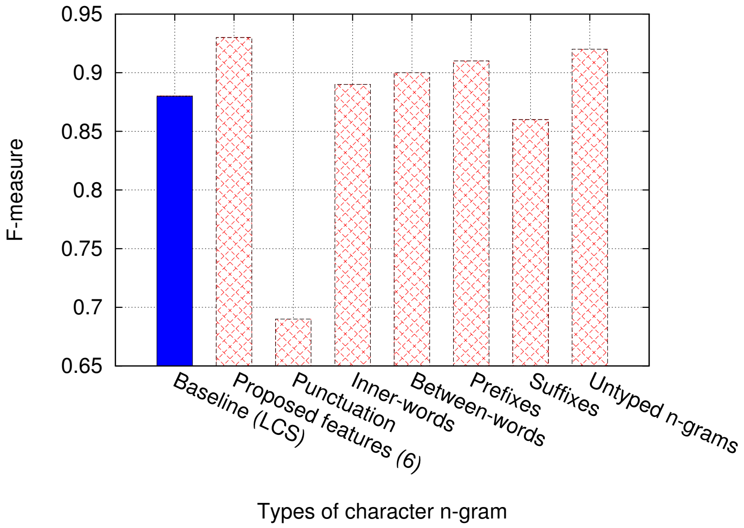

Fig. 3 Performance of the proposed method using different types of short character n-grams for representing text fragments.

An important aspect regarding the *punctuation* category, is that this type of n-grams did not perform very well ($F = 0.69$). This is due to the fact that such $n-$grams focus mostly on purely stylistic aspects. Regarding the *between-words* and the *suffixes* n-grams it is possible to observe that both categories represent a better trade off between content and style; consequently, results obtained using only these features are better than using just a stylistic feature, *i.e.,* $F = 0.90$ and $F = 0.86$ respectively. Although these results are closer to the baseline, it is important to point out that all these configurations use very small sets of n-grams for computing its respective representation vector. For instance, the number of n-grams employed by the *between-words* configuration represents only 32.8% of the text used by the LCS method.

Notice that the proposed method (*Proposed features (6)*) obtains the best overall performance ($F = 0.93$). The improvement by this configuration over the baseline method is statistically significant with a confidence level of 99 percent (i.e. $\alpha = 0.01$) in accordance with the paired *Student's t-test.* Contrary to our proposed method, the *Untyped n-grams* configuration employs only one similarity feature to represent a pair of passages where no distinction among n-grams categories is made. Even though the obtained F score is high ($F = 0.92$), our proposed method was consistently better across the performed experiments.

6 Conclusions

Paraphrase plagiarism identification consists in automatically recognizing document fragments that contain re-used text, which is intentionally hidden by means of some rewording practices such as semantic equivalences, discursive changes, and morphological or lexical substitutions. As established by our main hypothesis, the original author's writing style fingerprint prevails in the plagiarized text even when paraphrases practices are employed. Our main contribution relies on the proposal of a new representation that considers character-level features for identifying paraphrase plagiarism. In particular, this *representation models the similarity of a given pair of texts using different categories of short character n-grams.* Experimental results indicated that this representation is very effective for identifying paraphrase plagiarism because it is able to capture content and style information from texts fragments.

Additionally, we undertook the task of building an appropriate *corpus for evaluating automatic paraphrase plagiarism identification* methods. For this, we started from the corpus P4P, which contains several manually elaborated cases of paraphrase plagiarism. Our main contribution was the inclusion of not-plagiarism examples. The main goal was to come up with a realistic evaluation scenario, where difficulties of the posed task were reflected. Such released corpus, named P4PIN, contains both plagiarism and not-plagiarism cases, and represents a valuable resource to the NLP community interested in carrying out future research in this field.

The performed experiments allow us to formulate the following conclusions. Firstly, our initial set of experiments demonstrated that traditional plagiarism detection methods, *i.e.,* content, structural and knowledge based, face some difficulties for accurately identifying paraphrase plagiarism cases. Obtained results reinforced our claims about the greater complexity of paraphrase plagiarism identification against traditional plagiarism detection (see research question Q1 in Section 1.2). Our aim in research question Q2 was to investigate the impact of different preprocessing operations (techniques related to word-level features) and variations on the length of the character n-grams in preserving stylistic information. Particularly, the obtained results indicated that using small values of n without any preprocessing operations allows to obtain the best performance. In research question Q3 we investigate the robustness of our proposed style-based representation using different categories of short character n-grams across several paraphrase categories. Experiments confirmed that our approach is consistently effective over different paraphrases types. Finally, the research question Q4 aimed at understanding the contribution of the information being captured by each type of character n-gram on the paraphrase plagiarism identification process. For this, we proposed a categorization of the different n-grams based on the type of information that is being captured by them. Experiments using different types of n-grams showed that each category of short character n-grams conveys its own important information. Important findings from these experiments were: *i*) character n-grams capturing content information are very important in solving the posed task, *ii*) pure stylistic character n-grams did not perform well, and *iii*) short character n-grams that capture a combination of content and style characteristics tend to obtain better results.

Although the obtained results motivate us working on the same direction, it was possible to observe that the employed representation face some difficulties detecting the *semantic* paraphrase type. Such results were not surprising, since this particular type of paraphrase represents the most complicated paraphrase type. Our future work aims integrating additional information to the character n-grams in order to take into account *semantic* changes.

Acknowledgements This work is the result of the collaboration in the framework of the CONACYT Program of Thematic Networks and the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action. The first author was supported by CONACYT (scholarship 258345/224483). The second and third authors were partially supported by CONACyT (Project grants 258588 and 247870). The work of the fourth author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

References

1. Barrón-Cedeño, A., Rosso, P.: On automatic plagiarism detection based on n-grams comparison. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR), LNCS Vol. 5478, Springer-Verlag, pp. 696–700 (2009)
2. Barron-Cedeño, A., Vila, M., Martí, M.A., Rosso, P.: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics **39**(4), 917–947 (2013)
3. Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Degli Esposti, M.: A plagiarism detection procedure in three steps: Selection, matches and "squares". In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 2009), CEUR-WS Vol. 502. Donostia-San Sebastian, Spain (2009)
4. Biggins, S., Mohammed, S., Oakley, S.: University of shefield: Two approaches to semantic text similarity. In: First Joint Conference on Lexical and Computational Semantics (SEM at NAACL 2012), pp. 655–661. Montreal, Canada. (2012)
5. Burrows, S., Potthast, M., Stein, B.: Paraphrase acquisition via crowdsourcing and machine learning. ACM Trans. Intell. Syst. Technol. **4**(3), 43:1–43:21 (2013). DOI 10.1145/2483669.2483676. URL `http://doi.acm.org/10.1145/2483669.2483676`
6. Calvo, H., Segura-Olivares, A., García, A.: Dependency vs. constituent based syntactic n-grams in text similarity measures for paraphrase recognition. Computación y Sistemas **18**(3), 517–554 (2014)
7. Chien-Ying, C., Jen-Yuan, Y., Hao-Ren, K.: Plagiarism detection using rouge and wordnet. Journal of Computing **2**(3), 34–44 (2010)
8. Chong, M., Specia, L., Mitkov, R.: Using natural language processing for automatic detection of plagiarism. In: Proceedings of the 4th International Plagiarism Conference. Newcastle-upon-Tyne, UK. (2010)
9. Clough, P.: Old a new challenges in automatic plagiarism detection. In: National Plagiarism Advisory Service, pp. 391–407 (2003)
10. Clough, P., Gaizauskas, R., Piao, S.S., Wilks, Y.: Meter: Measuring text reuse. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia. (2002)
11. Courtney, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE at NAALC 2005), pp. 13–18 (2005)

12. Daelemans, W.: Explanation in computational stylometry. In: 14th International Conference on Intelligent Text Processing and Computational Linguistics (CIC-Ling 2013), Lecture Notes in Computer Science LNCS, vol. 7817, pp. 451–462 (2013)
13. Ehsan, N., Shakery, A.: Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. Information Processing and Management **http://dx.doi.org/10.1016/j.ipm.2016.04.006** (2016)
14. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing **22**(3), 251–270 (2007)
15. Hartrumpf, S., Brück, T.v.d., Eichhorn, C.: Semantic duplicate identification with parsing and machine learning. In: Eleventh International Conference on Text, Speech and Dialogue (TSD 2010) LNAI Vol. 6231, Springer-Verlag, pp. 84–92. Brno, Czech Republic (2010)
16. Hoad, T.C., Zobel, J.: Methods for identifying versioned and plagiarised documents. Journal of the American Society for Information Science and Technology **54**, 203–215 (2003)
17. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology (JASIST) **60**(1), 9–26 (2009). John Wiley and Sons, Inc. New York, NY, USA
18. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation **45**, 83–94 (2011)
19. Man, P.D.: Blindness and Insight: Essays in the Rhetoric of Contemporany Criticism, 2nd ed. edn., chap. Literature and Languege: A Commentary, pp. 277–89. Routtloedge (1983)
20. McNamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Information retrieval **7**(1-2), 73–97 (2004)
21. Oberreuter, G., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Approaches for intrinsic and external plagiarism detection. In: Notebook for PAN at CLEF'11. (2011)
22. Palkovskii, Y., Belov, A., Muzyka, I.: Using wordnet-based semantic similarity measurement in external plagiarism detection. In: Notebook for PAN at CLEF'11. (2011)
23. Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
24. Ravi, N.R., Gupta, D.: Efficient paragraph based chunking and download filtering for plagiarism source retrieval. In: Notebook for PAN at CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, PAN '15 (2015). URL `http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-papers-final/pan15-plagiarism-detection/ravi15-notebook.pdf`
25. Sapkota, U., Bethard, S., Montes-y Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT 2015), pp. 93–102 (2015)
26. Sapkota, U., Solorio, T., Montes, M., Bethard, S., Rosso, P.: Cross-topic authorship attribution: Will out-of-topic data help? In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1228–1237. Dublin City University and Association for Computational Linguistics (2014). URL `http://aclweb.org/anthology/C14-1116`
27. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: Local algorithms for document fingerprinting. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, SIGMOD '03, pp. 76–85. ACM, New York, NY, USA (2003). DOI 10.1145/872757.872770. URL `http://doi.acm.org/10.1145/872757.872770`
28. Sediyono, A., Mahamud, K.: Algorithm of the longest commonly consecutive word for plagiarism detection in text based document. In: Digital Information Management, ICDIM '08, pp. 253–259. IEEE (2008). DOI 10.1109/ICDIM.2008.4746827
29. Shivakumar, N., Garcia-Molina, H.: Scam: A copy detection mechanism for digital documents. In: Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (1995)
30. Si, A., Leong, H.V., Lau, R.W.H.: Check: A document plagiarism detection system. In: Proceedings of ACM Symposium for Applied Computing, SAC '97, pp. 70–77. ACM,

New York, NY, USA (1997). DOI 10.1145/331697.335176. URL http://dl.acm.org/citation.cfm?doid=331697.335176

31. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society For Information Science and Technology **62**(12), 2512–2527 (2011)
32. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy **21**(2), 421–439 (2013)
33. Stein, B., Potthast, M., Rosso, P., Barrón-Cedeño, A., Stamatatos, E., Koppel, M.: Fourth international workshop on uncovering plagiarism, authorship, and social software misuse. In: SIGIR Forum, vol. 45, pp. 45–48 (2011)
34. Sánchez-Vega, F., Villatoro-Tello, E., Montes-y Gómez, M., Villaseñor-Pineda, L., Rosso, P.: Determining and characterizing the reused text for plagiarism detection. Expert Systems with Applications **40**(5), 1804–1813 (2013)
35. Özlem Uzuner, Katz, B., Nahnsen, T.: Using syntactic information to identify plagiarism. In: Proc. 2nd Workshop on Building Educational Applications using NLP. Ann Arbor (2005)
36. Xu, W., Ritter, A., Dolan, W.B., Grishman, R., Cherry, C.: Paraphrasing for style. In: Proceedings of COLING 2012: Technical Papers, pp. 2899–2914. Mumbai (2012)
37. Zechner, M., Muhr, M., Kern, R., Granitzer, M.: External and intrinsic plagiarism detection using vector space models. In: SEPLN 2009, Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), pp. 45–55 (2009)