International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 12, December-

UNDERSTANDING PLAGIARISM FOR CONTEXTUAL FEATURES

DURGA BHAVANI DASARI[1], Dr. VENU GOPALA RAO. K[2]

[1]*Research Scholar, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India*

bhavani.dd@gmail.com, +91-9490988482

[2]*Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, India*

Kvgrao1234@gmail.com, +91-9849025342

Abstract: Plagiarism can be of many different natures, ranging from copying texts to adopting ideas, without giving credit to its originator. This paper presents a new taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism, from the plagiarist's behavioural point of view. The taxonomy supports deep understanding of different linguistic patterns in committing plagiarism. Different contextual features that characterize different plagiarism types are discussed. Systematic frameworks and methods of monolingual, extrinsic, intrinsic, and cross-lingual plagiarism detection are surveyed and correlated with plagiarism types, which are listed in the taxonomy.

Key words: Linguistic patterns, plagiarism, taxonomy, contextual features.

I. INTODUCTION

1.1 *Plagiarism:*

The problem of plagiarism has recently increased because of the digital era of resources available on the World Wide Web. Plagiarism detection in natural languages by statistical or computerized methods has started since the 1990s, which is pioneered by the studies of copy detection mechanisms in digital documents [2], [3]. Earlier than plagiarism detection in natural languages, code clones and software misuse detection has started since the 1970s by the studies to detect programming code plagiarism in Pascal and C [1], [4]–[5]. Algorithms of plagiarism detection in natural languages and programming languages have noticeable differences. The first one tackles different contextual features and diverse methods of detection. During the last decade, research on automated plagiarism detection in natural languages has actively evolved, which takes the advantage of recent developments in related fields like information retrieval (IR), cross language information retrieval (CLIR), natural language processing, computational linguistics, artificial intelligence, and soft computing.

This paper brings patterns of plagiarism together with contextual features for characterization of each pattern and computerized methods for detection.

II. PLAGIARISM TAXONOMY

There are no two humans, no matter what languages they use and how similar thoughts they have, write exactly the same text. Thus, written text, which is stemmed from different authors, should be different, to some extent, except for cited portions. If proper referencing is abandoned, problems of plagiarism and intellectual property arise. The existence of academic dishonesty problems has led most, if not all, academic institutions and publishers to set regulations against the offence. Borrowed content of any form require directly or indirectly quoting, in-text referencing, and citing the original author in the list of references [6].

A number of research works have addressed plagiarism in academia [6]–[7] and illustrated different types of plagiarism and available software for plagiarism detection. For example, a recent book [8], [9] provides an extensive linguistic analysis of plagiarism in academic writing. However, little research has related linguistic patterns of plagiarism with computerized contextual features and automated techniques for extracting and detecting such types.

The data were collected via several interviews with several faculty members with 10–20-year

teaching expertise at the university. The questions focused on different plagiarism practices by the students. The main output of the qualitative study is a new taxonomy of plagiarism that comprehensively relates different types, as shown in Fig. 2. The taxonomy divides plagiarism into two typical types: *literal plagiarism* and *intelligent plagiarism*, based on the *plagiarist's behaviour* (i.e., student's or researcher's way of committing plagiarism).

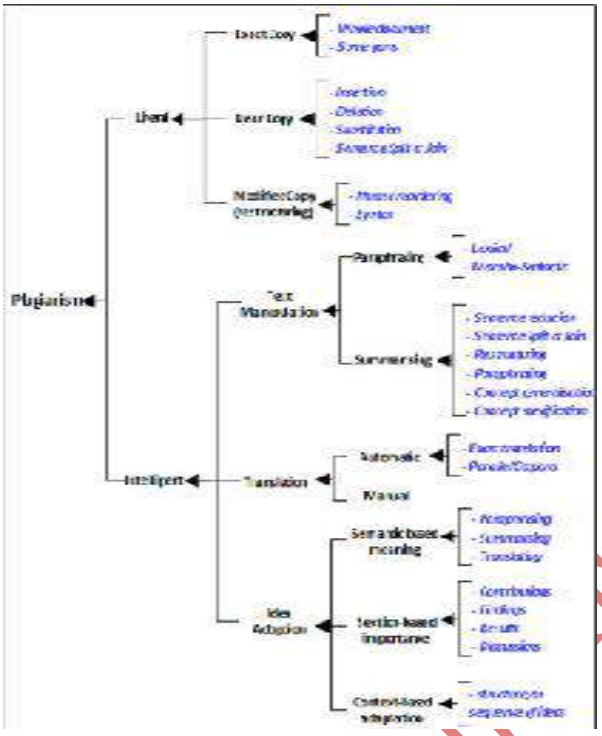

Fig. 2. Taxonomy of plagiarism.

III. CONTEXTUAL FEATURES

There are several textual features to quantify and characterize documents before applying a plagiarism detection method. This section discusses textual features needed in different frameworks: extrinsic, intrinsic, and cross-lingual.

A. Contextual Features for Extrinsic Plagiarism Detection

Contextual features to represent documents in extrinsic plagiarism detection include: *lexical* features, such as character *n*-gram and word *n*-gram; *syntactic* features, such as chunks, sentences, phrases, and POS, *semantic* features, such as synonyms and antonyms; and *structural* features that takes contextual information into account. Table I summarizes each types together with computational tools and resources required for their implementation. A detailed description of textual features for extrinsic plagiarism detection is given in the following.

1) Lexical Features: Lexical features operate at the character or word level. *Character-based n-gram (CNG)* representation is the simplest form whereby a document d is represented as a sequence of characters $d = \{(c1, d), (c2, d), \ldots, (cn, d)\}$, where (ci, d) refers to the ith character in d, and $n = d$ is the length of d (in characters). On the other hand, *word-based n-gram (WNG)* represents d as a collection of words $d = \{(w1, d), (w2, d), \ldots, (wn, d)\}$, where (wi, d) refers to the ith word in d, and $n = d$ is the length of d (in words) with ignoring sentence and structural bounds. Simple WNGs may be constructed by using *bigrams* (word-2-grams), *trigrams* (word-3-grams) or larger. *CNG* and *WNG* are commonly called *fingerprints* or *shingles* in text retrieval and plagiarism detection research. The process of generating fingerprints (or shingles) is called *fingerprinting* (or *shingling*). A document fingerprint can, therefore, identify the document uniquely as well as a human fingerprint does.

2) Syntactic Features: Syntactical features are manifested in part of speech (POS) of phrases and words in different statements. Basic POS tags include verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections. POS tagging is the task of marking up the words in a text or more precisely in a statement as corresponding to a particular POS tag. Sentence-based representation works by splitting the text into statements with the use of end-of-sentences delimiters, such as full stops, exclamation, and question marks. After splitting the text into sentences, POS and phrase structures can be constructed by using POS taggers. On the other hand, chunks is another feature that is generated by so-called windowing or sliding windows to characterize bigger text than phrases or sentences. POS could be further used in windowing to generate more expressive POS chunks. *Word order*, in a sentence or a chunk, could further be combined as a feature, and used as a comparison scheme between sentences.

3) Semantic Features: Semantic features quantify the use of word classes, synonyms, antonyms, hypernyms, and hyponyms. The use of thesaurus dictionaries and lexical databases, Word- Net, for instance, would significantly provide more insights into the semantic meaning of the text. Together with POS tagging,

semantic dependencies can be featured, and that would be very helpful in plagiarism detection.

4) ***Structural Features****:* Most plagiarism detection algorithms employ *flat* document features, such as lexical, syntactic, and semantic features. Very few algorithms have been developed to handle *structural* or *tree* features. Structural features reflect text organization and capture more document semantics. Documents can be described as a collection of paragraphs or passages, which can be considered as topical blocks. In many
cases, paragraphs that are topically related or discuss the same subject can be grouped into sections, i.e., structural features might characterize documents as *headers*, *sections*, *subsections*, *paragraphs, sentences,* etc. This type of features can be used in structured documents, such as HTML web pages and XML files, and semi structured documents, such as books, theses, and academic journal papers. Note that structural features are most likely to be stored as XML trees for easier processing. Structural features can be divided into *block-specific* and *content-specific*. In a recent study [10], *block-specific tree structured features* were used to describe a collection of web documents as blocks, namely, *document-page-paragraph*. Webpage
documents were divided into paragraphs by an HTML parser taking the advantage of different HTML tags, such as <p>,<hr>, and
 to segment each webpage. Then, paragraphs were grouped into pages, whereby a new paragraph is added to each page until a maximum threshold of word count is reached; otherwise, a new page is created. Because paragraphs are more likely to have topically related sentences than pages, a recent study [8] encoded documents features in a hierarchical multilevel representation *document-paragraph-sentence*. The existing structural implementations would be further improved, if the document features are encoded as *content-specific tree-structured features* by using semantically related blocks, such as *document-section-paragraph* or *class-concept-chunk*. The use of content-specific tree-structured features in combination with some *flat* features can be very useful in capturing the document's semantics and getting the gist of its sections/ concepts. The rationale of using *content-specific* is to segment the document into different ideas (i.e., semantic blocks) to allow for the detection of idea plagiarism, in particular. Besides, we can drill down or roll up through the *structural* representation to detect more or less plagiarism types patterns, which are mentioned in our taxonomy of Fig. 2.

B. Contextual Features for Intrinsic Plagiarism Detection

Stylometric features are based on the fact that each author develops an individual writing style. For example, authors employ, consciously or subconsciously, patterns to construct sentences, and use an individual vocabulary [11]. The stylometric features quantify various style aspects [12], [13], including 1) text statistics via various lexical features, which operate at the character or word level; 2) syntactic features, which work at the sentence level, quantify the use of word classes, and/or parse sentences into part of speech; 3) semantic features, which quantify the use of synonyms, functional words, and/or semantic dependencies; and 4) application-specific features, which reflect text organization, content-specific keywords, and/or other language-specific features. Table II summarizes the stylometric features together with computational tools and resources required for their implementation.

C. Contextual Features for Cross-Lingual Plagiarism Detection

Features that are based on lexical and syntactic types are improper in a cross-lingual setting, i.e., for cross-lingual text relatedness and plagiarism detection, syntactic features are usually combined with semantic or statistical features. Other features may be language-specific or content-specific keywords. Table III summarizes contextual features for cross-language plagiarism detection together with computational tools and resources required for their implementation.

IV . CONCLUSION

As plagiarists become increasingly more sophisticated, idea plagiarism is a key academic problem and should be addressed in future research. We also propose the use of structural features and contextual information with efficient STRUC-based methods to detect *section-based importance* and *context-based adaptation* idea plagiarism.

V. REFERENCES

[1]. A. Parker and J. O. Hamblen, "Computer algorithms for plagiarism detection," *IEEE Trans. Educ.*, vol. 32, no. 2, pp. 94–99, May 1989.

[2]. S. Brin, J. Davis, and H. Garcia-Molina, "Copy detection mechanisms for digital documents," in *Proc. ACMSIGMOD Int. Conf.Manage. Data*, New York, 1995, pp. 398–409.

[3] N. Shivakumar and H. Garcia-Molina, "SCAM: A copy detection mechanism for digital documents," in *D-Lib Mag.*, 1995.

[4] K. J. Ottenstein, "An algorithmic approach to the detection and prevention of plagiarism," *SIGCSE Bull.*, vol. 8, no. 4, pp. 30–41, 1977.

[5]. K. S. Marguerite, B. E. William, J. F. James, H. Cindy, and J. W. Leslie, "Program plagiarism revisited: Current issues and approaches," *SIGCSE Bull.*, vol. 20, pp. 224–224, 1988.

[6]. M. Roig, *Avoiding Plagiarism, Self-Plagiarism, and Other Questionable Writing Practices: A Guide to EthicalWriting*.NewYork: St. Johns Univ. Press, 2006.

[7]. K. R. Rao, "Plagiarism, a scourge," *Current Sci.*, vol. 94, pp. 581–586, 2008.

[8]. H. Zhang and T. W. S. Chow, "A coarse-to-fine framework to efficiently thwart plagiarism," *Pattern Recog.*, vol. 44, pp. 471–487, 2011.

[10]. T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree structured data for efficient document retrieval and plagiarism detection," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1385–1402, Sep. 2009.

[11]. M. zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections," in *Advances in Data Analysis*, 2007, pp. 359– 366.

[12]. E. Stamatatos, "A survey of modern authorship attribution methods," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, pp. 538–556, 2009.

[13]. B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," in *Language Resources & Evaluation*, 2010.

[14] P. M. McCarthy, G. A. Lewis, D. F. Dufty, and D. S. McNamara, "Analyzing writing styles with Coh-Metrix," in *Proc. Florida Artif. Intell. Res. Soc. Int. Conf.*, Melbourne, FL, 2006, pp. 764–769.

[15] M. Jones, "Back-translation: The latest form of plagiarism," presented at the 4th Asia Pacific Conf. Edu Integr., Wollongong, Australia, 2009.

[16] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan,"Stylistic text classification using functional lexical features: Researcharticles," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, pp. 802–822, 2007.

[17] L. Stenflo, "Intelligent plagiarists are the most dangerous," *Nature*, vol. 427, p. 777, 2004.

[18] M. Bouville, "Plagiarism: Words and ideas," *Sci. Eng. Ethics*, vol. 14,pp. 311–322, 2008.

[19] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, and D. Tambouratzis, "Discriminating the registers and styles in the modern greek language—Part 1: Diglossia in stylistic analysis," Lit. Linguist. Comput., vol. 19, no. 2, pp. 197–220, 2004.

[20] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, and D. Tambouratzis, "Discriminating the registers and styles in the modern Greek language—Part 2: Extending the feature vector to optimise author discrimination," *Lit. Linguist. Comput.*, vol. 19, pp. 221–242, 2004.

[21] C. K. Ryu, H. J. Kim, S. H. Ji, G. Woo, and H. G. Cho, "Detecting and tracing plagiarized documents by reconstruction plagiarism-evolution tree," in *Proc. 8th Int. Conf. Comput. Inf. Technol.*, Sydney, N.S.W., 2008, pp. 119–124.

[22] Y. Palkovskii, ""Counter plagiarism detection software" and "Counter counter plagiarism detection" methods," in *Proc. SEPLN*, Donostia, Spain, pp. 67–68.

[23] J. A. Malcolm and P. C. R. Lane, "Tackling the PAN'09 external plagiarism detection corpus with a desktop plagiarism detector," in *Proc. SEPLN*, Donostia, Spain, pp. 29–33.

[24] R. Lackes, J. Bartels, E. Berndt, and E. Frank, "A word-frequency based method for detecting plagiarism in documents," in *Proc. Int. Conf. Inf. Reuse Integr.*, Las Vegas, NV, 2009, pp. 163–166.

[25] S. Butakov and V. Shcherbinin, "On the number of search queries required for Internet plagiarism detection," in *Proc. 9th IEEE Int. Conf. Adv. Learn. Technol.*, Riga, Latvia, 2009, pp. 482–483.

