Arabic Text Copy Detection using Full, Reduced and Unique Syntactical Structures

Mohamed Taybe Elhadi
Depart. of Computer Science
Zawia University,
Zawia, Libya

ABSTRACT
This paper reports on work performed to investigate the use of a combined Part of Speech (POS) tagging and a minimum edit operations algorithm to determine the level of similarity between pairs of Arabic text documents. The level of similarity can be used as an indication of duplication in full or in part of the document's content. Text is first converted into POS tags that are then fed to the string similarity algorithm to determine the similarity of pairs of documents. A normalized score is calculated and used to rank documents. Documents ranked higher than some selected threshold are considered similar and can be near or complete duplicate. The performed experiments compare results based on the use of a set of selected common subsequences that are the results of translation of text into a sequence of syntactical units. The strings are first produced using full-text (FULL). These are further refined to produce a REDUCED; where repeated consecutive characters are reduced to a single character and a number, and more refined to produce a UNIQUE string; where all repeating characters are replaced by a single character. Syntactical features of the text were used as a structural representation of the documents' content. Results obtained from the experiments using the FULL, the REDUCED and the UNIQUE POS-strings showed a clear advantage over the use of the plain text in terms of reduced string size while maintaining the same discrimination power. In particular the unique (most-reduced) string has shown quite comparable results to the reduced, the full and the actual text string.

Keywords
Arabic text processing; syntactical structures; document similarity; reduction; edit-based string similarity; copy detection.

1. INTRODUCTION
Document analysis and re-use detections using some text similarity measures have become an important processing technique in light of the growth of the web and the emergence of digital libraries. The multiple lingualism of such data explosion further necessitates the need for more robust, efficient and universal tools. Most work in this area is based on actual text (characters or words) representations and processing. It spans many fields of research including, among many other applications, copy/ near-copy detection [1,2], plagiarism [3-5], Information Retrieval (IR) [6,7] and computational biology [8-10,12]. Many such applications employee a combination of techniques and apply to multidisciplinary fields [13,14,15].

This paper reports on a work that investigated how related Arabic documents can be treated as modified versions of one another using edit operations [16-18]. It compares results based on the use of a set of selected sequences that are the result of POS-tagging of Arabic text. The process uses a single sequence of the full document, and the more refined (reduced and unique) sequences extracted from the POS sequences.

Such transformation of text is the result of looking at documents variations as attempts to modify existing text whether maliciously (plagiarize) or purposefully (reducing or expanding on text articles) as a total cut-and-paste, insertion, deletion and/or substitution.

In this transformation, syntactical features of the text represented in terms of POS-tags were created and used as a structural representation of the document's text content. Secondly, each document was further compared using edit-based similarity [19] using the FULL POS-sequence and the refined (REDUCED and UNIQUE) sequences. The set of unique sequences in particular are shorter strings that are the result of removing any repeating subsequences. At the end a normalized score between 0 and 1 was calculated and used to rank the documents. Documents ranked higher than some determined threshold are considered similar if not duplicate.

Taking advantage of syntactic properties derived as POS tagged strings and using such strings instead of actual text in the similarity calculation has the advantages of representing text using meaningful, well-defined and clearly-represented set of units. This can also capture some semantics contained in the writing style of authors and the relationships defined by the used style, selection of word/phrase types and the ordered tag units. Moreover, reduction of text to its syntactical structures reduces the dimensionality of the document allowing us to deal with much shorter strings instead of the full text. Such reduction minimizes information loss when compared with processing based on mere text of characters or group of words as practiced by various n-gram, shingle-based techniques [20,21] and IR in general. Reduction in text representative to be used for comparison enables the efficient use of sequence compression algorithms such as LCS and other string approximation methods [22,16].

Experimental validation of the suggested procedure was performed using a human-created and classified corpus [23]. Results obtained showed a clear advantage of using the produced strings in terms of reduced size while maintaining the same discrimination power as that of the original text.

The rest of the paper is made up of sections: Section 2 is related work; section 3 is the proposed procedure, section 4 contains a description of the data set used and the experiments performed, section 5 contains a presentation and discussion of the results obtained and finally section 6 contains the conclusion and future work.

2. RELATED WORK

The combined use of syntactical POS tagging and text processing methods for the purpose of text similarity calculations and its applications is a recent endeavor. It is based on the idea that a realization of the intuition that similar (duplicate) documents would have similar (exact) syntactical structure. In particular, those documents that contain (reuse) other documents or parts of other documents would certainly contain similar structures. This is more certain when the production or refinement of new documents is the result of reduction, expansion, plagiarism or modifications. More on the use of syntactic properties and POS tagging to determine similarity of text by way of comparing POS strings can be found in [20,22,24-26]. A brief mention is provided next.

A major hurdle in comparing text can be attributed to the differences on the makeup of the strings on one hand and the lack of a theory that can be used for explaining this makeup on another. Different methods and approaches have been used to tackle the issue of similarities between documents using semantically and syntactically motivated approaches. Semantic approaches receive less attention due to the difficulties of representing semantics and the limitations on assessment coverage of user studies [27, 28]. Other mostly non-semantically oriented techniques have received more attention. These include fingerprinting [29], IR [6] and many hybrid techniques [5,13,14]. In information retrieval models, more emphasis is put on representing documents by their words and word frequencies. It uses indexing with an appropriate model to evaluate documents similarity [6].

Looking at a lump of text as a string made of meaningful, well defined and numerable units (alphabets), means that modified (and thus similar) text can be thought of as an intervention or application of edit operations commonly found in bio-sequences analysis of insertion, deletions and substitutions [16-18].

Work presented here is a continuation of previous experiments intended to take advantage of syntactical structures in finding similar text (particularly documents) [20-22,24,25,26].

3. THE PROPOSED PROCEDURE

To evaluate the proposed approach, a dataset taken from [23] was used without any previous knowledge of its contents. The used data set was made into a set of 1000 separate files. Each file contained text on a supposedly different subject indicated with one of Economics, Art, Politics, or Sports as labeled by the corpus providers. The corpus is compromised of 200 documents per topic or label. Table 1 has an example of a very short document.

In processing the documents the following steps were followed:

a) Text documents were first tagged using Stanford tagger (for Arabic Language) on the level of the whole document (FULL).

b) Tagged-text is then converted into a string of single-character symbols for ease and efficiency of processing.

c) The resulting full-document (FULL) based POS strings are further treated to produce what a refined or Reduced (REDUCED) and Unique (UNIQUE) set. Reduced set is the set produced by substituting the repeated consecutive characters with single character and a number reflecting the count of the repeats. The unique tag sequences where produced by the removal of all repeating characters and substituting them by the mere single repeated character.

d) The string similarity algorithm described in [31] was run on the all set of strings (Plain Text, POS-strings for PLAIN, FULL, REDUCED and UNIQUE).

e) A generalized final score between 0 and 1 based on the produced strings, is to be used for similarity ranking and comparison.

Obtained results were analyzed and compared to figure out the level of correctly classified documents on one hand and to compare the accuracy of using the different refined strings and substrings as compared to use of the actual text string. A brief description of the proposed procedure is shown in Figure 1. Over all phase of the used procedure are briefly described next:

1) Reduction of each document's text into a set of (POS) tags without exclusion of any stop words, stemming or removal of numbers, punctuation or special characters. Stanford Log-linear Part-Of-Speech Tagger [32,33], was adopted and used for this work.

2) Similarity algorithm is applied and a score is calculated for the text as is as well three types of POS strings (Full-text, Reduced and Unique).

3) Paired documents are then ranked based on the similarity score and analyzed.

In processing the documents the following steps were followed:

a) Text documents were first tagged using Stanford tagger (for Arabic Language) on the level of the whole document.

b) Tagged-texts are then converted into string of single-character symbols for easy and efficiency of processing.

c) The resulting full-document(FULL) based POS strings are further treated to produce the REDUCED and UNIQUE sets. Reduced set is the set produced by substituting the repeated consecutive characters with single character and a number reflecting the count of the repeats. The unique tag sequences where produced by the removal of all repeating characters and substituting them by the mere single repeated character.

d) The string similarity algorithm described in [31] was run on all the sets of strings: Pure text, its POS-string, the Reduced and Unique strings.

e) A generalized final score between 0 and 1 based on the produced strings, was used for similarity ranking and comparison.

The used string similarity algorithm as described in [31] roughly works by looking at the smallest number of edits to change of edits to change one string into the other. It calculates the similarity index of its two arguments. A value of 0 means that the strings are entirely different. A value of 1 means that the strings are identical. Everything else lies between 0 and 1 describes the partial similarity between the strings.

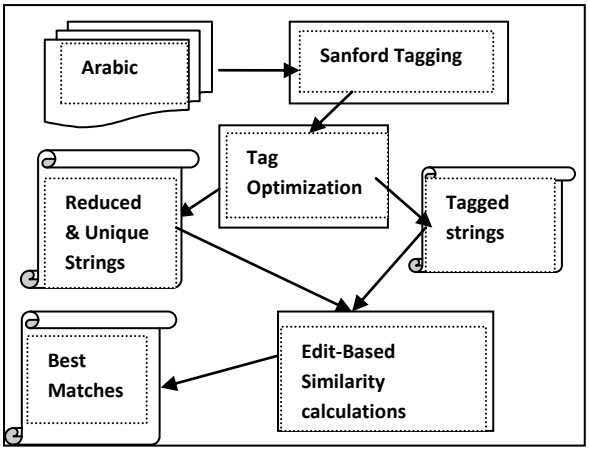

Fig 1: Overall depiction of the proposed procedure

4. DATASETS AND EXPERIMENTS

To evaluate the proposed approach, a dataset taken from [23] was used with no previous idea of its contents. The data set was divide into 1000 separate files. Each file contained text on a supposedly different subject indicated with one of Economics, Art, Politics, or Sports as labeled by the corpus providers. The corpus is compromised of 200 documents per topic or label. Table 1 has an example of a short document's text, POS-Tagged, and single-symbol strings along with the more refined (REDUCED AND UNIQUE) strings.

Obtained results were analyzed and compared to figure out the percentage of correctly classified documents on one hand and to compare the accuracy of using the different strings and as compared to use of the actual text string.

As described above in the proposed procedure, the conducted experiment included:

a) Tagging of documents and feeding the resulting strings of tags to mentioned similarity algorithm to compare and calculate the similarity score for the whole set of documents.
b) A set of new refined (REDUCED AND UNIQUE) strings were then created and used as in step (1) above in order to calculate similarity based on these new substrings.
c) The results from both previous steps were compared and analyzed. The results of each of the performed steps are explained next.

As documents were relatively short and the corpus was small, the similarity function was directly applied on the actual text without any preprocessing.

The results of this were used as a baseline in comparing the results of the tagged and refined strings.

5. RESULTS, ANALYSIS AND DISCUSSIONS

To investigate the utility of the proposed procedure in identifying meaningful different levels of similarity allowing for copy detection in Arabic-based text. Two approaches were used in pair-wise documents comparison.

Both approaches looked at the percentage of documents with high/low similarity scores (referred to as TOP/BOTTOM ranges from here on) across the three sequence types . The three types of tagged strings (Full, Reduced, and Unique) were also looked at and compared to the raw text-level strings of the contents of the document.

Table 1. Sample Document and Its Produced Strings

#516 Plain Text (Smallest text document) Size roughly 1000 Char: **مراكش تتحول عاصمة للفن السابع** **جمعت مراكش مشاهير السينما العربية والعالمية مهرجانهم السينمائي الدولي الرابع تم . اختيار اربعة عشر شريطا دوليا للمسابقة الرسمية تعرض للمرة الاولى تتغيب الافلام الفرنسية وتشارك المغرب بشريط واحد حضور نجوم السينما العرب اضفى المهرجان رونقا خاصا اذ تم تكريم المخرج المصري يوسف شاهين**
POS-Tagged(): مراكش/NNP تتحول/VBP عاصمة/NN للفن/NN السابع/ADJ **./PUNC جمعت/VBD مراكش/NNP مشاهير/NN السينما/DTNN العربية/DTJJ والعالمية/JJ مهرجانهم/JJ السينمائي/DTNN الدولي/DTJJ الرابع/ADJ تم/VBD اختيار/NN اربعة/CD عشر/NN شريطا/NN دوليا/JJ للمسابقة/NN الرسمية/DTJJ تتغيب/ADJ الاولى/NN للمرة/VBP تعرض/VBP الافلام/VBP الفرنسية/DTNN وتشارك/NNP المغرب/DTNNP بشريط/NN واحد/JJ حضور/NN نجوم/NN العرب/DTNN اضفى/VBD المهرجان/DTNN رونقا/NN خاصا/JJ السينما/DTNN المخرج/NN تكريم/VBD تم/IN اذ/NNP يوسف/DTJJ المصري/DTNN شاهين/NNP ./PUNC**
Single-Sym( ): sJJXRJABggABRJEJJgJBsJsABcJgJJABRAJgGRJABX
Reduced POS: sJ2XRJABg2ABRJEJ2gJBsJsABcJgJ2ABRAJgGRJABX
Unique POS():sJXRJABgABRJEJgJBsJsABcJgJABRAJgGRJABX

Since it has been shown that the vocabulary of written text can overlap by 50% [19], the investigation focused on the top 15% (TOP-Range) score results and bottom 15% (BOTTOM-Range) ones. It is believed that such percentages are sufficient to clearly show the utility of the idea. These cut-offs values were experimentally selected as was confirmed by the presence of clear gaps in the scores. For example, there was a jump in score from 0.900 to 0.666 in the actual text's and its POS-tagged strings. The calculated scores presented a convenient cut-off-value.

In the first approach a look at the classifications power of the procedure in the specified ranges (TOP/BOTTOM) to compare the three refined string representatives (FULL, REDUCED, UNIQUE) to those of the baseline text (Plain-Text). The idea is that similarly scored documents in the TOP (BOTTOM) range should contain less misclassification and vice versa.

In the second approach, a look at whether similarly scored documents do actually contain similar text. This required manual study of the contents of those documents scored in the specified ranges to confirm results.

Investigations of the correct classification in the specified ranges served as a validation of the results. It is worth noting that the 1000-document data set was used as-is except for one blank document which was removed.

Table 1 describes the obtained results showing identical results in TOP-range for all string types. As a matter of fact the full-set of 16 document pairs rated above 0.90 and all were exact duplicates. Even the 1 shown to have different topic (as per the corpus creators) it turn out to be the same text document that apparently mislabeled. This was apparently a mistake on part so the corpus creators.

The importance of such results is in the fact that all different string types included the refined (reduced and certainly most reduced unique) set to give same distinguishing power.

Table 2. Top/ Bottom range match and mismatch

String Type	Ave SIZE Percentage	TOP Range 0.85-1.00		BOTTOM Range 0.0-0.15	
PLAIN	2.41(100%)	1/16	99.94%	449/586	76.72%
TAGGED	0.18(7.3%)	1/16	99.94%	5/12	41.67%
REDUCED	0.17(6.97%)	1/16	99.94%	5/14	35.71%
UNIQUE	0.15.06%)	1/16	99.94%	0/6	0.00%

These were interesting results considering the fact that the refined (reduced and the unique) set were smaller than the plain tagged set and certainly much smaller than the original documents as is shown in Table 1. The total reduction of the size relative to the original text was 94%. Such reduction in size can translate to huge savings in string processing and manipulation time.

Analyzing the BOTTOM range proved a little complicated due to the large number of documents falling in that range. It could, however, be seen that a lot more pairs were classified as different. This means that the systems have done a good job in not misclassifying in the bottom range. It would have been bad if documents were matching (i. e. classified as similar or identical) were found in this rage.

In general, the following important observations can be made:

a) All three string types had the same set of pairs of documents ranked in the TOP range, in different orders though.

b) All the TOP ranked pairs were actual duplicates including one that was mistakenly mislabeled.

c) There was one pair that ranked a little lower (missed by the used bracket of 0.85 or more) in all categories of strings even though it is a very near duplicate.

d) Couple of more pairs were ranked closer to the top (with similarity above 0.7) that were verified as similar in the topic and the issues, personalities or situations discussed in them.

e) One case that is a very similar pair was picked up by the three resulting strings but not by the original text's string. This was an interesting case in that it was consistently shown similar (0.68) but was not at all picked up by the text string.

f) Other cases were also picked up at a lower similarity score by four different categories but never consistently and in all cases with other non-similar ones.

g) The BOTTOM bracket varied in the number of cases falling in that range with text-based string way more populated but in all categories the % of mismatched was much higher close to 40% with hardly any similar pairs.

It should be duly noted that non-text based categories have shown equivalent results in the detection of duplicates, near duplicates and partially similar ones.

6. CONCLUSIONS

This paper reported on work performed to investigate the use combined Part of Speech (POS) tagging and string similarity algorithm that use some minimum edit operations to detect duplication and similarity between pairs of Arabic text documents. Documents' text are converted into POS tags that are then fed to string similarity algorithms to determine the similarity when comparing any two documents. A normalized score is calculated and used to rank documents.

Syntactical features of the text were used as a structural representation of the documents' content. Each document is further compared using a set of tag sequences. The full document's tagged string (FULL) was further refined by removing any repeating subsequences (REDUCED) and by using an even smaller subset compromised of the unique characters (UNIQUE).

The obtained results based on the use of a set of selected sequences that were the result of translation of text into its sequence of syntactical unit compared to the more refined sequences (FULL, REDUCED and UNIQUE) were compared. The compression showed a clear advantage of using the refined strings in terms of reduced size while still maintaining the same discrimination power as that of the original text.

Overall results obtained by the performed experiments entailed encouraging results showing a clear advantage of using the refined representation using POS-Tags in terms of text reduced size while maintaining the same discrimination power as that of the original text.

Future work will include further evaluations of the efficiency of the proposed idea in terms of processing time. The evaluation will compare results of using POS-Tagging to other similarity calculation methods such as fingerprinting and n-grams. Other applications such as text categorization and clustering based on the POS-tagged strings will be investigated.

7. REFERENCES

[1] Grune, D, and M, Huntjens, Detecting copied submissions in computer science workshops, Vakgroep Informatica, Faculteit Wiskunde & Informatica, Vrije Universiteit, AMSTERDAM, 1989.

[2] D. M. Campbell, W. R. Chen and R. D. Smith, “Copy Detection Systems for Digital Documents”, IEEE, Washington, DC, USA, May, 2000, pp. 78-88.

[3] Clough, P., Old and new challenges in automatic plagiarism detection, Department of Information Studies, University of Sheffield, 2003.

[4] Bull, J., C. Collins, E. Coughlin and D. Sharp, Technical Review of Plagiarism Detection Software Report, Computer Assisted Assessment Centre, University of Luton, Luton, UK.

[5] Kang, N., A. Gelbukh and S. Han, PPChecker: Plagiarism Pattern Checker in Document Copy Detection, 2006.

[6] A. Singhal, “Modern Information Retrieval: A Brief Overview”, Google, Inc., IEEE, 2001.

[7] Poinçot, P., S. Lesteven and F. Murtagh, Comparison of Two “Document Similarity Search Engines”, ASP Conference Series, Vol. 153, 1998.

[8] L. Bergroth, H. Hakonen and T. Taita, "A Survey of Longest Common Subsequence Algorithms", In String Processing and Information Retrieval, 7th. International Symposium on, 27-29 Sept. 2000., pp. 39–48.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic Local Alignment Search Tool", J. Mol. Biol. Vol.215, Academic Press Limited, 1990, pp. 403-410.

[10] I. Yang, C. Huang and K. Chao, "A fast algorithm for computing a longest common increasing subsequence", Information Processing Letters, Vol.93(5), Elsevier B.V., 2004, pp. 249-253.

[11] Baral, C., Local Alignment: Smith-Waterman algorithm, CSE 591: Computational Molecular Biology Course, Arizona State University, 2004.

[12] M. S. Waterman, "General Methods of Sequence Comparison", Bull. Math. Biol.Vol(46), 1984, pp. 473-500.

[13] Y. Liu and L. Liang, "A Dual-method Model for Copy Detection", IEEE, IAT Workshops, 2006, pp. 634-7.

[14] K. Monostori, R. Finkel, A. Zaslavsky, G. Hodasz and M. Pataki, "Comparison of Overlap Detection Techniques", Intern. Conference on Computational Science, Amsterdam, Holand, 21-24 Apr., 2002, pp 51-60.

[15] Steinberger, R., B. Pouliquen and J. Hagman, Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC, Springer-Verlag Berlin Heidelberg, 2002.

[16] Esko Ukkonen (1983). On approximate string matching. Foundations of Computation Theory. Springer. pp. 487–495.

[17] Navarro, Gonzalo (1 March 2001). "A guided tour to approximate string matching" (PDF). ACM Computing Surveys.33 (1): 31–88. doi:10.1145/375360.375365. Retrieved19 March 2015.

[18] Daniel Jurafsky; James H. Martin. Speech and Language Processing. Pearson Education International. pp. 107–111.

[19] Finlay S (1999). CopyCatch, Masters Dissertation, University of Birmingham.

[20] Elhadi, M. Al-Tobi, M. "Detection of Duplication in Documents and WebPages Based Documents Syntactical Structures through an Improved Longest Common Subsequence", IJIPM: International Journal of Information Processing and Management, Vol. 1, No. 1, pp. 138 ~ 147, 2010.

[21] Mohamed Elhadi, Text Similarity Calculation Using Text and Syntactical Structures, 8th ICCIT: 2012 International Conference on Computer Sciences and Convergence Information Technology, December 3-5.2012, Seoul, Korea.

[22] Mohamed Elhadi and Amjad Al-Tobi Use of Text Syntactical Structures in Detection of Document Duplicates, Third IEEE International Conference on, Digital Information Management ICDIM 2008, University of East London, London, UK 2008.

[23] Bani-Ismail, B, Al-Rababah, K, Shatnawi, S., The effect of full word, stem, and root as index-term on Arabic information retrieval , Global Journal of Computer Science and Technology, 2011.

[24] Mohamed Elhadi and Amjad Al-Tobi Webpage Duplicate Detection Using Combined 2009 World Congress on Computer Science and Information Engineering (CSIE 2009), March 31 - April 2, 2009, Los Angeles/Anaheim, USA.

[25] Mohamed Elhadi and Amjad Al-Tobi Duplicate Detection in Documents and WebPages using Improved Longest Common Subsequence and Documents Syntactical Structures, 4th ICCIT: 2009 International Conference on Computer Sciences and Convergence Information Technology November 24-26, 2009, Seoul, Korea.

[26] Mohamed Elhadi and Amjad Al-Tobi, Part of Speech (POS) Tag Sets Reduction and Analysis using Rough Set Techniques, Twelfth International Conference on Rough Sets, Fuzzy Sets, Data Mining & Granular Computing RSFDGrC 2009, Indian Institute of Technology, Delhi, India, December 16-18, 2009

[27] A, G. Maguitman, F, Menczer, H. Roinestad and A. Vespignani, "Algorithmic Detection of Semantic Similarity", International World Wide Web Conference Committee, 2005, pp.107-116.

[28] Mihalcea, R., C, Corley and C, Strapparava, Corpus-based and Knowledge-based Measures of Text Semantic Similarity, American Association for Artificial Intelligence, Jul, 2006.

[29] S. Schleimer, D. S. Wilkerson and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting", International Conference on Management of Data, ACM, 2003, pp. 76–85.

[30] Mohamed Elhadi and Amjad Al-Tobi, Refinements of Longest Common Subsequence Algorithm, ACS/IEEE International Conference on Computer Systems and Applications. Hammamet, Tunisia, May 2010.

[31] Eugene Myers , "An O(ND) Difference Algorithm and its Variations", , Algorithmica Vol. 1 No. 2, 1986, pp. 251-266;

[32] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[33] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.