Review

N-gram Overlap in Automatic Detection of Document Derivation*

Siniša Bosanac
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
sbosanac@ffzg.hr

Vanja Štefanec
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
vstefane@ffzg.hr

Summary

Establishing authenticity and independence of documents in relation to others is not a new problem, but in the era of hyper production of e-text it certainly gained even more importance. There is an increased need for automatic methods for determining originality of documents in a digital environment.
The method of n-gram overlap is only one of several methods proposed by the literature and is used in a variety of systems for automatic identification of text reuse. Although the aforementioned method is quite trivial, determining the length of n-grams that would be a good indicator of text reuse is a somewhat complex issue. We assume that the optimal length of n-grams is not the same for all languages but that it depends on the particular language properties such as morphological typology, syntactic features, etc. The aim of this study is to find the optimal length of n-grams to be used for determining document derivation in Croatian language.
Among the potential areas of implementation of the results of this study, we could point out automatic detection of plagiarism in academic and student papers, citation analysis, information flow tracking and event detection in on-line texts.

Key words: Document derivation, text reuse, n-gram overlap, automatic plagiarism detection, string metrics

* The preliminary results of this research were in somewhat shorter form presented at the student linguistic conference StuLiKon, held on 6th – 8th May, 2011 in Belgrade, Serbia.

Introduction

The problems of originality and authenticity were always contemporary, but nowadays they demand even more attention. There are two principal factors that have made these problems so important: development of information and communication technology and commodification of intellectual products. Our objective was to deal with only a small segment of this problem that can be examined empirically – derivation of content. In this paper we will establish the basic theoretical framework of document derivation detection and describe our effort to test and modify an existing method for use on natural language texts in Croatian. Text reuse is a very common occurrence in the process of text production and some say that it is as old as storytelling itself.[1] In most cases the motivation to examine it came out of frustration caused by witnessing plagiarism and similar negative phenomena, and the same is valid for this research.[2]
Increased usage of ICT for producing texts and their dissemination has intensified this phenomenon, and also removed some of the technical limitations. Ethical and legal factors are also to be considered when analyzing the problem, but they will not be dealt with in this paper. An example supporting the claim that plagiarizing has become socially acceptable and moreover, profitable, is the existence of on-line markets of essays.[3] Automatic methods for detection of derivation have been researched to a far greater extent for use in the IT sector, where they are used to detect plagiarized programming source code, than for natural language texts.[4] Efforts made for detecting reuse in natural language text were aimed at protecting the content distributed by news-wire agencies, and for designing tools for detection and discouraging plagiarism in academia.[5]

Theoretical framework

Text reuse is a process by which literal content from a single source document is reused in the creation of a target document. Content is reused in the same context either word-for-word (verbatim) or paraphrased (rewritten).[6]
Derivation is the relationship between the two documents in which can be shown, with confidence, that the target document used the source document in its creation.[7]

---

[1] Wilks (2000), according to Clough (2001, p.3).

[2] The authors of this paper have independently dealt with plagiarism in their previous works: Bosanac *et al.* (2009), Štefanec (2010), Bosanac & Štefanec (2011).

[3] For examples refer to: http://seminarski.blog.hr/, http://www.maturskiradovi.net/.

[4] Clough (2001, p.12).

[5] Measuring Text Reuse project at the University of Sheffield was dealing with text reuse in news articles; Turnitin and Plagiarism Detect are tools popular in academia.

[6] Clough (2001, p.27).

We have listed, according to our experience, the most common examples of content derivation. We do not consider this list to be exhaustive.
Examples of content derivation generally considered desirable: 1) *Quoting* and using as sources texts from the domain of scientific and technical literature, literary fiction and journalism; 2) *Document updating*, i.e. adding new content to existing documents, but it can also refer to on-line journalist articles; 3) *Relaying the content* of press releases and content supplied by news agencies in media publications; 4) Automatic and manual *summarization*.
Examples of content derivation generally considered undesirable: 1) *Plagiarism* in all domains, but specifically in academia, journalism and literary fiction; 2) *Non-critical relaying* of content in media, e.g. disseminating information without accrediting the source; 3) *Journalistic theft* which occurs when a journalist copies the content from another media that has legally obtained it through subscription to a news agency's feed.
In our research, we have recognized the following methods of reusing text in a new document: *copy-pasting* (verbatim reuse) and paraphrasing (rewriting). Text considered to be reused verbatim has exactly the same word forms and structure as in the source document, shares the same context in both documents and can be aligned with the corresponding text in the source document.
Text that is reused by rewriting shares some similar word forms between the source and target documents, not all the words are exactly the same due to editing transformations, the context is the same and the rewritten text can be aligned with corresponding text from the source document.[8] For example this is a result that shows text reuse by paraphrasing:
Source document: "***U planinskom području Perua**, daleko od glavnog grada **Lime i daleko od bilo kojeg grada**, postoji **mreža seoskih knjižnica**.*"
Target document: "*U svom govoru Kay Raseroka spomenula je interesantan primjer razvoja ruralnih knjižnica **u planinskom području Perua**. **Daleko od glavnog grada Lime i daleko od bilo kojeg** drugog **grada** razvila se **mreža seoskih knjižnica**.*"[9]

[7] Clough (2001, p.28).

[8] Clough (2001, p.29).

[9] Source document: "In the mountainous area of Peru, far from the capital city Lima, and far from any city, there is a network of village libraries."
Target document: "In her speech, Kay Raseroka mentioned an interesting example of development of rural libraries in the mountainous area of Peru. Far from the capital city Lima, and far from any other city, a network of village libraries has developed."

Detection of document derivation

N-gram overlap

In linguistics, the term *n-gram* denotes a sequence of *n* successive language units. Depending on the application, those language units can be letters, syllables, morphemes, words, etc. In this case we will be dealing with n-grams of words. Using the n-gram overlap method we will try to determine the amount of n-grams common to two documents. This method is based on the assumption that overlapping of longer n-grams could be an indicator of text reuse. However, the most representative n-gram length for detecting derivation in a certain language should be determined experimentally because it depends on the morphological and syntactical features of the language. It is necessary to bear in mind that texts in the same language share a certain amount of lexis, and that amount increases substantially if the texts are on the same topic or belong to the same functional style. Therefore, if too short n-grams are taken into account, matching will be higher and nonrepresentational, while if too long n-grams are selected there is a risk that not all cases of derivation will be recognized. Some of the authors dealing with this problem in English are Lyon *et al*. (2001)[10] who showed that word trigrams could be used for discriminating independent from copied text within a collection of texts, Bloomfield[11] who claimed that 6-grams are more representative for identifying collusion between student assignments, Shivakumar and Garcia-Molina (1995) who find unigrams to be optimal for copy detection, etc. From such diverse results it is evident that the problem is far from simple and that we cannot even predict the interval in which our results could fall into. Besides finding the most representative n-gram length, we will try to determine the amount of common n-grams which can be used as a threshold for automatic derivation detection.

N-gram overlap is not the only method that can be used for detection of text reuse, but it is certainly one of the simplest. It lacks precision since it takes only fixed sized n-grams and performs simple binary comparison, but its low complexity makes it suitable to be applied over large collections of documents. In that way, this method can be used for identifying candidates that should be analyzed more thoroughly using some of the more precise and complex algorithms. For example, the widely used Longest Common Subsequence, or somewhat less familiar Greedy String Tiling algorithm. While the first is the basis of various file comparing functions and programs, the latter is specifically designed for detection of plagiarism in computer programs and other texts.[12]

[10] According to Clough (2003, p.117).

[11] *Ibid*., p.118.

[12] Wise (1993).

In this research we will perform measurements on a collection of texts in order to determine the relevant parameters for using n-gram overlap method in identifying derivation of documents in Croatian.

Measure

In order to be able to quantify the amount of text reuse in derived documents, it is necessary to introduce measure capable of expressing that kind of relation. Measure commonly used for this purpose is *resemblance*.

Resemblance is a symmetric measure which expresses the amount of common content within the total content of two documents. Resemblance will be calculated using the Jaccard similarity coefficient or Jaccard index. Since Jaccard index is a statistic for comparing the similarity of sample sets[13], our documents will be treated as sets of n-grams. That means that, due to the fact that a set consists only of unique elements, we will be dealing with types of n-grams instead of tokens, i.e. only one occurrence of n-gram will be recorded in the set.[14] Resemblance will then be computed as the size of the intersection divided by the size of the union of n-grams.

$$r(A,B)=\frac{|F(A)\cap F(B)|}{|F(A)\cup F(B)|}\quad[\cdot 100\%]$$

Although it is possible to use weighted resemblance instead, in which case higher weight is given to more frequent n-grams, it will not be used here.

Text collection

The text collection on which we conducted our measurement consisted in total of 238 documents of various sizes (69 – 34,397 words). They were taken from the digital repository of the Library of the Faculty of Humanities and Social Sciences, Web news sites, and other Web sources. Documents in the collection were classified by topic and functional style. The general topic of documents was determined according to classification used at the source, title, keywords, and additionally, for shorter texts, manually. The functional style was determined primarily according to the classification at the source. The collection consisted of 39 diploma papers, 42 scientific articles, 61 news articles, 61 literary columns, and 35 documents classified as "other" from the fields of library science and psychology.

Methodology

For the purpose of this research, we have composed a module in a dynamic programming language using which we have performed measurements over our

[13] Wikipedia contributors (2011)

[14] Clough (2001, pp.51-52)

text collection. The module was designed to combine all documents from the text collection into binomial combinations (they will be referred to as *derivation pairs*) and perform identification of common n-grams within them. In that way, out of 238 documents our module produced 28,203 derivation pairs which will serve as our training data set.

$$C_2^{238} = \binom{238}{2} = \frac{238!}{2!(238-2)!} = 28203$$

Documents in every derivation pair were compared in 10 subsequent iterations. In each iteration the comparison was performed with respect to n-gram of different length, starting on the level of 10-grams to the unigram level. In addition to calculating the resemblance in each iteration, the module generated an exhaustive summary containing the list of common n-grams for every derivation pair. After completing the measurements, every derivation pair was given a desired output value, namely, derivation pairs were marked either as derived or non-derived using a semi-automatic method. Specifically, derivation pairs consisted of documents for which there is a negligible likelihood of derivation were automatically marked as non-derived, while those for which non-derivation cannot be assumed, were classified manually by examining the generated summary. In overall, out of 28,203 pairs, derivation was established in 265 of them, while other 27,938 pairs were marked as non-derived. After that, all derivation pairs were ranked according to resemblance respectively for each n-gram length.

Further analysis of the results included finding the resemblance value that yields the maximum F1-measure for each respective n-gram length. Resemblance value will be used as a threshold for automatic classification of derivation pairs, and precision and recall scores will be obtained by comparing the output of the algorithm with the desired output value. F1-measure, as a harmonic mean of precision and recall, will show us for which resemblance value we have the optimal ratio between false-negatives and false-positives. Figure 1 gives an illustration of how F1-measure changes with the resemblance on the 3-gram level.

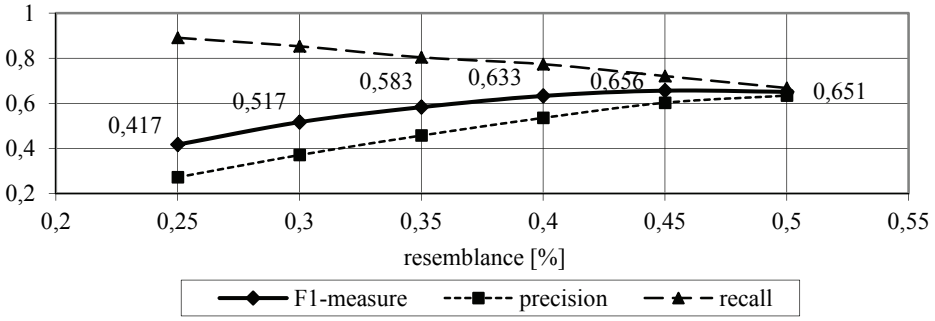

Figure 1: Trend of precision, recall and F1-measure on the level of 3-gram

After finding the F1-measure maxima at the level of each n-gram respectively, by comparing their values, it will be possible to determine at the level of which n-gram F1-measure reaches the highest value. That n-gram will then be considered as the most representative for detection of derivation, and the resemblance which yields that F1-measure value will be considered as a threshold for distinguishing derived from non-derived pairs. It is important to mention here that resemblance values on different n-gram levels are not mutually comparable, and cannot be directly converted.

Results

General results

The results are presented in the Figure 2. As we see, F1-measure reaches its maximum on the level of 6-grams with the value of around 0.82. Results in more detail are given in Table 1.

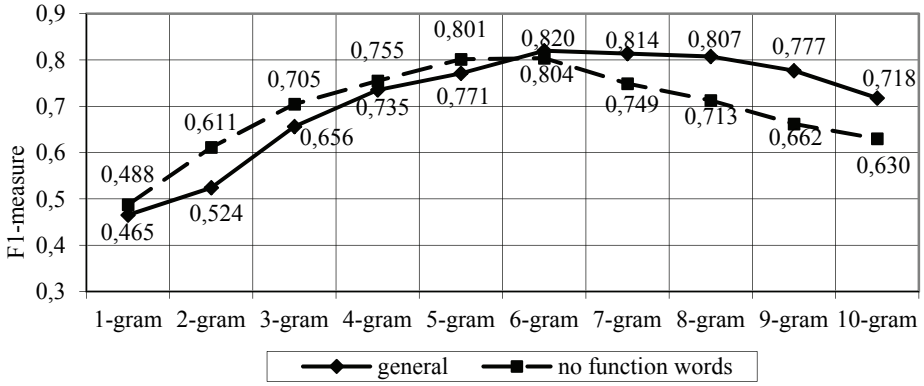

Figure 2: F1-measure maxima on the level of each n-gram respectively

Table 1: Detailed results of the analysis of measurements

	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram
resemblance [%]	0.45	0.15	0.06	0.025	0.015	0.01
precision	0.603	0.717	0.755	0.833	0.868	0.894
recall	0.721	0.755	0.789	0.808	0.766	0.736
F1-measure	0.656	0.735	0.771	0.820	0.814	0.807

Removing function words

Function words are words that have little lexical meaning, but instead serve to express grammatical relations. These are regularly high-frequency words and most of them are common to all texts written in the same language.[15] Our inten-

[15] Manning and Schütze (1999, p.20)

tion was to reduce the number of n-gram overlaps caused by these high-frequency words. We assumed that by removing function words from the text we shall increase the sensitivity of the method. Words that will be treated as function words and removed from the documents in our text collection are 50 most frequent words from the Croatian National Corpus[16]. The results are also shown in Figure 2.
From the results we can see that even though F1-measure increased on the level of shorter n-grams, for longer n-grams we have significantly worse results. Surprisingly, overall result remained practically the same; measurement over original documents gave just slightly better score.

Distinguishing functional styles
As we know, language does not always function in the same way, but in that many ways that the society needs.[17] Differences can be found in various language features, such as lexical choice, syntactic constructions, lexical and sentential semantics, etc. And functional style is a subsystem of a language that comprises all features characteristic for a certain function. These differences can manifest themselves in a way that certain features are avoided in some styles while favored in others, or in some styles occur more frequently than in others. By distinguishing functional styles in the analysis, we wanted to check whether the functional style of the documents affects the results in some way.
According to Silić and Pranjković (2005, p.375), in modern standard Croatian language, five functional styles can be distinguished, and these are: scientific (*znanstveni*), official (*administrativno-poslovni*), newspaper and publicistic (*novinsko-publicistički*), literary (*književnoumjetnički*) and colloquial (*razgovorni*) functional style. Most of the documents from our text collection could be associated with one of the two functional styles, namely, scientific and publicistic and newspaper style. We have focused on these two styles for several reasons. Documents written in these styles were easily available and because of the ethical implications mentioned earlier, the question of originality is especially emphasized when speaking of texts written in these styles. Derivation pairs consisted of documents written in the same functional style were filtered and analyzed respectively. The results are shown in Figure 3.
The analysis showed that although considerable differences can be seen in the trend of F1-measure between the scientific and publicistic and newspaper style, overall results showed no significant change. It is interesting to notice, though, how for pairs written in the scientific style F1-measure remains practically con-

[16] These words are: *i*, *je*, *u*, *se*, *na*, *da*, *za*, *su*, *od*, *s*, *o*, *a*, *će*, *koji*, *ne*, *iz*, *što*, *bi*, *to*, *nije*, *ili*, *te*, *kako*, *kao*, *do*, *koje*, *biti*, *koja*, *godine*, *ali*, *samo*, *sve*, *jer*, *još*, *sam*, *više*, *po*, *sa*, *može*, *prema*, *već*, *nakon*, *dana*, *bio*, *bilo*, *zbog*, *li*, *smo*, *pa* and *ni*.

[17] Silić and Pranjković (2005, p.375)

stant from the 4-gram level to the level of 9-grams (difference is barely 0.02). On the other hand, pairs written in publicistic and newspaper style show the strong tendency towards 6-grams as indicators of derivation.

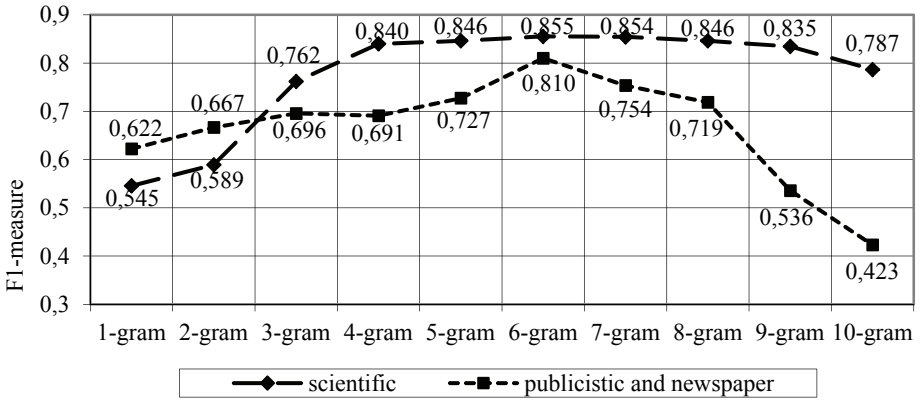

Figure 3: F1-measure maxima on the level of each n-gram respectively – differentiating functional styles

Conclusion

In this research, we have performed several measurements over the text collection to determine the most representative n-gram length for detecting reuse of text in Croatian language. We have taken 3 possible approaches to this problem: 1) measurements were performed over the original texts; 2) measurements were performed after the function words were removed from texts; and 3) derivation pairs consisted of documents written in the same functional style were analyzed separately. In every approach, we have proven that 6-gram is the optimal n-gram for detecting text reuse. Measurements by other authors on English texts have shown 7-gram to be the optimal one.[18]

Further on, we have concluded that removing function words as a text preprocessing method does not increase the sensitivity of the method, on the contrary, the quality of overall results decreases as n-grams are longer.

The first step in further research would be to enlarge the text collection and refine its system of text classification. The second would be to experiment with different kinds of text editing; e.g. POS tagging and extracting *hapax legomena*, stop words, labels or direct quotes. The third option would be to focus on a level of text other than words and observe it as a string of characters or sentences.[19]

Detection of derivation by direct translation from other languages is a challenge that goes beyond the scope of this paper, but should be very seriously consid-

[18] Clough (2003, p.185)

[19] As done in measurments on the METER corpus, Clough(2003, p.185)

ered when developing practical tools for use on documents in non-global languages.

We consider the results obtained by this research merely as a reference because the parameters which are to be used in a practical implementation highly depend on the requirements laid before it. Some implementations, for example in plagiarism detection systems, will probably favor recall over precision and thus choose shorter n-grams with lower resemblance threshold, while in others, such as identifying verbatim copies of news agencies articles, will most likely favor precision over recall and use longer n-grams with high resemblance threshold.

Nonetheless, we are confident that our findings will serve as a valuable contribution in developing methods and tools for automatic detection of derivation, which will find their application both in supporting academic integrity, and in aiding further research in the field of information sciences. The practical application of automatic methods for plagiarism detection should always be widely publicized so that it can help fight plagiarism by serving as a deterrent rather than a tool for enforcing justice and seeking punishment after the offense is already committed.

References

Bosanac, Siniša; Mandić, Bojana; Sprčić, Andrija. Objective Journalism or Copy-Pasted Press Releases: A Preliminary Media Content Analysis. // *INFuture2009: Digital Resources and Knowledge Sharing* / Stančić, H. *et al.* (ed.). Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2009, pp. 417-425.

Bosanac, Siniša; Štefanec, Vanja. Preklapanje n-grama pri automatskoj detekciji deriviranosti dokumenata. // *StuLiKon – Studentska lingvistička konferencija*. Beograd: Univerzitet u Beogradu, 2011 (in print)

Clough, Paul D. Measuring text reuse. PhD thesis. Sheffield: University of Sheffield, 2003. Available at: ir.shef.ac.uk/cloughie/papers/thesis.pdf.

Clough, Paul D. Measuring text reuse and document derivation. PhD transfer report. Sheffield: University of Sheffield, 2001. Available at: ir.shef.ac.uk/cloughie/papers/transfer.pdf.

Manning, Christopher D.; Schütze, Hinrich. Foundations of statistical natural language processing. Cambridge; London: The MIT Press, 1999

Shivakumar, Narayanan; Garcia-Molina, Hector. SCAM: A Copy Detection Mechanism for Digital Documents. // *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*. Austin, Texas, 1995. Available at: http://ilpubs.stanford.edu:8090/95/1/1995-28.pdf.

Silić, Josip; Pranjković, Ivo. Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta. Zagreb: Školska knjiga, 2005

Wikipedia contributors. Jaccard index. *Wikipedia, the free encyclopedia*. May 23, 2011. http://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=430488441 (Accessed July 6, 2011)

Wise, Michael J. String Similarity via Greedy String Tiling and Running Karp−Rabin Matching. Sydney: Department of Computer Science, University of Sydney, 1993. Available at: www.pam1.bcs.uwa.edu.au/~michaelw/ftp/doc/RKR_GST.ps.

Štefanec, Vanja. Automatska detekcija plagijata. Unpublished paper. Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb, 2010