OPEN ACCESS

1st International Electronic Conference on Entropy and Its Applications
3 - 21 November 2014

www.sciforum.net/conference/ecea-1

Conference Proceedings Paper – Entropy

Using of the Statistical Method for Authorship Attribution of the Text

Olga Kanishcheva [1,*]

[1] National Technical University "Kharkov Polytechnic Institute", ul. Frunze 21, Kharkov, Ukraine, 61002; E-Mails: kanichshevaolga@gmail.com (F.L.)

* Author to whom correspondence should be addressed; E-Mail: kanichshevaolga@gmail.com (F.L.); Tel.: +380-666-818-498.

Received: 4 September 2014 / Accepted: 21 October 2014 / Published: 3 November 2014

Abstract: This paper is devoted to the statistical methods for authorship attribution of the texts. The model was tested for the Russian language, but can also be applied for different languages. These methods are implemented using programming scripting language JavaScript.

Keywords: mathematical linguistics; statistical models; plagiarism; author's vocabulary, author's syntactics

1. Introduction

With the advent of the Internet, the plagiarism has gained significant features of the problem, because the knowledge contained in the texts and posted on its resources, becomes the common property. The rapid development of the Internet complicates the problem of copyright protection, sometimes it becomes impossible to identify the original author. Increasing of computer literacy in the society encourages the penetration of plagiarism into various spheres of human activity. So the plagiarism is a severe problem in education, industry and the scientific community.

Let’s give a definition of the phenomenon of "plagiarism". So, S.I. Ozhegov gives such an interpretation in an explanatory dictionary: "Pose the others' works as someone's own or illegal publication of someone else's work under someone's own name, the attribution of authorship". [1] In the explanatory dictionary of modern Russian language [2] is available a broader interpretation of the plagiarism: "The assignment of the fruits of another's creativity, the publishing of someone else's work under someone's own name without specifying the source or use without the transforming creative changes made by the person, who lends".

The analysis of the above definitions shows that the concept of "plagiarism" has no clearly defined meaning. These circumstances lead to the fact that it is not always possible to separate it, for example, from the imitation, co-authorship, borrowing, etc. In scientific and artistic achievements of the authors coincidence of ideas is not plagiarism, because they can not belong to the author. However, all the works of intellectual activity of a community based on previously created artifacts, which are their development or any other interpretation. However, the definitions are united by the fact that an act of misappropriation of authorship. In Ukraine, the protection of intellectual property is regulated by law. [3] However, the said law emphasizes that borrowing a theme or plot of a work, or scientific ideas that make up its contents, without borrowing a form of expression is not plagiarism.

Thus, the manifestation of this phenomenon may occur in art, science and technology, that the results of human intellectual activity or team. When considering legal action for theft of intellectual property judges find themselves in a difficult position, because it is necessary to separate the plagiarism of simulation, co-author or borrowing. This question can be answered only by specialists - professional philologists and linguists who conducted the examination.

Linguistic expertise is one of the types of linguistic research, which is assigned by an authorized person (body) in order to establish legally significant facts. In linguistic terms of linguistic expertise is the kind of research facilities that establishes the truth or falsity or the possibility or impossibility of descriptive statements about this object (objects). This linguistic aspects of the examination is based on currently existing theories of language and linguistics techniques developed in the study of linguistic objects.

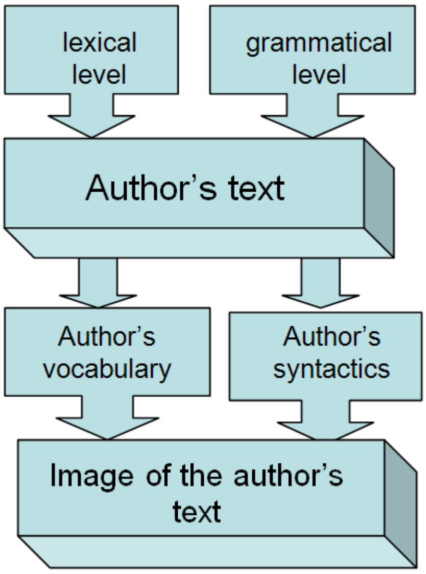
Due to this important area of research is the use of applied linguistics in law. In connection with this the topical area of research is the use of applied linguistics in law. It may be useful experience in applied linguistics concerning analytical processing of large volumes of textual information to solve the problems in the law [4]. For example, these capabilities of applied linguistics can be useful for solving the problems of the automation of linguistic analysis during holding of linguistic examinations of authorship attribution, holding of which is associated with the processing of large volumes of textual information. Thus, the integration of applied linguistics creates preconditions for the formulation and solution of the problems of intelligent processing of textual information in law, including authorship attribution.

The peculiarity of the problem of authorship attribution is the need to get the image of the stylistic presentation of the author's text. Based on the characteristics of the problem of authorship, let's look more closely at the composition of methods for holding of linguistic examination of authorship attribution. It is enough to use the associative-statistical approach and a combination of the methods of lexical and grammatical levels in the framework of this approach. The scheme of the combination of methods of lexical and grammatical levels is shown in Fig. 1.

2. The Statistical Model of Authorship Attribution of The Texts

The scheme of the combination of methods shows that it consists of lexical and syntactic levels. Using of the syntactic level suggests the computing of the linguistic relationships in combinations of words. Let's propose the model for constructing of the images of the author's style , which consists of the characteristic author's vocabulary and author's syntactics inherent to the author. To describe the syntax it is needed to use the formalized description of linguistic relations between lexical units of the phrase on the set-theoretic language. In the work [5] is propounded the formalized description of any text, but the formalized description of the linguistic relations between lexical units is not propounded. The formalized description of the text is also propounded in references [6, 7]. In the reference [6] the formalized text presentation was set up for the automation of the procedures of analysis of the scientific and educational texts with the object of definition of semantically meaningful fragments. In this work is propounded the set-theoretic description of the linguistic relations in phrases. Such models can be used for the description of the images of the author's vocabulary and author's syntactics, but they don't take into account the statistical information about frequency of the vocabulary and syntactics. The formalized description, which used for the analysis of the text of terminological dictionary with the view of building of the semantic network of its terms, is propounded in the reference [7]. However, the suggested model also doesn't stipulate the accounting of the statistical information about frequency of occurrence of vocabulary and syntactics.

Figure 1. The scheme of a combination of the methods of lexical and grammatical levels.



We can apply the formalized presentation of the text, as it was done in the reference [5], where text T is considered as the set of phrases composing given text: $T=\{\theta_\alpha\}$, where θ_α – are phrases of the text; $\alpha=\overline{1,n}$, n – ordinal number of the phrase in the text; n – the quantity of phrases in the text. Every phrase is described by the cortège from the sets of syntactic schemes and their mappings on the alphabet of allologs, which were used in the text: $\theta=\langle C_c,\psi\rangle$, where C_c – the syntactic scheme of the phrase, which is the cortège of two sets $C_c=\langle M,A\rangle$, where $M=\{x_i\}$ – the set of allologs x_i, which are the part of the phrase, i – ordinal number of the allolog in the phrase, $A=\{a_j\}$ – the set of the linguistic relations in the phrase; Ψ – mapping of the set M on the alphabet U, $\psi:M\to U$, where U – alphabet, the set of allologs, which are the part of the author's text.

Each phrase of the text consists of allologs, the sequence of which is strictly ordered and the grammatical form of which is determined according to the rules of natural language, creating the syntactic schemes of the phrases or their parts. The syntactic schemes reflects the linguistic relations $A=\{a_j\}$between the allologs of the phrase (concord, government, etc.).

Inflectional languages (Russian, Ukrainian, Belarusian and others) have five types of linguistic relations: the relation of consequence, concord, grammatical government, the relation of entry into the constituents and homogeneity. In the reference [1] is done the set-theoretic description of the above-cited linguistic relations, but we don't show it in given work.

Above-cited model doesn't stipulate the accounting of the statistic information about frequency of occurrence of vocabulary and syntactics, which is needed to construct the images of author's text. Therefore, you should modify the above-cited formalized representation of the text which take into consideration the probabilistic characteristic of author's vocabulary and syntactic schemes (syntactics) of the phrases in the text.

As was showed above the peculiarity of linguistic expertise is the necessity of composing the types of author's vocabulary and syntactics, constructing the author's text type. Appointed types have to contain probabilistic characteristics. Because of above-cited model is necessary to modify with the help of introduction statistical (probabilistic) characteristics. Then, as a result of modification, the set of allologs of author's text $M=\{x_i\}$ is transformed in $M=\{x_i,k_j\}$, where x_i – the allologs, meaning grammar category "substantive", in author's text, and k_j – the probability of such forms appearance in author's text. It's no need in determination the content of linguistic relations between the allologs, so linguistic relations set $A=\{a_j\}$ is modified to syntax constructions set in such way. Every element a_j is put in accordance with certainty of elements x_i, where index meaning x may change from 1 to 5. The choice of index interval is substantiate in reference [8] and means that allologs, surrounding the element x, are liable to syntax constructions x_i analysis. For example, if we are analyzing construction, where element x_i stands before element x_{i-1} in the text, this is corresponding to syntax construction a_{j-1}. If we are analyzing the construction with element x_i in the text, which corresponds to the syntax construction a_j, which is marked by substantive – one of the key word of author's vocabulary. Then the set of author's syntactic schemes (syntactics) modifies to following $A=\{a_j,k_f\}$,where a_j – author's syntax construction proceeding, k_f – probability of such construction appearance in author's text.

Proceeding from above-cited, we can show the author's text type by such cortège: $O=\langle M,A\rangle$, where $M=\{x_i,k_j\}$ – set of allologs with high rate of probability using in author's text; $A=\{a_j,k_f\}$ – set of author's syntax constructions (syntax items). The magnitude probability and appearance the key words in author's text and typical syntax constructions could be get with the help of author's text frequency analysis according to the Zipf's law.

3. The Statistical Method Using For Identification The Text Author

The problem of authorship recognition reduces to standard identification problem, which is solved in three phases: separating the signs, that characterize the recognizing object, dividing the signs set to classes and coming to the decision to the proper quarter of the object to one or another class.

The point of signs separating phase consists of the form of such signs, which allow computer processing of data. Solving the authorship recognizing problem, signs set are linguistic objects, consisting of author's words subset (author's vocabulary) and subset of indicative author's utterance (author's syntactics). It's enough to use statistic-associating approach to form such subset. This approach stipulates for combination of methods vocabulary and grammar levels. The statistical utilization for appointed subsets gives the opportunity to accomplish forming the author's type text and present it in format due to computer processing of data.

Methodological basis for statistical methods in linguistics is mathematical statistics, which methods are applied to linguistic objects. This allows accomplish calculation the probability the beginning of the events (words or speech measure) and their origin probability.

The basis of statistical methods is the laws formulated by Swiss statistic J. Zipf according to results of processing the great amount statistical data. He showed that the dividing of words in natural language comply with one simple law, the main point of which is in the following. The list is composed from allologs utilized in natural language text. Every allolog in this list is put in accordance with number of it's using in the text. The number of allolog utilization is called the frequency of utilization. In resulting list allologs are fall in decrease utilization order and enumerate. The ordinal number of allolog in the list is called it rank. Zipf's law holds that product of allolog's rank and it utilization frequency is constant value. Analytical formula Zipf's law is following: $C = f \times r$, where C – empiric constant (own to the each language); f – the number of utilizing the word in text (the frequency of utilizing); r – ordinal word number in frequency list (word rank). Later the American mathematic B. Mandelbrot [9] provides evidence of J. Zipf's empiric formula.

Due to the fact introduction of Internet network and necessity of solving the problems connected with searching text information recourses in this network, search system vendors base on appointed laws and some it modifications with the object of solving the searching text information problems and their clustering. As a consequence of method TF – IDF (from English TF – term frequency, and IDF – inverse document frequency) appears. TF – IDF is the statistical index utilized for validation of importance the words in document context, being the part of document collection or corpus [10]. Weight (significance) of the word is proportional to using magnitude of this word in document and inversely proportional to frequency of using this word in other documents of collection.

Thus foregoing method can be accommodated to solve the problem of authorship establishment. In such formulation of problem searching type (request) will be considered text or it part need to identification, and document collection – any author language glossary. Such assumption is making sense from following reasons. At first, author glossary is concentrated mark of author vocabulary and syntactics. At second, if text specimen (request), need to identification, is given in glossary format, i.e. using the glossary indexes, we can utilize one of plenty classification methods, e.g. support vector machine (SVM), request text to author language.

To solving the problem of authorship establishment request (text need to identification) and collection of texts should be set in common formats. This will supply the data statistical text model and will allow form the set $M = \{x_i, k_j\}$ for request text.

Certainty of text processing operations is following:

- All request text is to be processing graphemathically, separate lexical items and punctuation marks are to be selected;

- Application vocabulary array x_i is formed from normalized word linguistic usage, where *i* – number of word linguistic usage in require (text), $i=\overline{1,n}$ where n – quantity of words in text require;
- Application vocabulary array subject to frequency analysis. The result of the analysis will be list (r_j – lexical unit list number) application vocabulary x_i^z with application frequency $f_{x_i}^z$, ranking is made with decreasing of application frequency $f_{x_i}^z$ in text require;
- Application vocabulary list is limited to the number r_j lower bound application frequency $f_{x_i}^z$, separating the most significant words x_i^z and words forming vocabulary require. The lower bound (number r_j) is set empirically;
- Calculate probability value k_{x_i} of the appearance of each vocabulary unit x_i^z according to the following formula: $k_{x_i}^z = \frac{f_{x_i}^z}{n}$;
- To form the values multidimensional vector require coordinates in database. The values are set x_i^z with their emergency probability in text require k_{x_i}, i.e. $M_z = \left\{ x_i^z, k_{x_i^z} \right\}$.

On account of needs the statistical require text it's necessarily to form the syntactic require array in following certainty:

- Construct the used words pairs list x_i^z, x_{i+1}^z for every used word and following it in require text;
- The formed used words pairs list is subject to frequency analysis. The result of analysis will be the array used syntactical construction a_j^z, where *j* – syntactical construction number;
- The array used syntactical constructions rank in descending order frequency usage $f_{a_i}^z$;
- The constructions a_j^z having crest value frequency usage $f_{a_i}^z$ are considered to be significant syntactic require.

The procedure of building the following syntactic constructions and three-digit groups used words is accomplished similarly, i.e. $x_i^z, x_{i+1}^z, x_{i+2}^z$, processing according to the items 2 – 4 creates the syntactical require construction $j+1$. In the statistical text model is substantiated that necessary syntactic construction quantity is $j=\overline{1,5}$.

For resulting syntactic constructions require it's necessary to calculate the probability values $k_{a_j}^z$ appearance every syntactical construction a_j^z in text require according to next formula: $k_{a_j}^z = \frac{f_{a_i}^z}{m_j}$, where $f_{a_i}^z$ – the crest value frequency usage construction a_j; m_j – the general constructions quantity j type in text require.

The author syntactic array is formed from resulting data $A_z = \left\{ a_j, k_f^z \right\}$.

Therefore the database has to have two arrays characterizing require: the vocabulary array $M_Z = \left\{x_i^z, k_f^z\right\}$ and syntactic array $A_Z = \left\{a_j, k_f^z\right\}$ text need to identification. As well as database has to have the author text collection. I.e. corresponding arrays characterizing author language, it's vocabulary $M_a = \left\{x_i, k_{x_i}^a\right\}$ and syntactic $A_a = \left\{a_j, k_{a_j}^a\right\}$ are to be formed.

The elements appointed arrays are the multidimensional vectors coordinates. The arrays M_Z and A_Z create vector searching type require. The arrays M_a and A_a creates the author text vector set. Calculating membership vector require to author text vector set allow to come to the decision about the authorship text require.

4. Software Implementation

Software implementation includes the following blocks:

- Filling unit content of the lexical database GENERATOR;
- A block of text processing and computing accessories query text to the author's style – PARSER.

Software function generator unit is processing information from tables 1 and 2 and the automated generation of table relationships in a database. The input to the program served an array of of the author's vocabulary from the dictionary [11]. From this array is formed by an array of the author's vocabulary index_on table, where each lexical unit is assigned an index address and calculated the probability of copyright in the lyrics.

Table 1. Example of the structure of the array of the author vocabulary.

Rank lexical unit r_{x_i}	**The value of a lexical unit** x_i	**The probability of occurrence in the texts** k_j	**The number of texts in which usage of lexical items**
1	Скоро	0,125	33
2	Ль	0,05	96
3	У	0,072	143

Table 2. Example of the structure of the array of the author syntactics.

Rank structure $r_{x_i, x_{i+1}}$	**The value of construction** x_i, x_{i+1}	**Frequency of use in the texts in this sense** $f_{x_i, x_{i+1}}$	**The probability of occurrence in the texts** k_j	**The number of texts in which consuming design**
1	Скоро ль	3	0,053	33
2	Скоро ль будет	1	0,017	17
3	Скоро ль луга	1	0,017	11

Then processed table twos, i.e. the header and the word following it. In the same way to form arrays of threes and fours.

The operation of the PARSER formed with two stages. In the first stage to the input of the program unit is supplied text query that graphemic processed and formed arrays syntactic lexicon and, for each element is calculated probability of occurrence in the text.

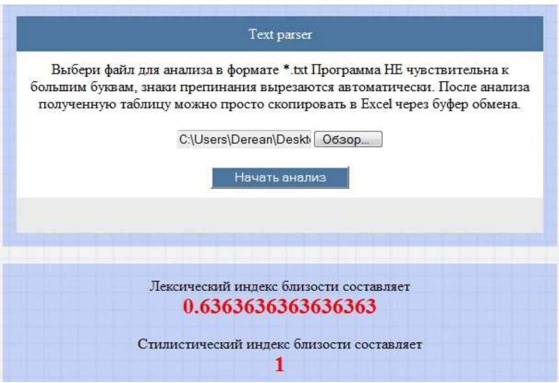
The second step is the conversion of arrays lexical database index_1, index_2, index_3 and index_4 that represent the author's language to the workspace. Computes the proximity style text query to the author's style.

In addition to the formation of arrays of vocabulary and syntactics query text block PARSER computes the index of the lexical and stylistic proximity to the query text.

Calculating these indices as follows. When the algorithm for finding the vicinity formed a list of lexical items matching the query text and the author's language. The index is the result of lexical proximity dividing the number of matching lexical units by the total number of lexical units of text query. In Fig. 2 the reduced screen form with the calculated values of the indices of lexical and stylistic affinity.

Thus, the software implementation of the developed statistical model of the text revealed the fundamental possibility of solving the problem of attribution and confirmed the adequacy of the proposed model.

Figure 2. Calculation logic and style indexes.



Text parser

Выбери файл для анализа в формате *.txt Программа НЕ чувствительна к большим буквам, знаки препинания вырезаются автоматически. После анализа полученную таблицу можно просто скопировать в Excel через буфер обмена.

C:\Users\Derean\Deskt

Лексический индекс близости составляет
0.6363636363636363

Стилистический индекс близости составляет
1

5. Conclusion

In this paper developed a linguistic model of the text, which takes into account the statistical characteristics of the author's text. To fill a lexical database of the author vocabulary developed a program generator which allows you to automate the process of establishing relationships between database tables. The algorithm for determining the proximity of texts. It calculates the indices of lexical and stylistic proximity of texts for adoption expert conclusion. Implemented the software implementation of the proposed method.

References and Notes

1. Ozhegov, S. I. The Russian language dictionary; The Russian language: Moscow, Russian, 1984; 797 p.
2. The explanatory dictionary of modern Russian language. http://slovari.yandex.ru/ (accessed on 01/02/2014).
3. The Law of Ukraine "On Copyright and Related Rights"; Kiev, Ukraine, 1994; Volume 13, Article 64.

4. Shirokov, V. A.; Bulgakov, A. V.; Hryaznuhina, T. A . Corpus Linguistics; Dovira, Kiev, Ukraine, 2005; 471 p.
5. Schröder, J. A. Equality, similarity, order; Nauka, Moscow, Russia, 1971; 254 p.
6. Fedorchenko, L. A. Stylized text fragments educational and methodological materials. Bulletin of the International Slavic University. Kharkov. Series "Engineering"; Kharkov, 2007; Volume 1, pp. 44–52.
7. Fedorchenko, L. A., Hayrova, N. F., Dounar, A. I., Bulgakov, S. O. The method of automatic construction of a semantic network of terms of discipline. Radioelektronni i komp'yuterni systems; Kharkov, 2011; Volume № 4, pp. 115–120.
8. Wilton, P. Basics JavaScript; Symbol-Plus, St. Petersburg, Russia, 2002; 430 p.
9. Mandelbrot, B. Fractal Geometry of Nature; Institute of Computer Science, Moscow, Russia, 2002; 656 pp.
10. Fedorovskyy, A. N., Kostin, M.Y. "Proceedings ROMYP'2005" Proceedings treteho Rossiyskogo the seminar on evaluation of methods is information search; I. S. Nekrestyanov; St. Petersburg State University Chemistry Research Institute: St. Petersburg, Russian, 2005; pp. 106-124.
11. The dictionary of Pushkin's language: 4 volumes. , 2nd ed.; Acad. Academy of Sciences of the USSR V.V . Vinogradov; Azbukovnyk: Moscow, Russia, 2000.