УДК 004.056.5

Informatics and Mathematical Methods in Simulation
Vol. 8 (2018), No. 2, pp. 121-128

APPROACH TO IDENTIFYING PLAGIARISM IN MULTILINGUAL TEXTS

V.G. Penko, I.H. Gafar Abdula

Odesa I.I. Mechnikov National University,
2, Dvoryanska Str., Odesa, 65082, Ukraine; e-mail: vpenko@onu.edu.ua

The task of identifying plagiarism between texts in different languages is an important variation of the general problem of identifying plagiarism. To solve this problem it is productive to calculate the degree of certain similarity of two texts which is called parallelism. In the article, the method of parallelism estimation based on Zipfian frequency distribution is studied. The key idea of the method is the construction of a linear regression model that compares the areas under the linearized Zipf curve for the corresponding documents. A computational procedure has been implemented to find the optimal classification parameters for such a model. To obtain a model more relevant to specific application conditions, computational experiments were performed to determine the optimal parameters corresponding to two classification metrics: the proportion of correct answers (accuracy) and F1-measure. The determination of the best classification parameters performed on the basis of the training subset of the corporal. To reliably estimate the model, classification metrics are recalculated on a test subset. The performed computational experiments using this approach showed limited applicability to language pairs composed of English, Russian and Ukrainian texts. To improve the filtering performance of parallel texts, a filter based on word frequencies in texts is proposed and implemented. To improve the quality of classification two directions have been formulated: an extension of the text corpora used in the model training, as well as methods for mutual using several classification filters.

Keywords: parallelism of multilingual texts, Zipfian frequency distribution, linear regression model, classification metrics, word frequency based filter, optimal classification parameters

Introduction

With the passage of time, the availability of information increases. This applies to both the volume and variety of information, and to tools that provide an increasingly convenient and easy way to access information of various activities and qualifications. In such circumstances, it becomes possible to create information-relevant documents, composed largely of fragments borrowed from other sources. This method considered unfair. It is customary to call this approach plagiarism. Under the conditions described, the task of identifying plagiarism is becoming more relevant.

The term "plagiarism" has an informal, but fairly stable definition. According to article 50 of the Law of Ukraine "On Copyright and Related Rights", plagiarism is the publication, in whole or in part, of someone else's work under the name of a person who is not the author of this work. Despite the brevity and simplicity of this formulation, in some cases this definition cannot be interpreted unequivocally.

In this situation, more and more sophisticated methods are required. It is important to understand the reason for the complexity of the task of detecting plagiarism. In our opinion, it is determined by the complexity of the objects that form the subject area (texts, language, knowledge). In fact, a full-fledged solution to this problem is an example of an extremely complicated task from the category of natural language processing (NLP). Other such tasks are machine translation, systems with a natural language interface, etc.

There is a large number of available automated systems for detecting plagiarism. Some of them are de facto standard for use within academic institutions. To gain access to the full functionality of such systems, commercial licenses are often required. An analysis of the standard functionality of antiplagiarism systems reveals their typical drawback. If an unscrupulous author gets access to the original material published, for example, in English, and then straightforwardly and qualitatively translates it into his native language, then from a legal point of view it is plagiarism. This phenomenon can be called a multi-lingual plagiarism. At the same time, available anti-plagiarism systems cannot detect it.

The proposed work is devoted to the development of automated means of detecting multilingual plagiarism. Such a task, obviously, will require the involvement of a diverse arsenal of methods offered for a broader range of NLP tasks. In this area, there are many methods. From our point of view, they can be divided into two categories:

- methods that analyze the internal structure of the document;
- methods that use generalized statistical information about texts.

It may seem that the first category of methods has a higher potential for qualitative detection of plagiarism. However, for the problem of multilingual plagiarism, their application is difficult, since it will require the development of non-trivial means of structural analysis of texts and at least partial implementation of machine translation.

The approach proposed in this article is the application of methods of the second category to solve the problem of multilingual plagiarism.

A common feature of all methods of this category is the ability to extract brief information about the text, which essentially characterizes its contents. The simplest example of such feature is the title of the text. Another way of compact representation of the characteristic features of the text arises from the application of the well-known Zipf's law.

Zipf's law is a statistical law, working in a wide range of subject areas. In the context of word processing, it is formulated as follows: the frequency of words in the text is inversely proportional to their rank. Rank - the ordinal number of the word, in the list of all words in descending order of their frequencies.

From now on, we shall call two texts parallel in the case when one of them is obtained as a result of the translation of the other. Obviously, this term is closely related to the identification of multi-lingual plagiarism.

The aim of this paper is to investigate the applicability of the analysis of the Zipf curve to identify multi-lingual plagiarism for "English-Russian", "English-Ukrainian" and "Russian-Ukrainian" language pairs.

The quantitative evaluation of the parallelism is determined in the form of a series of separate computational procedures.

Related works

Zipf's law is quite intensively used in solving NLP tasks. In [1] an attempt is made to justify Zipf's law, reliance on the features of human memory. This rationale is useful for understanding the hidden features that are universally manifested in virtually all large enough texts. These features allow us to apply the Zipf law in NLP problems. An earlier publication [2] investigated different characteristics of texts and text corpora that ensure the abidance of Zipf's law. This allows for a rough classification, relating the texts to one of two categories ("normal" and "abnormal"). An example of the practical application of the Zipf law in [3] is the identification of the functional significance of abbreviations and acronyms in English and German.

In [4] Zipf's law is used to identify the parameters of the Zipf curve characteristic for each language. As a result of computational experiments, the authors developed a filter that provided high performance when identifying parallel texts.

Preparatory stages of filter construction

Before applying to the text specific techniques that determine the degree of parallelism, each text must go through the stages of preliminary processing, presented in Fig. 1

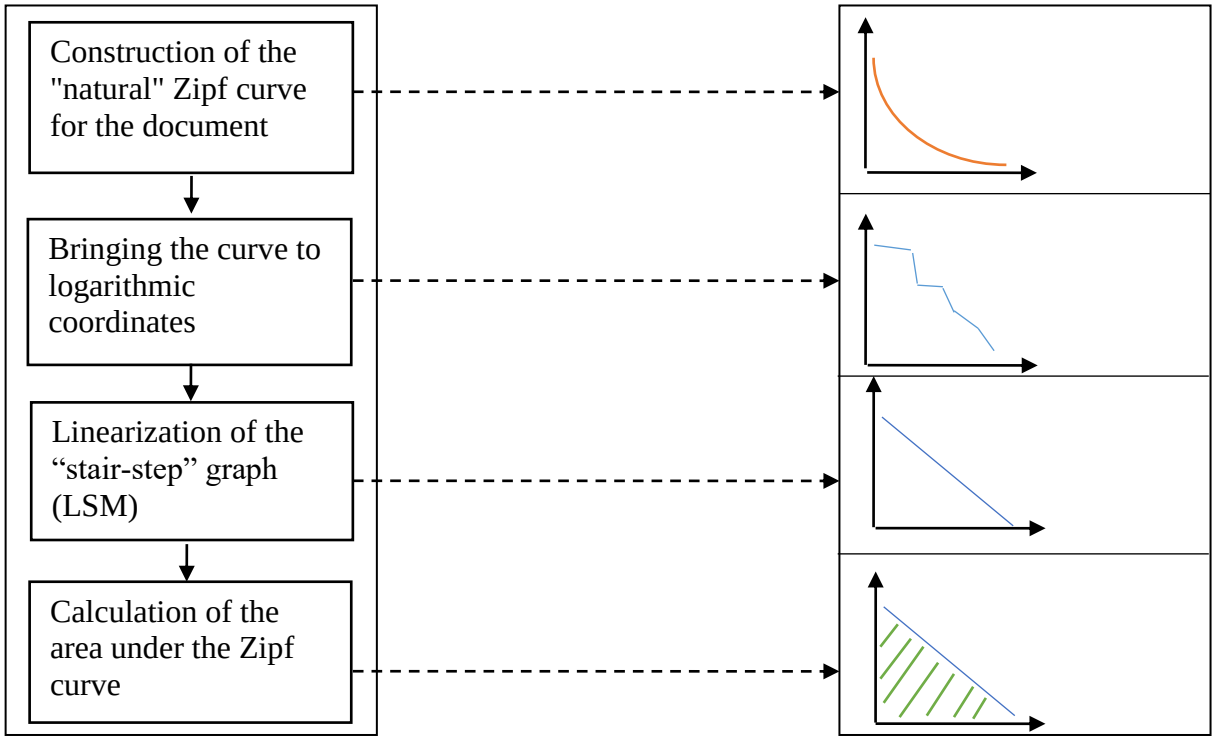

Fig. 1. The preliminary stages of Zipf-filter construction

The first step in the filter operation is the construction of the natural Zipf curve for one text. Further, the graph is reduced to logarithmic coordinates, which gives the natural hyperbolic Zipf curve the appearance of the "stepped" line. The next step is the linearization of the step graph using the least squares method (LSM). At the last stage, the area under the linearized Zipf curve is calculated.

The use of this area is the central idea in the construction of the Zipf filter. The relationship between the slope of the line and the area below it determines typical situations that allow one to conclude that the two texts represented by these lines are parallel. Consider the situations presented in Fig. 2.

a b c

Fig. 2. Various options for the relations between the frequency characteristics of texts in accordance with the Zipf law: a – equality of areas, different slope; b – different areas, different slope; c – different areas, equality of slope

From the point of view in [4], option a) (equality of areas, different slope) is treated as a sign of parallelism, and option c) (different areas, equality of slope) as its absence. The most interesting is option b) (different areas, different slope). At first glance, this situation can be interpreted as the lack of parallelism. However, our conclusion will change to the opposite if statistical analysis of the pairs of parallel texts for a certain language pair will reveal a stable pattern in the ratio between areas. This criterion is, to some extent, hypothetical and requires experimental confirmation. In this article, the authors presented the results of an experiment to confirm this hypothesis, and concluded that this criterion was competitive in comparison with several known ones. For a corpora of some dozen texts of several language pairs ("English-Spanish", "English-Dutch" and "English-Swedish"), a series of classifying models with different tuning parameters was built. The best values of the average harmonic characteristics of F1-measure with respective values of δ (parameter of classification model providing best result) are shown in Table 1. There is another useful classification metrics which is called accuracy. It was not represented in this research paper.

Table 1.

Classification metrics' results for three language pairs

English – Spanish		English - Dutch		English - Swedish	
δ	F1	δ	F1	δ	F1
6	0.57	3	0.58	4	0.74

We consider that obtained results show effectiveness of the described classification approach. It is the subject of interest to test this approach under specific conditions. In particular, it is important to apply this approach to the language pairs that are more relevant in our circumstances (English-Russian, English-Ukrainian, Russian-Ukrainian).

Training Zipfian filter

To build a Zipf filter, you need to pre-process rather large amount of parallel texts pairs, calculating the areas under the Zipf curve for them. Thus, each pair of parallel texts is represented by two values: x_i is the area under the Zipf curve of the i-th text in one language and y_i is the area under the Zipf curve of the i-th text in another language. To obtain a characteristic relation between these values within the bounds of two languages, based on these data, the linear regression model is constructed by using LSM: $y = a_0 + a_0 x$.

Having this model, we can formulate a criterion for determining the parallelism of two texts D_s and D_t.

$$Par_{zipf}(D_s, D_t) = \begin{cases} 1, if \ |\in| \le \delta, \\ 0, otherwise \end{cases} \quad (1)$$

where $\in$ is the model error for this text pair $\in_i = y_i - (a_0 + a_1 x_i)$.

The best result will be achieved when $\in = 0$, which means that the area obtained as a result of the model's operation coincides with the actual value. However, it is necessary to allow some degree of deviation of the actual area from the area obtained by using the model (threshold δ). The final stage of training the parallel text filter is to find the value of threshold δ which provides best classification results. The magnitude of this threshold can be found through the solution of the problem of optimizing the functioning of the classifier in accordance with one of the generally accepted metrics (accuracy, precision, recall, F1). This

problem can be solved empirically due to the linear nature of the classification model used. In this paper we used the following procedure.

Initially, the range within which the best threshold value is searched is empirically determined. Further, for each value from this range (with some step), the target classification metrics (accuracy or F1) are determined. At each step of this process, a subset of text pairs (the learning set) is used. The threshold value at which the best classification results are obtained is considered optimal. Another subset of text pairs (test set) is used to assess the effectiveness of the model with the calculated threshold. If the results of applying the obtained model to the test set are comparable with respect to the training set, the model can be considered sufficiently reliable.

Description of the computational experiments

To implement the approach described above, a software system was developed. Its object-oriented structure is shown in Fig. 3 as a class diagram.

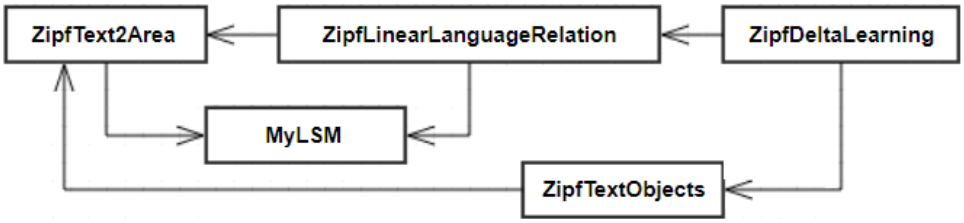

Fig. 3. Zipf filter class diagram

Description of classes in the diagram:

- ZipfText2Area - calculation of the area under the linearized logarithmic Zipf curve.
- ZipfLinearLanguageRelation - defines the relation between the areas of the texts for two languages using the linear regression model and calculates the deviation for two specific texts in terms of this model
- ZipfDeltaLearning - finds the best threshold value in accordance with the formula (1).
- MyLSM - least squares method for linearizing the Zipf curve and determining the parameters of linear regression of two languages.
- ZipfTextObjects - objects that store information about areas under the Zipf curve for a given set of texts.

The experiments in this work used relatively small text corpus consisting of parallel texts in three languages – 20 pairs (English, Russian), 12 pairs (English, Ukrainian) and 12 pairs (Russian, Ukrainian). Preparing this corpus in terms of the proposed research methodology, a significant shortage of lexical resources in the Ukrainian language was revealed. In order to avoid influencing the results with insignificant lexical information, each text has been cleaned from stop words.

The results of the experiments are summarized in Table 2.

Table 2.

Results of experiments with Zipfian filter

	Accuracy (training)	Best δ	Accuracy (testing)	F1 (training)	Best δ	F1 (testing)
EN-RU	0.9	0.03	0.89	0.37	0.43	0.35
EN-UK	0.95	0.32	0.81	0.84	0.32	0.66
RU-UK	0.91	0.16	0.81	0.76	0.21	0.66

The results in Table. 2 in some cases exceed the results from [4] with respect to the F1-measure. In addition, rather high values are obtained. Some instability in F1-measure can be associated with the limited size of the training corpora.

Implementing a filter based on the frequency profile of the text

The above results of applying the proposed Zipf filter do not allow to consider such a filter reliable for classifying parallel texts between language pairs English-Russian. The next step in the progress of this work was the development and implementation of a fairly simple parallel text filter, which in the future will be called a word frequency based filter. In this case, a characteristic feature of the text is the frequency distribution of words in the text. The hypothesis underlying this filter is that the frequency distribution of the most popular words in two parallel texts will show a high coincidence.

Formally, this can be represented as follows.

Let $w_1^1,\ldots\ w_n^1$ be the words of text 1 ordered from the frequency of occurrence in text1, and $w_1^2,\ldots\ w_m^2$ are the words of text 2 ordered from the frequency of occurrence in text2.

The sets $FW_1 = (w_1^1,\ldots, w_k^1)$ and $FW_2 = (w_1^2,\ldots, w_k^2)$ are the sets of k most popular words in texts 1 and 2

$\left|FW_1 \cap FW_2\right|$ – the number of matching words in FW_1 and FW_2 (taking into account the translation)

$$Par_{WordFreq}(D_s, D_t) = \begin{cases} 1, if\ \left|FW_1 \cap FW_2\right| \le \delta, \\ 0, otherwise \end{cases}$$

This filter was implemented using several procedures, among which we note the following:

− Automation of the translation from one language to another (performed using Google online services)

− The method of empirical selection of optimal parameters of the model for achieving the best classification results (similar to the one used to construct the Zipf filter)

The results of the experiments are summarized in Table 3.

Table 3.

Results of experiments with word frequency based filter

	Accuracy (training)	Best δ	Accuracy (testing)	F1 (training)	Best δ	F1 (testing)
EN-RU	0.99	0.2	0.94	0.9	0.2	0.8
EN-UK	0.97	0.21	0.94	0.86	0.21	0.86
RU-UK	0.89	0.03	0.75	0.69	0.02	0.66

Here we can see rather high values both in accuracy and F1-measure.

Conclusions

Two methods for determining parallel texts were developed and tested. The first method is based on the analysis of the frequency distribution of words in accordance with the Zipf law. The second is the overlapping of high-frequency words in documents. The results of experiments do not allow us to state that any of these methods cannot be used as a high

reliable classifier. However, given the different nature of these two methods, several schemes for their cooperative usage can be proposed, taking into account the specific requirements of the problem.

Thus, two directions are promising in the future:

− development of effective schemes for the joint use of several classification models;

− work on the expansion of linguistic resources corpora, applicable within the framework of the proposed approach.

References

1. Steven, T. Piantadosi Zipf's word frequency law in natural language: A critical review and future directions Psychon Bull Rev. 2014. — 21 (5). — Pp. 1112–1130. doi:10.3758/s13423-014-0585-6.
2. David, M.W. Powers. Applications and Explanations of Zipfs Law. In D.M.W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL, 1998. — Pp. 151-160.
3. Орехова, О.М. Реализация закона Ципфа на материале английского и немецкого языков Филологические науки. Вопросы теории и практики Тамбов: Грамота, 2014. — № 4 (34), Ч. I. — С. 161-163. ISSN 1997-2911.
4. Mehdi, M.P. Document Identification using Zipf's Law Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, LREC 2016, Portorož, Slovenia, May 23, 2016. — Pp. 21-25

ПІДХІД ДО ВИЯВЛЕННЯ ПЛАГІАТУ РІЗНОМОВНИХ ТЕКСТІВ

В.Г.Пенко, І.Х. Гафар Абдула

Одеський національний університет ім. І.І. Мечникова,
вул. Дворянська, 2, Одеса, 65082, Україна; e-mail: vpenko@onu.edu.ua

Завдання виявлення плагіату між текстами на різних мовах є важливим різновидом загального завдання виявлення плагіату. Для вирішення цього завдання плідним є можливість обчислювати ступінь подібності (паралельності) двох текстів. У статті вивчається метод оцінки паралельності на основі розподілу частот Зіпфа. Ключовою ідеєю методу є побудова лінійної регресійної моделі, що співставляє площі під лінеарізованими кривими Зіпфа для текстів, що співставляються. Реалізована обчислювальна процедура для знаходження оптимальних параметрів класифікації такої моделі. Для отримання моделі, яка більшою мірою відповідає конкретним умовам застосування проведені дві серії обчислювальних експериментів для визначення оптимальних параметрів, що відповідають двом класифікаційним метрикам: Accuracy і F1-міра. Визначення найкращих класифікаційних параметрів відбувається на основі навчального підмножини корпусу. Для надійної оцінки моделі класифікаційні метрики перераховуються на тестовому підмножині. Виконані обчислювальні експерименти з використанням цього підходу показали обмежену придатність що до мовних пар, складених з англійських, російських та українських текстів. Для поліпшення показників фільтрації паралельних текстів запропонований і реалізований фільтр, що базується на частотах слів в текстах. Сформульовано напрямки, що дозволяють поліпшити показники якості класифікації: розширення корпусу текстів, що використовуєтьмя при навчанні моделі, а також методи спільного використання декількох класифікаційних фільтрів.

Ключові слова: паралелізм різномовних текстів, розподіл частот по Зіпфа, лінійна регресійна модель, метрики класифікації, фільтр по частотах слів, оптимальні класифікаційні параметри

ПОДХОД К ВЫЯВЛЕНИЮ ПЛАГИАТА РАЗНОЯЗЫКОВЫХ ТЕКСТОВ

В.Г.Пенко, И.Х. Гафар Абдула

Одесский национальный университет им. И.И. Мечникова,
ул. Дворянская, 2, Одесса, 65082, Украина; e-mail: vpenko@onu.edu.ua

Задача выявления плагиата между текстами на разных языках является важной разновидностью общей задачи выявления плагиата. Для решения этой задачи плодотворным является возможность вычислять степень параллельности двух текстов. В статье изучен метод оценки параллельности на основе Zipfian frequency distribution. Ключевой идеей метода является построение линейной регрессионной модели, сопоставляющей площади под линеаризованной кривой Зипфа для соответсвущих документов. Реализована вычислительная процедура для нахождения оптимальных параметров классификации такой модели. Для получения модели, в большей степени соответствующей конкретным условиям применения проведены две серии вычислительных экспериментов для получения оптимальных параметров, соответствующих двум классификационным метрикам: доля правильных ответов и F1-мера. Определение наилучших классификационных параметров происходит на основе обучающего подмножества корпуса. Для надежной оценки модели классификационные метрики пересчитываются на тестовом подмножестве. Выполненные вычислительные эксперименты с использованием этого подхода показали ограниченную применимость к языковым парам, составленным из английских, русских и украинских текстов. Для улучшения показателей фильтрации параллельных текстов предложен и реализован фильтр, основанный на частотах слов в текстах. Сформулированы направления, позволяющие улучшить показатели качества классификации: расширение корпуса текстов, используемого при обучении модели, а также методы совместного использования нескольких классификационных фильтров.

Ключевые слова: параллелизм разноязыковых текстов, распределение частот по Зипфу, линейная регрессионная модель, метрики классификации, фильтр по частотам слов, оптимальные классификационные параметры