Plagiarism Detection System for Armenian Language

Gohar Tomeyan
National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: goharikyan93@gmail.com

Maria João Varanda Pereira
Polytechnic Institute of Bragança
Bragança, Portugal
e-mail: mjoao@ipb.pt

Gevorg Margarov
National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: gmargarov@gmail.com

ABSTRACT
In the academic context, it is very important to evaluate the uniqueness of reports, scientific papers and other documents that are everyday disseminated on the web. There are already several tools for this purpose but not for Armenian texts. In this paper, a system to analyze the similarity of Armenian documents is presented. The idea is to collect a set of documents of the same domain in order to identify keywords. Then, based on that information, the system receives two documents and compares them calculating the probability of plagiarism. For that, an approach based on several levels of analysis is implemented and some of those steps allow the user interaction choosing options or adding more information.

Keywords
Natural language processing, plagiarism detection, sinonymizer, document uniqueness, document similarity

1. INTRODUCTION
Internet is getting more widespread in our life and in our activities. However, visiting different Web sites, we see that all the found articles or other materials are very similar. Besides, there are many theses, term papers, research and other scientific works on the Internet. If formerly it was necessary for the students to take advantage of the published books and literature, now it is enough to write the name of the subject in the search engines and we can find thousands of items. The most common objects of plagiarism are texts, separate expressions, thoughts, inventions, facts described in novels. Scientific spheres include a large amount of ready works, course works and articles, in which we can make several changes and achieve results. That kind of change is considered plagiarism. In order to avoid these situations a plagiarism detection system is needed.

At first it will be defined what plagiarism means. There are many definitions of plagiarism. The scientific and educational sphere plagiarism is the form of deception, which means to appropriate other ideas, passages from another work or author. This is a forgery, generally in violation of copyright laws. Plagiarism is a steal and pass off the ideas of another as one's own, using another's manufacture without lending the source, present as a new and original idea taken from an existing source [1]. In the legal point, plagiarism is a direct privatization of the text. Legally plagiarism is text digestion, while the digestion of subjects and ideas can't be considered as plagiarism. The only thing, which is not allowed, is the whole copy of the text. But often the whole text is translated and presented as an original. Thus, the plagiarism, which is done by translation is widespread.

Usually in order to conceal plagiarism people carry out several steps, for example text morphological change, lexical change, reduction of the text up to some words, sentences, pictures or formulas, text syntactical changes, movement of sentences, punctuation marks change, spaces are replaced with transparent letters, and also create and use synonyms.

With the ever-increasing availability and accessibility of the Internet, students are able to access a multitude of resources in support of their studies. However, this has also led to an increase in their ability to cheat through plagiarizing text and claiming it as their own. So, one of the most important part of this work is to define plagiarism levels and what should be checked at each level. Then, create a tool implementing the plagiarism levels detection for a document written in Armenian.

There are already several tools for this purpose but not for Armenian text. Existing plagiarism detecting multilingual systems are not intended for the Armenian language.

2. RELATED WORK
There are many automatic systems to detect plagiarism; such systems are Antiplagiat.ru, eTXT, PlagScan, CheckforPlagiarism.net [2], Turnitin, etc. Here we describe a comparative analysis of some textual softwares.

The most famous online system is Antiplagiat [5]. The system searches from its own database. Therefore, the system has several disadvantages. At first, it isn't able to search on the Internet and there is a limit up to 3000-5000 words. The system Antiplagiat doesn't detect text morfological changes. If spaces are replaced with transparent letters, they will be visible to computer. The system Antiplagiat is able to detect, reduce, and replace words, sentences and paragraphs. The replacement of English letters with Russian is also detected. The change of punctuation marks has no influence on work of the system.

eTXT-Antiplagiat[6] gives the opportunity to search similar documents on the Internet. Matching parts of the text are indicated with the respective colors. It can easily detect non-unique texts. To avoid to be detected by the system we need to make changes in the text using synonyms, for instance.

PlagScan is a plagiarism detection software (available online and on-premises), used by academic institutions and businesses. PlagScan [2] offers teachers and professors to identify plagiarism and educate students on the appropriate usage of sources in academic works as well as protecting copyrights of texts. The main disadvantages of PlagScan are: it doesn't support synonym recognition, sentence structure checking and plagiarism detection over translated texts isn't supported.

Basically, all systems use the algorithm of shingle, which provides the highest correctness in detecting the copies.

In this Figure 2.1, the main characteristics of the tools are presented, allowing us to compare the facilities provided by each one. For example, there are tools that don't detect the use of synonym and only one tool can find plagiarism with translation only from English texts, that is the Turnitin [7].

Name	Compare in database	Compare on the Internet	Languages	Translation	Synonymize
Antiplagiat.ru	+	-	Russian	-	-
ETXT-Antiplagiat	+	+	Many	-	-
PlagScan	+	+	Germany French English Spain	-	-
Turnitin.com	+	+	Many	+	-

Figure 2.1. Comparative analysis

To avoid such situations, it was decided to develop a system that will automate the uniqueness analysis of the work done by students in the learning process and will allow teachers to detect quickly the existence of plagiarism. There are not this kind of systems for the Armenian language, this one will be used in a lot of universities and will be useful for lecturers.

The main features of this work are:

- Checking in the database: The program will check the students papers in the system database, where each year the research works done by students can be uploaded.
- Checking on the Internet: The system doesn't give an opportunity for searching on the web sources, but teachers can upload web documents to prevent the plagiarism based on the Internet.
- Checking the use of synonyms and sentence structure changes: The system will allow the following steps: normalization alphabet, keyword detection, stop word removal, stemmer, which will be used to search correct forms of words.
- Multiple Document Comparison: Our system will compare one document with more documents and will show the percentage of plagiarism possibility considering the keywords of the domain.
- Supported Languages: Armenian.
- Plagiarism with translation: The program will detect Russian and English text translations, and will compare with Armenian sources. The translation is based on the Google translator.

3. PLAGIARISM CHECKER SYSTEM: OUR PROPOSAL

To achieve the assigned goal, it is necessary to solve the following tasks:

- review the existing algorithms for detecting plagiarism in the texts,
- review the existing methods to conceal the fact of plagiarism, as well as methods of dealing with them,
- develop a method of searching plagiarism in Armenian texts that is resistant to possible text modifications,
- create a software tool based on the developed method, which provides plagiarism detection with the possibility of visualizing the borrowed pieces of text in the scanned document and in the source document. At the end, a percentage will be calculated in order to identify the document similarity level.

Below we will describe the main steps of our tools.

Natural Language Processing (NLP) techniques [8], are used to detect the possibility of plagiarism in Armenian texts. The main idea is to analyze similarity between two documents using those techniques of natural language processing [3].

The first step will be to compare the texts word by word but this work must go further. Everyone knows that the people that use the texts of other people change it a little bit to dissimulate the plagiarism.

Natural Language Processing includes semantic and syntactic changes, stop word removal, stemming, lemmatization, punctuation removal, etc., as part of the pre-processing stage. If the text has semantic and syntactic changes, the plagiarism detection systems do not work well. In order to detect such changes, linguistic techniques must be considered. It's important to detect intelligent plagiarism, when ideas are presented in different words, replacement with synonyms, translation, etc. Translation plagiarism is also very common, because students can also translate the text from one language into another without pointing the original source out. For example, we haven't many materials about Information Systems in Armenian language and students carry out translation from English or Russian texts including automatic translation (for example Google and another translators) and manual translation (which can be done by students who knows some languages).

So, one of the most important part of this work is to define plagiarism levels and what should be checked at each level. Then, create a tool implementing those plagiarism levels detection.

Possible modifications of the text plagiarism depends on the language used, and during the analysis of the text, we should take into account the specifics of the given language. Each language has different rules for sentence structure and different opportunities for synonym replacement.

Detecting plagiarism should be made by possible modifications when detecting, and the system must be able to allocate specific pieces of borrowed text, as well as the corresponding fragments of the source text. In order to process an algorithm, it is important to determine two aspects.

- standards of determining the similarity of texts (form and content),
- determining the level of similarity and its threshold value (when the text isn't a copy)

Technical uniqueness of text is a threshold value, which is usually measured by percentage. The text that has a 100 percent technical uniqueness, is not unique yet (de facto it can be unique also from about 0). For example, write off the thought of another person, and that is not unique, measured by other words. However, there are some exceptions too, factual unique texts can be technically unique for 50 percent. For example, the author's work is unique, when includes included exceptional materials that are written from scratch. A work is not unique when it includes citations, expressions, technical terminus, etc.

The main steps of this work are the search for exactly the same sentences, normalization of alphabet, keywords detection, stop-word removal, stemming, synonym recognition and search for plagiarism with translation. The steps of the plagiarism detection tools can be seen in Figure 3.1.

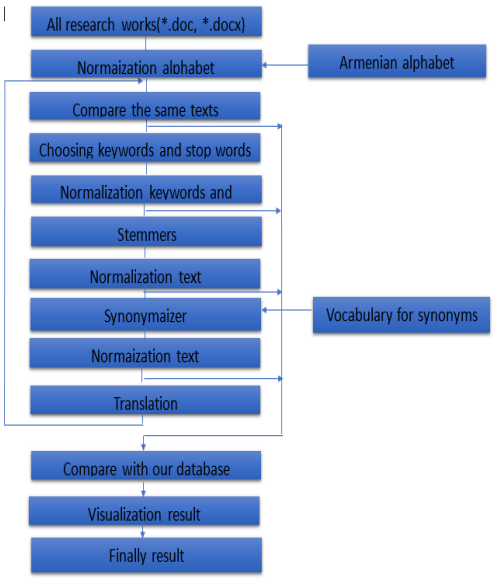

Figure 3.1. Steps of plagiarism detection

First step is to normalize alphabet and the second step is to find exactly the same texts. Then, removing the stop words and using a set of predefined keywords of the text domain the document should be compared again. Stop-words is a very frequent word but without any particular meaning. The usual way of determining what counts as a stop-word is just to use a dictionary that lists them [4]. And in our program we used the stemming system to delete endings in Armenian language. The motivation for using synonymy recognition comes from considering human behavior, whereby people may seek to hide plagiarism by replacing words with appropriate synonyms. The system also contains synonyms and steammers for the Armenian language.

The most significant principles are lexical analyses, as well as linguistic methods. To detect lexical changes we used steammers, which are based on Porters algorithm considering the features of Armenian language. The algorithm gives an opportunity to delete verb endings, noun ending, etc. In our program the idea of keyword is used, which gives an opportunity to organize searching in our database very quickly. Keywords have special meaning and they are chosen and formed according to each subject. To find synonyms we use Armenian vocabulary. Now the Shingles method is also used which will give an opportunity in the future to carry out search on the Internet. At the end, in our program we used the Google translator for finding plagiarism with translation.

3.1 The created system in detail

We will represent the main steps to find exactly the same sentences, choosing keywords, stop word removal, synonyms recognition and translation for finding the possibility of plagiarism. The program only compares *.doc, *.docx documents in our database. Database will be expanded by teachers uploading and checking the student documents, as well as translated documents. The system allows to carry out searching based on the works of previous years. The steps of the plagiarism detection tools can be seen in Figure 3.1.

Often students can replace the letters with another letters, for example, some systems are not able to detect if there is a Russian letter “a” instead of English letter “a”. We have some letters which are similar to another letters, for example Armenian ”h” seems like English ”h”. This program is able to find other letters and point out in another color.

In our program the alphabet is checked at first whether it is written in Armenian or not.

Checks are carried out through ASCII codes. If it is not Armenian, letters will be pointed out in red. The program includes Armenian letters, and letters are compared by the ASCII code. When the program points out letters, which are written in another language in red, the teacher will see the result, will be able to replace the letters with the Armenian ones, but manually. After these steps, the teacher can compare them. If the teacher does not replace them, sometimes the system will not be able to recognize and will consider them as other word. This is the main meaning of normalization alphabet.

First important part of plagiarism detection is to find exactly the same text. The program can find exactly the same text and show the percentage, if there are matching parts. The comparison is realized word by word. At first for comparing, we need to delete all characters except the “:”, which shows that the sentences are completed. This algorithm is used to split the source text into sentences. Separation is carried out by punctuation marks such as a point, exclamation mark, question mark, then the text is compared sentence by sentence and if there is a match it will indicate plagiarism existence otherwise continues to perform the next action. The program can compare two or more files. If we want to compare many documents, we will need only to choose the subject, after that keyword is extracted and we can see how the plagiarism is possible. We can compare two or more documents.

After identifying the most important elements of papers, we already have keywords for each subject, which are kept in Microsoft Word and saved by special name, for example name of the subject. The program allows teachers to upload a new file, which may contain its own keywords and synonyms of keywords or teachers can edit the already existing keyword files. The program is working like this, if we want to generate a keyword for any subject, first we need to put the password, and after choosing the name of the subject, or upload file. Depending on the fact that who will enter the password, the opportunity of the user will be different. As administrator, the user can add a new keyword with the help of the corresponding window, but a teacher doesn’t have that kind of solution. After that we need to choose the file, which we want to compare. The program will generate a new folder, and put there only those files that we have in our database and which have the same keywords. When we compare we can see the result presented by percentage. We do not need to compare all files; we only have to compare the text, which have the same keywords. Each subject has separately keywords that are kept in separate documents.

The program now includes 100 stop words for Armenian language and we can delete stop words, see frequency and compare. Stop words are saved in our database, and in the future, it will be extended.

One of such important and necessary things on computer linguistics is the operation of using steamers. Stemming is usually used in Information Retrieval systemms. Best way for determining steammer is just using the dictionary. The project Snowball contains the old version of Armenian suffix and prefix, but the Armenian language has endings, too, when we delete a suffix or prefix in Armenian, the words will change their meanings. But for the English language endings and suffixes have the same meanings. For example, if the words are finished in -ed, only in English we can delete the suffix “-ed” and the words will not change the meaning. In our program, teachers can see all endings. If the

teacher wants to compare two files and to know the possibility of plagiarism, after he/she can delete all endings, he/she needs only to choose the second “text endings” and compare. All endings will be deleted; therefore, the program already gives the percent without endings.

The most important thing for NLP is identification of synonyms. The main concept is to use synonyms but to keep the meaning of the text. After using stemming we can replace with synonyms. The program has an option which points out words in red color and replace with synonyms. Teacher has the opportunity to point out words in red color, choose the meaning, which corresponds to the context and save changes. After choosing the word in the right side appears a panel, where the user can see the meaning of synonyms and then choose the corresponding word, a comparison will be done and a result by percentage will be calculated.

Teachers can add and see the synonyms, which are existing in our database. The teacher can only delete synonyms, which he/she wrote. At first teacher needs to write a synonym, an explanation and choose the add button, after we can see all the explanation, if it is correct, the teacher can write in database.

As already mentioned, plagiarism can be performed translating the text from one language to another without reference to the original source. Translated plagiarism can include two types of translation: automatic and manual translation.

Plagiarism with translation is very difficult to detect. There are many problems: first is to translate words that have many meanings; the translator translates all words automatically, and the system has to find which one is the correct word. A word by word translation is not a good idea. Translation into Armenian is not working effectively, it’s enough only for understanding but not for detecting plagiarism, but Google gives a huge opportunity to make changes and optimize the texts returned by Google translator. We don’t have much information in Armenian language on the Internet, and students often translate the documents from Russian and English texts, and present as an own idea. Usually students use the already existing translators, especially Google translate. For that reason, we include Google translate in our program, because the translator allows the translation of the documents. And translation will work if the user has an Internet connection.

If we want to translate a document we need to choose the document, when program finished translation, teachers must copy and paste the text on the Microsoft Word, and after which upload that file to our database. And then the teacher can follow the same steps to detect plagiarism: choose keywords and compare with many documents or compare only two documents using stop word removal and synonym recognition.

4. SYSTEM IMPLEMENTATION

In order to implement the system a local database is used. Search for detecting should be carried out in the local database of documents. The implementation was done in the language C sharp, Windows Form Application for creating Desktop Application and Asp.net MVC for making Web Application. We used MSSQL (to work with the database) and Google translate for detecting plagiarism on the Internet with translation.

5. TESTING

We have already performed some tests with real users and some conclusions were drawn. We tested the system functionalities and also linguistic failures. The main disadvantage is retiled to hard interface of desktop application, which is very difficult to use without a user guide. Another disadvantage is the system has very few synonyms, which will be added in near future or we will use the synonymizer for Armenian texts.

More tests will be carried out in order to measure the effectiveness of the system.

6. CONCLUSION

This paper described the proposed plagiarism detection system for Armenian documents. The system compares two and more documents and allows the following steps: normalization alphabet, keyword detection, stop word removal, steamming, and it is able to detect the replacement by synonyms and find plagiarism with translation. Our plagiarism detection system compares the texts in directory, which is extended owing to teacher’s uploaded files. A Web application was also created, which will be extended and available not only for teachers but for all the users in the future.

7. ACKNOWLEDGEMENT

We should thank the Erasmus+ ICM project for supporting the research collaboration between IPB and NPUA.

REFERENCES

[1] Amalia, Performance evaluation of free anti-plagiarism software October 2-4, 2013

[2] Ali, A. M. E. T., Abdulla, H. M. D., & Snasel, V. (2011). Overview and Comparison of Plagiarism Detection Tools. In DATESO (pp. 161-172).

[3] Menai, M. E. B. (2012). Detection of plagiarism in Arabic documents. International journal of information technology and computer science (IJITCS), 4(10), 80.

[4] Ceska, Z., & Fox, C. (2011). The influence of text pre-processing on plagiarism detection. Association for Computational Linguistics.

[5] Мозгалева, П. И., Гуляева, К. В., & Замятина, О. М. (2013). Информационные технологии для оценки компетенций и организации проектной деятельности при подготовке технических специалистов. Информатизация образования и науки, (4), 30-46.

[6] ЧИРКИН, Е. С. (2013). Системы автоматизированной проверки на неправомерные заимствования. Вестник Тамбовского университета. Серия: Гуманитарные науки, (12 (128)).

[7] Weber-Wulff, D., Möller, C., Touras, J., & Zincke, E. (2013). Plagiarism detection software test 2013. Abgerufen am, 12, 2014.

[8] Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2), 133-149.