International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Stemming Implementation in Preprocessing Phase for Evaluating of Exams Using Data Mining Approach

Mehmet Balcı [1], Şakir Taşdemir *[2], Rıdvan Saraçoğlu [3]

Abstract: In educational activities, examinations are sometimes carried out in the form of multiple-choice tests or sometimes as open-ended long texts. When multiple-choice tests are performed, evaluating process is carried out either manual or computer-assisted. Exam questions prepared in the form of multiple choice tests are not suitable for every course. It may be necessary to use open-ended questionnaires in order for pupils to accurately measure their achievement in relation to the course. It can take a long time to evaluate examinations made with such questions. However, this process can create problems in terms of objective evaluation. Data mining, defined as the extraction of useful information from large quantities of data, can be used to process all kinds of data. The data mining method used in the processing of textual data is called text mining. In text processing studies, data is subject to preprocessing in order to obtain a high quality data set. The most important stage of preprocessing is stemming. In this study, stemming process is implemented to questions and correct answers taken from students. The results obtained in 2 different samples and 4 sentences are 71%, 69%, 86% and 78% correct. In order to be able to distinguish what the textual data written in the natural language really is, it is necessary to use the states of the words which are made up of construction and free from the suffixes. Therefore, in the pre-processing phase, stemming process is applied to the textual data in accordance with the grammar rules of the language they are written on, and stems of every word are found. Text processing is used in many areas of the natural language. Computer-aided solutions will be inevitable so that problems can be eliminated and open-ended questions can be quickly assessed. Despite the desirability of a computer aided solution for this measurement technique, studies of this solution are not included in the literature very much.

Keywords: *Preprocessing, Stemming, Data Mining, Exam Assessment*

1. Introduction

We witness the technological developments in the world at dizzying speed. Especially the developments of electronic technology, which started in the 20th century, have become almost too fast to catch up with in the 21st century. The fruits of this can be seen clearly all over the world particularly in the field of computer systems. Educational institutions in many countries spend big budgets for computer technologies so as to increase the quality in education and they equip their educational environments with computer systems. It is a fact that computer systems are used frequently not only in educational environments but also in the background of education for educational planning and assessment and evaluation.

Educational institutions use special software to prepare and evaluate the exams. There is various software about preparing and evaluating multiple choice tests especially. However, these software do not meet the needs every time. The reason is that every lesson or every subject in one lesson can't be evaluated by multiple choice questions. In such cases, the innovations brought with computer technologies are put away and old-style paper tests are preferred. The application of these kinds of exams costs much in terms of economy and time. In addition to this, classical exams applied on papers are quite difficult for the examinants both in application and evaluation phase. It is a fact that making use of computer systems is the best solution so as to overcome all these difficulties. Especially when students are not restricted for the questions that require textual answers, it becomes more complicated to evaluate these questions. In one of their studies, Akdağ and Çoklar (2009) described these kinds of questions as open-ended questions or open-ended surveys and they are among the techniques, which are used to collect data qualitatively [1].

Because of these negative reasons, the need for software studies is increasing each passing day so as to evaluate open-ended questions. In this study, a point of view was tried to be developed for the solution of this problem by using data mining methods that are important working areas of computer engineering.

2. Material and Methods

One of the studies in literature that put forward the importance of classical exams is the study named '' Physics, Chemistry and Mathematics Teachers' Approaches and Applications about Evaluation Instruments'' performed by Nazlıçiçek and Akarsu (2008) [2]. In this study, the teachers' opinions about evaluation instruments were taken from different perspectives. According to the findings obtained from this study, we meet these results when we compare tests to written exams (Table 1).

[1] *Computer Technologies Department, Higher School of Vocational and Technical Sciences, Selcuk University, Konya/Turkey*
[2] *Computer Engineering Department, Technology Faculty, Selcuk University, Konya/Turkey*
[3] *Electrical and Electronic Engineering Department, Faculty of Engineering and Architecture, Yüzüncü Yil University, Van/Turkey*
* *Corresponding Author: Email: stasdemir@selcuk.edu.tr*

Table 1. Teachers' opinions about the evaluation tools.

	Written Examination	Test Examination
Average of Knowledge Level (1-3)	3.00	2.95
Attach Importance (1-5)	4.40	3.78
Trusting (1-5)	4.16	3.45
Habit of Using (1-5)	4.09	3.57

It can be inferred from the Table 1 that as an assessment and evaluation instrument the written exams are more needed than the tests in terms of the knowledge level of the teachers, emphasis, reliance and usage habits.

According to the results obtained from another study, teachers use traditional methods that make them feel more competent to get to know the students and assess their achievements more. The four most important problems that are faced while using traditional methods are stated respectively as '' difficulty in preparing, crowded classes, lack of time and lack of parents' supports'' [3].

In a study [4] performed by Mintzes et al. (2001), they compared assessment and evaluation methods for new strategies so as to make the students understand the lesson better in biology science. It is also stated in their studies that another disadvantage of multiple choice tests is that they are inadequate to determine the students' opinions apart from specific confines because there are limited options.

When looking into the literature studies above, it is understood that software studies for the application of written exams are necessary because teachers will never give up applying written exams no matter how test techniques are developed.

2.1. Data Mining and Text Processing

To make a simple definition, data mining means reaching and mining the knowledge among large scale database or looking for the links that can make us predict about the future in big data stack by using computer programs [5]. Balci (2010) stated in one of his studies [6] that the database, which saves a vast quantity of textual information, is called document collection and the text mining, which is a data mining method, has emerged so as to get information from document collection. According to the same study, text mining is a field of data mining, so most of the methods used for data mining are valid for text mining as well. Balci declared that one of the most important of these methods is preprocessing.

2.2. Preprocessing and Stemming

Preprocessing is the first stage of data mining processes. In one of his studies, Türkeş (2007) state that preprocessing provides a faster process and lets us reach the suitable data quicker by performing housekeeping operations on the texts which will be processed [7]. For this housekeeping operation, the texts are divided into sentences first and then the sentences are divided into words. During this division, punctuations and space characters are used. The texts are saved as sentence index and sentences are saved as word index.

Stemming process is the process that makes a word pure and plain so that it contains its real meaning. This process changes depending on the structure of the language that a word is written in. For this reason, in the studies called natural language processing, the methods special to languages were developed by taking into account the structure of the language. In one of his studies [8] for Turkish texts, Kesgin (2007) says that although the meaning of a base or stem of a word does not change when it gets an inflectional suffix, it becomes a new word different from the old word because its spelling is changed. Balci (2010) states in his study [6] that the word gets rid of inflection suffixes in stemming process and derivational affixes must stay in the base of the word because they give new meanings to the words they are added to. For this reason, Balci thinks that when the word is purified from inflectional suffixes derivational affixes are not cleaned.

2.3. Methods

In the study, the longest matching algorithm from stemming methods was used in Turkish texts. This algorithm works with a dictionary that contains the words in base or in stem that have a meaning alone (derived by getting derivational affixes). We explained above that in Turkish the stemming process depends on the principle that the words are cleaned from inflectional suffixes and they keep the derivational affixes if there are any. The word that is wanted to be stemmed in the longest matching algorithm is first looked up in the dictionary that contains the stems. If it is not found, searching is repeated by deleting a letter in the end. The process ends when a stem is found or the word remains as one letter. The answer to the question "How work the Longest Matching Algorithm?" is shown in Figure 1.

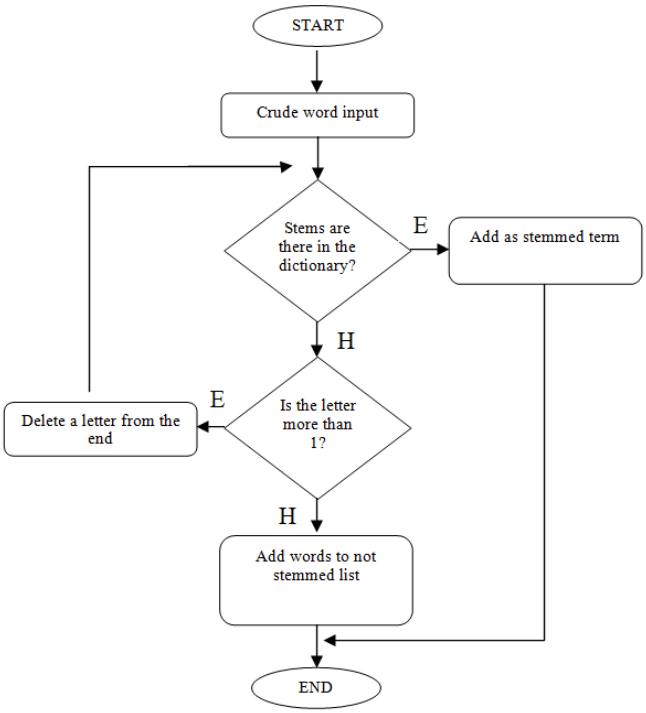

Figure 1. Longest Matching Algorithm Flow Chart

Flow chart is explained with an example in Table 2 below.

Example Raw Word: "incelendiğinde"

Table 2. An example of stemming process

Word	Is it in dictionary?	Is the letter more than 1?	Process
incelendiğinde	NO	YES	Delete a letter from the end.
incelendiğind	NO	YES	Delete a letter from the end.
incelendiğin	NO	YES	Delete a letter from the end.
incelendiği	NO	YES	Delete a letter from the end.
incelendiğ	NO	YES	Delete a letter from the end.
incelendi	NO	YES	Delete a letter from the end.
incelend	NO	YES	Delete a letter from the end.
incelen	NO	YES	Delete a letter from the end.
incele	YES		Add a stemmed term

This is one of the most understandable methods for the application; however it has a bad performance in terms of time because the searching is carried out many times in the dictionary for each term [6]. The fact that it has a bad performance in terms of time is regarded as a disadvantage that can be ignored when taking into account the advantage of using a dictionary. In addition to this, one of the biggest disadvantages of the method is that it can not control the sound events such as "Famous fall" and "Consonant softening" which can happen in Turkish. Therefore, this method fails in the stemming of some words.

In the study, all words in the right answer of a question saved in the database and all words in the answer given to this question by the student were subjected to stemming process.

3. Results and Discussion

In the application, a data set was created first. There are open-ended questions that require textual answers in the data set. When the system is active, data set is uploaded into the storage. After the question to be answered is selected, the question appears on the screen and the evaluation process is started after the answer is marked in the box on the screen. The evaluation process is to compare the right answer of the question saved in the data set to the answer given to the question. The first thing to be done for this process is to subject both data to stemming process after preprocessing. Two questions answered below are given as an example to this process and the results of the stemming are shown.

Example question 1:

İşveren Sağlık ve İşaretler Yönetmeliğine göre ne yapmamalıdır?

Right answer:

İşaretlerin anlamları ve bu işaretlerin gerektirdiği davranış biçimlerini, yazılı talimat haline getirmeden işçilere ve temsilcilerine vermemelidir.

Answer given by the students:

İşveren iş sağlığı ve güvenliği ile ilgili işaretleri işçilere vermemelidir.

According to the longest matching method, the results of the stemming process for the words of the right answer to the question obtained after preprocessing are seen at Table 2.

Table 3. The results of stemming of the correct answer for example 1

Word	True Stem	According to the stemming method	Is it correct?
işaretlerin	işaret	işaretle	No
anlamları	anlam	anlam	Yes
işaretlerin	işaret	işaretle	No
gerektirdiği	gerek	gerek	Yes
davranış	davranış	davran	No
biçimlerini	biçimle	biçimle	Yes
yazılı	yazı	yazı	Yes
talimat	talimat	talimat	Yes
haline	hal	hal	Yes
getirmeden	getir	getir	Yes
işçilere	işçi	işçi	Yes
temsilcilerine	temsilci	temsil	No
vermemelidir	ver	ver	Yes

In the right answer to the sample question above which is saved in the date set, 13 words that were considered to affect the meaning of the sentence were obtained after preprocessing. Each of these words was stemmed with the longest matching algorithm by means of the software that was prepared. The normal and right stems of the words and the stems obtained by the systems were given at Table 2 above comparatively. The graphic that shows the determination ratio of the suitable stem is shown in Figure 1.

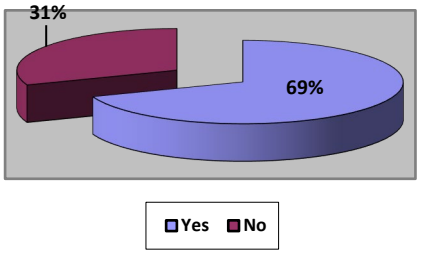

Figure 2. The detection rate of the appropriate stem in the correct answer for example 1

We processed the answer given to the question by the student in the same way and the obtained results are shown at Table 3.

Table 4. The stemming results of the answer given by the student for example 1

Word	True Stem	According to the stemming method	Is it correct?
işveren	işveren	işveren	Yes
iş	iş	iş	Yes
sağlığı	sağlık	sağ	No
güvenliği	güvenlik	güven	No
işaretleri	işaretle	işaretle	Yes
işçilere	işçi	işçi	Yes
vermemelidir	ver	ver	Yes

In the student's answer to the question, 7 words, which were considered to affect the meaning of the sentence, were obtained after preprocessing. Each of these words was stemmed with the longest matching algorithm by means of the software that was prepared. The normal and right stems of the words and the stems obtained by the systems were given at Table 3 above comparatively. The graphic that shows the determination ratio of the suitable stem is shown in Figure 2.

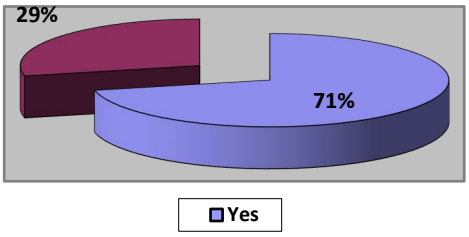

Figure 3. The detection rate of the appropriate stem in the answer given by the student for example 1

Example question 2:
Dil ile iletişim arasındaki ilişkiyi açıklayınız.
Right answer:
İnsan konuşma yetisine sahip bir varlık olduğu için en gelişmiş iletişim aracı dildir. Bu yüzden insanlar anlaşmak için dilini kullanır. Dille gerçekleştirilen iletişim diğer araçlarla gerçekleştirilen iletişimlere göre daha kullanılışlıdır.
Answer given by the students:
İnsanların arasındaki bilgi akışını sağlamanın en kolay ve en etkili yolu dildir.

Table 5. The results of stemming of the correct answer for example 2

Word	True Stem	According to the stemming method	Is it correct?
insan	insan	insan	Yes
konuşma	konuş	konuş	Yes
yetisine	yeti	yeti	Yes
sahip	sahip	sahip	Yes
varlık	varlık	var	No
olduğu	ol	ol	Yes
gelişmiş	geliş	geliş	Yes
iletişim	iletişim	iletişim	Yes
aracı	araç	aracı	No
dildir	dil	dil	Yes
insanlar	insan	insan	Yes
anlaşmak	anlaşma	anlaşma	Yes
dilini	dil	dil	Yes
kullanır	kullan	kullan	Yes
dille	dil	dil	Yes
gerçekleştirilen	gerçekle	gerçekle	Yes
iletişim	iletişim	iletişim	Yes
araçlarlar	araç	araç	Yes
gerçekleştirilen	gerçekle	gerçekle	Yes
iletişimlere	iletişim	iletişim	Yes
kullanışlıdır	kullanışlı	kullan	No

21 words, which were obtained crudely, were processed and the data that was obtained after this process is seen at Table 4. According to the Table 4, 18 stems were found suitable and the stems of 3 words were found wrong. The system's determination ratio for the suitable stem is shown in Figure 3 below.

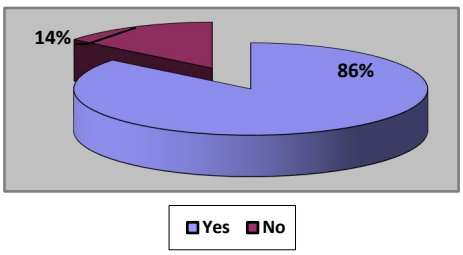

Figure 4. The detection rate of the appropriate stem in the correct answer for example 2

The data that we obtained after we processed the answer given to the question by the student are seen at Table 5.

Table 6. The stemming results of the answer given by the student for example 2

Word	True Stem	According to the stemming method	Is it correct?
insanların	insan	insan	Yes
arasındaki	ara	ara	Yes
bilgi	bilgi	bilgi	Yes
akışını	akış	akış	Yes
sağlamanın	sağlama	sağlam	No
kolay	kolay	kolay	Yes
etkili	etkili	etki	No
yolu	yol	yol	Yes
dildir	dil	dil	Yes

Totally 9 words were determined in the student's answer and suitable stems were established by the prepared system for 7 of these 9 words. The system's success to find the suitable stem is shown in Figure 4 below.

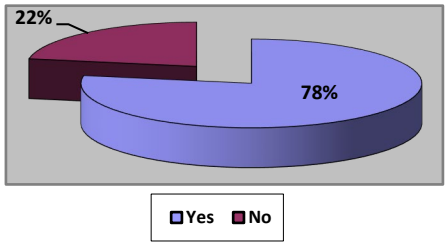

Figure 5. The detection rate of the appropriate stem in the answer given by the student for example 2

4. Conclusion

When we evaluate the results obtained above, we see that the system is considerably successful to find the stems of Turkish texts. On the other hand, it is seen that it is unsuccessful to stem some words. It is foreseen that this has two main reasons. The first reason is that the stem of the word that is searched does not exist in the dictionary. To get rid of this situation, the used dictionary needs to be extended so that it can contain the stems of more words. The other reason of the failure in stemming is the sound events in the structure of Turkish grammar. As is known, when words in Turkish get some suffixes some letters in the end of the word change. In the stemming method used in this study, these changing letters can't be determined. To clear up this problem, either a totally different stemming method must be used or this method must be developed so as to research grammatically as well.

Acknowledgements

The researchers express their sincere thanks to Selcuk University, Coordinatorship of Scientific Research Projects for their support with the project numbered 15401062.
A preliminary work of this study was presented at congress.

References

[1] Akdağ, H. and Çoklar, A.N., İlköğretim 6. ve 7. Sınıf Öğrencilerinin Sosyal Bilgiler Dersi Proje ve Performans Görevlerini Hazırlarken Yararlandıkları Kaynaklar, Internet'in Yeri ve Karşılaştıkları Güçlükler. Adiyaman

University Journal of the Institute of Social Sciences, Year 2, Issue 2, 2009, pp. 1-16.

[2] Nazlıçiçek, N. and Akarsu, F., Fizik, Kimya ve Matematik Öğretmenlerinin Değerlendirme Araçlarıyla İlgili Yaklaşımları ve Uygulamaları. Journal of Education and Science, Vol. 33, Issue 149, 2008, pp. 18-29.

[3] Gelbal, S. and Kelecioğlu, H., Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. Hacettepe University Journal of Education Faculty, 33, 2007, pp. 135-147.

[4] Mintzes, J. J., Wandersee, J. H. and Novak, J. D., Assessing Understanding in Biology, Journal of Biological Education, 35, 3, 2001, pp. 118-125.

[5] Wikipedia The Free Encyclopedia, Veri Madenciliği, Available link: https://tr.wikipedia.org/wiki/Veri_madencili%C4%9Fi, (Aug 09, 2016)

[6] Balci, M., Comparative Analysis Of The Long Match Algorithm In Computer Based Text Processing. Master Thesis, The Graduate School of Natural And Applied Science, Selcuk University, Konya, 2010

[7] Türkeş M.K., Phrase Based Indexing In Information Retrieval, Master Thesis, Graduate School of Natural and Applied Sciences, Istanbul Technical University, Istanbul, 2007

[8] Kesgin F., Topic Detection System For Turkish Texts, Master Thesis, Graduate School of Natural and Applied Sciences, Istanbul Technical University, Istanbul, 2007