Using Machine Translators in Textual Data Classification

Leila Rouka, Cemal K öse, and Parham Tofighi

Abstract—**In this paper, the effect of machine translators in the textual data classification is examined by using supervised classification methods. The developed system first analyzes and classifies an input text in one language, and then analyzes and classifies the same text in another language generated by machine translators from the input text. The obtained results are compared to measure the effect of the translators in textual data classification. The performances of the classification method used in this study are also measured and compared. The classification process can be described as training data preparation, feature selection, and classification of the input texts with/without translation. The obtained results show that Multinomial Naïve Bayes method is the most successful method, and that the translation has quite a small effect on the attained classification accuracy.**

Index Terms—**Text classification, machine translator, translated texts, textual data mining.**

I. Introduction

With the growth of online textual information, effective information retrieval is difficult without good indexing and summarization of document content. The textual data classification methods offer efficient solutions to this problem. Text classification (or categorization) is the process of structuring a set of documents according to a group structure to get a set of one or more categories [1]. In the literature, several methods are used for textual data classification, including, Bayesian classification, distance-based algorithms, decision tree-based methods [1]. In the Web domain, textual data classification is often done in English language because English is the most commonly used language in the mediums [1]. On the other hand, structures of languages are very different from each other. For example, the English language has 26 letters and whose morphological structures were relatively simple. In other languages with different structures, letters and writing types, for example Arabic languages consist of 28 letters, written from right to left and it has very complex morphology, the majority of words have a tri-letter root [1], [2]. On the other, hand Turkish languages consists of 29 letters, written from left to right, the Turkish belongs to the group of agglutinative languages and Turkish morphology is quite complex [2], [3]. In this paper, we investigate the same classification algorithms on the documents written in English and Turkish, and evaluate the efficiency of using translator in document classification. In the next sections we will describe the text classification process, the running experiment then the conclusion and appendix of application figures.

II. Text Data Classification

The system categorization is the task of determining the correct class or classes of invisible documents based on some learning examples. The classes are predetermined so a supervised learning algorithm is required to classify documents. Let, us given a description d □ X of a document, where X is the document space; and a fixed set of classes C= {c1, c2,…, cJ}. Classes are also called categories or labels. The text classification can be divided into four steps. The first is the documents preprocessing that is the most important step before the classification of textual data. Removing stop word and word weighting by using TF-IDF method, and then document tokenizing which means analyzing a text from symbols, and stemming of data are the process that applied in preprocessing of a document. Then, dimensionality reduction with feature selection, here we applied information gain (IG) as feature selection method in our study for improvements in the performance because feature selection plays an important role in text categorization [4]. In a text classification task, a text is represented as a vector of features from exclusive words occurring in documents but this vector can be high dimensional space. So, a good feature selection method reduces the feature space and can handle and contribute to high classification accuracy. The final step is the classification of the data. There are several methods for text classification including, Bayesian, distance-based curacy. In this study, we employed the text by Sequential Minimal Optimization, Naïve Bayes, Multinomial Naïve Bayes and J48 methods for classification algorithms and decision tree-based methods. Each of these methods presents different characteristic in the text classification. The differences between them are mostly based on the degree of classifying the textual data and measured the accuracy and Micro-averaged F- measure of the methods with/without translations.

In the application, two set of data with six categories are used for examining the effect of the machine translators in the textual data classifications. Firstly, the classification results are obtained for both of the data sets in all categories in the original languages. Then, the data sets are first translated to other language (the English to the Turkish and the Turkish to the English) by using the ACE translator of Google translators and then the translated documents are classified considering the categories.

To obtain knowledge from vast data and achieve high classification accuracy, the preprocessing of the document is a very important step before the categorization of documents. Removing stop word is the first step in the preprocessing. Elimination of stop words concerns omitting

Manuscript received February 27, 2012; revised April 23, 2012

The authors are with the Department of Computer Engineering, Faculty of Engineering, Karadeniz Technical University, and 61080 Trabzon, TURKEY.

non-meaningful words that they are having a very high frequency, and affecting the term weighting negatively. Then, documents will be consisting of meaningful words, and the system extracts candidate keywords from the document. Term Frequency-Inversed Document Frequency (TF-IDF) is one of the most popular weighting schemes in IR (information retrieval) [1]. TF-IDF computes weights for each word in a document. Thus, words with high TF-IDF means a strong association with the document they appear in. After pre-processing, stemming terms is an important step to get better classification accuracy [5], [6]. In the application, stemming gives us ways of finding morphological variants of each term Stemming also used to reduce the size of feature vectors. A simple language independent and task independent text categorization method based on character-level n-gram language model is given in [7]. And according to Guran [8], if the number of N-gram words increases in the feature space of a data collection sufficiently, its probability estimate will be better. Therefore, the uni-gram model is used in this application.

A. Algorithms Used for Textual Data Classification

In this study, Naïve Bayes, Multinomial Naive Bayes, J48 decision tree and Sequential Minimal Optimization classification methods are used to classify the textual data set before and after the translation. Hence, the most successful classification method is determined to measure the effect of the existing machine translator in data mining. The Naive Bayesian text classification algorithm is one of the commonly applied algorithms because it is a simple, fast and easily implementable algorithm. Here, it is assumed that all attributes of the examples are independent of each other given the context of the class. Naive Bayes classification (CNB) is expressed by equation (1) for the attribute values given as {x1,… xn}.

$$\mathrm{C_{NB}} = \arg\max_{c_j \epsilon C} \mathrm{P}(c_j) \prod_{i=1}^{|v|} \mathrm{P}(w_i \,|c_j) \qquad (1)$$

where P (wi|cj) and p(cj) are predictable from training documents with known classes. P (cj) is the probability of a class cj in training documents. For a class cj, p (wi|cj) is the conditional probability of word wi of a query document dq in training documents [8].

The Multinomial Naïve Bayes model is particularly appropriate for text classification [9]. In this approach, terms in a document are supposed to be drawn from an underlying multinomial distribution independently of each other. A document is represented by the number of occurrences of words or word weight in the document. Hence, the class CMNB for document dq= {w1, w2,…, w|v|} calculated by using the equation (2). Where p (wi|cj) is the relative frequency of term wi in documents depending to class cj.

$$\mathrm{C_{MNB}} = \arg\max_{c_j \epsilon C} \log \mathrm{P}(C_j) + \sum_{i=1}^{|v|} f_i \log \mathrm{P}(w_i|c_j) \quad (2)$$

In practice, the decision trees are most commonly used machine learning and text classification methods [8]. C4.5 decision tree or J48 algorithm builds a decision tree using a training data set. C4.5 as a recursive algorithm uses information gain ratio measure to determine an attribute that divides the data set into smaller data subsets. Here, the attribute with highest information gain ratio is chosen as a splitting attribute. At each step, the algorithm is applied on smaller subsets to form a decision tree by finding the remaining splitting attributes, recursively. The set of attributes found at each step make up the decision tree [8].

Another approach used in the textual data classification is Sequential Minimal Optimization (SMO) method. This method is a fast approach for training Support Vector Machine (SVM). Understanding and coding SMO approach is also easier because training a support vector machine requires the solution of a very large quadratic programming optimization problem. SMO method simply solves the quadratic programming problem by breaking it into several sub-problems [10].

B. The Classifications of Documents and Evaluations of Results

Two sets of documents in the English and the Turkish languages were collected from the Web medium. Then, two sets of documents are also generated by translating the documents from the English to the Turkish and the Turkish to the English. All stages of text classification tasks are applied to the data sets. For measuring classification performance, k-fold cross validation technique is applied to all of the English and Turkish data sets including the training and test data. The classification results are compared by using several metrics. These metrics are accuracy, precision, recall and F-measure [9], [11]. True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) rate is the percentage of correctly classified positive examples, negative examples incorrectly classified as positive, positive examples incorrectly classified as negative and correctly classified negative examples, respectively. Accuracy, expressed as equation (3), is measure of the correctly classified examples.

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \qquad (3)$$

The Precision and Recall are calculated by using equation (4) and equation (5), respectively

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (4)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (5)$$

$$\mathrm{F1} = \frac{2.\mathrm{Precision}.\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (6)$$

F-Measure accounted by recall and precision in the following equation (6).

Two metrics are used to measure the classification performance of the system. The first one is the accuracy measurement. The second is Micro-averaged F-measure, which is the average of the F-measure across all categories. The F-measure for each category is computed and the macro-averaged F-measure is obtained by taking the average of the F-measure values for each category. Here, M is the total number of categories and Fi is calculated for each class by using equation (7).

$$\mathrm{F_i} = \frac{2.\mathrm{Precision_i}.\mathrm{Recall_i}}{\mathrm{Precision_i}.\mathrm{Recall_i}}, \qquad (7)$$

$$\text{Macro F} - \text{measure} = \frac{\sum_{i=1}^{M} \mathrm{F_i}}{\mathrm{M}}$$

C. Data sets Used in the Evaluation in this study

TABLE I: NUMBER OF DOCUMENTS PER CATEGORY IN DATASET

Category of the document	Number of English documents	Number of Turkish documents
Sport	153	153
Economic	153	153
Medicine	153	153
Politic	285	285
Science	153	153
Religion	100	100
Total number of documents	997	997

In this study, two sets of data are collected from the Web to evaluate classification methods used in this study and measure effect of the machine translators in the textual data classifications. One of the set is in the English consisting of 997 documents in six categories, and other data set is in the Turkish consisting of 997 documents in same categories. Each document in the data sets was manually labeled and categorized into six categories as Sport, Economic, Medicine, Politic, Religion and Science, give in Table I.

III. RESULTS AND CONCLUSION

In this study, the text classification algorithms are first examined on the Turkish and the English data sets. The evaluation results show that the J48 classifier shows the worst performance among all the algorithms used in the classification. These results also show that the most successful classification method is the Multinomial Naive Bayes algorithm. According to the results given in the tables, the best classification results are obtained for the sport category and the worst classification results are obtained for the religion category of the Turkish data sets and economy category of the English data sets as given Table II and Table III.

TABLE II: CLASSIFICATIONS RESULTS OF ENGLISH DOCUMENTS AND THEIR TURKISH TRANSLATIONS FOR EACH CATEGORY

Category named	SMO		MNB		NB		J48	
	English data set	Translated EN to TR	English data set	Translated EN to TR	English data set	Translated EN to TR	English data set	Translated EN to TR
	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10
economy	0.875	0.867	0.914	0.885	0.9	0.857	0.693	0.688
medicine	0.954	0.947	0.957	0.947	0.947	0.916	0.857	0.817
political	0.914	0.906	0.926	0.926	0.908	0.871	0.755	0.718
religion	0.927	0.918	0.941	0.879	0.885	0.775	0.828	0.821
science	0.899	0.892	0.929	0.903	0.92	0.874	0.711	0.75
sport	0.961	0.968	0.977	0.98	0.967	0.964	0.873	0.828
Average	0.921	0.915	0.939	0.923	0.921	0.881	0.78	0.761

TABLE III: CLASSIFICATIONS RESULTS OF TURKISH DOCUMENTS AND THEIR ENGLISH TRANSLATIONS FOR EACH CATEGORY

Category named	SMO		MNB		NB		J48	
	Turkish data set	Translated TR to EN	Turkish data set	Translated TR to EN	Turkish data set	Translated TR to EN	Turkish data set	Translated TR to EN
	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10	C.V 10
economy	0.838	0.693	0.915	0.826	0.875	0.795	0.75	0.641
medicine	0.747	0.839	0.887	0.895	0.835	0.869	0.673	0.705
political	0.808	0.815	0.878	0.852	0.82	0.805	0.698	0.615
religion	0.701	0.814	0.809	0.85	0.716	0.759	0.644	0.698
science	0.788	0.842	0.862	0.862	0.782	0.816	0.698	0.644
sport	0.922	0.936	0.96	0.964	0.912	0.936	0.776	0.818
Average	0.807	0.822	0.888	0.873	0.83	0.83	0.69	0.677

The data sets are also translated from one language to another language by using the Google translator. Thus, original English and Turkish datasets are translated to the Turkish and the English, respectively. After the translation, the translated documents are classified again by using the same classification algorithms. These results exhibit that translations have a small effect on the classification accuracy of the documents as given in Table 2 and 3. Although textual data classification performance is slightly decreased after the machine translation, the results are still useful and meaningful as before the translation of textual data. For Turkish and English data set, the results obtained by using SMO algorithm differ dramatically.

In this application, the factor smoothing technique is not taken into account. In the future implementation of the system, the smoothing technique would be used for further improvement of the system. As another future work, the effect of other machine translators can also be examined in textual data mining. Therefore, the effect of machine translator in textual data mining can be revealed more precisely.

REFERENCES

[1] X. Yu, Y. Yuan, M. Tungare, M. P. Quinones, W. Yuan, and E. Fox, "Automatic syllabus classification using support vector machines," *IGI Global- USA,Virinia Tech*, pp. 61-74, 2009.

[2] T. Güngör, "Lexical and morphological statistics for turkish," *International XII.Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN*, pp.1-4, 2003.

[3] D. Torunoglu, E. Cakrman, M. C. Ganiz, S. Akyokus, M. Z. Gurbuz, "Analysis of preprocessing methods on classification of turkish texts," *IEEE - Department of Computer Engineering Doğuş University-Istanbul-Turkey*, pp. 112-117, 2011.

[4] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," In *Proc. of the Fourteenth International Conference on Machine Learning - San Francisco*, pp. 412-420, 1997.

[5] K. Oflazer, "Two-level description of turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp.137-148, 1994.

[6] H. Sak, T. Gungor, and M. Saralar, "Turkish language resources: morphological parser, morphological disambiguator and web corpus," *In GoTAL* 2008, vol. 5221, pp. 417-427, 2008.

[7] F. Peng, D. Schuurmans, and S. Wang, "Language and task independent text categorization with simple language models," in *Proc. of HLT – NAACL*, School of Computer Science, University of Waterloo-200 University Avenue West -Waterloo – Ontario - Canada, pp. 110-117, 2003.

[8] A. Guran, S. Akyokus, N. G. Bayazit, and M. Z. Gurbuz, "Turkish text categorization using N-Gram words," *International Symposium on Innovations in Intelligent Systems and Applications – Yildiz Technical University, Dogus University*, Yildiz Technical University, Yildiz Technical University, pp. 369-373, 2009.

[9] Y. Sasaki and R. Fellow, "The truth of the F-measure," *MIB -School of Computer Science*, University of Manchester, pp. 1-5, 2007.

[10] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," *MIT Press – Cambridge - MA*, pp. 185-208, 1999.

[11] N. Chinchor, "Evaluation metrics," In *Proc. of the Fourth Message Understanding Conference*, pp. 22–29, 1992.