Translation Based Arabic Text Categorization

M. K. Sankarapani
ICASA
New Mexico Tech
Socorro, USA
madhuk@cs.nmt.edu

R. Basnet
ICASA
New Mexico Tech
Socorro, USA
rbasnet@cs.nmt.edu

S. Mukkamala
ICASA
New Mexico Tech
Socorro, USA
srinivas@cs.nmt.edu

Andrew H. Sung
ICASA
New Mexico Tech
Socorro, USA
sung@cs.nmt.edu

B. Ribeiro
University of
Coimbra, Coimbra,
Portugal
bribeiro@dei.uc.pt

Abstract

This paper reports preliminary results of categorizing Arabic text into predefined categories by first translating the Arabic text into English using commercially available translators and then categorizing the English text using support vector machines (SVMs). An Arabic corpus from Leeds University of UK is used in the experiments.

Machine translation is prone to disfluencies and mistakes; even professional human translation often lacks precision. Text categorization, however, is a much easier task than translation and machine learning techniques have been utilized to obtain highly accurate automated categorization. This paper proposes that text categorization–especially the classification into a relatively small number of predefined categories such as Reuter's collection–relies only on lexical information and, therefore, it is feasible to categorize foreign-language texts using an automated translator (to translate texts into the target language, say English) and a trained classifier that categorizes texts in the target language.

Experiments described in this paper demonstrate that the combined use of a reasonable Arabic to English machine translator, and SVMs that are trained to perform English text categorization, results in accurate categorization of Arabic texts. Thus, our proposed translation-based methodology for foreign-language text categorization has obtained a validation, at least for Arabic texts.

Keywords

Translation based text categorization, Text categorization, Support vector machines, Arabic text categorization.

1. Introduction

The use of documents in digital form has widely increased over the years and Internet has played an important role. Retrieving data and information from large number of documents is a real problem. Good indexing and summarization techniques have to be developed since significant need on users to access documents in flexible way. In the last ten years automated content-based document management tasks have gained an important status in the information systems field. Text categorization is now being applied in many contexts like document indexing, document

Appearing in *Proceedings of the 2nd International Conference on Information Systems Technology and Management*, 2008.

filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of web resources and in general any application requiring document organization.

Text categorization or text classification (TC) refers to the process of classifying content-based documents, to pre-defined categories, giving it one or more category label. It is a useful technique to handle and organize massive data source available over the Internet [2].

Text categorization techniques are applied to novel domains such as web page categorization using hierarchical catalogues. These methods are also applied to a highly practical problem in biomedicine namely, Gene Ontology, which is one of the major activities in most model organism database projects [3]. Text categorization is useful in indexing documents for information retrieval and summarizing contents of documents of special interests.

Machine learning approaches have been applied to perform Automated Text Categorization (ATC). This ATC system builds classifiers that are capable of assigning a document to one or more labels which are predefined [4]. The advantages of ATC process are

- High classification accuracy, as good as human experts
- Considerable amount of savings in terms of manpower since no domain experts are needed for building the classifiers

2. Natural Language Processing

The goal of Natural Language processing (NLP) is to design and build software that will analyze, understand, and generate languages that humans use. NLP includes natural language understanding, natural language generation, speech synthesis, speech recognition and machine translation (translating one NL into another). Few applications of NLP are:

- Data analysis
- Data integration
- Better information retrieval
- Structure mining

English is one of the most widely used languages for natural language process, since it is very easy to train the learning machine for classification of content-based documents. This gives us an idea to translate Arabic text to English and then categorize them.

Arabic (العربية اللغة *al-luġatu l-arabiyyah* or just عربي *arabi*) is the largest living member of the Semitic language family (a family of languages spoken by more than 300 million people across) in terms of speakers.

There are many Arabic dialects. *Classical Arabic*, the language of the Qur'an, was originally the dialect of Mecca in what is now Saudi Arabia. An adapted form of this, known as *Modern standard Arabic*, is used in books, newspapers, on television and radio, in the mosques, and in conversation between educated Arabs from different countries. *Local dialects* vary, and a Moroccan might have difficulty understanding an Iraqi, even though they speak the same language. Arabic words are constructed from three-letter "roots" which convey a basic idea. Addition of other letters before, between and after the root letters produces many associated words [5].

There are 28 consonants and three vowel, a, i, and u. Alterations can be made to the basic meaning of a verb by adding some alphabets to the root. These

changes follow regular rules, giving ten possible verb forms (only three or four are most common) [5]. Example, the root k-s-r produces:

- Form I - kasara, "he broke"
- Form II - kassara, "he smashed to bits"
- Form VII - inkasara, "it was broken up"

In a few instances these different forms give different meaning for example *q-Atala* means “he fought” but *qatala* means, “He killed”.

3. Machine Translation

MT is a sub-field of computational linguistics. It is the translation of text from one language to other using translators. The various approaches in MT are

- Dictionary-based machine translation
- Statistical machine translation
- Example-based machine translation
- Inter-lingual machine translation

In this paper the Arabic documents are translated to English documents. These documents are preprocessed, followed by document indexing. Different classifiers can be applied for text categorization purposes, for examples, Support Vector Machines (SVMs), neural networks, etc.

In any predictive learning, such as classification, both a model and a parameter estimation method should be selected in order to achieve a high level of performance. Recent approaches allow a wide class of models of varying complexity to be chosen. Then the task of learning amounts to selecting the sought-after model of optimal complexity and estimating parameters from training data [6,7].

Within the SVMs approach, usually parameters to be chosen are (i) the penalty term C which determines the trade-off between the complexity of the decision function and the number of training examples misclassified; (ii) the mapping function Φ; and (iii) the kernel function such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. In the case of RBF kernel, the width, which implicitly defines the high dimensional feature space, is the other parameter to be selected [8].

We performed a grid search using 5-fold cross validation for each of the faults in our data set. We achieved the search of parameters C and γ in a coarse scale. Model selection results obtained through grid search using LIBSVM is given in figures 3-9.

4. Methodology

Figure 1 illustrates the steps carried out for TBTC. For our experiment we used Arabic corpus from University of Leeds, which consists of 386 documents in XML format [9]. These Arabic documents are translated to English. LingoWorld is used to translate the documents from Arabic to English [10]. Once the documents are translated they are converted back to their XML format. The conversion is done by using a program which takes the XML formatted document as input and builds a DOM structure. From this DOM structure appropriate tag elements are extracted and is built into a HTML formatted new file.

Our program creates DOM for a given XML file, and then appropriate values from each node are extracted and stored in HTML format. These HTML files are then used for translation from Arabic text to English text. Once the translation process is done, these HTML files are parsed, and corresponding contents are replaced in the DOM structure and converted back to XML file format. The translated XML files are processed to check for non-ASCII and non-translated Arabic characters,

which are then eliminated before using text categorizations methods.

As shown in Figure 2 the XML document is parsed for word extraction process. Tokenization is carried out by extracting words from the documents using suitable delimiters.

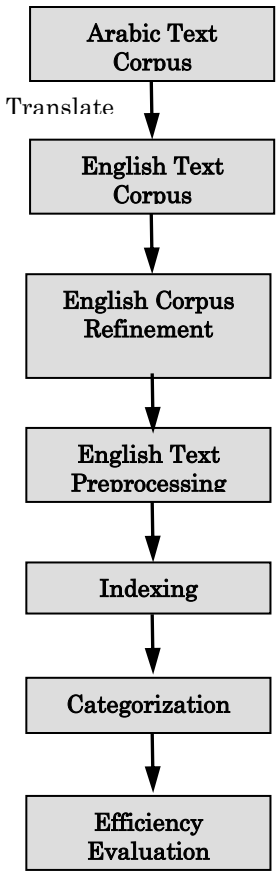

Figure 1 Methodology of translation based text categorization (TBTC)

Stop words are the most frequently occurring words which are not useful in text categorization process. In English language, there are about 570 stop words (Example: the, of, in, and, to etc). The number of stop words can also increase depending on the type of application being used. Removal of stop words contributes in improving the efficiency and effectiveness of the text categorization process, and hence these words are removed before classification.

Stemming is a process used to extract the root form of each word in the document. Stemming process differs from one language to another language [11]. Stemming finds the root of the word and replaces that word with the root. Once stemming is done, the resultant word is then stored in the database table 'Word'. Each unique word contains a unique ID. These unique words are considered as features. Use of database makes retrieval of information flexible and faster.

The document information is stored in a separate table 'Documents'. Here each document has a unique document id and other relative information such as author, topics, document name, which are nothing but the tag information. 'Word Count' column stores the total number of words (unique words) occurring in the particular document. The document's name and category id are also stored in the database. Then the frequency count of words (i.e. number of occurrences of a word in a document) is calculated.

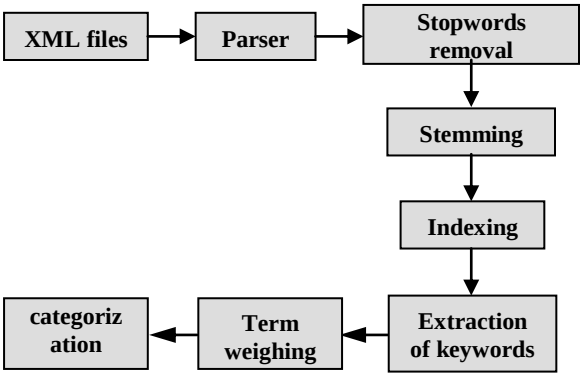

Figure 2 Stages of categorization process

Once this information is retrieved from the input documents, the Term Frequency (TF) and Inverse Document Frequency (IDF) values are calculated for each word in the document using the following formula:

Term Frequency (TF) given by

$$tf_i = n_i / \Sigma_k n_k$$

Where, n_i is the number of occurrences of a word 'i' in a document and $\Sigma_k n_k$ is total number of words in the document.

Inverse Document Frequency (IDF) is given by

$$idf_i = \log(|D| / |\{d:d \ni t_i\}|$$

Where, $|D|$ is total number of documents in the corpus and $|\{d: d \ni t_i\}|$ is number of documents where the term t_i appears (i.e. $n_i \neq 0$) [12].

5. Results

The corpus obtained contains 386 documents of the following categories given in Table 1.

ID	No of Documents	Categories
1	173	Arts
2	30	Autobiography
3	36	Science
4	58	Social science
5	8	Leisure
6	21	Public
7	9	Politics
8	16	Religion
9	35	Application science

Table 1 Number of documents in each category

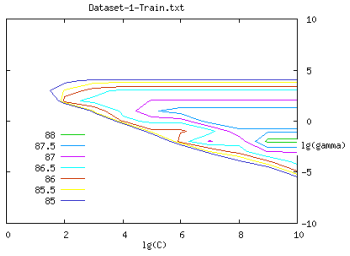

Figure 3 Model file for category 1

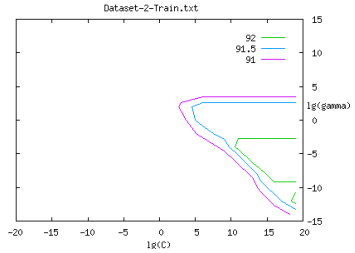

Figure 4 Model file for category 2

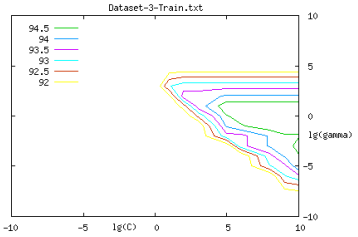

Figure 5 Model file for category 3

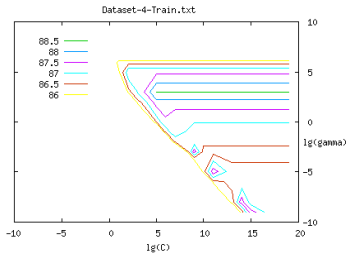

Figure 6 Model file for category 4

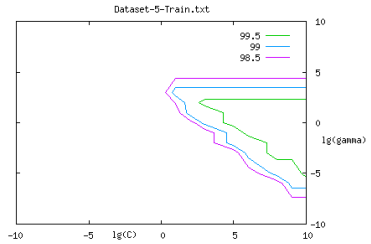

Figure 7 Model file for category 5

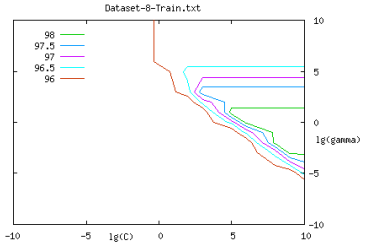

Figure 8 Model file for category 8

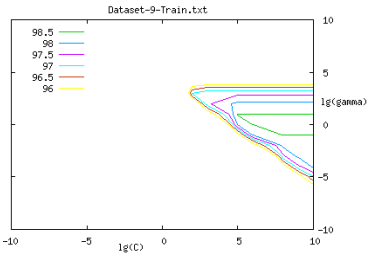

Figure 9 Model file for category 9

Table 2 gives the classification accuracies of SVMs:

Category	Train Accuracy	Test Accuracy	C	Gamma
1	88.15	89.65	512.0	0.3
2	92.22	97.41	64.0	1.0
3	94.81	89.59	64.0	1.0
4	88.52	92.24	32.0	8.0
5	99.62	97.35	32.0	1.0
6	95.18	93.90	64.0	1.0
7	98.90	98.25	256.0	1.0
8	98.15	97.25	64.0	1.0
9	98.52	93.96	64.0	1.0

Table 2 Training and testing accuracy of each category with corresponding C and Gamma Value

6. Conclusions

A simple method for foreign-language text categorization using commercial translators and SVMs is proposed in this paper. Categorization of Arabic text with accuracy better than 90% has been achieved, thus validating, preliminarily, our proposed methodology which relies on the assumption that text categorization can be performed based solely on lexical information and, therefore, an accurate translation is not required.

Although the results are promising, future work will include larger datasets from multiple languages with larger number of categories. The goal of our research is to integrate automated translators and English-text categorizers in order to develop automated and accurate foreign-language text categorization, without the need to train classifiers in each native language or to first obtain accurate translations.

Acknowledgement

The authors would like to acknowledge the support received from ICASA (the Institute for Complex Additive Systems Analysis, a division of New Mexico Tech).

References

[1]. Jones, D. A., Shen, W., Weinstein, C. (2005) "New Measures of Effectiveness for Human Language Technology", Lincoln Laboratory Journal, 15 (2) 341-345.

[2]. Sebastiani, F. (2002) "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 34 (1) 1–47.

[3]. Seki, K., Mostafa, J. (2005) "An Application of Text Categorization Methods to Gene Ontology Annotation", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp 138-145.

[4]. Silva, C. (2007) "On Text-based Mining with Active Learning and Background Knowledge using SVM", Journal of Soft Computing-A Fusion of Foundations, Methodologies and Applications, 11 (6), 519-530.

[5]. http://www.al-ab.com/arab/language/lang.htm

[6]. V. Cherkassy, V. (2002) "Model Complexity Control and Statistical Learning Theory", Journal of Natural Computing 1 (1), 109–133.

[7]. Cristianini, N., J. S. Taylor., J. S. (2000) "Support Vector Machines and Other Kernel-based Learning Algorithms", Cambridge, UK: Cambridge University Press.

[8]. Chang, C. C., Lin, C. J. (2001) "LIBSVM: A Library for Support Vector Machines", Department of Computer Science and Information Engineering, National Taiwan University.

[9]. Arabic Corpus University of Leeds: http://www.comp.leeds.ac.uk/eric/latifa/research.htm

[10]. http://www1.myworldlingo.com

[11]. Porter M. F. (1997) "An Algorithm for Suffix Stripping, Readings in Information Retrieval", Morgan Kaufmann Publishers Inc.

[12]. TF AND IDF http://en.wikipedia.org/wiki/Term_Frequency_Inverse_Document_Frequency