Translation Based Foreign Language Text Categorization

Ram Basnet, Guadalupe Janoski Torres,
Andrew H. Sung, Srinivas Mukkamala
Computer Science Department
New Mexico Tech
Socorro, NM 87801, USA
{rbasnet| silfalco| sung| srinivas}@cs.nmt.edu

Bernardete Ribeiro
CISUC- Computer Science Department
University of Coimbra
Coimbra, Portugal
bribeiro@dei.uc.pt

***Abstract*— This paper reports results of translating foreign-language text into English utilizing machine translation and then performing categorization of the English text by using support vector machines (SVMs).**

Machine translation is prone to disfluencies and mistakes; even human expert 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–relies only on lexical information and it is therefore feasible to categorize foreign-language texts using an automated translator (to translate texts into English) and a trained classifier that categorizes texts in English language by exploiting the fact that English language has proven and much mature automated text categorization process than any other spoken languages.

We experimentally demonstrate that the combined use of reasonable Arabic, Chinese, and Portuguese texts to English machine translators, and SVMs that are trained to perform English text categorization, results in highly accurate categorization of the native texts.

***Keywords*— Translation based text categorization, Text categorization, Support vector machines, Arabic, Chinese, and Portuguese text categorization.**

I. INTRODUCTION

The use of digital documents has widely increased over the years thanks to popularity of the Internet. Nevertheless, retrieving data and information from large numbers of documents is a challenging problem. The problem becomes worse when it comes to retrieving similar documents written in various foreign languages. As a result, good translating, indexing, and summarization techniques are essential since there is tremendous need for users to retrieve documents efficiently written and published in any languages on the earth.

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 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. A document may receive one or more category labels, or even a relevance value. TC is a useful technique to handle and organize the massive data source available over the Internet [3].

TC is useful in indexing documents for information retrieval (IR) and summarizing contents of documents of special interests. Text categorization techniques have been applied to several novel domains. One such domain is web page categorization which utilizes hierarchical catalogues. Another application is biomedicine namely, Gene Ontology, which is one of the major activities in most model organism database projects [4].

A very limited literature can be found regarding the TC in foreign languages. It becomes mind-numbingly difficult to do TC in other language without having the knowledge of that language domain. Minimally, one needs to know how to properly parse and extract features in order to represent the document in vector space model that is commonly used in IR and TC. On the other hand, English language has very mature and proven techniques for natural language processing. There is handful of automated machine translators available as software package or service. Machine translation is becoming impressively fluent and accurate. We can, therefore, exploit the rich techniques and capabilities available in English language processing in order to easily perform TC of foreign texts by first translating them into English using various commercial or free automated translation services. Above all, languages won't be of any barrier in IR, TC, or any text mining processes when they all can be first translated into one common language, English for instance.

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 [5]. The advantages of ATC process are:

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

II. 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 recognition and synthesis, and machine translation (translating one NL into another). A few applications of NLP are:

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

English is one of the most widely used languages, and much research has been done on English text categorization. This gives us the idea to translate Arabic, Chinese, and Portuguese text to English and then categorize them.

III. Related Works

M. Sankarpani et al have done the initial work on the Arabic language. They have translated Arabic text corpus into English using automated machine translation and then classified the translated English text into predefined categories using Support Vector Machines [1]. In this paper, we adopt the same process for Chinese and Portuguese texts and produce the experiment results for all three languages. Google offers Translated Search [16]. Users can provide words and phrases in their preferred language and ask the service to search pages written in one of about 23 other languages. The results pages are translated into the user's preferred language. It certainly lacks the capability to search in many languages simultaneously.

IV. Machine Translation

Machine Translation (MT) is a sub-field of computational linguistics in which text in one language is translated to other using machine translators. The various approaches in MT are:

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

In our experiments, we used Google Translate, a service provided by Google Inc. to translate Chinese and Portuguese texts into English. Unlike other translation service such as Babel Fish, AOL, and Yahoo which uses Systran, Google now uses its own translation software after switching from Systran.

Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach and require a great deal of work by linguists to define vocabularies and grammars. Several research systems, including Google's, take a different approach: they feed the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of example of human translations between the languages. Statistical learning techniques for building a translation model are then applied. [11].

V. Support Vector Machines

Support vectors machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis *h* for which we can guarantee the lowest error. The true error of *h* is the probability that *h* will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of hypothesis h with the error of h on the training set and the complexity of *H*, the hypothesis space containing *h* [13].

T. Joachims and several other authors have compared several machine learning algorithms for TC. T. Joachims has provided theoretical and empirical evidence that SVMs are very well suited for TC as SVMs acknowledge the particular properties of text: a) high dimensional feature space, b) few irrelevant features (dense concept vector), and c) sparse instance vectors [12]. The experimental results in [12] show that SVMs consistently achieve good performance on TC tasks, outperforming other existing methods such as neural networks, naïve Bayes classifier, a distance weighted k-nearest neighbor classifier (k-NN), the Rocchio algorithm, and the C4.5 decision tree/rule learner. This is the primary reason why we chose SVM as the classifier for our translation based TC tasks.

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. The task of learning then amounts to selecting the sought-after model of optimal complexity and estimated parameters from training data [6, 7].

Within the SVMs approach, the usual parameters to be chosen are (i) the penalty parameter on the training error, 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) kernel specific function values. In the case of RBF kernel, the width, which implicitly defines the high dimensional feature space, is the other parameter to be selected [8].

Despite the high dimensional space spanned by feature data in all our test cases (Arabic, Chinese, and Portuguese) the best results were achieved with Gaussian kernel. We performed a grid search using 10-fold cross validation to find parameters c and g and used those values to generate ROC curves using LIBSVM tools and to generate F1-measures using svm-light.

VI. Methodology

Fig.1 below shows the steps involved in our proposed

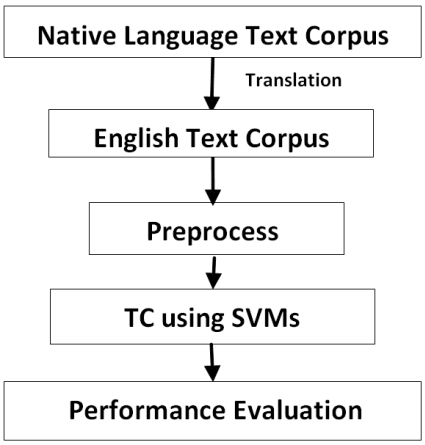

Figure 1. Methodology of translation based text categorization (TBTC).

Translation Based Text Categorization (TBTC).

The Portuguese CETEM Publico corpus consisting of 1.5 million extracts, more than 225 million tokens are excerpts of Portuguese newspaper Publico [14]. There are a total of 10 different categories. For this experiment 8 of the categories were used, which are shown in Table I. The "nd" category which is short for "not defined" and pol-soc (politics-society) categories were excluded from our experiments. 1000 randomly chosen documents with at least 75 tokens were extracted for each category.

The Lancester Corpus of Mandarin Chinese (LCMC) was used as the source for Chinese text documents. LCMC is a one million word balanced corpus of written Mandarin Chinese [15]. Table II displays the categories covered by LCMC and the number of samples available in each category.

Chinese and Portuguese text documents were translated into English using the Google Translate service. For the Chinese corpus native characters/words in Chinese that were not translated were either removed or manually translated by native speakers.

We used an Arabic corpus from Leeds University of UK as the source for Arabic text documents. The corpus consists of 386 documents with 9 different categories the distribution of which is listed below in Table III.

TABLE I. GENRES COVERED IN THE CETEMPUBLICO CORPUS

ID	Category	Samples	Proportion
1	Culture	1000	12.5%
2	Culture-Society	1000	12.5%
3	Technology	1000	12.5%
4	Sports	1000	12.5%
5	Economics	1000	12.5%
6	Opinions	1000	12.5%
7	Politics	1000	12.5%
8	Society	1000	12.5%
Total:		8000	100%

TABLE II. GENRES COVERED IN THE LCMC CORPUS

ID	Category	Samples	Proportion
1	Press reportage	44	8.8%
2	Press editorials	27	5.4%
3	Press reviews	17	3.4%
4	Religion	17	3.4%
5	Skills, trades and hobbies	38	7.6%
6	Popular lore	44	8.8%
7	Biographies and essays	77	15.4%
8	Miscellaneous (re-ports and official documents)	30	6%
9	Science (academic prose)	80	16%
10	General fiction	29	5.8%
11	Mystery and detective fiction	24	4.8%
12	Science fiction	6	1.2%
13	Adventure and martial arts fiction	29	5.8%
14	Romantic fiction	29	5.8%
15	Humor	9	1.8%
Total:		500	100%

TABLE III. GENRES COVERED IN THE ARABIC CORPUS

ID	Category	Samples	Proportion
1	Arts	173	44.4%
2	Autobiography	30	7.7%
3	Science	36	9.2%
4	Social science	58	14.9%
5	Leisure	9	2.3%
6	Public	23	5.9%
7	Politics	9	2.3%
8	Religion	17	4.4%
9	Application science	35	8.9%
Total:		390	100%

Tokenization was carried out by using suitable delimiters such as any non-alpha-numeric characters. Stop words or functional words such as article, prepositions, etc. that are not useful in the text categorization process were removed during preprocessing. Stemming was used to extract the root form of each word in the document. Since stem word as features performs better than single words and noun-phrase [17], we applied the popular and publicly available Porter Stemmer algorithm to stem translated English words.

Though there are various term weighting schemes such as BINARY, TF, LOGTF, LOGTFIDF, IDF, TF-CHI, TF-RF [10, 17], we have used the traditional but popular weighting scheme burrowed from the field of information retrieval, TF-IDF (term frequency-inverse document frequency) which is also one of the best, performance-wise. Term Frequency (TF) infers that a term that appears many times within a document is likely to be more important than a term that appears only once. However, some term may appear a lot more times in longer documents despite the actual importance of that term in the document. In order to prevent a biased situation like this TF is usually normalized. TF of each stemmed word is calculated using the formula in (1).

$$TF(w_{ij}) = n_{ij} / |d_i| \quad (1)$$

where, n_{ij} is the number of occurrences of a term w_j in a document d_i and $|d_i|$ is total number of all the terms in the document d_i.

Inverse Document Frequency (IDF) on the other hand says that a term that occurs in a few documents is likely to be a better discriminator than a term that appears in most or all documents. This notion is very intuitive in both information retrieval and text categorization.

$$IDF(w_{ij}) = \log(n / n_j) \quad (2)$$

where, n is total number of documents in the corpus and n_j is total number of documents where the term w_j appears. Multiplying TF with IDF we get the weight X_{ij} of a feature term w_j in a document i as:

$$X(w_{ij}) = TF(w_{ij}) * IDF(w_{ij}) \quad (3)$$

As a result, TF-IDF filters out the common terms by giving high weight to the term having high term frequency (in the given document) and the low document frequency of the term in the whole corpus.

VII. EXPERIMENTS

Since the LCMC and Arabic Corpus did not contain large enough samples for each category, 10 fold-cross-validations was done using LIBSVM tools. Using one-against-all

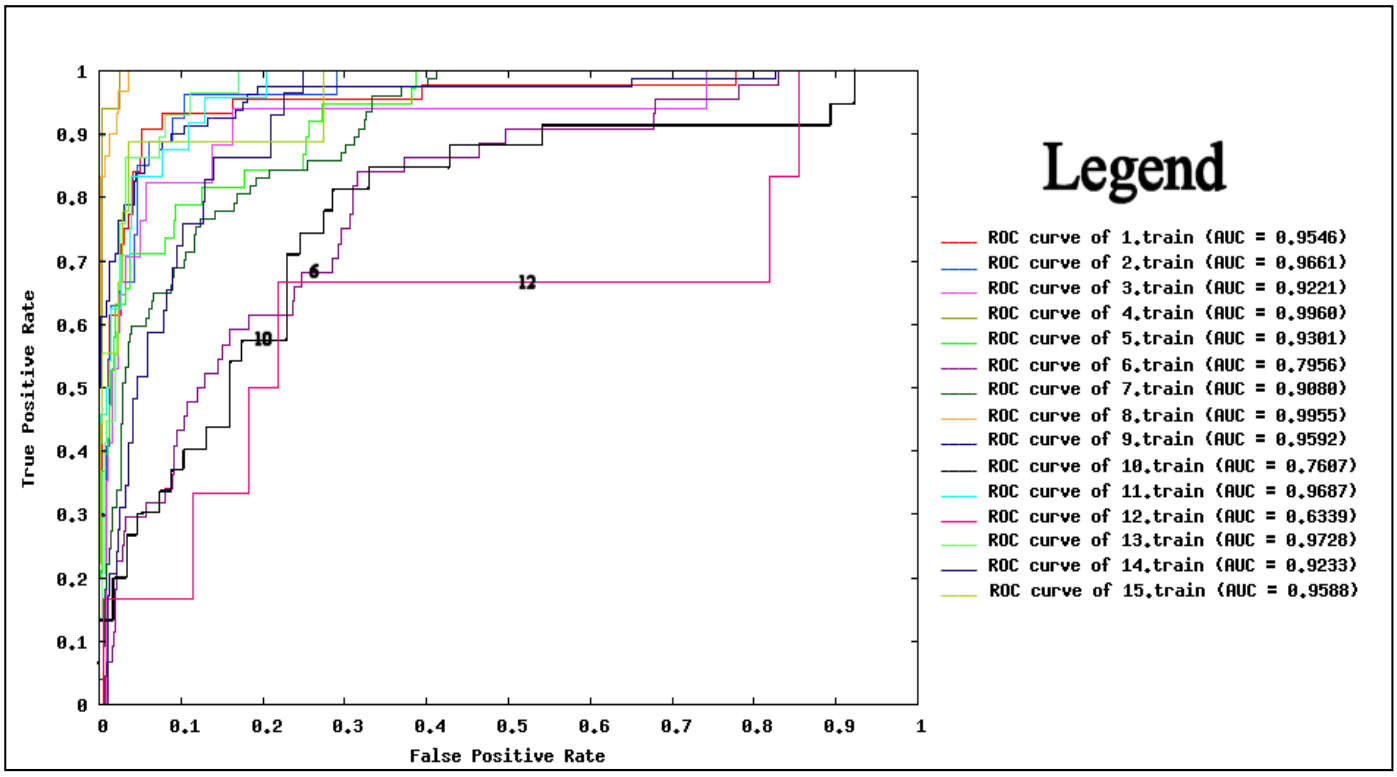

Figure 2. ROC curves on LCMC corpus.

approach, the grid search tool was used to find the best c (penalty parameter), g (gamma) for the RBF kernel. The best value pair was then used to train the SVM classifier with 10 fold-cross validation to produce the category wise ROC curves shown in section Fig.2 and Fig. 3.

For the Portuguese corpus, each of the training and testing data set was composed of 4000 samples of which 500 came from each category. The results for Portuguese test data set are shown in Tables V and VI.

Explain..Arabic stuff…. Translation service from WorldLingo [20] was used to translate Arabic documents into English.

VIII. Performance Evaluation

Since accuracy doesn't account for false negatives and false positives and assumes equal cost for both kinds of errors, Receiver Operating Characteristic (ROC) curves or F1-measure is usually generated to evaluate the performance of a binary classifier on a given dataset. ROC is a plot of true positive rate (sensitivity) against the false positive rate (1-specificity) for the different possible cut-off points of a test [19]. The area under the curve is a measure of text accuracy i.e. the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the result is. Normally, an area of .90-1 represents an excellent result and .80-.90 represents good result and the rest follows similar to the traditional academic point system. F1-measure can be interpreted as a weighted average of the precision and recall, where the value of 1 is considered the best and 0 the worst accuracy.

A. LCMC Corpus

TABLE IV. Accuracy for 15 categories

ID	Accuracy (%)
1	95.46
2	96.61
3	92.21
4	**99.60**
5	93.01
6	79.56
7	90.80
8	99.55
9	95.92
10	76.06
11	96.87
12	**63.39**
13	97.28
14	92.33
15	95.88
Average	**90.97**

The ROC curves in Fig. 2 show performance on one-against-all categories on the LCMC corpus. The AUC (Area Under Curve) for most categories tested is high, indicating good performance of the classifier. Categories 6, 10 and 12 do show a poorer performance. The low accuracies for category 6 (Popular Lore) and category 10 (General Fiction) are most likely caused by the general nature of their subject matter. The low accuracy of category 12 (Science Fiction), however, is probably due to its low sample size. The difference of their performance versus the other category classifiers is quite

noticeable when all the ROC curves are compared at once, as seen in Fig. 2.

B. CETEMPublico Corpus

The one-against-all classification accuracy, precision, recall and F1-Measure for the Portuguese corpus by using SVM-light are given below in Table V.

TABLE V. ACCURACY VALUES FOR 8 CATEGORIES

ID	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)
1	95.95	93.78	72.4	81.72
2	**94.20**	90.12	60.20	72.18
3	95.97	90.48	75.85	82.52
4	**99.02**	98.73	93.40	95.99
5	98.67	98.48	90.80	94.48
6	98.72	99.78	89.98	94.63
7	98.47	98.67	89.00	93.59
8	94.58	92.49	61.60	73.95
Average	**96.95**	**95.32**	**79.15**	**86.13**

The values in the diagonal of the confusion matrix in Table VI are total True Positive cases out of 500 positive samples.

TABLE VI. CONFUSION MATRIX FOR 8 CATEGORIES

		Predicted Category							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Actual Category	**1**	**362**	15	3	3	2	0	0	1
	2	9	**301**	27	0	3	0	1	14
	3	1	4	**380**	1	1	0	0	0
	4	1	0	0	**467**	0	0	0	0
	5	0	0	7	0	**454**	0	1	3
	6	3	0	2	2	1	**449**	3	3
	7	3	0	0	0	0	0	**445**	4
	8	7	14	1	0	0	1	1	**308**
False Positive		**24**	**33**	**40**	**6**	**7**	**1**	**6**	**25**

C. Arabic Corpus

The 10-fold cross-validation training accuracy on the translated Arabic corpus is given in Table VII. The ROC curves are put together in one single chart in Fig. 3 for better comparison and visualization.

TABLE VII. ACCURACY VALUES FOR 9 CATEGORIES

ID	Accuracy (%)
1	89.49
2	93.59
3	94.87
4	**88.21**
5	**98.97**
6	94.10
7	98.72
8	98.46
9	96.15
Average	**94.72**

IX. CONCLUSION

A simple method for foreign-language text categorization using publicly available translators and SVMs proposed in this paper has demonstrated excellent results. Categorization of Arabic text has achieved more than 90% accuracy in average, Chinese more than 90%, and Portuguese more than 95%. These results have preliminarily validated our proposed methodology of TBTC (translation based text categorization) which relies on the assumption that text categorization can be performed based solely on lexical information and, therefore, an accurate translation is not required.

With the promising initial results, future work will include expansion through the use of larger datasets from various languages such as Korean, Farsi, Urdu, etc. and expansion with the numbers of categories to work with.

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.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support received from ICASA (the Institute for Complex Additive Systems Analysis, a division of New Mexico Tech). The authors would also like to thank Messrs. D. Chen, R. Lopes, R. Koduru, S. Mungala, K. Bondili, M. Batta for their help in translating foreign language documents into documents into English.

REFERENCES

[1] M. K. Sankarapani, R. Basnet, S. Mukkamala, A. H. Sung, and B. Ribeiro, "Translation Based Arabic Text Categorization," Proceedings of Second International Conference on Information Systems Technology and Management, 2008.

[2] D. A. Jones, W. Shen, C. and Weinstein, "New Measures of Effectiveness for Human Language Technology," Lincoln Laboratory Journal, 2005, vol. 15, No. 2, pp. 341-345.

[3] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, 2002, vol 34, No. 1, pp. 1-47.

[4] K. Seki, J. Mostafa, "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, 2005, pp. 138-145.

[5] C. Silva, B. Ribeiro, "On Text-based Mining with Active Learning and Background Knowledge using SVM," Journal of Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2007, vol. 11, No. 6, pp. 519-530.

[6] V. Cherkassy, "Model Complexity Control and Statistical Learning Theory," Journal of Natural Computing, 2002, vol. 1, pp. 109-133.

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

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

[9] M. F. Porter, “An Algorithm for Suffix Stripping, Readings in Information Retrieval,” Morgan Kaufmann Publishers Inc, 1997.

[10] M. Lan, S.-Y Sung, H.-B. Low, and C.-L. Tan, “A Comparative Study on Term Weighting Schemes for Text Categorization,” vol. 1, pp. 542-545.

[11] O. Franz, “Official Google Research Blog: Statistical machine translation live,” <http://googleresearch.blogspot .com/2006/04/statistical-machine-translation-live.html>, 2006.

[12] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Machine Learning: ECML-98, Tenth European Conference on MachineLearning, 1998, pp. 137-142.

[13] V. N. Vapnik, “The Nature of Statistical Learning Theory,” Springer, New York, 1995.

[14] “Linguateca,” <http://www.linguateca.pt/Repositorio/>, 2007.

[15] Z. Xiao, A. McEnery, P. Baker, and A. Hardie, “Developing Asian language corpora: standards and practice,” Proceedings of the 4th Workshop on Asian Language Resources, 2004, pp. 1-8.

[16] “Google Translate,” <http://translate.google.com/translate_t>, 2008.

[17] C. Liao, S. Alpha, and P. Dixon, “Feature Preparation in Text Categorization,” Procedings of the Autralasian Data Mining Workshop, Canberra, Australia, 2003.

[18] T. Joachims, “SVM-Light Support Vector Machine” <http://svmlight.joachims.org/>, 2004.

[19] “The Area Under an ROC Cuve,” <http://gim.unmc.edu/dxtests/roc3.htm>, 2008.

[20] “myWorldLingo – your world of translations,” <http://www1.myworldlingo.com>, 2007.

[21]