Kwan Yi & Jamshid Beheshti
Graduate School of Library and Information Studies, McGill University
Montreal, Quebec

A text categorization model based on Hidden Markov models

Abstract: The Hidden Markov model (HMM) has been successfully used for speech recognition, part-of-speech tagging, and pattern recognition. In this study, we apply the HMM to automatically categorize digital documents into a standard library classification scheme. In the proposed framework, a HMM-based system is viewed as a model to generate a list of words and each document is seen as output of such a model. Experiments with a sample of documents from Digital Dissertations showed that the performance of our model surpasses that of the Naïve Bayes model, which has been used extensively in text categorization.

Résumé: Les modèles cachés de Markov (MCM) ont été utilisés avec succès dans le cadre de la reconnaissance de la parole, du taggage du discours partiel et de la reconnaissance des formes. Dans cette étude, nous appliquons les MCM pour catégoriser automatiquement les documents numériques dans un schéma de classification normal de bibliothèque. Dans la structure proposée, un système basé sur les MCM est considéré comme un modèle qui génère une liste de mots et chaque document est considéré comme une donnée d'un tel modèle. Des expériences tentées à l'aide d'un échantillon de documents provenant de *Digital Dissertations* ont démontré que la performance de notre modèle surpasse le modèle de *Naïve Bayes* qui est utilisé fréquemment dans la catégorisation textuelle.

1. INTRODUCTION

Over the past decade, an enormous amount of digital information has been generated by private and public sectors and transmitted through the Internet and intranets. The rapid growth and availability of information poses new challenges for finding information of interest, and requires the development of new retrieval methods. An estimated three billion Web pages are currently available on the Internet with 1.5 million new pages added daily. In addition to search engines, subject directories such as Yahoo!, LookSmart, and the Open Directory have been developed to assist users to browse through a vast amount of information. With the exception of Open Directory, which relies upon hundreds of volunteers, all the other portals employ professionals to manually classify and index Web sites into hierarchical subject structures. Two factors, however, hamper the indexing or categorization efforts of the Web portals: cost and time. Hiring professionals is an expensive method of organizing digital information, and using volunteers results in inconsistent indexing practices. Manual classification has also meant that less than two percent of all the digital information on the Web is currently covered by the directories.

Text Categorization or Text Classification (TC) is one way to cope with the information glut. The major task of TC is to assign a document into a category from a pre-defined set of categories with the objective of the rapid classification and therefore fairly easy access to documents. Since 1990s, the machine learning approach has placed TC projects in the main stream of information

retrieval research. In this approach, the classification rules for document categorization are learnt from a collection of pre-classified documents. Machine learning based TC applications have been investigated as means of learning classification rules. These techniques include decision trees (Crawford et al., 1991), Bayesian model (Lewis & Ringuette, 1994), Neural networks (Ng et al., 1997), and Support Vector Machines (Joachims, 1998).

Designing new classification rules, however, can be challenging and time consuming. On the other hand, librarians have used classification schemes to organize physical items in libraries for decades. Since the early 1980s, conventional library classification schemes have been also considered as a potential retrieval tool for subject access to digital information (Svenonius 1983; Markey 1985). Conventional library classification schemes such as the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) may be considered for TC, comparable to the subject directories used in the Web, where specific classification schemes designed for the topical representation of digital information on the Web are employed. The TC applications using library classification schemes have been considered recently as a viable option to designing custom classification (Finni & Pauson, 1987; Guenther, 1992). As one of the pioneering works on automatic classification based on a library classification scheme, Larson's research (1992) attempted to classify a set of machine readable cataloging (MARC) records into LCC, based on the information embedded in their title and subject heading, by creating clustering vectors for the LC subclasses of interest. Two other significant TC projects using library classification schemes are Scorpion (Schafer, 1997; Schafer & Fausey, 1997; Scorpion, 1999) and Pharos (Dolin et al., 1999).

In this study, we propose a novel method of automatic classification combining a statistical model, the Hidden Markov Model (HMM), with LC classification and LC Subject Headings (LCSH). HMM has been successfully applied in other text-related tasks, including information extraction, information retrieval, and text segmentation, but few in text classification. This study is the first attempt to extend the use of HMM to automatic classification of digital documents using library classification and subject headings. The data set based on the catalogue records provides the LC classification-LC subject heading associations necessary for HMM system training. Once the HMM is constructed, the model's performance as a classifier of digital documents is evaluated and measured using a sample of the ProQuest digital dissertations database as a test bed, and is compared to a text classification system based on Naïve Bayesian.

2. RELATED WORKS

Since the eighties, HMM has been used as a major statistical model for sequential process in the applications of speech processing and pattern matching (Rabinar, 1989). More recently the model has been extended to the text-related tasks such as information retrieval (Elke & Schäuble, 1994; Miller et al., 1999), information extraction (Leek, 1997; Seymore et al., 1999; McCallum et al., 2000), and text summarization (Conroy & O'Leary, 2001). Frasconi (Frasconi et al., 2002) used HMM to classify each page of a multi-paged document into a document format such as title, preface, and table of content. His HMM trained to learn the sequence of the document structure using various types of documents and find the most probable document format as a

page category. Our study focuses on the classification of the content rather than the structure of a document.

Miller et al. (1999) proposed an information retrieval (IR) model using HMM. Given a query Q and a set of documents D, a retrieval system tries to find a document d that is relevant to the query Q. In the expression of probability, the situation can be formulated as follows: $\max_{d \in D} P(d/Q)$. By applying Bayes' rule, $P(d/Q) = \frac{P(Q/d)P(d)}{P(Q)}$ can be achieved. Miller et al. tried to model $P(Q/d)$ using HMM, by interpreting the IR model as a generator of a query Q. In this study the proposed TC model uses a similar approach, with the exception that document and query correspond to target subject category and relevant document, respectively. Each target category is implemented by a TC model, and selection of relevant document relies on the similarity produced by category model. The structure and the parameters of the two HMMs, and the theoretical assumptions underlying their approach are also different.

Over the years, librarians have developed effective and efficient classification schemes, and procedures for storage and retrieval of documents. Only a few projects have experimented with applying classification schemes designed for printed materials to the classification of digital documents. Scorpion (1999) uses DDC to classify information that it has extracted through automatic indexing. The outcome of the project shows that while Scorpion cannot match manual classification, it may theoretically produce relatively acceptable results under certain conditions. Pharos (Dolin et al., 1999) utilized cataloguing records as a training set to classify information in the newsgroups based on LCC. More than 1.5 million records were used to construct a vector space between LCC and subject terms. In this project, we use a selected number of LCC categories to test the proposed HMM.

3. MODEL DESCRIPTIONS

3.1. Overview

The conceptual process of text classification may be described as a process of finding a relevant category *c*, for a given document *d*. The conceptual process of TC can be divided into sub-processes by applying the Bayes' rule as follows:

$$P(c/d) = \frac{P(d/c)P(c)}{P(d)} \qquad (1)$$

As a component of the expression (1), the probability of a given document *d*, $P(d)$, is constant for any category. The probability of a category, $P(c)$, is the prior probability of a category *c*. In this study, we provide a TC model implementing the probability of a document *d* given a category *c* to approximate the probability of a category given a document. Each category is represented by a HMM-based TC model. The TC model generates a relevance rank list for each document for a specific category. In this model, a document is treated as an equivalent to a list of words, and the

TC model for a category is viewed as a generator of a sequence of words. The most probable document generated by a TC model for a category is characterized by the expression (2).

$$\max_{d \in D} \quad P(d / c) \qquad (2)$$

3.2. TC Model based on HMM

A HMM $M = (I, E, T, O, S)$ is characterized by the five major components: initial probabilities I, output symbol emission probabilities E, state transition probabilities T, a set of output symbols O, and a set of states S. There are important assumptions underlying HMM. First, an emission probability of a symbol is only subject to the current state. Second, the current state is dependent on only one previous state, instead of the history of all previous states. More details on the principles and algorithms of HMM can be found in Rabiner's article (1989).

Figure 1 shows the HMM framework for the proposed TC model, which is applied to all the TC models for different categories. According to the structure, there are two dummy states and two states used for the model. The two dummy states indicate the "Start" and the "End" states. The two internal states appearing inside the model are designed to represent different sources of information. The current model incorporates two different information sources (ISs) into the system, with the future possibility of adopting more diverse ISs. The first IS, *Classified Documents,* denotes information collected from a set of abstracts from ProQuest digital dissertations, which is classified in a preset category. The second IS, *Subject-specific Documents,* symbolizes the descriptions from LCSH field in MARC records for the same category as used in the first IS.

In designing the system, we assume that various information sources may convey different aspects of information for a particular subject category. Based on this assumption, the HMM has been designed to reflect the idea that the IS from ProQuest digital dissertations (*Classified Documents* state) can provide specific information and the IS from LCSH (*Subject-specific Documents* state) can deliver more general information, for a target subject. Therefore, the set of output symbols O used in this model is defined as all the words from the two ISs included in the model.

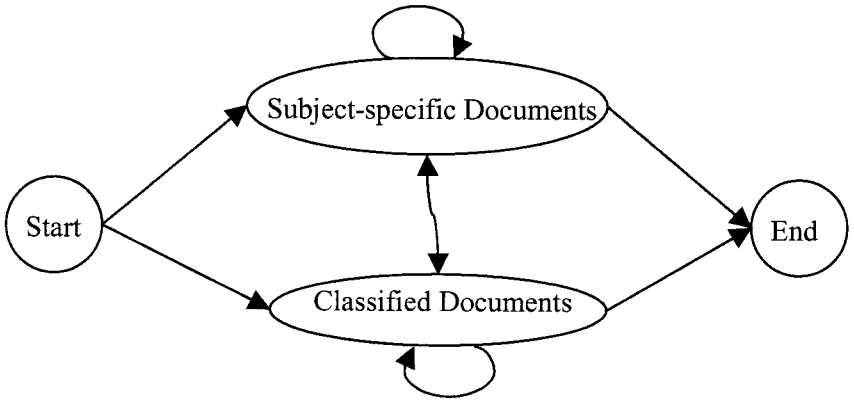

Figure 1: An architecture of HMM with two different information sources for a subject.

3.2. Model Training

Once the structure of HMM consisting of a set of states and transition flows is determined, the next procedure is to parameterize the model variables for emission probabilities and state transition probabilities.

For the training data set for our HMM, 500 records were selected from the OCLC Cat CD pertaining to English-language thesis type. These records were then traced in the ProQuest digital dissertations database (PQDD) and the abstracts were downloaded. The subject descriptions from the 650 field in MARC records were also collected from the 102,650 records of the OCLC CatCD for Windows.

In the general situation with learning models, especially with HMM, the Baum-Welch algorithm as a special case of Expectation-Maximization (EM) method (Dempster et.al., 1977) is used for training the variables of HMM with incomplete training data (Rabiner, 1989). The training data used in this study are manually collected to include all the labels with corresponding categories. With all labeled data, the emission probabilities of output symbols in a state can be estimated by the ratio of the number of occurrences of a symbol given the total number of all the output symbols (as shown in (3)) that indicates emission probability of the symbol W_i at the state S_j where V is a set of distinct symbols in the state.

$$P(W_i / S_j) = \frac{N(W_i, S_j)}{\sum_{k=1}^{|V|} N(W_k, S_j)} \qquad (3)$$

The problem with the formula (3) for emission probability is that it generates zero probability if a symbol does not appear in training data. We adopt the n-estimate probability (Mitchell, 1997) for our model to provide a constant probability by a ratio of the number of total symbols for new symbols (4),

$$P(W_i / S_j, C_c) = \frac{1 + N(W_i, S_j, C_c)}{|V| + \sum_{k=1}^{|V|} N(W_k, S_j, C_c)} \qquad (4)$$

where $N(W_i,S_j,C_c)$ is the total number of word occurrences W_i in training documents whose target class is c and target information source is j. $|V|$ is the total number of distinct terms appearing in the training documents for class c and information source j.

In general, the estimation of state transition probability in HMM can be made by simply counting transition occurrences between states with labeled data, or can be obtained by maximum likelihood process led by EM (Rabiner, 1989) to deal with unlabeled data. In either case, the methods involve the estimation of parameters for the observation of the relation between different states. However, we argue that such parameters will not be effective in our model because the relation among ISs (which are *states* with this model) do not support subject category information. Our method consists of a new role for transition parameters as measured by the average amount of information a word in a target state holds. First, we consider the initial transition probabilities that shift from the start state to the other states to be the probabilities of the number of different information sources (IS). Second, we interpret the transition probabilities between states as the probabilities of the amount of information each IS has at the level of category. In our model, an information quantity for each IS is measured by the formula based on the standard term frequency (TF) / inverse document frequency (IDF) (Salton & Buckley, 1988) as follows:

$$I(C_i) = \sum_{w \in Ci} I(w) = \sum_{w \in Ci} TF(w, C_i) IDF(w)$$

$$I_{normal}(C_i) = I(C_i) / |C_i|$$

$$P(C_i) = \frac{I_{normal}(C_i)}{\sum_{\forall j} I_{normal}(C_j)} \qquad (5)$$

The sequence of expressions in (5) denotes the process of obtaining the probability of shifting from the start state to the state i. As the formula shows, the initial probability is the normalized amount of information an IS holds in terms of TF/IDF.

$$I(C_i^k) = \sum_{w \in C_i^k} I(w) = \sum_{w \in C_i^k} TF(w, C_i^k) IDF(w)$$
$$I_{normal}(C_i^k) = I(C_i^k) / \left| C_i^k \right|$$
$$P(C_i^k) = \frac{I_{normal}(C_i^k)}{\sum_{\forall j} I_{normal}(C_j^k)} \qquad (6)$$

The states transition probability formulated in (6) is similar to the initial probability except that the training data are divided into different categories. In summary, for the estimation of initial probabilities, the quantity of information among different ISs is measured, whereas for the state transition probabilities, the amount of information of different ISs for the same category is measured. In either case, they are normalized in two ways: (1) size of each IS, and (2) various types of ISs.

For the use of TF/IDF algorithm in our TC system, we adopt the modified version of TF/IDF (Roberson et al., 1995) that was implemented by several IR systems (Ponte & Croft, 1998; Miller et al., 1999), shown below.

$$TF(t,d) = \frac{tf(t,d)}{tf(t,d) + 0.5 + 1.5 \dfrac{l(d)}{AveNum}}$$
$$IDF(t) = \frac{\log\left(\dfrac{N + 0.5}{df(t)}\right)}{\log(N+1)}$$

$TF(t,d)$ = the modified version of term frequency
$tf(t,d)$ = number of the term t in document d
$l(d)$ = total number of terms appearing in doument d
$AveNum$ = average number of terms of a document in the corpus
$IDF(t)$ = the standard version of inversed document frequency
N = total number of documents in the corpus
$df(t)$ = number of documents in the corpus having the term t

3.3. Document Classification

So far we have discussed the design of a TC model and the building of the TC system based on HMM. Now, we will describe the principles and procedures of document classification.

A set of documents based on the ProQuest digital dissertations (*test documents*) are prepared for testing our model. Each document in the test dataset, already classified by a professional, is labeled with a LCC category. For testing the proposed TC model, given a test document *d*, a HMM, designed to represent a LCC category *c*, produces a probability. This probability is then

used to measure the relevance between *d* and *c*, as described in section 3.1. The same procedure is used to calculate relevancy probabilities between documents and LC categories with other HMMs. The result is a list of relevant probabilities for all the documents and categories.

There are two questions that should be answered in the classification process: how to define the relevancy between a subject category and a document, and how to measure it. First, as previously explained in section 3.1, in our model a document is viewed as a list of words, and a TC system is considered to be a model, which generates words. Relevancy is defined as the probability of similarity between the terms in the document and the predefined subject categories. The similarity between a document and a subject category is measured by the probability of a list of HMM generated terms in the document. The TC model produces a sequence of words along with the probabilities corresponding to the words. In the HMM, given a list of output symbols, there are numerous possible paths of states producing the same output symbols. Therefore, more formally, the probability of a HMM TC model given a list of terms is the summation of all the probabilities from different paths that was taken to produce the terms.

Second, the estimation of relevance is the process of how to obtain a probability indicating the relevance between a document and a category. Given N being the number of states and T being the number of terms in a document, the direct method of calculating the probability requires $\Theta(N^T)$ multiplications (Rabiner, 1989), which poses logistical problems even with small number of N and T. For the estimation of the probability, our model uses an efficient method based on dynamic procedure referred to as the *forward* algorithm (Baum & Egon, 1967; Rabiner, 1989).

4. EXPERIMENTAL RESULTS

4.1 Experimental Background

This HMM-based TC model was built and tested on two data sets. The first set consists of a sample of documents from ProQuest digital dissertations, produced by UMIProQuest. LC number and LCSH for the selected abstracts are found in OCLC FirstSearch Database-WorldCat database. The second set consists of MARC records from the OCLC WorldCat CD database. A subset of information in the cataloguing records containing topical subjects and their descriptors is used as the training data set in the proposed model.

We have implemented a HMM-based prototype of TC model for three major LC classes: class Q (Science), class S (Agriculture), and class N (Fine Arts). The decision for selecting the three disparate classes was based on the nature of the disciplines, which represent various "hard" and "soft" disciplines, on the spectrum suggested by McGrath (1978) and Harter (1986). The assumption here is that the semantic ambiguities of words or phrases are inherently dependent on the discipline that they represent, and hence may affect the TC model.

Generally, the problem of TC is recognized as a task of learning concepts or categories. Among the machine learning algorithms, the naïve Bayes classifier is known as one of the most effective algorithms for such learning tasks (Mitchell, 1997, p. 155) and is widely used in TC (Lewis & Ringuette, 1994; Koller & Sahami, 1997; Baker & McCallum, 1998). For the evaluation of the

performance of our model, a TC model based on simple naïve Bayes classifier was built. The standard naïve Bayes classifier algorithm was implemented using Mitchell's description, and substituting Roberson's version of TF/IDF (described in section 3.2) for the m-estimate equation (Mitchell, 1997) as an estimator of the probability terms.

4.2. Results and Analysis

To demonstrate the effectiveness of our HMM, two different TC models are built for the experiments: One is based on the HMM theory and the other is based on Naïve Bayes (NB) algorithm. To observe the effect of our new approach for transition probability to the system performance, two different HMM-based TC models are designed for the experiments. The first type of HMM is implemented with our new approach on transition probability described in Section 3.2, and the second type is a more straightforward version of HMM without transition probabilities. The three different models are trained by two different sizes of training data to find out the variation of the effectiveness of the experimental models due to the different amount of training data sets.

Figures 2-4 show the performances of the models trained with 10 documents, and Figures 5-7 show the performances with 20 documents. In this experiment, each of the 5 test documents was investigated with all TC systems modeled for all categories (25 subcategories consisting of 7 for N class, 12 for Q class, and 6 for S class). A probable rank of categories for the tested documents is generated. In the figures, each square marker indicates the average rank of the target categories for 5 documents tested.

As the figures indicate, in all different experimental settings, the HMM-based TC models outperform NB classifier. Regardless of the size of training data and the discipline, the average rank of the documents tested on the HMM-based TC models is below 5, whereas the average rank for the NB classifier is about 22. In addition, the two HMM-based TC models produce more or less similar results for different settings. Since the same method is adopted for estimating transition parameters and emission parameters, it may be simpler to use the model without the transition probabilities.

In testing the HMM in different subjects or disciplines, the result shows that the calculated ranks have a relatively narrow range. Figures 2 and 5 show the performance lines for the N class chosen as a representative for a "soft" discipline, and Figures 3 and 6 indicate the results for the Q class selected to represent a "hard" discipline. The figures show that the average ranks of the documents tested for the two different disciplines are approximately identical. This finding is somewhat surprising given the nature of the two disciplines. Common wisdom dictates that documents from "soft" subjects are more difficult to classify correctly than those from "hard" disciplines. As shown in Figures 2 and 5, the average ranks for the "soft" discipline are evenly distributed over all the range of the target categories, whereas for the "hard" discipline, the output ranks fluctuate over the categories. One possible explanation for these results is that the terms from "hard" disciplines are less ambiguous, and thus if they are found in the training data the TC model finds the correct category for a document containing those terms. If the key terms are not found, other terms in the document hinder classification and increase the value of the

ranks. The results also show that the ambiguous property of terms from the "soft" discipline, on the other hand can help the process of document classification.

Determining the correct size of training data for a TC model can be problematical. In this experimental setting, the average ranks for all the disciplines were similar for training sets with 10 documents and 20 documents, with the exception of a couple of subcategories.

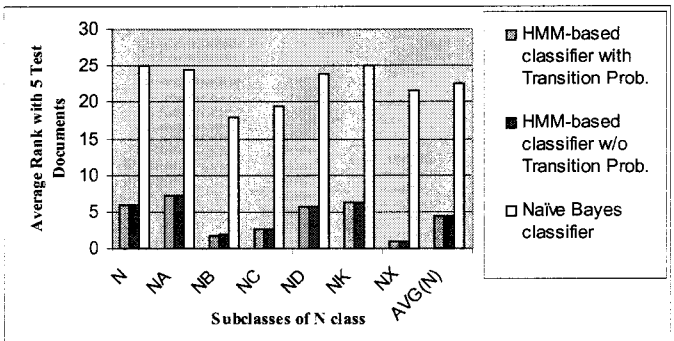

Figure 2: Comparison of TC models trained with 10 documents for N class (Fine Arts)

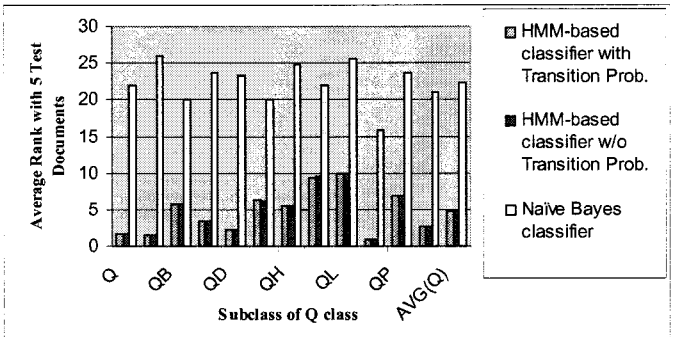

Figure 3: Comparison of TC models trained with 10 documents for Q class (Science)

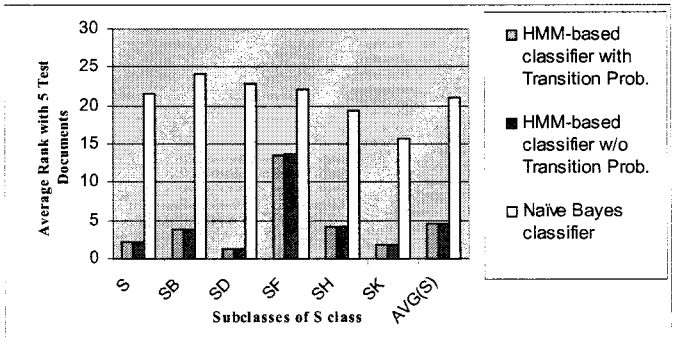

Figure 4: Comparison of TC models trained with 10 documents for S class (Agriculture)

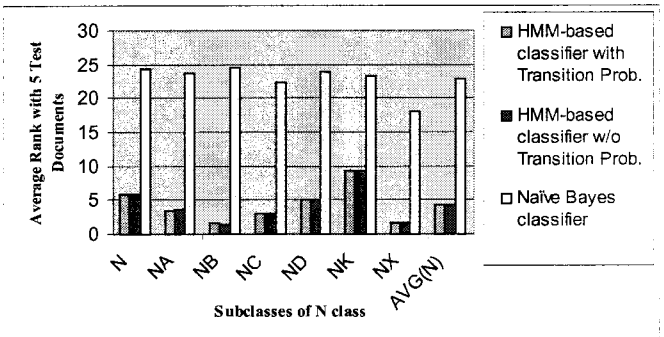

Figure 5: Comparison of TC models trained with 20 documents for N class (Fine Arts)

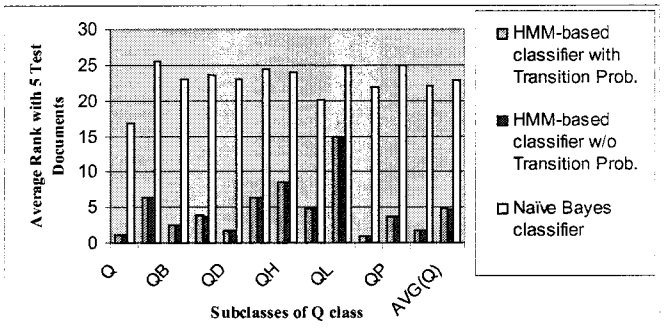

Figure 6: Comparison of TC models trained with 20 documents for Q class (Science)

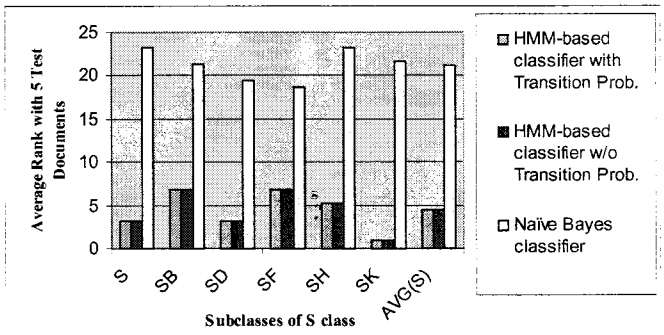

Figure 7: Comparison of TC models trained with 20 documents for S class (Agriculture)

5. CONCLUSIONS AND FUTURE RESEARCH

We have proposed a novel TC model based on HMM for the document classification. The model was tested on a set of documents derived from ProQuest digital dissertations. Text categorization was accomplished by using the HMM to classify the documents based on the Library of Congress classification scheme. While the HMM is based on the use of the model in areas other than TC, it has been modified for classification purposes. A new feature of the model is a modified version of the transition probabilities. Our model implemented a new type of transition probability and was compared to one without state-transition probability. The performances of the two different models are in generally similar, with a few exceptions.

Another feature of the HMM is its structural flexibility. In this model, two different information sources (ISs) were incorporated to reflect the concept of a target subject category with the assumption that different ISs can hold different aspects of the information for the same subject. This is an unverified assumption in this experiment, and remains to be tested in the future research by introducing more different ISs.

A number of experiments were conducted to determine the effectiveness of the proposed HMM-based TC models. The experimental results show that the performance of our model is better in automatically classifying documents than the widely used Naïve Bayes model under all experimental settings. In this experiment, we proposed two new methods: (1) the concept of IS (2) a new view on transition probability as information quantity. Our model may be improved by incorporating more different ISs and by investigating various methods of measuring the amount of information in the ISs.

We plan to provide a number of extensions to our model in the future. First, our models were trained using a limited number of documents and terms. Most classifiers built on other models are trained with a much larger data set, ranging between 10,000 and 15,000 documents (Dumais et al., 1998; Joachims, 1998; Lewis, 1992). To overcome the overfitting problem with a small training data set, a K-fold cross-validation technique can be used (Breiman et al., 1993, pp. 12; Mitchell, 1997, pp. 112) to test the validity and reliability of the system. Second, we can explore other methods such as Mutual Information and Information Entropy for measuring the amount of information in the ISs.

Second, although this experiment produced similar results for both models (with and without transition probability), further research is needed to verify these results. Finally, the TC-based HMM should be compared to other proposed classification models (Sebastiani, 2002), under various experimental conditions and with different data sets to determine the most effective and efficient method of automatic classification of digital information.

ACKNOWLEDGMENTS

We would like to thank Online Computer Library Center, INC (OCLC) for allowing us to use data from OCLC Cat CD database for this research under cooperative agreement. The financial support of the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

REFERENCES

Baker, L. Douglas & McCallum, Andrew K. (1998). Distributed clustering of words for text categorization. *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 96-103.

Baum, L. E. & Egon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions for a Markov process and to a model for ecology. *Bull. Amer. Meteorol. Soc.*, 73, 360-363.

Breiman, Leo, et al. (1993). *Classification and regression trees*. Chapman & Hall: New York.

Conroy, John M. & O'Leary, Dianne P. (2001). Text summarization via hidden Markov models. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 406-407.

Crawford, S.L. et al. (1991). Classification trees for information retrieval. The 8th International Workshop on Machine Learning. Evanston, IL. Northwestern University. 245-249.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39, 1-38.

Dolin, R., Agrawal, D. & El Abbadi, A. (1999). Scalable collection summarization and selection. *Proceedings of the 4th ACM International Conference on Digital Libraries*, 49-58.

Dumais, S. T., Platt, J., Heckerman, David, & Sahami, Mehran. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management*, 148-155.

Elke, M. & Schäuble, Peter. (1994). Document and passage retrieval based on hidden Markov models. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318-327.

Finni, John J. & Pauson, Peter J. (1987). The dewey decimal classification enters the computer ages: developing the DDC Database™ and editorial support system. *International Cataloguing*, 16, (4), 46-48.

Frasconi, P., Soda, Giovanni & Vullo, Alessandro. (2002). Hidden Markov models for text categorization in multi-page documents. *Journal of Intelligence Information Systems*, 18, (2), 195-217.

Guenther, Rebecca S. (1992). The development and implementation of the USMARC format for classification data. *Information Technology and Libraries*, 11, (2), 120-131.

Harter, Stephen P. (1986). *Online Information Retrieval.* Orlando, FL: Academic Press.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning,* 137-142.

Koller, Daphne & Sahami, Mehran. (1997). Hierarchically classifying documents using very few words. *Proceeding of the 14th International Conference on Machine Learning*, 170-178.

Larson, Ray R. (1992). Experiments in automatic library of congress classification. *Journal of the American Society for Information Science*, 43, (2), 130-148.

Leek, T. R. (1997). *Information extraction using hidden Markov models*. Master thesis, University of California at San Diego, CA.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 37-50.

Lewis, D. D. & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *Proceeding of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, 81-83.

Markey, K. (1985). Subject-searching experiences and needs of online catalogue users: Implications for library classification. *Library Resources and Technical Services*, 29, 34-51.

McCallum, Andrew, Freitag, Dayne, & Pereira, Fernando. (2000). Maximum entropy Markov models for information extraction and segmentation. *Proceedings of the 17th International Conference on Machine Learning*, 591-598.

McGrath, William E. (1978). Relationship between hard/soft, pure/applied, and life/nonlife disciplines and subject bookuUse in a university library. *Information Processing and Management*, 14, (1), 17-28.

Miller, D. R. H., Leek, T. & Schwartz, R. M. (1999). A hidden Markov model information retrieval system. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 214-221.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw Hill: New York.

Ng, H.T., Goh, W.B., & Low, K.L. (1997). Feature selection, perception learning, and a usability case study for text categorization. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 67-73.

PQDD. *Proquest Digital Dissertations database*. Accessed February 2003, from UMI ProQuest Web site through McGill University catalogue : http://wwwlib.umi.com/dissertations/.

Ponte, Jay M. & Croft, W. Bruce. (1998). A language modeling approach to information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275-281.

Rabiner, Lawrence R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, (2), 257-286.

Roberson, S.E. et al. Okapi at TREC-3. (1995). *Proceedings of the 3rd Text Retrieval Conference*, 109-126.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24, (5), 513-523.

Shafer, Keith. (1997). *A brief introduction to Scorpion*. Accessed January 10, 2001, from Online Computer Library Center, Inc. Web site: http://orc.rsch.oclc.org:6109/bintro.html.

Shafer, K, Subramanian, S., & Fausey, J. (1997). *Measures for Evaluating Automatic Subject Assignment of Electronic Resources*. Accessed October 15, 000, from Online Computer Library Center, Inc. Web site: http://orc.rsch.oclc.org:6109/measures.html.

Scorpion. (1997). Scorpion Project. Accessed March 31, 2003, from Online Computer Library Center Inc. Web site: http://orc.rsch.oclc.org:6109.

Sebastiani, Fabrizio. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34, (1), 1-47.

Svenonius, E. (1983). Use of classification in online retrieval. *Library Resources and Technical Services*, 27, 76-80.

Seymore, Kristie, McCallum, Andrew, & Rosenfeld, Ronald. (1999). Learning hidden Markov model structure for information extraction. *Proceedings of the 16th National Conference on Artificial Intelligence: Workshop on Machine Learning for Information* Extraction, 37-42.