HMM-based Passage Models for Document Classification and Ranking

Ludovic Denoyer
LIP6, University of Paris 6
Case 169, 4 Place Jussieu, F - 75252
Paris cedex 05, France. denoyer@poleia.lip6.fr

Hugo Zaragoza
Microsoft Research
1 Guildhall Street
Cambridge CB2 3NH, U.K. hugoz@microsoft.com

Patrick Gallinari
LIP6, University of Paris 6
Case 169, 4 Place Jussieu, F - 75252
Paris cedex 05, France. Patrick.Gallinari@lip6.fr

16th March 2001

Abstract

We present an application of Hidden Markov Models to supervised document classification and ranking. We consider a family of models that take into account the fact that relevant documents may contain irrelevant passages; the originality of the model is that it does not explicitly segment documents but rather considers all possible segmentations in its final score. This model generalizes the multinomial Naive Bayes and it is derived from a more general model for different access tasks.

The model is evaluated on the REUTERS test collection and compared to the multinomial Naive Bayes model. It is shown to be more robust with respect to the training set size and to improve the performance both for ranking and classification, specially for classes with few training examples.

1 Introduction

Statistical sequence models have been used in a variety of settings in the field of natural language processing such as part of speech tagging, language , translation, etc. More recently, it has been proposed to use such models for handling different information access (IA) tasks ranging from document classification to summarization and information extraction (IE). These models can be used to capture word dependencies and to perform inference on sequences, be they whole documents or short passages. They allow to extend the classical paradigms of information retrieval (IR) by considering sequences of text elements instead of the classical bag-of-words representation. This opens the way for new models and applications which go beyond traditional IR.

The need for extended IR relevance paradigms and models arises partly due to the evolution of document collections and user needs. An increasing amount of textual resources comes from the Internet, a source which lacks the homogeneity (in length, content, formats, style, etc.) of corpora for which traditional IR models were developed. This is for example the case for Web pages, web directories, FAQs, HOWTOs, e-mails, newsgroup discussions, etc. These document sources are very heterogeneous, and documents are loosely structured, differently formatted, ungrammatical, etc. Documents may contain many different sections relevant to different topics, or irrelevant information such as headings, signatures, etc. Users, on the other hand, wish to complete more complex tasks with this information; new tasks such as filtering, detection of associations, document summarization and information extraction need to be addressed simultaneously to document retrieval.

In an earlier work, we introduced a general probabilistic model for IA based on sequence models and we showed its potential for handling a variety of tasks [25] [1] ranging from document retrieval to information extraction. The

proposed model was defined using two levels of relevance within a document: *term labels*, associated to words, and *document classes*, associated to entire documents. A term label reflects the relevance of that term with regard to a specific topic, whereas a document class defines the category to which the document belongs. This distinction permits to build quite complex document classes. Suppose that terms may be labeled with respect to semantic classes, as for example in the case of information extraction tasks. For example, an e-mail containing a conference announcement may be judged relevant because: i) it presents a conference on a topic we are interested in, ii) it gives the date and the location of the conference, iii) it includes the program. While it may be impossible to learn the combination of all these classes separately, one may consider these sub-topics as term labels and use this information to construct a document class.

In this paper we are investigating the benefits of using such sequence models for performing classical document classification and ranking with respect to a category. For these tasks all the available knowledge is a set of documents labeled relevant or irrelevant for a particular information need. These documents are labeled *as a whole*, even though not all parts of relevant documents are necessarily relevant. We therefore consider only two possible term labels: relevant and irrelevant for a category, and we will consider that within a document there may be relevant and irrelevant segments.

The question is then to i) estimate the *probability of relevance of terms* for any category and ii) to define the *probability of relevance of documents with respect to the relevance of its terms*. The implicit assumption is that the relevance of a document depends solely on the relevance of its smaller units and the type of information need; the relevance of the smaller units may depend on a wide range of factors such as their distribution, their context, etc. Problem (i) is similar to the estimation of weights in probabilistic approaches. Problem (ii) is open and its solution should allow for different *a priori* knowledge about the nature of the documents and the task. Going from term labels to document classes we perform a form of soft on-line segmentation. *Soft* because many alternative segmentations are considered in parallel and weighted in a probabilistic manner. *On-line* because the segmentation process is bound to the computation of the retrieval function and is done *a posteriori* of preprocessing, indexing and model training.

2 Related work

Key concepts for our work are the development of sequence models for text processing, text segmentation and retrieval with regard to passages, probabilistic IR models. We will review below related work for these different aspects.

Sequence models, mainly HMMs have been recently proposed for handling different tasks in IR or IE. [15] have been the first to propose a generative model based on HMMs for document retrieval and highlighting. Generative methods estimate the conditional distribution $P(x|t)$ using a model $P(x|t,\theta)$ where θ represents the model parameters to be estimated. Recently [12] applied a similar approach - each document is modeled by a HMM - for the ad-hoc task on the large document collections of TREC6 and TREC7. In the field of IE, HMMs have been used for named entity extraction by [3], their experiments show that simple ergodic models where each state is associated with a topic reach surprisingly good performances. [11] uses HMMs for extracting information in the form of simple binary relations between two entities on a limited domain in biology. This model has been carefully hand-crafted for this task. [6] consider the extraction of document entities such as document title, abstract, citations etc,. They build a discrete HMM to extract this information where each state represents a particular label. This tool has been used for document indexing in search engines. [22] use recurrent neural networks - which are another formalism for sequence modeling - for routing.

The use of passages or other document parts for retrieving whole documents has been advocated by several authors. Among the early work on this subject, [20] have shown that using sentences helps to determine the relevance of documents and [23] proposes to combine the score of document sections for computing the document score and performs tests on the second TREC collection (TREC 2). [4] introduces different types of passages and discusses their role for long documents retrieval, he performed tests on the Tipster collection [7]. The recent paper by [10] provides a thorough discussion and evaluation of passage retrieval, they performed tests on the TREC 5 collection . Note that all these works focus on ad-hoc retrieval and rely on the adaptation of classical IR document ranking techniques for taking into account text passages instead of whole documents. Text segmentation has also been considered from an IE perspective. [21] uses decision trees for the extraction of keywords or key-phrases (two or more words) from text,

framing the extraction problem as a classification problem on pre-segmented text. A large amount of work has also been dedicated to text segmentation into coherent passages, e.g. [8][17][2].

The models introduced in this paper can be considered as extensions of multinomial Naive Bayes classifiers which have often been used by the machine learning community for text classification [13]. Closed query tasks like e.g. document ranking and routing have been extensively studied by the IR community. A classical benchmark for that is the Reuters collection [19]. Recently, different machine learning techniques have been tested on these problems. Some papers like [9], [5], [24] compare different techniques on the Reuters collection.

3 Terminology and Notation

Let us define a *document* d as a *sequence of words*. Let $\mathcal{D}$ be the set of documents considered. We assume that there is some unknown process generating documents $d \in \mathcal{D}$. Consider now that there are some sets of documents $\mathcal{R}_r \subset \mathcal{D}$ in which we are interested (r indexes these sets). We will say that a *document is relevant to* $\mathcal{R}_l$ if it belongs to the set $\mathcal{R}_l$. Because we deal here with each set in an independent manner, we can drop the index r and consider only one set at a time. We will refer to this as the relevant set (R) and to its complement as the irrelevant set (I).

A mapping function must be used to map a document into some simplified representation. Let us denote $\mathbf{d}$ the document representation of d. Different documents belonging to different relevance sets can be mapped to the same representation; therefore, one cannot determine the relevance of a document in a deterministic manner. We will denote $P(R|\mathbf{d})$ the *probability that* $\mathbf{d}$ *represents a document which belongs to the set* R.

In this paper we will consider that documents are preprocessed and translated into a *sequence of terms* (by terms we mean some normalized form of words). Let L be the set of all terms (the *lexicon*). We will represent a document d as the sequence $\mathbf{d} = (w_1, w_2, \ldots, w_{|d|})$ where $w_i \in L$ is the term corresponding to the ith word in the document d. We will not consider a bag-of-words representation of documents here as our models will make use of the ordering of the words in a document.

4 Proposed Model

The model proposed here relies on the observation that many documents have a sequential structure which could be exploited for IR. Some documents have a generic structure, this is the case for example of journal or conference papers which are composed of a title, abstract, introduction, etc. In documents composed from distinct parts, the distribution of relevant terms may be different for each part, or one might like to weight parts differently according to their position. Many documents also carry small parts of relevant information in the middle of irrelevant information.

A priori information of this type could be easily taken into account with sequence models and encoded via deterministic or stochastic grammars.

In (Fig.1 top), the successive parts of a document are modeled with different successive states, i.e. the different parts are supposed to be generated successively through these different states. During training, term statistics for the different states are evaluated on the different parts of the documents. This is possible if we know the structure of the documents. Since the relevance will be computed as a summation over authorized paths, the sequence model structure encodes the a priori knowledge on the sequential structure of the document.

A simpler case is illustrated in (Fig.1, bottom). Here the document is supposed to contain blocks of relevant information separated by passages of non relevant information. From a generative viewpoint, the user who writes such a document might have in mind different items, and among them the one we are interested in, he then writes about them in different passages, he might come back to a previous item and so on. Each time he writes about a particular item, his mind switches to a specific vocabulary or term distribution. Since we are interested here into computing the relevance with regard to a particular item, the distribution of terms relevant for this item will be modeled as a particular state -denoted s_R- in the figure, while the distribution for all other items will be modeled by a common -garbage- state denoted s_I. Note that for each of these examples, more sophisticated models could be easily developed along the same line when it is needed and when the a priori information is available.

In the following, we will build on the idea of computing the relevance of a document by focusing on relevant passages, and show different ways to do that. Given our task, we will consider only two types of passages: relevant

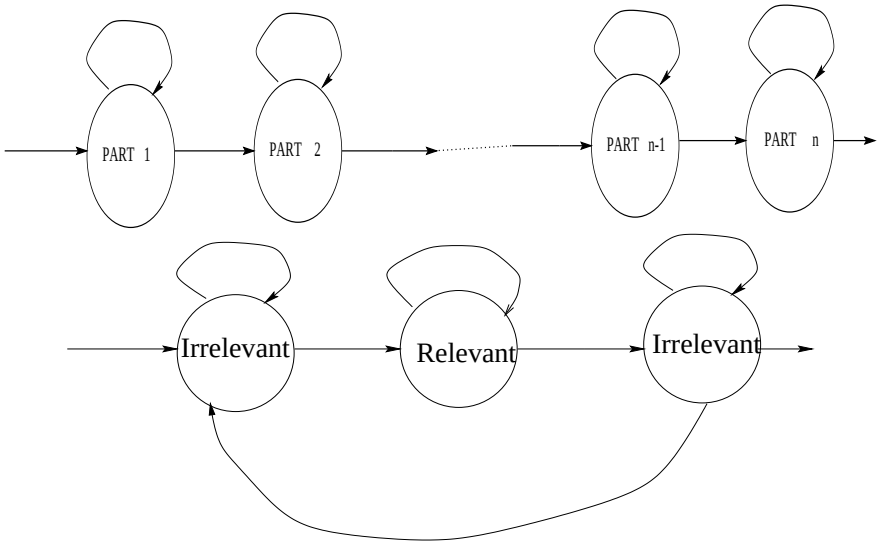

Figure 1: Two sequence models for information retrieval.

and irrelevant. We will consider the binary classification case, where documents themselves are either Relevant or Irrelevant.

Generative Model

We first consider that documents belonging to the each of the classes relevant (R) and irrelevant (I) are generated by two different generative models θ_R and θ_I respectively. Given a new document, its probability of belonging to the relevance set R can be written as:

$$P(R|d) \quad = \quad P(d|R)\frac{P(R)}{P(d)} \tag{1}$$

Now, modeling $P(d|R)$ and $P(d|I)$ by $P(d|\theta_R)$ and $P(d|\theta_I)$ respectively, we may write:

$$P(R|d) = \frac{P(\mathbf{d}|\theta_R)P(R)}{P(\mathbf{d}|\theta_R)P(R) + P(\mathbf{d}|\theta_I)P(I)} = \left(1 + \frac{P(\mathbf{d}|\theta_I)}{P(\mathbf{d}|\theta_R)}\frac{P(I)}{P(R)}\right)^{-1} \tag{2}$$

Alternatively, the usual *log-odds score* can be written as:

$$log\frac{P(R|d)}{P(I|d)} = log\left(\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)}\frac{P(R)}{P(I)}\right) \tag{3}$$

In what follows we will concentrate on the estimation of $\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)}$ using a generative model for the sequence of terms in the document. $\frac{P(R)}{P(I)}$ can be estimated accurately by maximum likelihood, dividing the number of relevant documents by the number of irrelevant ones.

Hidden Markov Models

We will develop a particular family of generative sequence models, first order Hidden Markov Models (HMMs) [18]. HMMs have a set of states $\mathcal{S} = \{\mathtt{s}_1, \ldots, \mathtt{s}_S\}$ which emit symbols with different probabilities. In our case each state is supposed to emit a set of words, and the whole model generates the sequence of words that constitutes a document. Recall that a document is represented by the sequence of its terms $\mathbf{d} = (w_1, w_2, \ldots, w_{|\mathbf{d}|})$ where w_i is the term associated to the ith word in document d. We will note s_i the state generating w_i and $\mathbf{s}$=$(s_1, s_2, \ldots, s_{|d|}), s_i \in \mathcal{S}$

a sequence of states corresponding to document d. A HMM is governed by the probabilities of emission of all the words in the lexicon for each of the given states, $P(w_i|s_j), j \in [1..S], w_i \in L$, and the probabilities of state transition $P(s_{t+1}|s_t)$, as well as the initial state probabilities $P(s_j|s_0)$ and exit probabilities $P(s_E|s_{|\mathbf{d}|})$. s_0 and s_E are non-emitting states which do not generate words. Such an HMM can generate documents with different lengths. The probability of a particular observed sequence $\mathbf{d}$ being produced by a HMM (call it θ) can be computed as follows:

$$P(\mathbf{d}|\theta) = \sum_{\mathbf{s}} P(\mathbf{d}|\theta, \mathbf{s}) P(\mathbf{s}|\theta) \tag{4}$$

$$= \sum_{\mathbf{s}} \left(P(s_E|s_{|\mathbf{d}|}, \theta) \prod_{t=1}^{|\mathbf{d}|} P(s_t|s_{t-1}, \theta) P(w_t|s_t, \theta) \right) \tag{5}$$

where the sum is over all possible state sequences s which produce $\mathbf{d}$ with nonzero probability under the model θ. To simplify equations, we will omit the conditioning on the model.

In the simplest case, only one state is used to generate documents (this model can be seen as a multinomial model as will be seen later in this paper). We will use such a model for the irrelevant documents since there is usually no a priori structure on a these documents. Let us denote $\rho_\emptyset$ the probability $P(s_\emptyset|s_\emptyset)$. Because this model can only stay in the state (and generate a word) or exit, we denote $(1 - \rho_\emptyset)$ the exit probability. Then we have:

$$P(\mathbf{d}|\theta_I) = (1 - \rho_\emptyset)\rho_\emptyset^{(|\mathbf{d}|-1)} \prod_{t=1,\ldots,|d|} P(w_t|\mathtt{s}_\emptyset) \tag{6}$$

For any given HMM model for relevant documents, we have from (4) and (6) :

$$\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)} \propto \left(\rho_\emptyset^{-|\mathbf{d}|}\right) \sum_{\mathbf{s}} \left(P(s_E|s_{|\mathbf{d}|}) \prod_{t=1}^{|\mathbf{d}|} \frac{P(w_t|s_t)}{P(w_t|s_\emptyset)} P(s_t|s_{s-1}) \right) \tag{7}$$

where the proportionality symbol $\propto$ indicates equality up to a constant term independent of the document d. The complexity of this computation, i.e the cost of classifying a document, is of the order $|\mathbf{d}|S^2$ where S is the number of states in the relevant model. Let us now look into some special cases of this model.

4.1 Irrelevant-Relevant-Irrelevant (IRI) models

Let us constrain the relevant model θ_R to two states: $S = \{\mathtt{s}_I, \mathtt{s}_R\}$, which we call relevant and irrelevant states. Furthermore, let us constrain the possible state sequences to be of the form $< \mathtt{s}_I+, \mathtt{s}_R+, \mathtt{s}_I+ >$, where $+$ indicates one or more repetitions of the preceding symbol. That is, we consider that relevant documents contain a single relevant segment of arbitrary length, similarly to **[15].** Note that the first and last word of every document is always labeled irrelevant by this model. This may seem strange but has a negligeable effect in practice and simplifies the equations presented in this paper. The model can be easily modified to allow irrelevant passages of size 0 by allowing transitions between the initial and final states and the relevant state. We will denote $\mathbf{s}_{IRI}$ such a sequence of states.

Let T_R^s be the number of states $\mathtt{s}_R$ in the sequence s. Denote: $\rho_R = P(\mathtt{s}_R|\mathtt{s}_R)$, and $\rho_{RI} = P(\mathtt{s}_I|s_0)P(\mathtt{s}_R|\mathtt{s}_I)P(\mathtt{s}_I|\mathtt{s}_R)P(s_E|\mathtt{s}_I)$. We can now rewrite equation (4) as:

$$P(\mathbf{d}|\theta_R) = \rho_{RI} \sum_{\mathbf{s}_{IRI}} \left(\rho_R^{(T_R^s-1)} \rho_I^{(|\mathbf{d}|-T_R^s-2)} \left[\prod_{t|s_t=\mathtt{s}_R} P(w_t|\mathtt{s}_R) \right] \left[\prod_{t|s_t=\mathtt{s}_I} P(w_t|\mathtt{s}_I) \right] \right) \tag{8}$$

Now, we further make the assumption that terms in irrelevant sections of relevant documents follow the same distribution as terms in irrelevant documents, that is $P(w|\mathtt{s}_I) = P(w|s_\emptyset)$, which is a reasonable approximation and greatly simplifies computations. Finally, we can rewrite (7) as:

$$\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)} \propto \rho_\emptyset^{(-|\mathbf{d}|)} \sum_{\mathbf{s}_{IRI}} \left[\rho_R^{(T_R^s-1)} \rho_I^{(|\mathbf{d}|-T_R^s-2)} \prod_{t|s_t=\mathtt{s}_R} \frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)} \right] \tag{9}$$

Note that this new expression depends only on the *ratio of the conditional probabilities of terms within the relevant sections*, as well as the length of the sequence and on the length of the relevant passage.

This score takes into account the probability of all possible document segmentations, many of which have a probability close to 0. Indeed, if the segmentation considers an irrelevant passage as relevant, the probability of each word w_t from this passage $P(w_t|\mathtt{s}_R)$ will be lower than $P(w_t|\mathtt{s}_I)$, so the quotient $\frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)}$ will be less than 1 and the product $\prod_{t|s_t=\mathtt{s}_R} \frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)}$ will be very close to 0. The score of such a segmentation does not contribute much to the final score. On the other hand, segmentations where relevant labels fall on truly relevant regions of the document will account for the main part of the final score.

4.1.1 Fixed sized window

If we further constrain the model to have relevant sections of a fixed size $T_R^s = K$ for all documents, the previous equation simplifies to:

$$\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)} \propto \rho_\emptyset^{(-|\mathbf{d}|)} \rho_I^{|\mathbf{d}|-K-2} \sum_{i=1}^{(|\mathbf{d}|-K+1)} \left[\prod_{t=i}^{(i+K-1)} \left(\frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)} \right) \right] \tag{10}$$

In this case the dependence on constants T_R^s, and ρ_R disappears, and all that one needs to estimate are the emission probabilities of the model. Furthermore, note the computational complexity of this model is now reduced to $2K|\mathbf{d}|$.

4.1.2 Transition probabilities

The values of $\rho_\emptyset$, ρ_R and ρ_I are related to the average length of the relevant and irrelevant sections and documents. For the remaining of this paper we will consider these three terms equal. This is in general not true (in general $\rho_\emptyset > \rho_R > \rho_I$) and should lead to a decrease in performance if the generative model chosen was accurate. In fact, the generative model chosen implicitly assumes an exponential model for document length, because terms such as $\rho_\emptyset^{|\mathbf{d}|}$ and $\rho_I^{|\mathbf{d}|}$, and this is not an appropriate model for document length. By setting $\rho_\emptyset = \rho_R = \rho_I$ we will see that the previous models become independent of these constants.

For the general IRI model in $\rho_\emptyset = \rho_R = \rho_I$ we obtain:

$$\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)} \propto \sum_{\mathbf{s}_{IRI}} \prod_{t|s_t=\mathtt{s}_R} \frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)}$$

and for the fixed sized window model :

$$\frac{P(\mathbf{d}|\theta_R)}{P(\mathbf{d}|\theta_I)} \propto \sum_{i=1}^{(|\mathbf{d}|-K+1)} \left[\prod_{t=i}^{(i+K-1)} \left(\frac{P(w_t|\mathtt{s}_R)}{P(w_t|\mathtt{s}_I)} \right) \right]$$

Not that these two quantities depend now only on sums of ratios between relevant and irrelevant word probabilities, which can be easily computed from a training corpus.

4.1.3 Parameter estimation

Word probability emissions are estimated as usual in the multinomial Naive Bayes model by the following smoothed maximum likelihood estimator:

$$P(w_l|C) = \frac{occ(w_l, C) + 1}{\sum_{m=1}^{|L|} occ(w_m, C) + |L|} \tag{11}$$

where L is the lexicon and $|L|$ its size. C is any set of documents, in our case Relevant and Irrelevant documents. Note that this is actually an approximation in our model, since relevant documents are not assumed to be completely relevant. In fact we should use the EM algorithm to appropriately estimate the emission probabilities $P(w_l|\mathtt{s}_R)$, but this is computationally much more expensive, and $P(w_l|\mathtt{s}_R)$ proved to be a reasonable approximation.

5 Relationship to the Multinomial Model

The multinomial model (also referred to as Naive-Bayes) has recently gained attention in machine learning due to its successful application to a wide range of tasks. This model is theoretically very simple and sound and yields good results in general. It has been extended and applied with success to new and difficult IR problems such as multi-class document classification and learning with unlabelled data [16]. We will briefly present the model and show that the model presented in this paper is indeed a generalization of this model.

Consider the same document representation used so far in the paper $\mathbf{d} = (w_1, w_2, \ldots, w_{|d|})$ where $w_i \in L$ is the term corresponding to the ith word in document d. Consider furthermore that there is a one-to-one mapping from documents to classes and consider that each class is generated from a single word probability distribution (a *component* of the resulting generative model). Consider furthermore that document length is evenly distributed across classes so that $P(|\mathbf{d}||R) = P(|\mathbf{d}|)$. Then, the probability of a relevant document can be written down as:

$$P_{NB}(\mathbf{d}|R) = \frac{P(|\mathbf{d}|)}{Z} \prod_{t=1}^{|\mathbf{d}|} P(w_t|R)$$

where Z is a constant independent of the class. The probability of relevance of a document is then :

$$P_{NB}(R|\mathbf{d}) = \left(1 + \prod_{t=1}^{|\mathbf{d}|} \left(\frac{P(w_t|I)}{P(w_t|R)} \frac{P(I)}{P(R)}\right)\right)^{-1}$$

We can see that this model is similar to the model presented in section (4), where both θ_I and θ_R are fixed to be monogram models and transition probabilities $\rho_\emptyset$, ρ_R and ρ_I are considered equal. In this sense, our model extends the Naive Bayes model with more sophisticated generative models for the relevant class.

Note that this is quite different from the multi-class extension of the multinomial Naive Bayes model presented in [14], in which it is considered that irrelevant documents are generated from a mixture of classes; this model continues to operate in a bag-of-word representation of documents and as such cannot take into account the ordering of words or the notion of relevance of *passages.*

6 Results

6.1 Experimental Setting

We used the Reuters-21578 text categorization test collection for evaluation purposes [19]. This corpus is very heterogeneous, and this is one of the reasons why it was chosen. Some classes are very large and some very small (ranging from more than 2700 documents to zero). Some classes are quite ambiguous while others depend basically on the presence or absence of one or two key words. Some sets of classes partly overlap (meaning the same documents belong to several classes), others form partitions of larger classes, and some classes are disjoint from the rest. Finally, certain classes deal with very specific topics containing documents that use very technical language while others cover broader areas and make use of much more general terms and expressions.

We used the ModApte split of the Reuters-21578 test collection. We eliminated all the classes that had zero training or test documents. This leads to a set of 90 classes and 12902 documents (this is the same set used in evaluations such as [24][9][5]).

Choosing an adequate evaluation measure is difficult. Because documents can belong to several classes in Reuters, we evaluate our models using break-even points which is independent of the classification decision. Break-even point is defined as the point where precision equals recall in a precision-recall curve. This point is linearly interpolated from the two closest points when necessary. *Macro-average break-even points* are obtained by averaging break-even points for every class considered. *Micro-average break-even points* are obtained by weighting the average by the relative size of each class. Because in the Reuters corpus most documents belong to a small set of classes, the difference between micro and macro averages tells us how models' performance degrades on small classes.

Model	MultiN		R1	R3	R5	R10	R20
earn	94,6		89,4	93,1	93,6	93,7	86,5
acq	90,0		53,7	85,5	87,2	88,5	88,5
money-fx	63,9		37,2	61,0	64,7	69,0	70,0
grain	58,9		36,7	74,1	73,9	73,4	71,3
crude	67,0		55,7	79,5	78,4	77,2	76,6
trade	70,0		44,5	64,2	60,5	62,2	58,7
interest	58,7		29,7	62,1	65,3	68,7	67,7
ship	77,7		55,7	79,0	80,2	79	75
wheat	50,8		25,5	65,9	68,7	64,2	64,1
corn	36,5		28,8	65,7	61,3	53,8	46,1
micro-average	71,7		56,6	77	77,2	76,5	73,2
macro-average	28,9		27,2	48,9	46,1	40,9	38,1

Table 1: Micro and macro average break-even point and break-even points for the 10 largest classes. See text for notations.

Finally we report the probability of good classification averaged over the documents (micro-PGC) and averaged over the classes (macro-PGC). This is the number of correctly classified relevant documents when the highest scoring class is attributed to the document. When a document belongs to n classes the n highest scoring classes are attributed to the document; since this would not be possible on real data, one must regard this PGC as an upper bound or an *optimistic* approximation which can be reach with a *perfect* classification decision. This measure is interesting because some models, such as the multinomial Naive Bayes model, are designed to classify documents with respect to a set of classes, and are not good at ranking documents with respect to a single class. The micro and macro PGC measures give us an insight on the performance of systems for the task of document classification. Again, large classes mostly dominate the micro average result and small classes the macro average.

6.2 Evaluation

We will evaluate fixed-window IRI models of varying window size (1, 3, 5, 10 and 20 words) as well as the baseline multinomial model. We refer to these models as R*n*, where *n* is the particular size of the window used. For all models, data is preprocessed in the following manner: words are stemmed with the Porter algorithm and a stop-list of 350 words is used to eliminate the most common empty words. Finally, words appearing less than 3 times in the (training) corpus are eliminated. No other feature selection is carried out for any of the models.

Table 1 presents the break-even point micro and macro averages as well as break-even points for several classes (for illustration purposes). First, let us note that the baseline multinomial model yields reasonably high micro-average performances for this task (71%), but a very low macro-average (28.9%). Best result reported for this task as far as we know is the linear kernel SVM models (micro: 86 %, macro not given) [9]. This implies that break-even points for small classes are very poor.

Results for the HMM models vary depending on the window size, but are better than the baseline for all except the R1 model. The R3 model is 5% greater in micro average and 20% higher in macro average than the baseline model. The R5 model performs similarly. Performance decreases as the window size is increased, both in micro and macro averages. We see that the greater gain in performance is clearly over macro-averages, indicating that the model is outperforming the baseline model specially in small classes.

Several conclusions can be drawn from this. Firstly, the HMM are more robust than the multinomial model with respect to the size of the training corpus. The basic problem with small classes is the difficulty of estimating term relevance. This is usually resolved by applying feature selection techniques; however, these rely on heuristics and are difficult to apply in a consistent manner. The good performance of the HMMs on small classes indicates that

	micro-PGC	macro-PGC
Multinomial	78	27
R1	84	47
R3	83	45
R5	82	40

Table 2: Micro and macro PGC comparison

feature selection is less critical for these models. Secondly, small window sizes (e.g. 3,5) seem to perform better than larger ones (e.g. 10, 20). This may seem counterintuitive but in fact denotes that even in very relevant documents many words are in fact more probably irrelevant than relevant. Because a document score is made up of all possible segmentations, small window sizes seem to be an advantage for this model. Nevertheless, trivially small window sizes (e.g. 1) lead to performances lower than the baseline model (table 2).

Looking at micro and macro PGC results we find that again HMM models perform consistently better than the baseline model. Even the R1 model outperforms the baseline model under this performance measure; in fact, it is the model with highest micro and macro PGC, 6% and 20% higher than the baseline model respectively. This model is obviously a good candidate for document classification but does not perform well for document ranking. This means that this model produces higher precision values than the rest for certain recall values, but that precision-recall curves it produces deteriorate very quickly before the break-even point. This also indicates that this model is a good candidate for low-recall applications. Macro-PGC deteriorates fast as the window size is increased (45% for R3 and 40% for R5) while micro average is quite stable (83% for R3 and 82% for R5). This supports our previous claim that small classes require small window-sizes.

7 Conclusion

We have presented a new family of generative models for information retrieval based on probabilistic sequence models. We motivated such models in the light of increasing complexity of textual data and tasks, showing how they could extend the classical notion of document retrieval. We then developed a HMM implementation for the task of document ranking and classification, and detailed several possible derivations stating explicitly their assumptions. We also showed that the proposed models generalizes the classic multinomial Naive Bayes models, taking into account the existence of non-relevant passages within relevant documents.

We evaluated these models on the Reuters data set and compared them to a baseline multinomial Naive Bayes model. We could show a systematic improvement of performances over the baseline model. Furthermore, by the use of several performance measures we could gain some insights on the weaknesses and strengths of the proposed models.

References

[1] Amini M.R., Zaragoza H.,Gallinari P. Learning for Sequence Extraction Tasks. Content-Based Multimedia Information Access, RIAO'2000

[2] Beeferman D., Berger A., Lafferty J. Statistical Models for Text Segmentation. Machine Learning 1999; 34:177

[3] Bikel D. M., Schwartz R., Weischedel R. M. An algorithm that Learns What's in a Name. Machine Learning 1999; 34:211-231

[4] Callan J. Characteristics of text. 1997

[5] Dumais S. , Platt J., Heckerman D., Sahami M. Inductive Learning Algorithms and Representations for Text Categorization, In Proceedings of ACM-CIKM98, 1998, 148-155.

[6] Freitag D., McCallum A. Information extraction with HMMs and shrinkage, Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.

[7] Harman D. The DARPA Tipster Project, SIGIR Forum, 1992; 26(2):26-28.

[8] Hearst M. A., Plaunt C. Subtopic structuring for full document access, SIGIR93, 1993.

[9] Joachims T. Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998, pp. 137–142, Chemnitz, Germany.

[10] Kaszkiel M., Zobel J., Davis Sacks-R. Efficient passage ranking for document databases, ACM Trans. on information systems 1999, 17(4):406-439.

[11] Leek T. R. Information Extraction using Hidden Markov Models, Master thesis, University of California, San Diego, 1997

[12] Miller D., Leek R.H., Schwartz T. R. M. BBN at TREC7: Using hidden Markov Models for Information retrieval. Proceedings of TREC7, 1999

[13] McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification, AAAI-98 Workshop Learning for text categorization, 1998.

[14] McCallum A., Nigam K., Thrun S., Mitchell T. Text Classification from Labeled and Unlabeled Documents using EM, Machine Learning, 39(2):103-134, 2000

[15] Mittendorf E., Schauble P. Document Passage Retrieval Based on Hidden Markov Models, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'94), 1994, 318:327

[16] Nigam K., McCallum A., Thrun S., Mitchell T. Text classication from labeled and unlabeled documents using EM. Machine Learning, 39:103-134, 2000.

[17] Ponte J. M., Croft W. B. Text segmentation by topic. Proceedings of the First European Conference on Research and Advanced Technology for Digitial Libraries. 120-129, 1997

[18] Rabiner L. R. A Tutorial on Hidden Markov Models and Seleceted Applications in Speech Recognition, Proceedings of the IEEE, 1989, 77(2):257-285

[19] Reuters Collection , http://www.research.att.com/ lewis/reuters21578.html

[20] Salton G., McGill M. J. Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983

[21] Turney P. D., Learning Algorithms for Keyphrase Extraction, Information Retrieval, 2(4):303-336, 2000

[22] Wermter S., Arevian G., Panchev C. Recurrent neural networks for text routing, International Conference on Neural Networks,Internation Conference of Artificial Neural Networks (ICANN'99), 898-903, 1999

[23] Wilkinson R. Effective Retrieval of structured documents,Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'94) , 1994

[24] Yang Y. An evaluation of statistical approaches to text categorization, Information Retrieval, 1(2):69-90

[25] Zaragoza H., Modèles Dynamiques d'Apprentissage Numérique pour l'Accès à l'Information Textuelle, PhD thesis, University of Paris 6, 1999