SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON
MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 4, ISSUE 1
ISSN 2054-7390

An Information Reinstatement Dealing with Machine Learning

Firoj Parwej[1] and Hani Alquhayz[2]

*[1]Department of Computer Science, College of Science, Al-Zulfi, Majmaah University,
Majmaah, Kingdom of Saudi Arabia (KSA)*
[1]f.hussain@mu.edu.sa; [2]h.alquhayz@mu.edu.sa

ABSTRACT

Information retrieval using probabilistic techniques has attracted significant attention on the part of researchers in information and computer science over the past few decades. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. Machine learning programs detect patterns in data and adjust program actions accordingly. In this paper, we are exploring the use of machine learning techniques for information retrieval and we are using machine learning algorithms that can benefit from limited training data in order to identify a ranker likely to achieve high retrieval performance over unseen documents and queries. This problem presents novel challenges compared to traditional learning tasks, such as regression or classification. We are investigating the discriminative learning of ad-hoc retrieval models. For that purpose, we propose different models based on kernel machines or neural networks adapted to different retrieval contexts. The proposed approaches rely on different online learning algorithms that allow efficient learning over large collection and finally approaches rely on discriminative learning and enjoy efficient training procedures, which yields effective and scalable models.

Keywords: Soft Computing, Machine Learning, Information Retrieval (IR), Gaussian Mixture Model (GMM), Unsupervised learning, Supervised learning.

1 Introduction

The history of Information Retrieval (IR) parallels the development of libraries. The first civilizations had already come to the conclusions that efficient techniques should be designed to fully benefit from large document archives.

As early as 5,000 years ago, the Sumerian librarians were already describing and categorizing official transaction records, legends and theological documents in indexes. Information retrieval (IR) systems were [1] originally developed to help manage the huge scientific literature that has developed since the 1940s. Today, numerous university, corporate, and public libraries now use information retrieval systems to provide access to books, journals, and other documents.

The first automated information retrieval systems were introduced in the 1950s and 1960s. By 1970 several different techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s. In 1992, the US Department of Defense along with the National

DOI: 10.14738/tmlai.41.1691
Publication Date: 6th January, 2016
URL: http://dx.doi.org/10.14738/tmlai.41.1691

Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program [2]. The objective of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even ahead. Information Retrieval has primordially changed with the advent of computers. Digital technologies give a unified infrastructure to store, exchange and automatically process big document collections. The search for information consequently developed from the manual examination of brief document abstracts within [3] predefined categories to algorithms searching through the whole content of each archived document. Nowadays, automatic retrieval systems are widely used in several application domains (e.g. web search, book search or video search) and there is a constant need for improving such systems. The main tasks of information retrieval, the so-called ad-hoc retrieval task which target at finding the documents episodic to submit queries.

Machine Learning proposes and studies algorithms that allow computer systems to automatically improve through experience, i.e. from training data. Learning systems are commonly used for various perception tasks, [1] such as automatic face detection or automatic speech recognition. In this paper firstly our task corresponds to a ranking problem, which insinuates that the performance for a given query cannot be formalized as a sum of a measure of performance estimated for each corpus document. Secondly, most retrieval queries, current a highly disorganized setup, with a set of relevant documents, accounting only for a very small fraction of the corpus [4]. Thirdly, ad-hoc retrieval corresponds to a kind of dual generalization problem, since the learned model should not only confrontation new documents but also new queries. Finally, our task also presents stimulating efficiency compellable, since ad-hoc retrieval is typically applied to enormous corpora.

2 The Machine Learning

Machine Learning is the field of study that intends to give the computers or the machines the ability to learn from data without being explicitly programmed and act according to this information learnt [5]. It is considered a branch within Artificial Intelligence, which reckon with the idea of giving human-like intelligence to software-defined machines. Machine learning also helps us find solutions too many problems in vision, speech recognition, and robotics. Machine learning is programming computers to optimize a performance criterion using example data or past experience [6]. We have a model defined up for some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions on the forthcoming, or descriptive to gain knowledge from data, or both. The range of applications of Machine Learning is very comprehensive including.

- Prediction, where some current variable outputs are computed according to prior acquired results of the same variables. One area where prediction is applied is to make weather forecasts.
- Classification, were given an input point the algorithm tries to classify it into one of the groups of the model previously defined. It might be used to build classifiers for spam or fraud detection for instance.

- Pattern recognition, with applications in many fields such as text recognition (OCR), object recognition in computer vision, medicine (ECG measurements), face recognition, etc.
- Recommender platforms, which intend to predict the preference of a user with respect to a certain product, based on latest purchases or liked the items. They have become extremely famous recently, with the exponential evolution of Internet advertising companies or Amazon-like websites.
- Data mining, where the goal is to extract valuable information (patterns) out of large data sets. It is a widely used term to refer to many applications, with search engines and customer data extraction being two of the most famous.

There are several ways in which machine learning can be organized. The machine learning techniques classify techniques and algorithms based on whether the desired output is known. If this is the case, we call the problem a supervised learning problem. In this case, we possess a number of training examples, for which we know the desired output. If the desired output is not known, the problem is an [7] unsupervised learning problem. In this case, we ask the machine to optimize some function, but we do not really know what the output should be.

2.1 Supervised learning

Supervised learning problems are typical problems that humans can solve or know the answer to, but don't really know how to solve. Examples may be speech recognition or weather prediction. These techniques are often classified according to their purpose. We can discriminate between two great families of supervised machine learning techniques based on this distinction. The classification techniques occasionally, the purpose of a machine is to discriminate between a finite number of classes. For example, we may want an intelligent car to be able to discriminate between pedestrians and road signs, or a quality control robot to discriminate between correctly manufactured items and manufacturing failures, etc. This is a broad class of problems, and many different machine learning techniques focused only on these. Next regression techniques in other settings, the purpose of a machine will be to output continuous values. For example, we may want an intelligent car to be able to control the steering in such a way as to stay on the road, or to control speed so as to bring us to our destination in [8] a speedy, comfortable and safe way, or we may want a machine to be able to predict the amount of rainfall for tomorrow. The problem of supervised learning can be approached from different angles, which again leads to a classification of methods for non-parametric methods as we mentioned, it is not feasible to store all the measurement values that might ever occur, and to store the desired answer for that measurement value. However, we do have some example values with corresponding label, we could store those and perform classification or regression on new measurements based on how similar those measurements are to the stored values. These methods are called non-parametric methods.

In discriminant functions another possible approach is to select a function that will provide an output of the required format for a given input of the measurement's format. For example, in a two-class classification problem, we may choose a function that returns f(x) = 0 for a measurement x in class B1 and f(x) = 1 for a measurement in class B2. Learning then consists of finding values for the parameters of this function, so that it performs correctly on the largest possible number of examples. The model-based approaches Instead of finding a function that will optimistically provide us with the right output given some input and we could also look at what the inputs look like for each possible output.

We build a model of the data, and use it to perform the task at hand. We need to assume that the data's distribution can be delineated by a probability density function of certain family and we then need to estimate the parameters of that distribution. When we know the distribution of the data for each possible class, we can compute the probability that a new data point should be seen if it belonged to any of the classes, and based on that we can compute the probability that a data point belongs [9] to any given class. The massive mileage of this method is that it is possible to not only provide the most likely classification, but also a assuredness estimate of that classification.

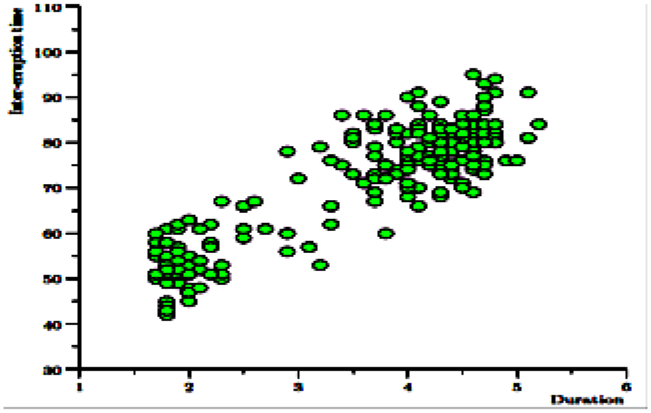

Figure 1. The some measurements that might be used for clustering. Humans will readily detect two clusters in this data. But how did we do that? Why don't we see three clusters in there?

2.2 Unsupervised learning

The unsupervised learning techniques are given the measurements, but no information as to what we expect to obtain from this data. Such techniques can be divided into two broad categories, firstly clustering and secondly dimensionality reduction.

In clustering deals with the problem of trying to group data by recognizing some structure in the data, when no corresponding labels are known. This is something that humans excel at, but that is very hard to evaluate objectively. Finding the correct number of clusters is an especially challenging task. Clustering can be extremely utilized as a form of data compression. It is often not necessary to know what the measurement was in order to perform a given task knowing which cluster it belonged to is often enough. For example, consider the task of predicting a plane's speed from its size. It is perhaps enough to know that a plane's size falls in the tourism, fighter jet or commercial transport category in order to be able to make the prediction. The precise size measurements are not occasional.

In dimensionality reduction it is over and over again possible to make many measurements simultaneously, crummy very high-dimensional data. As an example, consider a digital camera. Such a camera may have a few million sensors, each registering how much light fell on them during a given time span. Every picture taken with such a camera may therefore be seen as a very high-dimensional vector of measurements. We cannot list all the possible images that such a sensor could take, because there are far too many possibilities even though this number is finite. We do know that actual images will have characteristics that limit the number of possible valid images. For example, most pairs of neighboring pixels in an actual image will have very similar color and intensity. We can therefore detract the dimensionality of the data point without renouncing any occasional information. These automated techniques to do this are called dimensionality reduction techniques.

3 The Information Retrieval

An information retrieval system is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. The form of an object in an information retrieval system is diverse the text aspect has been the only data type that lent itself to fully functional processing. The other data types have been treated as highly informative sources, but are primarily linked for retrieval based upon a search of the text.

An information retrieval system consists of a software program that facilitates a user in finding the information the user needs [10]. The system may use standard computer hardware or specialized hardware to support the search sub function and to transform non-textual sources to a searchable media (e.g., transcription of audio to text). The gauge of the success of an information system is how well it can minimize the overhead for a user to find the needed information. From a user’s perspective is the time required to find the information needed, excluding the time for actually reading the relevant data. Hence search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead. The general objective of an information retrieval system is to minimize the overhead of a user in locating needed information. The favorable outcome of an information system is very subjective, based upon what information is needed and the willingness of a user to accept overhead.

The total information storage and retrieval system is collected of four major functional processes: Item Normalization, Selective Dissemination of Information (i.e., “Mail”), archival Document Database Search, and an Index Database Search along with the automatic file build process that supports index files. The commercial systems have not integrated these capabilities into a single system, but supply them as [11] independent capabilities. In figure 2 shows the logical view of these capabilities in a single integrated information retrieval system. The boxes are used in the diagram to represent functions while disks describe data storage.

3.1 Item Normalization

The first step in any integrated system is to normalize the incoming items to a criterion format. In addition to translating multiple external formats that might be received into a single consistent data structure that can be manipulated by the functional processes, item normalization provides logical restructuring of the item. Additional operations during item normalization are needed to create a searchable data structure identification of processing tokens (e.g., words), [12] characterization of the tokens, and stemming (e.g., removing word endings) of the tokens.

3.2 Selective Dissemination of Information

The selective dissemination of information process provides the potential to dynamically differentiation newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the item. The Mail process is collected in the search process, user statements of interest and user mail files. The every item is received, it is processed against every user’s profile. A profile contains a typically broad search statement along with a list of the user mail files that will receive the document if the search statement in the profile is satisfied. The user search profiles are different than ad hoc queries in that they contain notably more search terms and cover a wider range of interests.

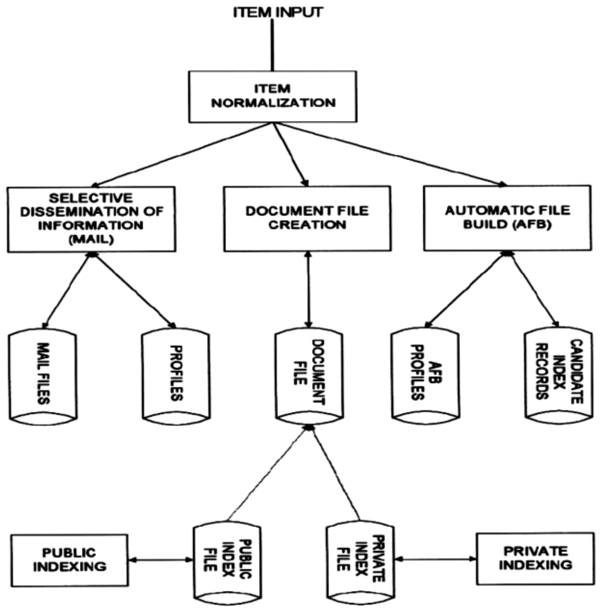

Figure 2. The Total Information Retrieval System

3.3 Document Database Search

The document database search process provides the impressibility for a query to search against all items received by the system. The document database search process is composed of the search process, the user entered queries and the document database which contains all items that have been received, processed and stored by the system. It is the retrospective search source for the system. If the user is on-line, the selective dissemination of information system delivers to the user items of interest as soon as they are processed into the system. Whichever search for information that has already been processed into the system can be considered a retrospective search for information. This does not prevent the search to have search statements constraining it to items received in the last few hours.

3.4 Index Database Search

When an item is determined to be of interest, a user may want to save it for future reference. This is in effect filing it. In an information system this is versed via the index process. In this process the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item [13]. It is also possible to have index records that do not reference an item, but contain all the substantive information in the index itself.

3.5 Multimedia Database Search

From a system perspective, the multimedia data is not logically its own data structure, but an augmentation to the existing structures in the information retrieval system. It will consist almost entirely in the area described as the document database [14]. The specialized indexes allow search of the multi-media (e.g., vectors representing video and still images, text created by audio transcription) will be augmented search structures. The original source will be kept as a normalized digital real source for access possibly in their own specialized retrieval servers.

4 The related work

The core objective of keyword spotting is to discriminate between the segments of the signal belonging to a keyword utterance and the others. The first approaches based on dynamic time warping (DTW) proposed to compute the alignment distance between a template utterance of the keyword and all possible subsequences of the test signal [15]. The keyword is considered as detected for the subsequences for which the distance is below some predefined threshold. Such approaches are, however greatly affected by speaker mismatch and varying recording conditions between the template sequence and the test signal. The discrete HMMs were introduced to ASR [16], and then for keyword spotting [17]. A discrete HMM assumes that the observations of a sequence of discrete events are independent conditioned on a hidden state variable that follows a Markov process. This type of model introduces several advantages compared to dynamic time warping based approaches, including an improved robustness to speak and channel changes, when several training utterances of the targeted keyword are available.

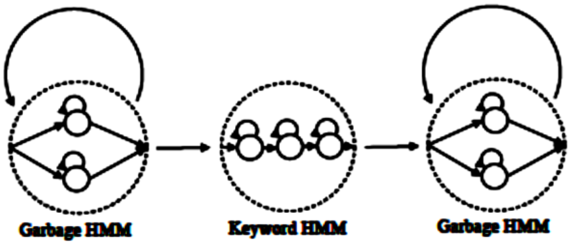

Figure 3. The HMM topology for keyword spotting with a likelihood of the sequence given the keyword is uttered

The HMMs remove the need for acoustic vector quantization, as the distributions associated with the HMM states are continuous densities, generally Gaussian Mixtures. The learning of both the Gaussian Mixture parameters and the state transition probabilities is performed in a single integrated framework, maximizing the likelihood of the training data given its transcription through the Expectation-Maximization algorithm [18]. It is now the most widely used approach for both ASR and keyword spotting.

To gain some robustness, likelihood ratio approaches have been proposed [19]. In this case, the confidence score outputted by the keyword spotter corresponds to the likelihood ratio estimated by an HMM requiring an occurrence of the keyword, and an HMM excluding it, see figure 3 and 4. The performed by comparing the outputted score to a predefined threshold. The fewer studies have proposed discriminative parameter training approaches to circumvent this weakness [19]. The maximize the likelihood ratio between the keyword and garbage models for keyword utterances and to minimize it over a set of false alarms generated by a first keyword spotter. [20] We are proposing to apply minimum classification error (MCE) to the keyword spotting problem. Other discriminative approaches have focused on combining different HMM-based keyword detectors. For instance, trains a neural network to combine likelihood ratios from different model [21].

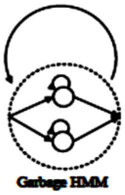

Figure 4. The HMM topology for keyword spotting with a likelihood of the sequence given the keyword is not uttered

The support vector machines to combine different averages of phone-level likelihoods. Both of these approaches propose to minimize the error rate, which equally weights the two possible spotting errors, false positive and false negative in other words missing a keyword occurrence, often called keyword deletion. This measure is however barely used to evaluate keyword spotters [22], due to the unbalanced nature of the problem. The maximization of the AUC would hence be an appropriate learning objective for the discriminative training of a keyword spotter.

5 The Information Retrieval Using Machine Learning

We focus on the learning of ranking functions for information retrieval systems. This means that we are interested in identifying a ranking function f from a set of training data, such that its expected performance on a new ranking problem is high. The two main types of learning approaches have been applied in this context, supervised and unsupervised learning. In the case of supervised learning, the training data consists of both documents and queries along with the corresponding relevance assessments [23]. This means that the learning procedure should generalize to a new ranking problem, while having access to the desired output on the training data. In the unsupervised case, the training data simply consist in a set of documents, without queries and relevance assessments. As the learning procedure has no access to examples of the desired strategy [24], it should discover some hidden structure of the data, from which it is possible to identify an effective ranking function.

In latent semantic analysis (LSA) aims at modeling term correlation [25], to overcome the term mismatch problem. For instance, one of LSA's goals is to assign a high RSV to a document which does not use any query term, but only related terms or synonyms. For that purpose, LSA assumes that the vocabulary-sized vectors actually originate from a lower dimensional space (k < T, the vocabulary-size), to which orthogonal noise has been added. Given a set of n training documents, represented as a matrix

$$D = [d_1, \dots, d_n] \in \mathbb{R}^{T \times n},$$

LSA solves the least square problem,

$$D^k = \mathrm{argmin}_{X:\mathrm{rank}(X)=k} \|D - X\|_2^2.$$

The replaces D with $D^k = [d^k_1, \dots, d^k_n]$ as the denoised representation of documents. The substitution of D with Dk actually projects each document to a k dimensional subspace, and LSA hence assumes that the term mismatch problem can be solved through linear projection. The probabilistic latent semantic analysis, PLSA [26], proposes a probabilistic interpretation of the notion of topics in text documents to address the term mismatch problem. The documents can be decomposed as a mixture of aspects, where each aspect defines a multinomial over the vocabulary terms. In this model, documents and terms are considered as the observation of two discrete random variables D and T. The occurrence of a term t in a document d corresponds to the observation of the pair (t, d), which is modeled by the joint probability.

$$P(t,d) = \sum_{i} P(z_i)P(t|z_i)P(d|z_i),$$

Where the discrete random variable Z, of values $z_1, \ldots, z_k$, is called the aspect variable. The term variable T is conditionally independent from the document variable D, given the aspect variable Z. The parameters of the model, i.e. $P(z_i)$, $P(t|z_i)$, $P(d|z_i)$ for all aspects z_i, all vocabulary terms t and all corpus documents d are learned to maximize the likelihood of the pairs (t, d) occurring in the training corpus, relying on the expectation maximization (EM) algorithm [26]. The pair classification formalizes the learning of ranking functions as a binary classification problem. Given a query q and a document d, the ranking function f should determine whether (q, d) is a positive pair, i.e. d is relevant to q, or a negative pair, i.e. d is not relevant to q. Inter-query discrimination refers to an intrinsic problem of the pair classification framework, which presents a more difficult problem to the classifier than the actual retrieval task. In this framework, the classifier should output a positive score f (q, d) > 0 for any positive pair (q, d) and a negative score f (q0, d0) < 0 for any negative pair (q0, d0).

6 The Proposed Information Reinstatement with Machine Learning

In the keyword spotting task, we are provided with a speech utterance along with a keyword k, and we should determine whether k is uttered in¯x. The keyword spotter f can be evaluated relying on the receiver operating curve (ROC). This curve plots the true positive rate (TPR) as a function of the false positive rate (FPR). The TPR measures the fraction of keyword occurrences correctly spotted, while the FPR measures the fraction of negative utterances yielding a false alarm. The points on the curve are obtained by sweeping the threshold b from the largest value outputted by the system to the smallest one. These values, hence correspond to different trade-offs between the two types of errors a keyword spotter can make, i.e. missing a keyword utterance or rising a false alarm. In order to evaluate a keyword spotter over various trade-offs, it is common to report the area under the ROC (AUC). The AUC can be written as,

Where | · | refers to set cardinality and ‖· refers to the indicator function. The A_k hence estimates the probability that the score assigned to a positive utterance is greater than the score assigned to a negative utterance. This quantity is also referred to as the Wilcox on Mann Whitney statistics. where w is a vector of importance weights, ϕ(¯x , ¯p^k , ¯s) is a feature vector, measuring different characteristics related to the confidence that ¯p^k is pronounced in ¯x$\overline{x}$ with the segmentation ¯s. In other words, our keyword spotter outputs a confidence score by maximizing a weighted sum of feature functions over all possible segmentations. This maximization corresponds to [27] a search over an exponentially large number of segmentations. Nevertheless, it can be performed efficiently by selecting decomposable feature functions, which allows the application of dynamic programming techniques, like for HMMs our keyword spotter f is parameterized as

$$f_{\mathbf{w}}(\bar{x},\bar{p}^{k}) = \max_{\bar{s}} \mathbf{w}\cdot\phi(\bar{x},\bar{p}^{k},\bar{s}),$$

Our objective now is to identify the vector w minimizing a regularized version of the loss $L(f_w)$ to avoid over fitting,

$$L^{\text{Reg}}(f_{\mathbf{w}}) = \|\mathbf{w}\|^2 + C \sum_{(k,\bar{x}^+,\bar{x}^-)\in T_{\text{train}}} \beta_k \; l(\mathbf{w};\bar{p}^k,\bar{x}^+,\bar{x}^-),$$

Where

$$l(\mathbf{w};\bar{p}^k,\bar{x}^+,\bar{x}^-) = \left|1 - \max_{\bar{s}} \mathbf{w}\cdot\phi(\bar{x}^+,\bar{p}^k,\bar{s}) + \max_{\bar{s}} \mathbf{w}\cdot\phi(\bar{x}^-,\bar{p}^k,\bar{s})\right|_+$$

The C is a hyper parameter setting the importance of the training loss versus the regularized. One can note that $L^{\text{Reg}}(f_w)$ is not a convex function of w. This section explains why the minimization of $L_{\text{Reg}}(f_w)$ corresponds to a large margin approach. $L^{\text{Reg}}(f_w)$ combines two terms, the regularized and the loss. The loss sums lc $(f_w;\bar{}p^k, \bar{}x+, \bar{}x-)$ over the training triplets $(\bar{}p^k, \bar{}x+, \bar{}x-)$ ε T_{train}. For each term, the lowest possible value $(f_w;\bar{}p^k, \bar{}x+, \bar{}x-)=0$ is reached when,

$$\mathbf{w}\cdot\phi(\bar{x}^+,\bar{p}^k,\bar{s}^+) - \max_{\bar{s}} \mathbf{w}\cdot\phi(\bar{x}^-,\bar{p}^k,\bar{s}) > 1,$$

Which is equivalent to

$$\forall \bar{s}, \quad \mathbf{w}\cdot\phi(\bar{x}^+,\bar{p}^k,\bar{s}^+) - \mathbf{w}\cdot\phi(\bar{x}^-,\bar{p}^k,\bar{s}) > 1.$$

These inequalities can be rewritten as,

$$\forall \bar{s}, \quad \mathbf{u}\cdot\phi(\bar{x}^+,\bar{p}^k,\bar{s}^+) - \mathbf{u}\cdot\phi(\bar{x}^-,\bar{p}^k,\bar{s}) > \frac{1}{\|\mathbf{w}\|},$$

7 The Experiment Result

We conducted two types of experiments to evaluate the proposed discriminative approach. First, we learned the parameters of our model over the training set of TIMIT dataset, and compared its performance against an HMM baseline over the test set of TIMIT dataset. The corpus provides manually aligned phoneme and word transcriptions for each utterance. It also provides a standard split into training and testing data. From the training part of the corpus, we extract three disjoint sets consisting of 1500, 300 and 200 utterances.

7.1.1 The TIMIT Experiments

The GMM (Gaussian Mixture Model) [28] corresponds to a Bayes classifier combining one GMM per class and the phoneme prior probabilities, both learned from the training data. In this case, the log posterior of a phone given the frame vector is used as the function g. We compare the results of both models against an HMM baseline, in which each phoneme is modeled with a left-right HMM of 5 emitting states. The density of each state is modeled with a 40-Gaussian GMM.

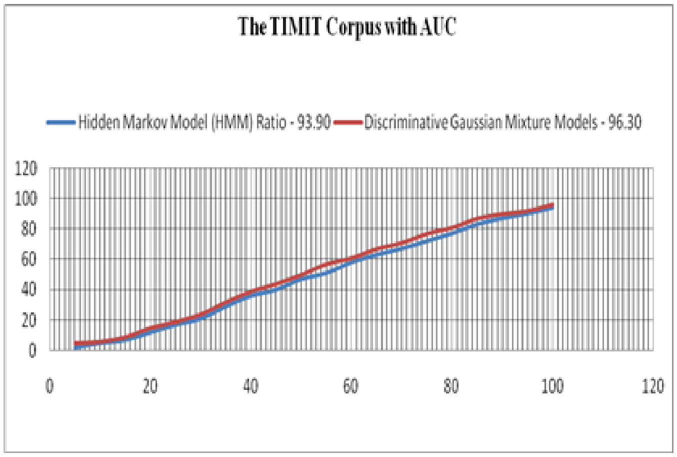

Figure 5. The TIMIT Corpus with AUC Using Discriminative Gaussian Mixture Model and HMM

Training is performed over the whole TIMIT training set. Embedded training is applied, i.e. after an initial training phase relying on the provided segmentation, a second training phase which dynamically determines the most likely segmentation is applied. The hyper parameters of this model are selected to maximize the likelihood of a held-out validation set. The evaluation of discriminative and HMM-based models is performed over 80 keywords, randomly selected among the words occurring in the test set of TIMIT. This random sampling of the keyword set aims at evaluating the expected performance over any keyword. The table 1 reports the AUC results, averaged over the 80-word test set, for the evaluated models. These results show the advantage of our approach. The two HMM based solutions are outperformed by the keyword spotters relying on our discriminative learning approach. The improvement introduced by our discriminative training algorithm can be observed when comparing the performance of the Discriminative Gaussian Mixture Model to the performance of the HMM spotters.

Table 1. The Area under the Curve (AUC) over the TIMIT Corpus

Model	The TIMIT Corpus with AUC
Hidden Markov Model Ratio	93.90%
Discriminative Gaussian Mixture Model	96.30%

8 Conclusion

Today, scenario information retrieval (IR) is a multidisciplinary field. The Humans retrieve information every time they ask a question and receive a response that addresses their question and adds to their knowledge about the queried topic. Information requests vary widely in their complexity and in the quantity of potentially relevant material that can be retrieved, as well as in the effort required to retrieve satisfactory information. Today, much of our business and cultural information is being recorded on electronic media and stored in multiple electronic databases. In order to make this information available to an information seeker, there must be an electronic information retrieval system that facilitates location and retrieval of documents that are relevant to the information seeker's question. Since information can be recorded on various media types, such as tables, images, text, audio and video, the retrieval system must be able to retrieve information from varying media representations. This work proposed a learning algorithm, which aims at maximizing the AUC over a set of training spotting problems. Our strategy is based on a large margin formulation of the task, and relies on an efficient iterative training procedure. The resulting model contrasts with standard approaches based on Hidden Markov Models (HMMs), for which the training procedure does not rely on a loss directly related to the spotting task.

Compared to such alternatives, our model is shown to yield significant improvements over various spotting problems on the TIMIT and the WSJ corpus. For instance, the best HMM configuration over TIMIT reaches 93.90% AUC, compared to 96.30% for the best Discriminative Gaussian Mixture Model spotter.

REFERENCES

[1] Frakes, William B. (1992). Information Retrieval Data Structures & Algorithms. Prentice-Hall, Inc. ISBN 0-13-463837-9.

[2] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35(12):29–38, 1992.

[3] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.

[4] STANFILL, C. (1990a). Information Retrieval Using Parallel Signature Files. IEEE Data Engineering Bulletin, 13 (1), 33-40.

[5] C. M. Bishop, “Pattern Recognition and Machine Learning (Information Science and Statistics),” Aug. 2006.

[6] Dr. Yusuf Perwej, (2015), “An Evaluation of Deep Learning Miniature Concerning in Soft Computing” , International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol 04, Issue 02, pp 10 – 16, 28, ISSN (Print) 2319-5940, ISSN (Online) 2278-1021, with Impact Factor = 2.117 DOI : 10.17148/IJARCCE.2015.4203

[7] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[8] C. Ji and S. Ma, “Performance and efficiency: Recent advances supervised in learning,” Proc. IEEE, vol. 87, pp. 1519–1535, Sept. 1999.

[9] S. Zribi Boujelbene, D. Ben Ayed Mezghani, and N. Ellouze, “Support Vector Machines approaches and its application to speaker identification”, IEEE International Conference on Digital Eco-Systems and Technologies DEST-09, Turkey, pp. 662-667, Jun 2009.

[10] Belkin, N. and W. Croft, “Retrieval Techniques”, in Annual Review of Information Science and Technology, Elsevier Science publishers, New York, 1989, pages 109-145.

[11] Card, K., “Visualizing Retrieved Information: A Survey”, IEEE Computer Graphics and Applications, Vol. 16, No. 2, March 1996, pages 63-67.

[12] Chalmers, M. and P. Chitson, “Bead: Explorations in Information Retrieval”, Proceedings of SIGIR 92, Copenhagen, Denmark, June 1992, pages 330-337.

[13] Crew, B. and M. Gunzburg, “Information Storage and Retrieval”, U.S. Patent 3, 358, 270, December 12, 1967.

[14] Leek, T., Miller, D. and R Schwartz, "A Hidden Markov Model Information retrieval Ssystem", In Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pages214-221.

[15] J. S. Bridle. An efficient elastic-template method for detecting given words in running speech. In British Acoustic Society Meeting, pages 1–4, London, UK, April 1973.

[16] L. R. Bahl, P. F. Brown, P. de Souza, and R. L. Mercer. ,“Maximum mutual information estimation of hidden markov model parameters for speech recognition” , In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 49–52, Tokyo, Japan, April 1986.

[17] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman. ,“Automatic recognition of keywords in unconstrained speech using hidden markov models”, IEEE Transactions on Acoustics, Speech and Signal Processing (TASSP), 38 (11):1870–1878, 1990.

[18] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, International Computer Science Institute, Berkeley, CA, USA, 1998.

[19] M. Weintraub. LVCSR log-likelihood ratio scoring for keyword spotting. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 297–300, Detroit, MI, USA, May 1995.

[20] E. D. Sandness and I. Lee Hetherington. Keyword-based discriminative training of acoustic models. In International Conference on Spoken Language Processing (ICSLP), volume 3, pages 135–138, Beijing, China, October 2000.

[21] Y. Benayed, D. Fohr, J. P. Haton, and G. Chollet. Confidence measures for keyword spotting using support vector machines. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 588–591, Hong Kong, China, April 2003.

[22] R. A. Sukkar, A. R. Seltur, M. G. Rahim, and C. H. Lee. Utterance verification of keyword strings using word-based minimum verification error training. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pages 518–521, Atlanta, GA, USA, May 1996.

[23] Dr. Yusuf Perwej, “The Bidirectional Long-Short-Term Memory Neural Network based Word Retrieval for Arabic Documents” Transactions on Machine Learning and Artificial Intelligence (TMLAI) which is published by Society for Science and Education, Manchester, United Kingdom (UK), Volume 03, No.01, Pages 16 – 27, 02 February 2015, ISSN 2054 - 7390, DOI : 10.14738/tmlai.31.863

[24] Forsyth, R. and R. Rada, "Adding an Edge", in Machine Learning: application in expert systems and information retrieval, Ellis Horwood Ltd., 1986, pages 198-212.

[25] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 6(41):391–407, 1990.

[26] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1):177–196, 2001.

[27] C. Cortes and M. Mohri. Confidence intervals for the area under the roc curve. In Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, December 2004.

[28] Reynolds, D.A., Rose, R.C.: Robust, “Text-Independent Speaker Identification using Gaussian Mixture Speaker Models”, IEEE Transactions on Acoustics, Speech, and Signal Processing 3(1) (1995