2006

A Hierarchical, HMMbased Accuracy for a Digital Library of Books

Shaolei Feng
University of Massachusetts - Amherst

A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books

Shaolei Feng and R. Manmatha*
Multimedia Indexing and Retrieval Group
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts,Amherst
[slfeng, manmatha]@cs.umass.edu

ABSTRACT

A number of projects are creating searchable digital libraries of printed books. These include the Million Book Project, the Google Book project and similar efforts from Yahoo and Microsoft. Content-based on line book retrieval usually requires first converting printed text into machine readable (e.g. ASCII) text using an optical character recognition (OCR) engine and then doing full text search on the results. Many of these books are old and there are a variety of processing steps that are required to create an end to end system. Changing any step (including the scanning process) can affect OCR performance and hence a good automatic statistical evaluation of OCR performance on book length material is needed. Evaluating OCR performance on the entire book is non-trivial. The only easily obtainable ground truth (the Gutenberg e-texts) must be automatically aligned with the OCR output over the entire length of a book. This may be viewed as equivalent to the problem of aligning two large (easily a million long) sequences. The problem is further complicated by OCR errors as well as the possibility of large chunks of missing material in one of the sequences. We propose a Hidden Markov Model (HMM) based hierarchical alignment algorithm to align OCR output and the ground truth for books. We believe this is the first work to automatically align a whole book without using any book structure information. The alignment process works by breaking up the problem of aligning two long sequences into the problem of aligning many smaller subsequences. This can be rapidly and effectively done. Experimental results show that our hierarchical alignment approach works very well even if OCR output has a high recognition error rate. Finally, we evaluate the performance of a commercial OCR engine over a large dataset of books based on the alignment results.

*This work was done while both authors were visiting Google.

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms, Documentation

Keywords

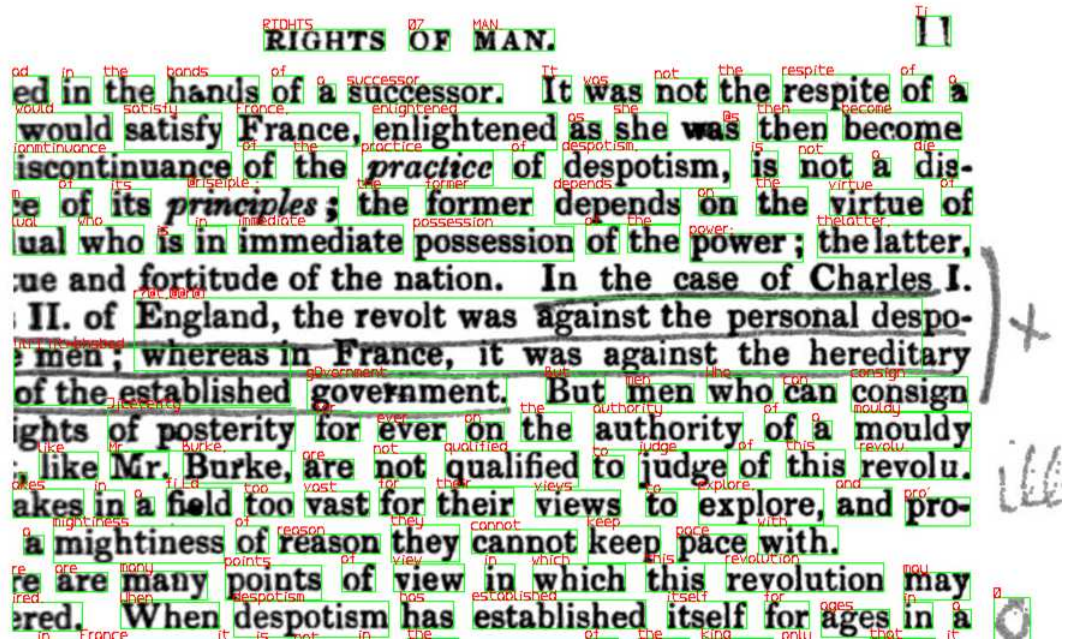
OCR Evaluation, Book Alignment, Digital Libraries

1. INTRODUCTION

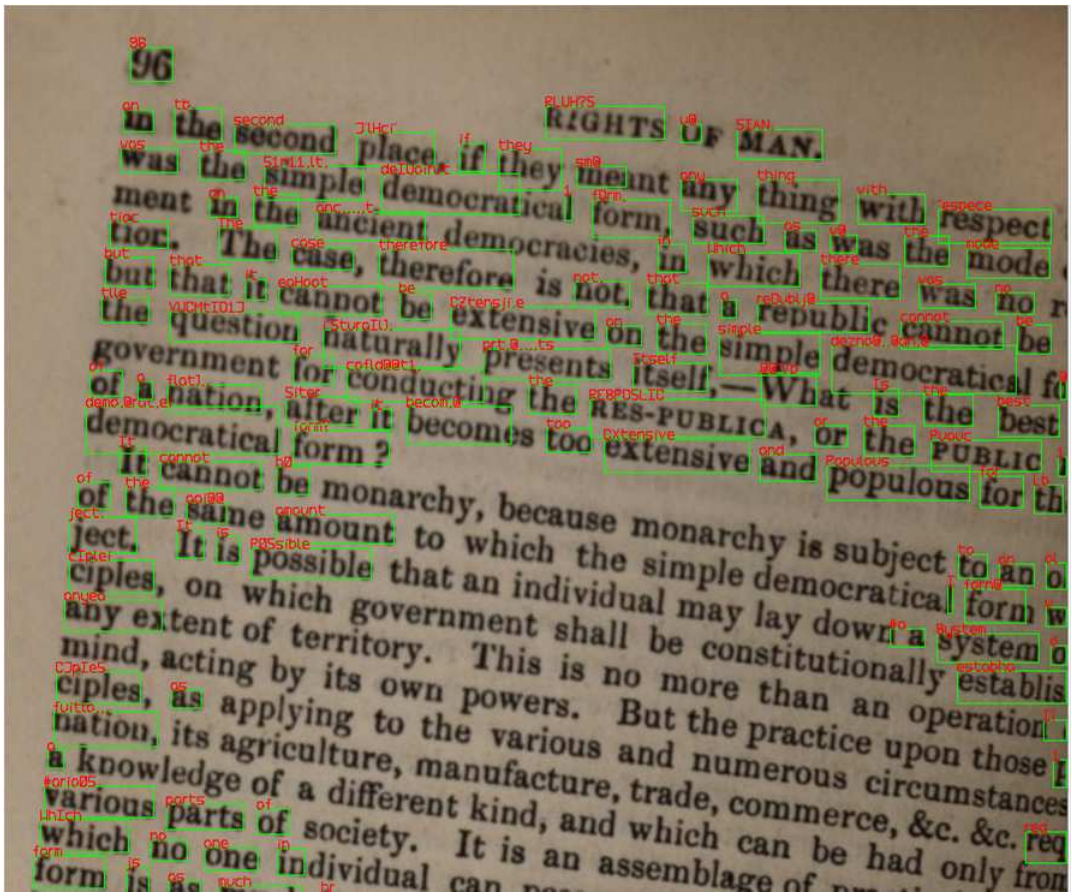
Efforts like Google Books, the Million Book Project and similar projects from Yahoo and Microsoft aim to provide searchable digital libraries of printed books. The aim of these digital libraries is to provide easy access to library material. The basic system involves rapidly scanning large amounts of printed books, processing the scanned images, converting the imaged text into ASCII using an optical character recognition system (OCR) and then indexing and retrieving the OCR output using a text retrieval system. Many of these books are old (out of copyright material) with a variety of different problems and a number of processing steps are required to before OCR can be run effectively on them. Such problems include noise, variable ink, bleedthrough, markings by users which cause erroneous OCR results, books with tight bindings so that the edge of the printed material is not scanned properly. In a large digital library, books have varied and sometimes complicated layouts introducing further errors. OCR engines are not very effective when the background [20, 2] is colored especially if the color is not uniform. This color may be inherent or may arise because the page has become colored and faded with age. The scanning process may itself introduce errors. To create large digital libraries with reasonable costs and in a reasonable amount of time requires rapid scanning which can cause blurred, cropped or skewed pages as well as missed or duplicated pages. Sometimes there may be 10 or 15 pages in sequence which may have been missed or duplicated.

To obtain the highest possible recognition accuracies in a large robust digital library, many processing steps must be carried out before the OCR engine is actually applied. A few examples include image rectification, cleanup, deskewing and deblurring. [1]. Any modifications in even a single

[1]While sometimes commercial OCR packages include these



(a)



(b)

Figure 1: Examples of OCR outputs on scanned images. The word printed on the top right corner of each rectangle is the OCR output for that word image

one of these processing steps may change the recognition results. It is not difficult to evaluate a processing step on its own. For example, how does one quantify the amount of blur left? Visual evaluation is not sufficient to always evaluate a processing step on its own since small amounts of blur hurt OCR performance although they don't seem to be visually significant. Thus, a processing step which appears to enhances a few selected pages when they are examined visually may actually hurt OCR performance. Usually, OCR engines are not trained on this material because of the difficulty of obtaining ground-truthed training material from books (more on that later). Thus, the performance of the OCR engine on this material is of interest and often different from the nominal numbers specified by the OCR engine maker. One may also be interested in determining whether an alternative OCR engine is better or whether an OCR engine trained differently would improve performance on this material. Therefore, *OCR evaluation on this material is a good proxy for determining how well the system and its different components including the OCR perform.* It is not possible to evaluate an OCR engine accurately by examining individual pages manually. Instead, one needs statistical results obtained automatically on a large number of pages, which requires automatically aligning every OCR output character with it's corresponding character in the ground truth.

There are a number of challenges to automatically evaluating the accuracy of the OCR over book length material. Ground truth is very difficult to come by and would be expensive to create for this purpose. OCR engines are usually evaluated by creating a page electronically, adding synthetically generated noise (see for example [8, 5]) and then evaluating the results. However, these noise models do not accurately reflect what happens when recognizing old books and this approach is, therefore, not a good idea in this case. Although authors have provided their manuscripts in electronic format for at least a decade, till recently many publishers (surprisingly) have discarded most of their electronic versions. Instead, publishers have often created electronic versions of books by scanning paper copies and embedding the scanned images in pdfs [2]. In any case such electronic texts are rarely available for out of copyright texts - the ones which cause the most problems for the system. The only easily available source we were able to identify were the Gutenberg texts [4] available on line. They are created by either typing the entire book or by first scanning, then recognizing the text using an OCR and finally manually proof reading and correcting mistakes. Thousands of electronic books are available on line in Gutenberg. While the Gutenberg books are freely available - since they are created from out of copyright texts - there are significant challenges in using them as ground truth for evaluating OCR output.

The Gutenberg texts do not preserve line or page breaks. Thus, the ground truth text and the OCR text need to be first aligned over the entire length of the book. This may be viewed as similar to the sequence alignment problem discussed in a number of fields, like genomic alignment in bioinformatics [13, 17], parallel corpus alignment in statistical machine translation [3] or aligning parallel corpora for machine translation [9], aligning synthesized speech with speech [12] and the alignment of speech recognition output with video captions [7].

In this paper we present a hierarchical Hidden Markov Model (HMM) [14] based-algorithm to align the ground truth text and the OCR text. We demonstrate using experiments that the approach can evaluate book length material rapidly and accurately. The technique is language independent (the program uses Unicode encoding). Besides evaluating the performance of the OCR and the stages prior to that, the algorithm has a number of other possible applications. For example it may be used to obtain training data for OCR for old books. The hierarchical HMM approach could also potentially be modified for use in aligning book length parallel corpora in different languages or for obtaining ground truth in handwritten data.

On the face of it, the book alignment problem seems straightforward but it is actually very challenging. OCR and scanning errors, long sequences of missing or duplicated pages make the alignment problem here challenging. Noise such as stain, marks in the original books also cause a lot of OCR errors. Figure 1 shows some recognition results of passing scanned book pages through one commercial OCR system, in which the OCR output for each word image is printed as a red word on its right-hand corner. We can see that because of the marks on the book by readers and the skewness of the scanned page, the OCR engine makes a lot of recognition errors. Although skewness is usually automatically corrected, for old books the algorithms sometimes fail. The Gutenberg ground truth text may have errors in it. In addition, the Gutenberg text may be of a different edition than the one scanned (in practice it turned out to be very difficult to determine edition information despite having bibliographic data from publishers and the library). Often the difference between the two editions consisted of an additional preface or introductory section. Bound books can cause printed material to be cropped at the edge. In the most extreme example encountered, almost every line in a sequence of 80 pages had one word cropped. Given that a book with 500 pages may contain more than 180,000 words or a million characters all of these problems make the sequence alignment problem challenging.

The hierarchical alignment is necessary since directly aligning an entire book not only computationally intensive (a book with 500 pages can contain more than 180K words and 1M characters) but is also prone to generating alignment errors. Theoretically, the number of possible alignments between two sequences is exponential in the length of the sequences. State purging techniques like beam search could help reduce the computation but impair the alignment precision a lot when directly aligning long sequences. Furthermore, when large chunks of books are missed or duplicated in the OCR output, directly aligning long sequences can mess up the whole alignment. To reduce the computation and make the alignment robust, we propose a hierarchical scheme for book alignment which divides the whole problem into a set of smaller alignment problems and also supports parallel computing. Our hierarchical alignment method basically works at three levels: at the top level, we first align anchor words (which are unique words in ground truth and OCR output after filtering.) over the whole book; at the second level, the contents between anchor words are aligned

steps they may not be accurate or consistent enough for a large digital library and one often needs to create new preprocessing algorithms

[2]The electronic material often went directly to the printer and was presumably discarded after printing

at word level; at the bottom level, the contents between exactly matched words are aligned character by character. The higher level alignment allows one to detect large chunks of books missed or duplicated in the OCR output so the whole alignment is more robust.

At each level, we use a HMM-based algorithm to align two text sequences. Compared with other alignment algorithms such as edit distance, the HMM-based alignment algorithm constrains the alignment based on both similarities between the two texts and also the likelihood of certain transitions occurring. That is, there is a generative probability which accounts for the similarity and there is a transition probability accounting for which characters are most likely to follow the current sequence. One of the challenges in using the HMM is that the model must be robust to rough estimates of the generative probabilities since the actual OCR confusion matrix is not available to us.

To verify our alignment algorithm, we establish a noise model to generate synthesized OCR documents from original documents, meanwhile recording the real alignment between them according to each operation. Then we align synthesized OCR documents with the original ones using our algorithm and compare the alignment results with the real alignment in order to evaluate our alignment algorithm. We then evaluate the performance of the algorithm on the OCR output of a large number of books and show that the average character and word error rates are 0.98 and 0.92 respectively.

The rest of this paper is laid out as follows. The next section discusses the previous work done on alignment and on OCR evaluation. Section 2 gives a detailed description of the hierarchical alignment and the HMM based alignment model. Section 3 describes a noise model for testing the alignment with synthetic data followed by experiments on synthetic and real data and the conclusion.

1.1 Related Work

Sequence alignment has been widely applied in various domains to study the similar and different properties of sequences from the same resource, for example, aligning protein sequences or DNA sequences in bioinformatics and aligning sentences from different languages in machine translation. Dynamic programming is the core of many sequence analysis methods, e.g. dynamic time warping, edit distance [19] and linear HMM [11]. Alshawi et al. [1] proposed an alignment algorithm to search pairings of words from bitexts (source language sentences with their translations) for machine translation, which makes use of dynamic programming to learn a mapping function minimizing the total costs of a set of pairings. Needleman-Wunsch algorithm [13] and Smith-Waterman algorithm [17] are well-known pairwise sequence alignment algorithms for protein and DNA alignments, both of which are extensions of edit distance with a predefined linear gap penalty and a similarity matrix to specify the scores for aligned characters. Hobby [6] created ground truth for OCR's by using a machine readable description to print the document and then matching character bounding boxes with bounding boxes derived from a scanned image of the document. Xu et al. [21] aligned an imperfect transcript obtained from a scanned image of a printed page with the characters in unsegmented text image. Neither of these are really appropriate since we do not have the approximate mapping that is required nor are we aligning images with text. HMM is a model widely used for alignment tasks in different domains, e.g. for sequence alignment in speech recognition [16], the alignment of synthesized speech with speech [12], machine translation [3], aligning parallel corpora in machine translation [9], the alignment of speech recognition output with captions in video [7].

Krogh et al. [11] proposed to use a linear HMM as a structure generating protein sequences by a random process. It is basically a hidden Markov chain with three kinds of state nodes: match, insert and delete, in which all transitions and character distance costs are position-dependent, i.e. different distributions are associated with the same kind of states or transitions at different positions.

Unlike Krogh's linear HMMs, the HMM at each level of our hierarchical alignment approach directly takes positions as states and calculates the probability of generating a sequence of OCR output given any possible sequence of positions in the ground truth. That is, there is a state corresponding to every position in the ground truth sequence. This structure is very similar to the HMM model proposed by J. Rothfield et al. [15] for word by word alignment of scanned handwritten document images with ASCII transcripts. This model is not hierarchical and is not practical for aligning large sequences. In this paper we seek to align text to text not text to images. Given the problem and domain differences, the transition probabilities and generative probabilities have to be and are defined differently. The details are given in section 2.1. For the same task of handwriting alignment, Kornfield employs dynamic time warping (DTW) [10] to align feature sequences extracted from word image series with ASCII transcripts, which is essentially an edit distance based global alignment method with deletion, insertion and match costs uniformly defined as the dissimilarity between corresponding items from two time series. Compared with edit distance based alignment algorithm, the HMM based alignment allows one to learn the domain knowledge through training over aligned or even unaligned sequences and formulate the probabilities of alignments using arbitrary distributions and is more flexible and powerful.

2. HIERARCHICAL ALIGNMENT

In this section, we describe the details of our hierarchical alignment scheme as well as the HMM-based alignment model for text sequences. The HMM-based alignment model doesn't explicitly deal with the case of extra OCR output (ie. OCR text not found in the ground truth). In the latter half of this section, we will discuss the behavior of our alignment model when encountering extra OCR text and introduce heuristic rules to deal with it.

The ground truth data does not have structural information for books(e.g. line or page information) available for hierarchical alignment. For example, the lines and pages in the Gutenberg text do not correspond to the lines and pages obtained from the scanned book. A straightforward hierarchical scheme is to follow the natural structure of a language, i.e. the sentence, word and character level. One can first align the OCR output with the ground truth sentence by sentence, then word by word, and finally character by character. An implementation of this approach revealed a number of problems. First, recognition errors on delimiters for sentences (a set of punctuation) make accurate determination of sentences difficult This leads to incorrect alignments at the higher (sentence) level which are difficult to

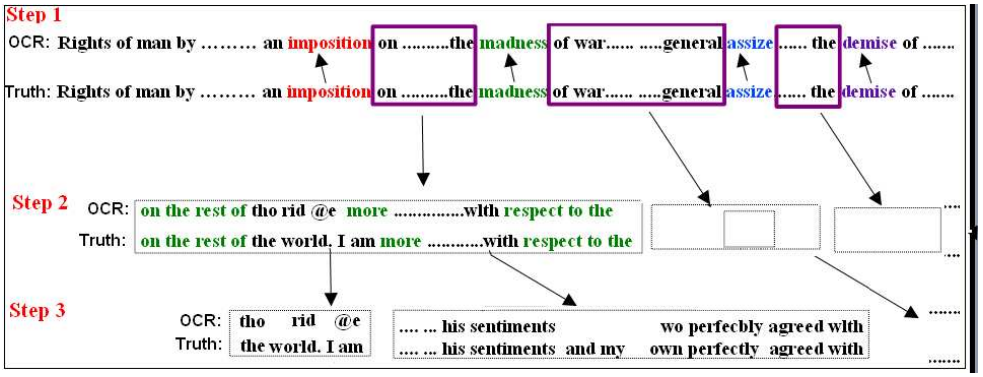


Figure 2: The diagram of our hierarchical alignment framework. Step 1: align anchor words over the whole book; Step 2: align text between anchors at word level; Step 3: align text between exactly matched words at character level.

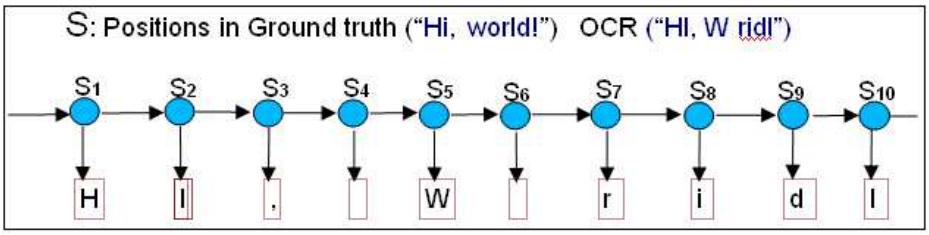


Figure 3: Illustration of the HMM-based Alignment Model

undo at the lower level alignment. Second, costs for similarity calculation between sentences are expensive considering the huge number of sentences ($\approx 30,000$) in a book. Reducing this cost by trying to find matching sentences doesn't always work. In at least one book which was bound very tightly, there so many OCR errors that it was difficult to find even a single pair of matching sentences in 80 pages. So instead of aligning sentences at the top level, we first align **anchor words** to partition a book into smaller portions for alignments at lower levels.

Figure 2 shows our hierarchical framework for book alignment. The alignment at the upper level aims at providing a rough alignment between two sequences on a larger scale and allows us to break up the original problem of aligning long sequences into the problem of aligning much shorter subsequences. These subsequences are aligned at a lower level. Given the ground truth and the OCR output for a book, the hierarchical approach works as follows:

1. At the top level, we look for and align a set of unique words in order to partition an entire book into small portions. It is done in 3 steps.
   (a) we first extract all the unique words in the ground truth, each of which occurs only once in the book, and create a word list **A** which is sorted according to the order that they appear in the book. According to the Zipf's law on the distribution of word frequencies in a natural language document, almost half of the distinct words are unique.
   (b) For each unique word in the ground truth, we look for the same words in the OCR output. Because of OCR errors and duplicate pages, it is possible that a unique word has no correspondence or more than one correspondence in the OCR output. We, therefore, filter out those words from the list **A**, which have no correspondences in the OCR output and whose immediate neighbors do not match. The words in the OCR text which correspond to the filtered outputs in A form a sorted word list **B** which is ordered according to their position in the OCR text.
   (c) Using our alignment model(section 2.1), we filter out those repeated correspondences caused by redundant texts in OCR output from list **B** and finally get a one to one mapping from the unique words in the ground truth to those in the OCR output. The unique words after filtering and alignment are called **anchor words**.
2. At the middle level, we use anchor words as boundaries to partition the OCR output and the ground truth of the whole book into smaller corresponding subsequences. Using our alignment model, we align each pair of subsequences at the word level.
3. After word alignment, exactly matched words are directly mapped to the character level. Using exactly matched words as boundaries, we align the texts between every pair of these boundaries at character level.

The first step in the hierarchical alignment framework is quiet robust. Even if there are large chunks of texts missed, reduplicated, or wrongly recognized, the anchor words can be correctly located and aligned. After these three steps of alignment, we finally get the character by character alignment between OCR output and ground truth . The next

subsection describe the details of our HMM-based alignment model at each level of the hierarchical framework.

2.1 HMM-based Alignment Model

Hidden Markov Models(HMMs) are widely applied to sequence data analysis. Here, we formulate the sequence alignment at each level of our hierarchical framework as an inference problem in a HMM. For the sake of convenience, we use the word "term" to denote the elements to be aligned in the sequences, which maybe words or characters according to whether we are performing word level alignment or character level alignment. Given two sequences, one of which is the OCR output and the other the ground truth, we try to find the position sequence traversed in the ground truth which has the highest probability of generating the OCR output. In this HMM-based alignment model, observations are OCR terms. The state space is defined as the positions of all the terms in the ground truth sequence. Let $G =< g_1, g_2, \ldots, g_m >$ represent the ground truth sequence, $O =< o_1, o_2, \ldots, o_n >$ the OCR output sequence, and $S =< s_1, s_2, \ldots, s_n >$ a hidden position sequence which is a series of indices of ground truth terms in charge of generating the OCR sequence. So each item in S is basically an integral index to a term in the ground truth and for $\forall s_i \in S$, $s_i \leq m$. For example, if $s_6 = 10$, that means the 6-th OCR output term o_6 is generated by the 10-th ground truth term g_{10}. Note n and m, i.e. the lengths of the OCR output sequence and the ground truth sequence, can be different. The HMM-based alignment model estimates the joint probability of the OCR sequence and the hidden position sequence $P(O, S)$ as:

$$P(O, S) = \prod_{i=1}^{n} P(s_i|s_{i-1})P(o_i|s_i) \tag{1}$$

where $P(s_i|s_{i-1})$ is the transition probability which indicates the possibility of transition from one position s_{i-1} to another s_i in the ground truth, and $P(o_i|s_i)$ the generative probability which indicates the possibility of generating the current OCR term o_i by the ground truth term at the hidden position s_i.

Inference in the HMM-based alignment model requires finding the $\tilde{S}$ maximizing $P(O, S)$, i.e.:

$$\tilde{S} = arg\max_{S} P(O, S) \tag{2}$$

In our alignment model the transition probability simulates the possibility of an OCR system skipping or repeating ground truth terms, which is defined as a distribution related to the number of skipped terms when jumping from position s_{i-1} to s_i in the ground truth. This distribution should be subject to these facts: OCR never traverses the ground truth backwards; OCR seldom repeats a ground truth term; The longer the chunk of text in the ground truth that is missed, the smaller the transition probability. According to these constraints, the transition probability $P(s_i|s_{i-1})$ is defined as:

$$P(s_i|s_{i-1}) = \begin{cases} 0 & s_i < s_{i-1} \\ k_1 & s_i = s_{i-1} \\ k_2 & s_i - s_{i-1} = 1 \\ \lambda e^{-\lambda(s_i - s_{i-1})} & s_i - s_{i-1} > 1 \end{cases} \tag{3}$$

where k_1 and k_2 are two constants. k_1 represents the probability of two consecutive OCR output terms corresponding to the same ground truth term (e.g. caused by over-segmentation of one word into two separate parts when aligning at the word level) . k_2 is the probability the two consecutive ground truth terms are correctly recognized. That is, probability that two consecutive OCR output terms are identical to the two consecutive ground truth terms. Since OCR accuracies are fairly high, this is the commonest case. So $k_2 \gg k_1$. When $s_i - s_{i-1} > 1$ we assume that the transition probability is subjected to an exponential distribution to accommodate the fact that the more ground truth terms missed by the OCR, the smaller the transition probability. Also note that when $s_i < s_{i-1}$, the probability is zero because of the fact that OCR never traverses the ground truth backwards.

Since we don't have aligned data from which we can learn the distributions of transition probabilities, we empirically select the parameters by visually checking the alignment results on two selected books from Gutenberg texts. One book which is 145000 characters long (about 27000 words) has relatively good OCR results, while the other book with 530000 characters (about 100,000 words) has relatively bad OCR results. In our experiments, $k_1 = 0.001$, $k_2 = 0.8$ and $\lambda = 0.5$. We also found the alignment results are not sensitive to the values of k_1 and k_2 as long as the above constraints are satisfied.

The generative probability $P(o_i|s_i)$ in our alignment model simulates the possibility of OCR wrongly recognizing a ground truth term. This probability may be modeled using a monotonic function of the similarity between the OCR term o_i and the ground truth term g_{s_i} at position s_i. One possibility is to make it a function of edit distance or the ratio of the number of common elements with the length of the longer term. Using a function of edit distance makes the algorithm very slow. For simplicity and speed, we only consider whether these two terms are exactly matched or not for word level and character level alignment, making the generative probability a simple function, defined as:

$$P(o_i|s_i) = \begin{cases} \mu_1 & o_i = g_{s_i} \\ \mu_2 & o_i \neq g_{s_i} \end{cases} \tag{4}$$

where μ_1 is a constant representing the probability of an OCR term exactly matching the aligned ground truth term and μ_2 a constant for the probability of not matching. $\mu_1 \gg \mu_2$ should hold to give a large penalty for recognition errors. In the similar way for transition probabilities, we empirically select $\mu_1 = 0.99$ and $\mu_2 = 0.0001$ through visually checking the alignment results on two selected books. Also the alignment results are not sensitive to the specific values of μ_1 and μ_2 as long as $\mu_1 \gg \mu_2$ holds.

Although theoretically both the transition probability and generative probability should be normalized to 1, a constant factor for these probabilities doesn't affect the choice of the optimal alignment $\tilde{S}$ in equation 2.

The Viterbi algorithm [18] is used to determine the most likely state sequence $\tilde{S}$ through decoding over the OCR sequence. Once equation 2 is solved, we get a sequence of positions in the ground truth with the same length as the OCR output sequence. For each OCR term, the assigned position value indicates the ground truth term from which it is generated. Figure 3 shows a simple example of how HMM works for alignment at character level, where "Hi,

world!" is the ground truth sequence, and "Hl, Wridl" the OCR output. State sequence $\{s_1, \ldots, s_{10}\}$ represents the positions in the ground truth in charge of generating the OCR output. Through Viterbi decoding on this graphical model, one should get a position sequence of "1 2 3 4 5 7 8 9 10". The Viterbi algorithm will find a path in the ground truth with the least total costs for missing, repeating ground truth terms and making recognition errors.

However, the alignment model doesn't explicitly deal with extra text in the OCR output, which may be caused by repetitively scanned pages, omitted comments and annotations in ground truth or/and some other reason. The state space is defined on positions of the ground truth, so for each term in the OCR output some ground truth term is force aligned with it. When there is extra text in the OCR output, the model tends to align them with the ground truth term which corresponds to the OCR term prior to the extra text in OCR output - this is due to constraints from the Viterbi algorithm on the terms before and after the extra text. In this case, there will be a series of repeated numbers (positions) appearing in the alignment results.

To detect the extra text in the OCR output, a heuristic-based post-processing step is performed after each alignment at each level. When a continuous section of the OCR output is aligned to the same term in the ground truth sequence, heuristic rules are used to determine which term in this section is the real correspondence of the assigned ground truth term and designate the others as extra materials. The heuristic rules are as follows: If there are some terms in this section of OCR which are exactly matched with the assigned ground truth term, select the first exact match as the real correspondence and label all the others as extra. If there are no exact matches in this section with the aligned ground truth term, it is necessary to calculate the similarities between each term in this section of OCR output with the assigned ground truth term and the neighbors of the assigned ground truth term. If the similarities are lower than some predefined threshold, this OCR term is extra.

3. VERIFICATION USING NOISE MODELS

Ground truth for the alignment between the OCR output and book contents is difficult to acquire in the real world. To evaluate our alignment approach, we build a noise model which allows us to create synthesized OCR documents which are aligned with the original documents.

We first select one electronic book as our original document and keep a sequence of indices from 1 to the length of the original document. The noise model repetitively does three basic operations over the original document, which are deletion, replacement and insertion, until the amounts of deleted, replaced and inserted characters reached the predefined criteria respectively. Following each operation, the noise model records the real alignment between the updated document and the original one through adjusting the indices of the characters of the updated document in the original sequence, i.e. after each deletion on the text document, the noise model also deletes the indices of the deleted characters; for each inserted character, it inserts -1's corresponding to the new characters in the index sequence; it keeps the index unchanged for each replacement. The position where each operation is implemented in the document is randomly generated by the model. Finally, we make a synthesized document and also keep track of the real alignment of this document with the original one. By aligning synthesized documents with the corresponding original ones using our alignment approach and then comparing the alignment results with the real alignment, we evaluate the performance of our alignment approach.

4. EXPERIMENTAL RESULTS

In this section, we report the experimental results of verifying our alignment approach through aligning synthesized documents, as well as the evaluation of the performance of one OCR system through aligning this OCR system's output with the ground truth for books.

4.1 Results of Alignments on Synthesized Documents

We select one electronic book downloaded from the Gutenberg website as our original document. This book contains about 550K characters including white space. For simplicity, we set the numbers of deleted, replaced and inserted characters to be equal for the noise model described in section 3. We test our alignment approach on two sets of numbers for the three kinds of operations, which are respectively 10% and 5% of the total number of character in the original document. For each of these two parameter settings, we generate 5 synthesized documents and record the real alignments between them and the original document. After aligning the synthesized documents and the original document, we calculate the average accuracy rate of the alignment results for each parameter setting. The results are shown in Table 1, from which we see that even with high error rates (30% and 15% in total respectively) between the synthesized documents and the original one, our alignment approach still works very well.

4.2 Evaluation of OCR Performance based on Alignment

We now use real data and align the OCR output with ground truth from the Gutenberg texts and evaluate the performance of the OCR system using this alignment.

4.2.1 OCR Performance Metric

According to the alignment results, each character in the OCR output is labeled as correct, wrong, or extra and each character in the ground truth can be labeled as correctly recognized, wrongly recognized, or missed. For some purposes, OCR evaluation on character level may be not sufficient, e.g. book retrieval is usually done at word level So we also provide OCR evaluations at the word level. As for characters, OCR words can also be labeled as correct, wrong, or extra (if and only if all the characters in that word are labeled as extra), and ground truth words can be labeled as as correctly recognized, wrongly recognized, or missed (if and only if all characters in that word are labeled as missed). We defined two criteria to evaluate OCR performance for both characters and words:

1. **Accuracy Rate** The ratio of the number of characters/words in the OCR output labeled as correct to the number of characters/words detected by the OCR (the sum of the number of correctly recognized characters/words and the number of wrongly recognized characters/words).

Deletion %	Replacement %	Insertion %	Total Error %	Accuracy Rate
10	10	10	30	0.953
5	5	5	15	0.980

Table 1: Performance of the alignment approach on a synthesized document. Column 4 is the sum of the first three columns

Num Samples	Average Missing Rate	Average Accuracy Rate
Chars 74M	0.0546	0.9796
Words 16M	0.0490	0.9206

Table 2: OCR Performance Evaluation based on Alignment Results. Note that because of the way the numbers are defined, the sum of columns 3 and 4 is not 1.

2. **Missing Rate** The ratio of the number of characters/words in ground truth sequence labeled as missed to the total number of characters/words in ground truth.

4.2.2 Results of OCR Performance Evaluation

Our ground truth consists of ebooks downloaded from the Gutenberg website `http://www.gutenberg.com`, which contains up to 17,000 free electronic books manually typed by hundreds of volunteers. All these ebooks are in plain text files without any layout, line or page information.

Our dataset for OCR performance evaluation consists of 147 electronic books downloaded from the Gutenberg website and the outputs of the OCR engine on scanned books which have the same author name and title with the downloaded electric books. After aligning every book with their corresponding OCR outputs, we evaluate the OCR performance using measurements defined in 4.2.1. Table 2 shows the performance of the OCR engine on the 147 books, from which we can see that even when the character accuracy is very high, about 5 words are missed by this OCR engine for every 100 ground truth words, and about 8% wrongly recognized within those detected.

Figure 4 shows some snippets from the alignment results for the book "The Rights of Man" (written by "Thomas Paine"), which correspond to the examples of OCR output showed in figure 1. The alignment approach works very well even when there are a lot of recognition errors.

5. CONCLUSION

In this paper, we proposed a hierarchical alignment method for aligning OCR output and ground truth for books. Our hierarchical alignment approach partitions the alignment problem for an entire book into the problem of aligning many shorter subsequences. A HMM-based model is employed for alignment at each level. Experimental results show that even on OCR output with high error rate, our alignment method works very well.

Acknowledgements

This work was done while the two authors were visiting Google. We would like to thank Google for support. We would also like to thank Alan Eustace for inviting us to Google, Chris Uhlik and Dan Clancy for encouragement and support, Toni Rath for discussions on handwriting alignment and Luc Vincent, Dar Shyang Lee and Igor Krivokon for discussions. Any opinions, findings, conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of Google or the University of Massachusetts, Amherst.

6. REFERENCES

[1] H. Alshawi, S. Bangalore, and S. Douglas. Learning phrase-based head transduction models for translation of spoken utterances. In *Proceedings of the fifth International Conference on Spoken Language Processing (ICSLP98)*, Sydney, 1998.

[2] X. Chen and A. Yuille. Detecting and reading text in natural scenes. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, pages 366–373, Washington, DC, USA, 2004.

[3] Y. Deng and W. Byrne. Hmm word and phrase alignment for statistical machine translation. In *Proceedings of HLT-EMNLP*, 2005.

[4] Gutenberg Website:. http://www.gutenberg.com.

[5] T. Ho and H. Baird. Evaluation of ocr accuracy using synthetic data. In *Proceedings of 4th UNLV Symp. on Document Analysis and Information Retrieval*, Las Vegas, Nevada, USA, April 1995.

[6] J. Hobby. Matching document images with ground truth. *International Journal on Document Analysis and Recognition*, 1(1):52–61, 1997.

[7] P. Jang and A. Hauptmann. Learning to recognize speech by watching television. *IEEE Intelligent Systems*, 14:51–58, 1999.

[8] T. Kanungo, R. Haralick, H. Baird, W. Stuezle, and D. Madigan. A statistical, nonparametric methodology for document degradation model validation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(11):1209–1223, 2000.

[9] M. Kay and M. Roscheisen. Text-translation alignment. *Computational Linguistics*, 19:121–142, 1993.

[10] E. Kornfield, R. Manmatha, and J. Allan. Text alignment with handwritten documents. In *Proceedings of Document Image Analysis for Libraries (DIAL)*, pages 195–209, 2004.

[11] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[12] F. Malfrre, O. Deroo, and T. Dutoit. Phonetic alignment: Speech synthesis based vs. hybrid

```
O : despotism,  were  still  ha RIOH TS 07 MAN. Ii ad in  the  bands of  a  successor.
T : despotism,  were  still  liable  to be re   vived in  the  hands of  a  successor.

O : It  was  not  the  respite  of  a            would  satisfy  France,  enlightened
T : It  was  not  the  respite  of  a  reign that would  satisfy  France,  enlightened

O : as  she   @s then  become    ia         nmtinuance of  the  practice  of  despotism,
T : as  she  was then  become. A casual discontinuance of  the  practice  of  despotism,

O : is  not  a  die          m of  its  @riseiple.; the  former  depends  on  the
T : is  not  a  discontinuance of  its  principles: the  former  depends  on  the

O :virtue  of            lual who  is  in  immediate  possession  of  the  power;
T :virtue  of  the individual who  is  in  immediate  possession  of  the  power;

O :the latter, @                   r?@t
T :the latter, on the virtue and fortitude of the nation. In the case of Charles I.

O :                         ,@ @   r@)    @   !  i               l       i'I'
T :and James II. of England,  the revolt was against the personal despotism  of the

O :             i          it* b           h               s                  b    ed
T :men; whereas in France, it  was against the hereditary despotism of the established

O :gOvernment  But  men  Who  can  consign    @ J jce              teiity for  ever
T :Government. But  men  who  can  consign  over  the rights of posterity for  ever

O : on  the  authority  of  a  mouldy           , like  Mr.  Burke,  are  not  qualified
T : on  the  authority  of  a  mouldy  parchment, like  Mr.  Burke,  are  not  qualified

O : to  judge  of  this  revolu    .  @  akes in  a  fi.Ld too  vast  for  their  views
T : to  judge  of  this  Revolution. It takes in  a  field too  vast  for  their  views

O : to  explore,  and  pro'          a  mightiness  of  reason  they  cannot  keep
T : to  explore,  and  proceeds with a  mightiness  of  reason  they  cannot  keep
```

(a)

```
0 :no@,.@ an,@ @i tlle VUCMtI01J i'Sturall. prt.@,.,,ts Itself  @@"Vb Is  the  best
T :form;  an d    t he quest ion  natura lly pr   esents itself,  What is  the  best

0 :                  for  cofld@@t1,  the  RE8 P  OSLIC or  the  Puau   c 1 demo.@rat,ei
T : form of government for  co nducting the  Res-Publica, or  the  Pu ublic   Bus     iness

0 :form? of  a  flat)  ,, Siter it  becom,@ too  CXtensive and  Populous  for   Lb
T :      of  a  n ation,, after it  becomes too  extensive and  populous  for  the

0 :                         It  cannot          h0                              to
T :simple democratical form? It  cannot  be monarchy, because monarchy is subject to

0 : an  ol       of  the   aai@@ amount I                                  form@ w
T : an  objection of  the  saa me amount   to which the simple democratical for m was

0 :   ject. It  is  POSsible #                                  a 8ystem o      cIple
T :subject. It  is  possible  that an individual may lay down a system of principles,

0 :     i                              estab   ha     any ea     I      i
T :on which government shall be constitutionally established to any extent of territory.

0 :1    CJp I     e5
T :This is  no more  than an operation of the mind, acting by its own powers. But the

0 :                              as     fui   tlo, ,,    r           eq
T :practice upon those principles, as  applying t o  the various and numerous circumstances

0 :   a  #ario05
T :of a  nation, its agriculture, manufacture, trade, commerce, etc., etc., a knowledge

0 :                                                          parts  of
T :of a different kind, and which can be had only from the various parts  of  society.

0 :                                    WhIch  no  on                e    in
T :It is an assemblage of practical knowledge, which  no  individual can possess; and

0 :                         form  js as  much        hr
T :therefore the monarchical form  is as  much  limited, in useful practice, from the

0 :                          as  Was  The
T :incompetency of knowledge, as  was  the  democratical form, from the multiplicity
```

(b)

Figure 4: Snippet of the Alignment Results for One Book.

hmm/ann. In *Proceedings of the ICSLP*, pages 1571–1574, 1998.

[13] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–53, 1970.

[14] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, January 1986.

[15] J. Rothfeder, T. Rath, and R. Manmatha. Aligning transcripts to automatically segmented handwritten manuscripts. In *to appear in Proceedings of the Seventh International Workshop on Document Analysis Systems, DAS'06*, Nelson, New Zealand, 2006.

[16] D. Roy and C. Malamud. Speaker identification based text to audio alignment for an audio retrieval system. In *ICASSP '97*, Munich, Germany, 1997.

[17] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(3):195–197, 1981.

[18] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13:260–267, April 1967.

[19] R. Wagner and M. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.

[20] V. Wu, R. Manmatha, and E. Riseman. Textfinder: An automatic system to detect and recognize text in images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(11):1224–1229, 1999.

[21] Y. Xu and G. Nagy. Prototype extraction and adaptive ocr. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(12):1280–1296, 1999.