A Statistical Language Modeling Approach to Lattice-Based Spoken Document Retrieval

Tee Kiah Chia[†] **Haizhou Li**[‡] **Hwee Tou Ng**[†]
[†]Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
{chiateek,nght}@comp.nus.edu.sg

[‡]Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
hli@i2r.a-star.edu.sg

Abstract

Speech recognition transcripts are far from perfect; they are not of sufficient quality to be useful on their own for spoken document retrieval. This is especially the case for conversational speech. Recent efforts have tried to overcome this issue by using statistics from speech lattices instead of only the 1-best transcripts; however, these efforts have invariably used the classical vector space retrieval model. This paper presents a novel approach to lattice-based spoken document retrieval using statistical language models: a statistical model is estimated for each document, and probabilities derived from the document models are directly used to measure relevance. Experimental results show that the lattice-based language modeling method outperforms both the language modeling retrieval method using only the 1-best transcripts, as well as a recently proposed lattice-based vector space retrieval method.

1 Introduction

Information retrieval (IR) is the task of ranking a collection of documents according to an estimate of their relevance to a query. With the recent growth in the amount of speech recordings in the form of voice mails, news broadcasts, and so forth, the task of spoken document retrieval (SDR) – information retrieval in which the document collection is in the form of speech recordings – is becoming increasingly important.

SDR on broadcast news corpora has been "deemed to be a solved problem", due to the fact that the performance of retrieval engines working on 1-best automatic speech recognition (ASR) transcripts was found to be "virtually the same as their performance on the human reference transcripts" (NIST, 2000). However, this is still not the case for SDR on data which are more challenging, such as conversational speech in noisy environments, as the 1-best transcripts of these data contain too many recognition errors to be useful for retrieval. One way to ameliorate this problem is to work with not just one ASR hypothesis for each utterance, but multiple hypotheses presented in a *lattice* data structure. A lattice is a connected directed acyclic graph in which each edge is labeled with a term hypothesis and a likelihood value (James, 1995); each path through a lattice gives a hypothesis of the sequence of terms spoken in the utterance.

Each lattice can be viewed as a statistical model of the possible transcripts of an utterance (given the speech recognizer's state of knowledge); thus, an IR model based on statistical inference will seem to be a more natural and more principled approach to lattice-based SDR. This paper thus proposes a lattice-based SDR method based on the statistical language modeling approach of Song and Croft (1999). In this method, the *expected word count* – the mean number of occurrences of a word given a lattice's statistical model – is computed for each word in each lattice. Using these expected counts, a statistical language model is estimated for each spoken document, and a document's relevance to a query is computed as a probability under this model.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 810–818, Prague, June 2007. ©2007 Association for Computational Linguistics

The rest of this paper is organized as follows. In Section 2 we review related work in the areas of speech processing and IR. Section 3 describes our proposed method as well as the baseline methods. Details of the experimental setup are given in Section 4, and experimental results are in Section 5. Finally, Section 6 concludes our discussions and outlines our future work.

2 Related Work

2.1 Lattices for Spoken Document Retrieval

James and Young (1994) first introduced the lattice as a representation for indexing spoken documents, as part of a method for vocabulary-independent keyword spotting. The lattice representation was later applied to the task of spoken document retrieval by James (1995): James counted how many times each query word occurred in each phone lattice with a sufficiently high normalized log likelihood, and these counts were then used in retrieval under a vector space model with $tf \cdot idf$ weighting. Jones et al. (1996) combined retrieval from phone lattices using variations of James' method with retrieval from 1-best word transcripts to achieve better results.

Since then, a number of different methods for SDR using lattices have been proposed. For instance, Siegler (1999) used word lattices instead of phone lattices as the basis of retrieval, and generalized the $tf \cdot idf$ formalism to allow uncertainty in word counts. Chelba and Acero (2005) preprocessed lattices into more compact Position Specific Posterior Lattices (PSPL), and computed an aggregate score for each document based on the posterior probability of edges and the proximity of search terms in the document. Mamou et al. (2006) converted each lattice into a word confusion network (Mangu et al., 2000), and estimated the inverse document frequency (idf) of each word t as the ratio of the total number of words in the document collection to the total number of occurrences of t.

Despite the differences in the details, the above lattice-based SDR methods have all been based on the classical vector space retrieval model with $tf \cdot idf$ weighting.

2.2 Expected Counts from Lattices

A speech recognizer generates a 1-best transcript of a spoken document by considering possible transcripts of the document, and then selecting the transcript with the highest probability. However, unlike a text document, such a 1-best transcript is likely to be inexact due to speech recognition errors. To represent the uncertainty in speech recognition, and to incorporate information from multiple transcription hypotheses rather than only the 1-best, it is desirable to use expected word counts from lattices output by a speech recognizer.

In the context of spoken document search, Siegler (1999) described expected word counts and formulated a way to estimate expected word counts from lattices based on the relative ranks of word hypothesis probabilities; Chelba and Acero (2005) used a more explicit formula for computing word counts based on summing edge posterior probabilities in lattices; Saraclar and Sproat (2004) performed word-spotting in speech lattices by looking for word occurrences whose expected counts were above a certain threshold; and Yu et al. (2005) searched for phrases in spoken documents using a similar measure, the expected word relevance.

Expected counts have also been used to summarize the phonotactics of a speech recording represented in a lattice: Hatch et al. (2005) performed speaker recognition by computing the expected counts of phone bigrams in a phone lattice, and estimating an unsmoothed probability distribution of phone bigrams.

Although many uses of expected counts have been studied, the use of statistical language models built from expected word counts has not been well explored.

2.3 Retrieval via Statistical Language Modeling

Finally, the statistical language modeling approach to retrieval was used by Ponte and Croft (1998) for IR with text documents, and it was shown to outperform the $tf \cdot idf$ approach for this task; this method was further improved on in Song and Croft (1999). Chen et al. (2004) applied Song and Croft's method to Mandarin spoken document retrieval using 1-best ASR transcripts. In this task, it was also shown to

outperform $tf \cdot idf$. Thus, the statistical language modeling approach to retrieval has been shown to be superior to the vector space approach for both these IR tasks.

2.4 Contributions of Our Work

The main contributions of our work include

- extending the language modeling IR approach from text-based retrieval to lattice-based spoken document retrieval; and
- formulating a method for building a statistical language model based on expected word counts derived from lattices.

Our method is motivated by the success of the statistical retrieval framework over the vector space approach with $tf \cdot idf$ for text-based IR, as well as for spoken document retrieval via 1-best transcripts. Our use of expected counts differs from Saraclar and Sproat (2004) in that we estimate probability models from the expected counts. Conceptually, our method is close to that of Hatch et al. (2005), as both methods build a language model to summarize the content of a spoken document represented in a lattice. In practice, our method differs from Hatch et al. (2005)'s in many ways: first, we derive word statistics for representing semantics, instead of phone bigram statistics for representing phonotactics; second, we introduce a smoothing mechanism (Zhai and Lafferty, 2004) to the language model that is specific for information retrieval.

3 Methods

We now describe the formulation of three different SDR methods: a baseline statistical retrieval method which works on 1-best transcripts, our proposed statistical lattice-based SDR method, as well as a previously published vector space lattice-based SDR method.

3.1 Baseline Statistical Retrieval Method

Our baseline retrieval method is motivated by Song and Croft (1999), and uses the language model smoothing methods of Zhai and Lafferty (2004). This method is used to perform retrieval on the documents' 1-best ASR transcripts and reference human transcripts.

Let $\mathcal{C}$ be the collection of documents to retrieve from. For each document $\mathbf{d}$ contained in $\mathcal{C}$, and each query $\mathbf{q}$, the relevance of $\mathbf{d}$ to $\mathbf{q}$ can be defined as $\Pr(\mathbf{d} \mid \mathbf{q})$. This probability cannot be computed directly, but under the assumption that the prior $\Pr(\mathbf{d})$ is uniform over all documents in $\mathcal{C}$, we see that

$$\Pr(\mathbf{d} \mid \mathbf{q}) = \frac{\Pr(\mathbf{q} \mid \mathbf{d}) \Pr(\mathbf{d})}{\Pr(\mathbf{q})} \propto \Pr(\mathbf{q} \mid \mathbf{d});$$

This means that ranking documents by $\Pr(\mathbf{d} \mid \mathbf{q})$ is equivalent to ranking them by $\Pr(\mathbf{q} \mid \mathbf{d})$, and thus $\Pr(\mathbf{q} \mid \mathbf{d})$ can be used to measure relevance (Berger and Lafferty, 1999).

Now express $\mathbf{q}$ as a series of words drawn from a vocabulary $\mathcal{V} = \{w_1, w_2, \cdots w_V\}$; that is, $\mathbf{q} = q_1 q_2 \cdots q_K$, where K is the number of words in the query, and $q_i \in \mathcal{V}$ for $1 \leq i \leq K$. Then given a unigram model derived from $\mathbf{d}$ which assigns a probability $\Pr(w \mid \mathbf{d})$ to each word w in $\mathcal{V}$, we can compute $\Pr(\mathbf{q} \mid \mathbf{d})$ as follows:

$$\begin{aligned} \Pr(\mathbf{q} \mid \mathbf{d}) &= \Pr(q_1 q_2 \cdots q_K \mid \mathbf{d}) \\ &= \prod_{i=1}^{K} \Pr(q_i \mid \mathbf{d}) \\ &= \prod_{\substack{w \in \mathcal{V}, \\ C(w|\mathbf{q})>0}} \Pr(w|\mathbf{d})^{C(w|\mathbf{q})} \quad (1) \end{aligned}$$

where $C(w \mid \mathbf{q})$ is the word count of w in $\mathbf{q}$.

Before using Equation 1, we must estimate a unigram model from $\mathbf{d}$: that is, an assignment of probabilities $\Pr(w \mid \mathbf{d})$ for all $w \in \mathcal{V}$. One way to do this is to use a maximum likelihood estimate (MLE) – an assignment of $\Pr(w \mid \mathbf{d})$ for all w which maximizes the probability of generating $\mathbf{d}$. The MLE is given by the equation

$$\Pr{}_{\text{mle}}(w \mid \mathbf{d}) = \frac{C(w \mid \mathbf{d})}{|\mathbf{d}|}$$

where $C(w \mid \mathbf{d})$ is the number of occurrences of w in $\mathbf{d}$, and $|\mathbf{d}|$ is the total number of words in $\mathbf{d}$. However, using this formula means we will get a value of zero for $\Pr(\mathbf{q} \mid \mathbf{d})$ if even a single query word q_i is not found in $\mathbf{d}$. To overcome this problem, we smooth the model by assigning some probability mass to such unseen words. Specifically, we adopt

a two-stage smoothing method (Zhai and Lafferty, 2004):

$$\Pr(w \mid \mathbf{d}) = (1-\lambda)\frac{C(w \mid \mathbf{d}) + \mu \Pr(w \mid \mathcal{C})}{|\mathbf{d}| + \mu} + \lambda \Pr(w \mid \mathcal{U}) \quad (2)$$

Here, $\mathcal{U}$ denotes a background language model, and $\mu > 0$ and $\lambda \in (0, 1)$ are parameters to the smoothing procedure. This is a combination of Bayesian smoothing using Dirichlet priors (MacKay and Peto, 1984) and Jelinek-Mercer smoothing (Jelinek and Mercer, 1980).

The parameter λ can be set empirically according to the nature of the queries. For the parameter μ, we adopt the estimation procedure of Zhai and Lafferty (2004): we maximize the leave-one-out log likelihood of the document collection, namely

$$\ell_{-1}(\mu \mid \mathcal{C}) = \sum_{\mathbf{d} \in \mathcal{C}} \sum_{w \in \mathcal{V}} C(w \mid \mathbf{d}) \log\left(\frac{C(w \mid \mathbf{d}) - 1 + \mu \Pr(w \mid \mathcal{C})}{|\mathbf{d}| - 1 + \mu}\right) \quad (3)$$

by using Newton's method to solve the equation

$$\ell'_{-1}(\mu \mid \mathcal{C}) = 0$$

3.2 Our Proposed Statistical Lattice-Based Retrieval Method

We now propose our lattice-based retrieval method. In contrast to the above baseline method, our proposed method works on the lattice representation of spoken documents, as generated by a speech recognizer.

First, each spoken document is divided into M short speech segments. A speech recognizer then generates a lattice for each speech segment. As previously stated, a lattice is a connected directed acyclic graph with edges labeled with word hypotheses and likelihoods. Thus, each path through the lattice contains a hypothesis of the series of words spoken in this speech segment, $\mathbf{t} = t_1 t_2 \cdots t_N$, along with acoustic probabilities $\Pr(o_1 \mid t_1)$, $\Pr(o_2 \mid t_2)$, $\cdots$ $\Pr(o_N \mid t_N)$, where o_i denotes the acoustic observations for the time interval of the word t_i hypothesized by the speech recognizer. Let $\mathbf{o} = o_1 o_2 \cdots o_N$ denote the acoustic observations for the entire speech segment; then

$$\Pr(\mathbf{o} \mid \mathbf{t}) = \prod_{i=1}^{N} \Pr(o_i \mid t_i)$$

We then rescore each lattice with an n-gram language model. Effectively, this means multiplying the acoustic probabilities with n-gram probabilities:

$$\Pr(\mathbf{t}, \mathbf{o}) = \Pr(\mathbf{o} \mid \mathbf{t}) \Pr(\mathbf{t}) = \prod_{i=1}^{N} \Pr(o_i \mid t_i) \Pr(t_i \mid t_{i-n+1} \cdots t_{i-1})$$

This produces an expanded lattice in which paths (hypotheses) are weighted by their posterior probabilities rather than their acoustic likelihoods: specifically, by $\Pr(\mathbf{t}, \mathbf{o}) \propto \Pr(\mathbf{t} \mid \mathbf{o})$ rather than $\Pr(\mathbf{o} \mid \mathbf{t})$ (Odell, 1995). The lattice is then pruned, by removing those paths in the lattice whose log posterior probabilities – to be precise, whose $\gamma \ln \Pr(\mathbf{t} \mid \mathbf{o})$ – are not within a threshold Θ of the best path's log posterior probability (in our implementation, $\gamma = 10000.5$).

Next, we compute the expected count of each word in each document. For each word w and each document $\mathbf{d}$ comprised of M speech segments represented by M acoustic observations $\mathbf{o}^{(1)}$, $\mathbf{o}^{(2)}$, $\cdots$ $\mathbf{o}^{(M)}$, the expected count of w in $\mathbf{d}$ is

$$\mathrm{E}[C(w \mid \mathbf{d})] = \sum_{j=1}^{M} \sum_{\mathbf{t}} C(w \mid \mathbf{t}) \Pr(\mathbf{t} \mid \mathbf{o}^{(j)})$$

where $C(w \mid \mathbf{t})$ is the word count of w in the hypothesized transcript $\mathbf{t}$. We can also analogously compute the expected document length:

$$\mathrm{E}[|\mathbf{d}|] = \sum_{j=1}^{M} \sum_{\mathbf{t}} |\mathbf{t}| \Pr(\mathbf{t} \mid \mathbf{o}^{(j)})$$

where $|\mathbf{t}|$ denotes the number of words in $\mathbf{t}$.

We now replace $C(w \mid \mathbf{d})$ and $|\mathbf{d}|$ in Equation 2 with $\mathrm{E}[C(w \mid \mathbf{d})]$ and $\mathrm{E}[|\mathbf{d}|]$; thus

$$\Pr(w \mid \mathbf{d}) = (1-\lambda)\frac{\mathrm{E}[C(w \mid \mathbf{d})] + \mu \Pr(w \mid \mathcal{C})}{\mathrm{E}[|\mathbf{d}|] + \mu} + \lambda \Pr(w \mid \mathcal{U}) \quad (4)$$

In addition, we also modify the procedure for estimating μ, by replacing $C(w \mid \mathbf{d})$ and

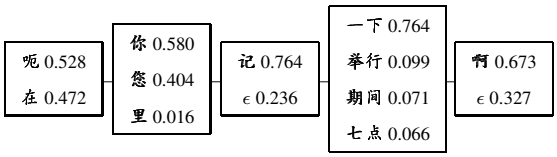

Figure 1: Example of a word confusion network

$|\mathbf{d}|$ in Equation 3 with $\left\lfloor \mathrm{E}[C(w \mid \mathbf{d})] + \frac{1}{2} \right\rfloor$ and $\sum_{w \in \mathcal{V}} \left\lfloor \mathrm{E}[C(w \mid \mathbf{d})] + \frac{1}{2} \right\rfloor$ respectively. The probability estimates from Equation 4 can then be substituted into Equation 1 to yield relevance scores.

3.3 Baseline $tf \cdot idf$ Lattice-Based Retrieval Method

As a further comparison, we also implemented Mamou et al. (2006)'s vector space retrieval method (without query refinement via lexical affinities). In this method, each document $\mathbf{d}$ is represented as a word confusion network (WCN) (Mangu et al., 2000) – a simplified lattice which can be viewed as a sequence of confusion sets $c_1, c_2, c_3, \cdots$. Each c_i corresponds approximately to a time interval in the spoken document and contains a group of word hypotheses, and each word w in this group of hypotheses is labeled with the probability $\Pr(w \mid c_i, \mathbf{d})$ – the probability that w was spoken in the time interval of c_i. A confusion set may also give a probability for $\Pr(\epsilon \mid c_i, \mathbf{d})$, the probability that no word was spoken in the time of c_i. Figure 1 gives an example of a WCN.

Mamou et al.'s retrieval method proceeds as follows. First, the documents are divided into speech segments, lattices are generated from the speech segments, and the lattices are pruned according to the path probability threshold Θ, as described in Section 3.2. The lattice for each speech segment is then converted into a WCN according to the algorithm of Mangu et al. (2000). The WCNs for the speech segments in each document are then concatenated to form a single WCN per document.

Now, to retrieve documents in response to a query $\mathbf{q}$, the method computes, for each document $\mathbf{d} \in \mathcal{C}$ and each word $w \in \mathcal{V}$,

- the "document length" $|\mathbf{d}|$, computed as the number of confusion sets in the WCN of $\mathbf{d}$;
- the "average document length" $avdl$, computed as

$$avdl \;=\; \frac{1}{|\mathcal{C}|} \sum_{\mathbf{d}' \in \mathcal{C}} \left|\mathbf{d}'\right|;$$

- the "document term frequency" $C^*(w \mid \mathbf{d})$, computed as

$$C^*(w|\mathbf{d}) = \sum_{c \in occ(w,\mathbf{d})} (b_{rank(w|c,\mathbf{d})} \cdot \Pr(w|c,\mathbf{d}))$$

  where $occ(w, \mathbf{d})$ is the set of confusion sets in $\mathbf{d}$'s WCN which contain w as a hypothesis, $rank(w \mid c, \mathbf{d})$ is the rank of w in terms of probability within the confusion set c, and $(b_1, b_2, b_3, \cdots) = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 0, 0, \cdots)$ is a boosting vector which serves to discard all but the top 10 hypotheses, and gives more weight to higher-ranked word hypotheses;

- the query term frequency $C(w \mid \mathbf{q})$, which is simply the word count of w in $\mathbf{q}$; and
- the "inverse document frequency" $idf(w)$, computed as

$$idf(w) \;=\; \log \frac{O}{O_w}$$

  where

$$O_w \;=\; \sum_{\mathbf{d} \in \mathcal{C}} \sum_{c \in occ(w,\mathbf{d})} \Pr(w \mid c, \mathbf{d})$$

$$O \;=\; \sum_{w' \in \mathcal{V}} O_{w'}$$

With these, the relevance of $\mathbf{d}$ to $\mathbf{q}$ is computed as (Carmel et al., 2001)

$$rel(\mathbf{d}, \mathbf{q}) \;=\; \frac{\sum_{w \in \mathcal{V}} C^*(w \mid \mathbf{d}) \cdot C(w \mid \mathbf{q}) \cdot idf(w)}{\sqrt{0.8 \cdot avdl + 0.2 \cdot |\mathbf{d}|}}$$

4 Experiments

4.1 Document Collection

To evaluate our proposed retrieval method, we performed experiments using the Hub5 Mandarin training corpus released by the Linguistic Data Consortium (LDC98T26). This is a conversational telephone speech corpus which is 17 hours long, and

contains recordings of 42 telephone calls corresponding to approximately 600Kb of transcribed Mandarin text. Each conversation has been broken up into speech segments of less than 8 seconds each.

As the telephone calls in LDC98T26 have not been divided neatly into "documents", we had to choose a suitable unit of retrieval which could serve as a "document". An entire conversation would be too long for such a purpose, while a speech segment or speaker turn would be too short. We decided to use $\frac{1}{2}$-minute time windows with 50% overlap as retrieval units, following Abberley et al. (1999) and Tuerk et al. (2001). The 42 telephone conversations were thus divided into 4,312 retrieval units ("documents"). Each document comprises multiple consecutive speech segments.

4.2 Queries and Ground Truth Relevance Judgements

We then formulated 18 queries (14 test queries, 4 development queries) to issue on the document collection. Each query was comprised of one or more written Chinese keywords. We then obtained ground truth relevance judgements by manually examining each of the 4,312 documents to see if it is relevant to the topic of each query. The number of retrieval units relevant to each query was found to range from 4 to 990. The complete list of queries and the number of documents relevant to each query are given in Table 1.

Test queries		
Topic	Keywords	# relevant documents
Contact information	电话, 号码, 地址, 联系, 姓	103
Chicago	芝加哥	15
The weather	天气, 冷, 热, 暖和, 风, 凉快, 雨, 空调, 干燥, 潮湿, 气候, 温度	117
Housing matters	房子, 家, 住, 房租, 家具, 搬, 厨房, 卧室, 水电, 房东, 院子	354
Studies, academia	毕业, 学位, 考试, 修, 读, 托福, 念书, 课, 学分, 进修, 学费, 同学	990
Litigation	法律, 律师, 打官司, 起诉	31
Raising children	小孩, 孩子, 生育, 儿子, 幼儿园, 玩, 玩具, 女儿	334
Christian churches	教会, 神, 主, 礼拜, 教堂, 活动, 圣经, 团契	78
Floods	洪水, 淹, 堤, 水	4
Clothing	衣服, 皮衣, 帽子, 裤子, 高统袜, 袜子, 牛仔裤, 西服, 穿	28
Eating out	吃, 餐馆, 外卖, 中餐, 请客, 饭店	57
Playing sports	打球, 活动, 橄榄球, 排球	24
Dealings with banks	银行, 支票, 钱, 开户, 贷款	54
Computers and software	电脑, 计算机, 软件	175
Development queries		
Topic	Keywords	# relevant documents
Passport and visa matters	护照, 签证, 入境, 手续, 绿卡, 移民	143
Washington D. C.	华盛顿	15
Working life	活儿, 钱, 打工, 税, 工作, 老板, 出差, 公司, 挣, 工资, 上司, 同事, 忙, 职业	509
1996 Olympics	奥运会, 亚特兰大	8

Table 1: List of test and development queries

4.3 Preprocessing of Documents and Queries

Next, we processed the document collection with a speech recognizer. For this task we used the Abacus system (Hon et al., 1994), a large vocabulary continuous speech recognizer which contains a triphone-based acoustic system and a frame-synchronized search algorithm for effective word decoding. Each Mandarin syllable was modeled by one to four triphone models. Acoustic models were trained from a corpus of 200 hours of telephony speech from 500 speakers sampled at 8kHz. For each speech frame, we extracted a 39-dimensional feature vector consisting of 12 MFCCs and normalized energy, and their first and second order derivatives. Sentence-based cepstral mean subtraction was applied for acoustic normalization both in the training and testing. Each triphone was modeled by a left-to-right 3-state hidden Markov model (HMM), each state having 16 Gaussian mixture components. In total, we built 1,923 untied within-syllable triphone models for 43 Mandarin phonemes, as well as 3 silence models. The search algorithm was supported by a loop grammar of over 80,000 words.

We processed the speech segments in our collection corpus, to generate lattices incorporating acoustic likelihoods but not n-gram model probabilities. We then rescored the lattices using a backoff tri-

gram language model interpolated in equal proportions from two trigram models:

- a model built from the TDT-2, TDT-3, and TDT-4 Mandarin news broadcast transcripts (about 58Mb of text)
- a model built from corpora of transcripts of conversations, comprised of a 320Kb subset of the Callhome Mandarin corpus (LDC96T16) and the CSTSC-Flight corpus from the Chinese Corpus Consortium (950Kb)

The unigram counts from this model were also used as the background language model $\mathcal{U}$ in Equations 2 and 4.

The reference transcripts, queries, and trigram model training data were all segmented into words using Low et al. (2005)'s Chinese word segmenter, trained on the Microsoft Research (MSR) corpus, with the speech recognizer's vocabulary used as an external dictionary. The 1-best ASR transcripts were decoded from the rescored lattices.

Lattice rescoring, trigram model building, WCN generation, and computation of expected word counts were done using the SRILM toolkit (Stolcke, 2002), while lattice pruning was done with the help of the AT&T FSM Library (Mohri et al., 1998).

We also computed the character error rate (CER) and syllable error rate (SER) of the 1-best transcripts, and the lattice oracle CER, for one of the telephone conversations in the speech corpus (`ma_4160`). The CER was found to be 69%, the SER 63%, and the oracle CER 29%.

4.4 Retrieval and Evaluation

We then performed retrieval on the document collection using the algorithms in Section 3, using the reference transcripts, the 1-best ASR transcripts, lattices, and WCNs. We set $\lambda = 0.1$, which was suggested by Zhai and Lafferty (2004) to give good retrieval performance for keyword queries.

The results of retrieval were checked against the ground truth relevance judgements, and evaluated in terms of the non-interpolated mean average precision (MAP):

$$\text{MAP} = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{1}{R_i}\sum_{j=1}^{R_i}\frac{j}{r_{i,j}}\right)$$

Retrieval method	Retrieval source	MAP for development queries	MAP for test queries
Statistical	Reference transcripts	0.5052	0.4798
Statistical	1-best transcripts	0.1251	0.1364
Vector space $tf \cdot idf$	Lattices, $\Theta = 27,500$	0.1685	0.1599
Statistical	Lattices, $\Theta = 65,000$	0.2180	0.2154

Table 2: Summary of experimental results

where L denotes the total number of queries, R_i the total number of documents relevant to the ith query, and $r_{i,j}$ the position of the jth relevant document in the ranked list output by the retrieval method for query i.

For the lattice-based retrieval methods, we performed retrieval with the development queries using several values of Θ between 0 and 100,000, and then used the value of Θ with the best MAP to do retrieval with the test queries.

5 Experimental Results

The results of our experiments are summarized in Table 2; the MAP of the two lattice-based retrieval methods, Mamou et al. (2006)'s vector space method and our proposed statistical retrieval method, are shown in Figure 2 and Figure 3 respectively.

The results show that, for the vector space retrieval method, the MAP of the development queries is highest at $\Theta = 27,500$, at which point the MAP for the test queries is 0.1599; and for our proposed method, the MAP for the development queries is highest at $\Theta = 65,000$, and at this point the MAP for the test queries reaches 0.2154.

As can be seen, the performance of our statistical lattice-based method shows a marked improvement over the MAP of 0.1364 achieved using only the 1-best ASR transcripts, and indeed a one-tailed Student's t-test shows that this improvement is statistically significant at the 99.5% confidence level. The statistical method also yields better performance than Mamou et al.'s vector space method – a t-test

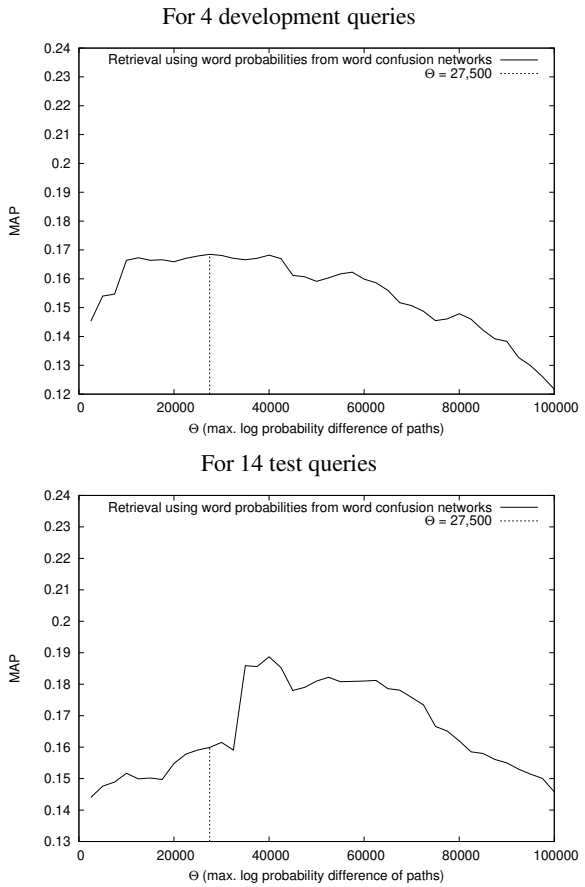

Figure 2: MAP of Mamou et al. (2006)'s vector space method for lattice-based retrieval, at various pruning thresholds Θ

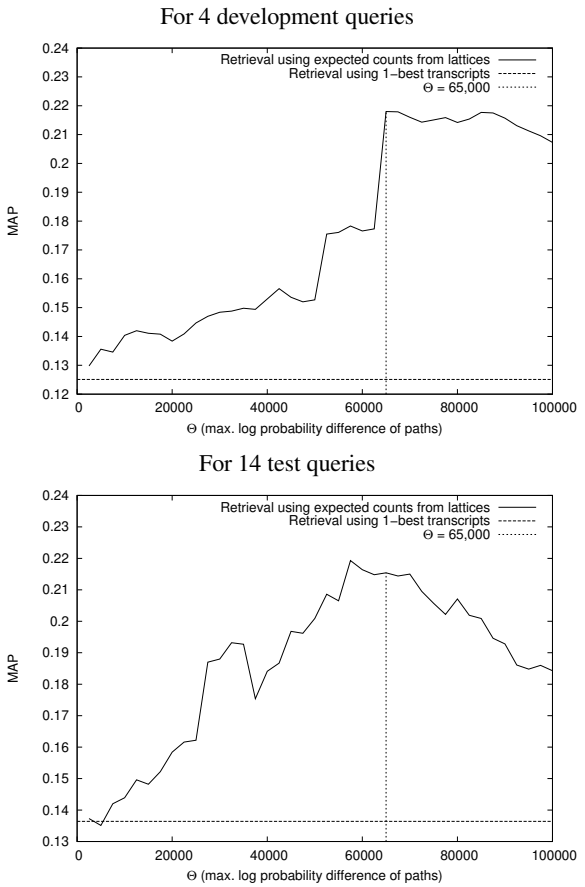

Figure 3: MAP of our proposed statistical method for lattice-based retrieval, at various pruning thresholds Θ

shows the performance difference to be statistically significant at the 97.5% confidence level.

6 Conclusions and Future Work

We have presented a method for performing spoken document retrieval using lattices which is based on a statistical language modeling retrieval framework. Results show that our new method can significantly improve the retrieval MAP compared to using only the 1-best ASR transcripts. Also, our proposed retrieval method has been shown to outperform Mamou et al. (2006)'s vector space lattice-based retrieval method.

Besides the better empirical performance, our method also has other advantages over Mamou et al.'s vector space method. For one, our method computes expected word counts directly from rescored lattices, and does not require an additional step to convert lattices lossily to WCNs. Furthermore, our method uses all the hypotheses in each lattice, rather than just the top 10 word hypotheses at each time interval. Most importantly, our method provides a more natural and more principled approach to lattice-based spoken document retrieval based on a sound statistical foundation, by harnessing the fact that lattices are themselves statistical models; the statistical approach also means that our method can be more easily augmented with additional statistical knowledge sources in a principled way.

For future work, we plan to test our proposed method on English speech corpora, and with larger-scale retrieval tasks involving more queries and more documents. We would like to extend our method to other speech processing tasks, such as spoken document classification and example-based spoken document retrieval as well.

References

Dave Abberley, David Kirby, Steve Renals, and Tony Robinson. 1999. The THISL broadcast news retrieval system. In *Proceedings of ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pages 14–19.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of SIGIR 1999*, pages 222–229.

David Carmel, Einat Amitay, Miki Herscovici, Yoelle Maarek, Yael Petruschka, and Aya Soffer. 2001. Juru at TREC 10 – Experiments with index pruning. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 228–236.

Ciprian Chelba and Alex Acero. 2005. Position specific posterior lattices for indexing speech. In *Proceedings of ACL 2005*, pages 443–450.

Berlin Chen, Hsin-min Wang, and Lin-shan Lee. 2004. A discriminative HMM/n-gram-based retrieval approach for Mandarin spoken documents. *ACM Transactions on Asian Language Information Processing*, 3(2):128–145.

Andrew O. Hatch, Barbara Peskin, and Andreas Stolcke. 2005. Improved phonetic speaker recognition using lattice decoding. In *Proceedings of IEEE ICASSP 2005*, 1:169–172.

Hsiao-Wuen Hon, Baosheng Yuan, Yen-Lu Chow, S. Narayan, and Kai-Fu Lee. 1994. Towards large vocabulary Mandarin Chinese speech recognition. In *Proceedings of IEEE ICASSP 1994*, 1:545–548.

David Anthony James and Steve J. Young. 1994. A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of ICASSP 1994*, 1:377–380.

David Anthony James. 1995. *The Application of Classical Information Retrieval Techniques to Spoken Documents.* Ph. D. thesis, University of Cambridge.

Frederick Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397.

Gareth J. F. Jones, Jonathan T. Foote, Karen Spärck Jones, and Steve J. Young. 1996. Retrieving spoken documents by combining multiple index sources. In *Proceedings of SIGIR 1996*, pages 30–38.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164.

David J. C. MacKay and Linda C. Bauman Peto. 1994, A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19.

Jonathan Mamou, David Carmel, and Ron Hoory. 2006. Spoken document retrieval from call-center conversations. In *Proceedings of SIGIR 2006*, pages 51–58.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 1998. A rational design for a weighted finite-state transducer library. *Lecture Notes in Computer Science*, 1436:144–158.

National Institute of Standards and Technology. 2000. TREC-9 SDR Track web site. www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm.

Julian James Odell. 1995. *The Use of Context in Large Vocabulary Speech Recognition*. Ph. D. thesis, Cambridge University Engineering Department.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*, pages 275–281.

Murat Saraclar and Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *Proceedings of HLT-NAACL 2004*, pages 129–136.

Matthew A. Siegler. 1999. *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance.* Ph. D. thesis, Carnegie Mellon University.

Fei Song and W. Bruce Croft. 1999. A general language model for information retrieval. In *Proceedings of CIKM 1999*, pages 316–321.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of ICSLP*, 2:901–904.

Andy Tuerk, Sue E. Johnson, Pierre Jourlin, Karen Spärck Jones, and Philip C. Woodland. 2001. The Cambridge University multimedia document retrieval demo system. *International Journal of Speech Technology*, 4(3–4):241–250.

Peng Yu, Kaijiang Chen, Lie Lu, and Frank Seide. 2005. Searching the audio notebook: keyword search in recorded conversations. In *Proceedings of HLT/EMNLP 2005*, pages 947–954.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.