Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Speech Communication 42 (2004) 93–108
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Statistical language model adaptation: review and perspectives

Jerome R. Bellegarda *

Spoken Language Group, Apple Computer, Inc., MS-302-2LF, 2 Infinite Loop, Cupertino, CA 95014, USA

Abstract

Speech recognition performance is severely affected when the lexical, syntactic, or semantic characteristics of the discourse in the training and recognition tasks differ. The aim of language model adaptation is to exploit specific, albeit limited, knowledge about the recognition task to compensate for this mismatch. More generally, an adaptive language model seeks to maintain an adequate representation of the current task domain under changing conditions involving potential variations in vocabulary, syntax, content, and style. This paper presents an overview of the major approaches proposed to address this issue, and offers some perspectives regarding their comparative merits and associated trade-offs.

1. Introduction

Language modeling plays a pivotal role in automatic speech recognition. It is variously used to constrain the acoustic analysis, guide the search through multiple (partial) text hypotheses, and/or contribute to the determination of the final transcription (Bahl et al., 1983; Jelinek, 1985; Rabiner et al., 1996). Fundamentally, its function is to encapsulate as much as possible of the syntactic, semantic, and pragmatic characteristics for the task considered.

1.1. Focus

In the search, the successful capture of this information is critical to help determine the most likely sequence of words spoken, because it quantifies which word sequences are acceptable in a given language for a given task, and which are not. In that sense, language modeling can be thought of as a way to impose a collection of constraints on word sequences. Since, generally, many different such sequences can be used to convey the same information, these constraints tend to be statistical in nature (Gorin, 1995). Thus, regularities in natural language are governed by an underlying (unknown) probability distribution on word sequences. The ideal outcome of language modeling, then, would be to derive a good estimate of this distribution.

In the unrestricted case, however, carrying out this task is not feasible: some simplifications are necessary to render the problem tractable. The standard approach is to constrain allowable word sequences to those that can be parsed under the control of a probabilistic context-free grammar (PCFG), a somewhat crude yet well-understood model of natural language (Church, 1987). Unfortunately, because parser complexity tends to be nonlinear, at the present time context-free parsing is simply not practical for any but the most

* Tel.: +1-408-9747647; fax: +1-408-9740979.
E-mail address: jerome@apple.com (J.R. Bellegarda).

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2003.08.002

rudimentary applications. The problem is then restricted to a subclass of PCFGs, strongly regular grammars, which can be efficiently mapped into equivalent (weighted) finite state automata with much more attractive computational properties. This has led to a large body of literature exploiting such properties in finite state transducers (Mohri, 2000).

Situations where such stochastic automata are especially easy to deploy include relatively self-contained, constrained vocabulary tasks (Pereira and Riley, 1997). This is often the case, for example, of a typical dialog state in a dialog system. At that point, given one or more input strings as input, the goal is to reestimate state transition probabilities pertaining only to the input set, so input string matching on a finite automaton is a convenient solution. In dictation and other large vocabulary applications, however, the size and complexity of the task complicates the issue of coverage. Generic stochastic automata indiscriminately accepting variable length sequences become unsuitable. To maintain tractability, attention is further restricted to a subclass of probabilistic regular grammars, stochastic n-gram models. Such models have been most prominently used with $n = 2$ and 3, corresponding to classical statistical bigrams and trigrams (Jelinek, 1985).

This leads to the focus of this paper. Statistical n-grams, of course, can also be represented by equivalent (n-gram) finite state automata. In practice, the difference between the stochastic automaton and the original n-gram representation is largely a matter of implementation. Many systems in use today, especially for complex dialog applications, are based on the former (cf., for example, Riccardi and Gorin, 2000; Zue et al., 2000), while the latter is more prevalent amongst transcription systems (see e.g., Adda et al., 1999; Ohtsuki et al., 1999). In what follows, since the discussion is essentially unaffected by implementation details, the terminology "statistical language model" (SLM) will refer to the general concept of a stochastic n-gram.

1.2. Why adaptation?

Natural language is highly variable in several aspects.

First, language evolves as does the world it seeks to describe: contrast the recent surge of the word "proteomics" to the utter demise of "ague" (a burning fever, from Leviticus 26:16, King James translation of the Bible). The effective underlying vocabulary changes dynamically with time on a constant basis.

Second, different domains tend to involve relatively disjoint concepts with markedly different word sequence statistics: consider the relevance of "interest rate" to a banking application, versus a general conversation on gaming platforms. A heterogeneous subject matter drastically affects the underlying semantic characteristics of the discourse at topic boundaries.

Third, people naturally adjust their use of the language based on the task at hand: compare the typical syntax employed in formal technical papers to the one in casual e-mails, for example. While the overall grammatical infrastructure may remain invariant, syntactic clues generally differ from one task to the next.

And finally, people's style of discourse may independently vary due to a variety of factors such as socio-economic status, emotional state, etc. This last effect, of course, is even more pronounced on spoken natural language.

As a result of this inherent variability, the lexical, syntactic, or semantic characteristics of the discourse in the training and recognition tasks are quite likely to differ. This is bad news for n-gram modeling, as the performance of any statistical approach always suffers from such mismatch. SLMs have indeed been found to be extremely brittle across domains (Rosenfeld, 2000), and even within domain when training and recognition involve moderately disjoint time periods (Rosenfeld, 1995). The unfortunate outcome is a severe degradation in speech recognition performance compared to the ideal matched situation.

It turns out, for example, that to model casual phone conversation, one is much better off using two million words of transcripts from such conversations than using 140 million words of transcripts from TV and radio broadcasts. This effect is quite strong even for changes that seem trivial to a human: a language model trained on Dow–Jones newswire text sees its perplexity *doubled* when

applied to the very similar Associated Press newswire text from the same time period (Rosenfeld, 1996, 2000).

In addition, linguistic mismatch is known to affect cross-task recognition accuracy much more than acoustic mismatch. For instance, results of a cross-task experiment using Broadcast News models to recognize TI-digits, recently reported in (Lefevre et al., 2001), show that only about 8% of the word error rate increase was due to the acoustic modeling mismatch, while 92% was imputable to the language model mismatch. In a similar experiment involving ATIS, these figures were approximately 2% and 98%, respectively (Lefevre et al., 2001). Analogous trends were observed in (Bertoldi et al., 2001) for different tasks in a different language.

1.3. Organization

The above discussion makes a strong case for SLM adaptation, as a means to reduce the degradation in speech recognition performance observed with a new set of operating conditions (Federico and de Mori, 1999). The various techniques that have been proposed to carry out the adaptation procedure can be broadly classified into three major categories. Where a particular technique falls depends on whether its underlying philosophy is based on: (i) model interpolation, (ii) constraint specification, or (iii) meta-information extraction. The latter category refers to knowledge about the recognition task which may not be explicitly observable in the word sequence itself. This includes the underlying discourse topic, general semantic and syntactic information, as well as a combination thereof.

The paper is accordingly organized as follows. The next section poses the adaptation problem and reviews the various ways to gather suitable adaptation data. Section 3 covers interpolation-based approaches, including dynamic cache models. In Section 4, we describe the use of constraints, as typically specified within the maximum entropy framework. Section 5 gives an overview of topic-centered techniques, starting with adaptive mixture n-grams. Alternative integration of semantic knowledge, i.e., triggers and latent semantic analysis, is discussed in Section 6. Section 7 addresses the use of syntactic infrastructure, as implemented in the structured language model, and Section 8 considers the integration of multiple knowledge sources to further increase performance. Finally, in Section 9 we offer some concluding remarks and perspectives on the various trade-offs involved.

2. Adaptation framework

The general SLM adaptation framework is depicted in Fig. 1. Two text corpora are considered: a (small) *adaptation* corpus A, relevant to the current recognition task, and a (large) *background* corpus B, associated with a presumably related but perhaps dated and/or somewhat different task, as discussed above.

2.1. Adaptation problem

Given a sequence of N words w_q $(1 \leqslant q \leqslant N)$ consistent with the corpus A, the goal is to compute a suitably robust estimate of the language model probability:

$$\Pr(w_1, \ldots, w_N) = \prod_{q=1}^{N} \Pr(w_q | h_q), \tag{1}$$

where h_q represents the history available at time q. In the general case the expression (1) is an exact equality. For an n-gram model, the Markovian assumption implies

$$h_q = w_{q-n+1}, \ldots, w_{q-1} \tag{2}$$

and the expression (1) becomes an approximation. In any event, the estimation of $\Pr(w_1, \ldots, w_N)$

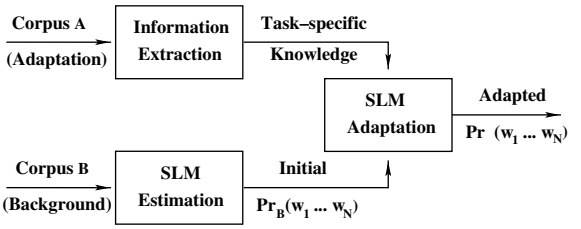

Fig. 1. General framework for SLM adaptation.

leverages two distinct knowledge sources: (i) the well-trained, but possibly mismatched, background SLM, which yields an initial estimate $\Pr_B(w_1, \ldots, w_N)$, and (ii) the adaptation data, which is used to extract some specific information relevant to the current task. As we will see shortly, this information may take the form of cache counts, marginal constraints, topic identity, etc. The general idea is to dynamically modify the background SLM estimate on the basis of what can be extracted from A.

In what follows, we concentrate on how to carry out this adaptation procedure, and thus assume, in particular, that all SLM probabilities are appropriately smoothed. For example, if $C_B(h_q)$ and $C_B(h_q w_q)$ denote the frequency counts of the word sequence h_q and $h_q w_q$ in B (i.e., the number of times they respectively appeared in the background corpus), then the well-known maximum likelihood estimate

$$\Pr_B(w_1, \ldots, w_N) = \prod_{q=1}^{N} \frac{C_B(h_q w_q)}{C_B(h_q)} \tag{3}$$

may lead to overfitting. It is therefore necessary to smooth the resulting probability (see e.g., Chen and Rosenfeld, 2000) in order to obtain a robust background SLM.

Since the adaptation procedure depends critically on the quality of the available adaptation data, we first review the various ways to gather a suitable corpus A.

2.2. Adaptation data

In some cases, the corpus A may already be available. For instance, in cross-domain adaptation, a small amount of domain-specific text may have been collected for some other purpose, but can readily serve as adaptation data. Alternatively, in the course of deploying dialog systems, it is customary to conduct "wizard of Oz" experiments to fine tune certain dialog states (cf. Riccardi and Gorin, 2000), which can often lead to relevant adaptation material.

If the corpus A is not available a priori, or deemed too small, appropriate text needs to be gathered. When the recognition task is covered by a grammar, the grammar itself can be used to generate expected user utterances. Monte Carlo methods are then used to randomly choose rules at the branching points of the grammar and thereby create an artificial corpus A, from which, for example, an n-gram can be trained. If a small amount of adaptation data is available, this can be used to weigh the different rules and thereby obtain a more realistic corpus, as in the case of the bootstrap SLM (Kellner, 1998).

If it is not practical to either collect or artificially generate the corpus A, the only other solution is to accumulate it during the recognition process itself. One approach that turns out to be quite efficient is to use multiple sentence hypotheses from an N-best list as adaptation material. For that, each hypothesis gets a weight derived from its a posteriori likelihood such that the weights add up to 1. Every sentence now contributes to the corpus A according to its weight. The idea behind this approach is that well-understood parts of a sentence occur in most hypotheses of an N-best list, whereas for misrecognitions different candidates usually appear in different hypotheses. Thus, the effect of a recognition error is distributed over several competing hypotheses and does not result in a strong error reinforcement (Souvignier et al., 2000). A variation on the same concept was recently presented in (Gretter and Riccardi, 2001), based on exploiting sausages rather than N-best lists.

Finally, once enough adaptation material has been accumulated to suitably characterize the new domain, it can be the basis for gathering "similar" data. For example, it is possible to leverage document retrieval techniques to increase the size of A by dynamically searching online databases or the world wide web (Berger and Miller, 1998; Iyer and Ostendorf, 1999; Zhu and Rosenfeld, 2001). Alternatively, data augmentation can also be performed at the n-gram count rather than the word level: a preliminary investigation along these lines was recently reported in (Janiszek et al., 2001). Ultimately, some form of data updating has to be entertained, to counteract the inability of any corpus to reflect events which postdate it.

3. Model interpolation

In interpolation-based approaches, the corpus A is used to derive a task-specific (*dynamic*) SLM, which is then combined with the background (*static*) SLM. This appealingly simple concept provides fertile grounds for experimentation, depending on the level at which the combination is implemented.

3.1. Model merging

The most straightforward combination is at the model level, which entails the estimation of $\Pr_A(w_1, \ldots, w_N)$, the language model probability derived from A for the word sequence under consideration. Or course, because of the extremely limited amount of data involved, the dynamic SLM tends to be poorly trained, most of the time resulting in a rather inaccurate estimate. Only for certain idiosyncratic word sequences, particularly frequent in the current task, may the dynamic model outperform the initial estimate $\Pr_B(w_1, \ldots, w_N)$ obtained from the background SLM. This provides the justification for merging the two estimates, in order to take advantage of the new information as appropriate.

The simplest way to do so is via linear interpolation. Given estimates for word w_q denoted by $\Pr_A(w_q|h_q)$ and $\Pr_B(w_q|h_q)$, each term in the right hand side of (1) can be expanded as follows:

$$\Pr(w_q|h_q) = (1-\lambda)\Pr_A(w_q|h_q) + \lambda \Pr_B(w_q|h_q), \quad (4)$$

where $0 \leqslant \lambda \leqslant 1$ serves as the interpolation coefficient (sometimes also referred to, for reason to become clear shortly, as the mixture coefficient). This parameter may be a function of the word history h_q, or, for better performance, some equivalence class thereof. It is typically estimated on held-out data from (a subset of) A. This can be done empirically, or under the maximum likelihood criterion using the EM algorithm (see e.g., Federico and de Mori, 1999); in the absence of held-out data, the EM algorithm can also be applied in leaving-one-out mode.

Alternatively, it is possible to back-off from the dynamic estimate to the static one depending on the associated frequency count. This is referred to as the fill-up technique (Besling and Meier, 1995), one implementation of which is

$$\Pr(w_q|h_q) = \begin{cases} \Pr_A(w_q|h_q) & \text{if } C_A(h_q w_q) \geqslant T; \\ \beta \Pr_B(w_q|h_q) & \text{otherwise,} \end{cases} \quad (5)$$

where T is an empirical threshold, and the back-off coefficient β is calculated to ensure that $\Pr(w_q|h_q)$ is a true probability.

These two closely related approaches display trade-offs that are identical to those observed with the well-known techniques of interpolation and back-off for (static) language model smoothing (Chen and Rosenfeld, 2000). More generally, this framework has given rise to a large number of variants, depending on the way the corpus A is accumulated, the specific form of either $\Pr_A(w_q|h_q)$ or $\Pr_B(w_q|h_q)$, and the particular method selected for estimation and/or smoothing. For example, it is common to use class-based models (see e.g., Jardino, 1996) as well as variable-length models (cf. Niesler and Woodland, 1996) for either $\Pr_A(w_q|h_q)$ or $\Pr_B(w_q|h_q)$. A large body of work has also experimented with various hybrids, for example by combining n-grams with probabilistic finite-state automata (e.g., Galescu and Allen, 2000; Nasr et al., 1999).

3.2. Dynamic Cache models

A special case of linear interpolation, widely used for within-domain adaptation, deserves special mention: dynamic cache memory modeling (Kuhn and de Mori, 1990). Cache models exploit self-triggering words inside the corpus A to capture short-term (dynamic) shifts in word-use frequencies which cannot be captured by the background model. In other words, they correspond to the unigram case ($n = 1$) of the general model merging strategy just discussed.

In an effort to propagate the power of the method to higher order cases, the cache memory concept has been extensively applied in conjunction with a (usually syntactic) class model of the form:

$$\Pr(w_q|h_q) = \sum_{\{c_q\}} \Pr(w_q|c_q) \Pr(c_q|h_q), \quad (6)$$

where $\{c_q\}$ is a set of possible classes for word w_q, given the current history h_q. The language model probability thus comprises a class n-gram component—$\Pr(c_q|h_q)$—and a class assignment component—$\Pr(w_q|c_q)$. The class n-gram component is assumed to be task independent, and is therefore taken from the background SLM, i.e., $\Pr(c_q|h_q) = \Pr_B(c_q|h_q)$. The class assignment component, on the other hand, is subject to dynamic cache adaptation, resulting in

$$\Pr(w_q|c_q) = (1-\lambda)\Pr_A(w_q|c_q) + \lambda \Pr_B(w_q|c_q), \quad (7)$$

where λ is estimated in the same manner as before. Further improvements have been reported when this parameter is made a function of recency (Clarkson and Robinson, 1997).

3.3. MAP adaptation

More recently, it has been argued that the combination should be done at the frequency count level rather than the model level. The underlying framework is the maximum a posteriori (MAP) criterion, which can be regarded as a more principled way to combine static and dynamic SLM information (Chen and Huang, 1999; Federico, 1996; Masataki et al., 1997). In this approach, the MAP-optimal model $\boldsymbol{M}^*$ is computed as

$$\boldsymbol{M}^* = \arg\max_{\boldsymbol{M}} \Pr(A|\boldsymbol{M}) \Pr(\boldsymbol{M}), \quad (8)$$

where $\Pr(\boldsymbol{M})$ is a prior distribution over all models in a particular family of interest, whose role is to penalize models by the amount from which they diverge from the background model.

This framework leads to a solution of the form:

$$\Pr(w_q|h_q) = \frac{\varepsilon C_A(h_q w_q) + C_B(h_q w_q)}{\varepsilon C_A(h_q) + C_B(h_q)}, \quad (9)$$

where ε is a constant factor which is estimated empirically. This factor was originally introduced to reduce the bias of the estimator (Federico, 1996). It was later made a function of the word history h_q, and used as a tuning parameter to exploit various domain-specific properties (Chen and Huang, 1999).

4. Constraint specification

In approaches based on constraint specification, the corpus A is used to extract features that the adapted SLM is constrained to satisfy. This is arguably more powerful than model interpolation, since in this framework a different weight could presumably be assigned separately for each feature.

4.1. Exponential models

Historically, constraint-based methods have been associated with exponential models trained using the maximum entropy (ME) criterion. This leads to minimum discrimination information (MDI) estimation (Della Pietra et al., 1992). Typically, features extracted from the training corpus are considered to be constraints set on single events of the joint probability distribution (such as, for example, a word and a history), in such a way that the constraint functions obey the marginal probabilities observed in the data. (Additional constraints for the never observed words may then be added for each history, as necessary to satisfy optimization conditions.)

To illustrate, rather than deriving the conditional probability $\Pr(w_q|h_q)$ directly, consider the associated event of the joint probability distribution, corresponding to the particular word $w = w_q$ and the history $h = h_q$. Assume further that this joint distribution is constrained by K linearly independent constraints, written as

$$\sum_{\{(h,w)\}} I_k(h,w) \Pr(h,w) = \alpha(\hat{h}_k \hat{w}_k), \quad 1 \leqslant k \leqslant K, \quad (10)$$

where I_k is the indicator function of an appropriate subset of the sample space (selecting the appropriate feature $\hat{h}_k \hat{w}_k$), and $\alpha(\hat{h}_k \hat{w}_k)$ denotes the relevant empirical marginal probability. Constraints like (10) are usually set only where marginal probabilities can be reliably estimated, i.e., for

well-observed events $\hat{h}_k\hat{w}_k$. Similar constraints can be placed using all lower order n-gram histories $h = w_{q-i}, \ldots, w_{q-1}$, $1 \leqslant i \leqslant n-2$.

It can be shown (Darroch and Ratcliff, 1972) that the joint distribution $\Pr(h, w)$ satisfying the constraints (10) belongs to the exponential family (Mood et al., 1974). It has the parametric form:

$$\Pr(h, w) = \frac{1}{Z(h, w)} \prod_{k=1}^{K} \exp\{\lambda_k I_k(h, w)\}, \tag{11}$$

where λ_k is the MDI parameter associated with the kth linear constraint in (10), and $Z(h, w)$ is a suitable normalization factor. The λ parameters are typically trained using the generalized iterative scaling (GIS) algorithm (Darroch and Ratcliff, 1972). This algorithm converges to a unique solution, provided the constraints (10) are consistent. Recent improvements on the GIS algorithm have been reported in (Della Pietra et al., 1997). The reader is referred to Lafferty and Suhm (1995) and Rosenfeld (1996) for additional information on MDI estimation.

4.2. MDI adaptation

Exploiting exponential models in an adaptation context is referred to as MDI adaptation. In that approach, the features extracted from A are considered as important properties of the adaptation data, that the joint n-gram distribution $\Pr(h, w)$ is requested to match, in the same manner as before. But, in addition, the solution has to be close to the joint background distribution $\Pr_B(h, w)$, taken as prior distribution. This is achieved by minimizing the Kullback–Leibler distance from the joint background distribution:

$$\min_{Q(h,w)} \sum_{\{(h,w)\}} Q(h, w) \log \frac{Q(h, w)}{\Pr_B(h, w)}, \tag{12}$$

while simultaneously satisfying the linear constraints:

$$\sum_{\{(h,w)\}} I_k(h, w) Q(h, w) = \alpha_A(\hat{h}_k\hat{w}_k), \quad 1 \leqslant k \leqslant K, \tag{13}$$

where the notation α_A emphasizes the fact that the relevant empirical marginal probabilities are now obtained from the adaptation corpus A.

It can be shown (Darroch and Ratcliff, 1972) that the solution of (12) subject to the constraints (13) has the parametric form:

$$\Pr(h, w) = \frac{\Pr_B(h, w)}{Z(h, w)} \prod_{k=1}^{K} \exp\{\lambda_k I_k(h, w)\}, \tag{14}$$

where λ_k is now the MDI parameter associated with the kth linear constraint in (13). Note that without the initial probability distribution $\Pr_B(h, w)$, the solution reduces to the maximum entropy model (11) for the features $I_k(h, w)$. The reader is referred to (Federico, 1999; Rao et al., 1997; Reichl, 1999) for further details on MDI adaptation. For a discussion of computational considerations in the context of log-linear models, see also Kneser et al. (1997) and Peters and Klakow (1999).

4.3. Unigram constraints

An important special case is MDI adaptation with unigram constraints. Given the typically small amount of adaptation data available, it is often the case that only unigram features can be reliably estimated on the adaptation corpus A. This situation is the MDI counterpart of the usual dynamic cache implementation discussed in the previous section. In that case, the constraints become

$$\sum_{\{(h,w)\}} I_k(h, w) Q(h, w) = \alpha_A(\hat{w}_k), \quad 1 \leqslant k \leqslant K, \tag{15}$$

where $\alpha_A(\hat{w}_k)$ now represents the empirical unigram probability obtained from A for the feature $\hat{w}_k$.

Dropping the dependence on the word history considerably simplifies the GIS algorithm, and in fact it can be shown that the resulting solution reduces to the closed form (Federico, 1999):

$$\Pr(h, w) = \Pr_B(h, w) \frac{\alpha_A(w)}{\Pr_B(w)}. \tag{16}$$

Note that this is effectively the direct solution to the context-independent optimization problem, for

which there is no need to resort to the maximum entropy framework. In that case, the adapted SLM is simply the background SLM scaled by the scaling factor $\alpha_A(w)/\Pr_B(w)$, which can be interpreted as the adjustment in unigram probability which is necessary to match the adaptation data. Interestingly, performance improvements have been reported by exponentially smoothing this scaling factor, so as to reduce the effect of the adaptation distribution $\alpha_A(w)$ (Kneser et al., 1997). This is analogous to what occurs in the Bayesian adaptation framework (Chen and Huang, 1999; Federico, 1996; Masataki et al., 1997) described in Section 3.3.

5. Topic information

In approaches exploiting the general topic of the discourse, the corpus A is used to extract information about the underlying subject matter. This information is then used in various ways to improve upon the background model based on semantic classification.

5.1. Mixture models

The simplest approach is based on a generalization of linear interpolation (4) to include several pre-defined domains. Consider a (possibly large) set of topics $\{t_k\}$, usually from a hand-labelled hierarchy, which covers the relevant semantic space of the background corpus B. Assume further that the background n-gram model is composed of a collection of K sub-models, each trained on a separate topic. One of these models is often the "general" n-gram model itself, trained on the entire corpus B. The others are smaller models trained on the relevant portions of B (Kneser and Steinbiss, 1993).

Mixture SLMs linearly interpolate these K n-grams in such a way that the resulting mixture best matches the adaptation data A. This is close in spirit to the original mixture approach used in acoustic modeling (see e.g., Bellegarda and Nahamoo, 1990). The SLM probability is obtained as

$$\Pr(w_q|h_q) = \sum_{k=1}^{K} \lambda_{A,k} \Pr_{B,k}(w_q|h_q), \tag{17}$$

where $\Pr_{B,k}$ refers to the kth pre-defined topic sub-model, and the notation $\lambda_{A,k}$ for the interpolation coefficients reflects the fact that they are estimated on A, typically using maximum likelihood as mentioned previously. Since the parameters $\lambda_{A,k}$ can be reliably estimated using only a comparatively small amount of data, this approach enables a better estimation of the overall probability. In essence, mixture models provide the framework necessary to extend the dynamic cache memory technique to the general case $n > 1$.

There are many variations on this basic concept, depending on how the topic clusters are formed, and how the interpolation coefficients $\lambda_{A,k}$ are estimated (cf. e.g., Adda et al., 1999; Donnelly et al., 1999; Seymore and Rosenfeld, 1997). For example, the interpolation can be performed at the sentence rather than the word level (Iyer et al., 1994). Adaptive mixtures have also been used in addition to the original cache framework (Clarkson and Robinson, 1997). For an insightful review of mixture and cache language models (see Iyer and Ostendorf, 1999).

5.2. Practical considerations

It turns out that, in actual usage, the mixture SLM (17) is less practical than a single SLM, in part because it complicates smoothing (Adda et al., 1999). To address that issue, it is possible to simply merge the n-gram counts from the mixture model and train a single SLM on these counts. When some pre-defined topics are more appropriate than others for the recognition task at hand, the n-gram counts can be empirically weighted using some held-out data. While this can be effective, it has to be done by trial and error and cannot be easily optimized (Adda et al., 1999).

Another approach is to merge the different SLM components of the SLM mixture to create a single SLM for use in the decoder. In this solution, there are as many n-grams in the resulting SLM as there are distinct n-grams in the individual topic SLMs trained on the separate portions of B. In

addition, the single merged SLM is amenable to proper optimization and smoothing. Experimental results did not show any difference between the original SLM mixture implementation and the SLM mixture merging alternative (Adda et al., 1999).

While these solutions address implementation efficiency, the biggest drawback of the adaptive mixture approach is the inherent fragmentation of the training data which occurs when partitioning the corpus B into different topics. This has sparked interest in a different way to exploit topic information.

5.3. Explicit topic models

Mixture modeling includes topic information rather indirectly, by way of the individual topic SLMs comprising the overall background model. Another solution is to express the topic contribution more directly. Consider the language model probability (cf. Gildea and Hoffman, 1999; Schwartz et al., 1997):

$$\Pr(w_q|h_q) = \sum_{k=1}^{K} \Pr(w_q|t_k)\Pr(t_k|h_q), \quad (18)$$

where t_k is one of the K topics above. This approach is less restrictive than topic mixtures, since there is no assumption that each history belongs to exactly one topic cluster. The probabilistic interpretation of topic information, however, typically requires a conditional independence assumption on the words and the topics present in the training data (Hofmann, 1999a; Hofmann, 1999b). In addition, it is clear that (19) does not capture the local structure of the language, so it must be combined with the standard n-gram in some way (see, for example Martin et al., 1997).

Setting aside these issues, the language model probability now comprises two components: a topic n-gram—$\Pr(t_k|h_q)$—and a topic assignment – $\Pr(w_q|t_k)$. The topic n-gram is assumed to remain unaffected by new material, and is therefore taken from the background SLM, i.e., $\Pr(t_k|h_q) = \Pr_B(t_k|h_q)$. The topic assignment, on the other hand, is adapted as

$$\Pr(w_q|t_k) = (1-\lambda)\Pr_A(w_q|t_k) + \lambda \Pr_B(w_q|t_k), \quad (19)$$

where λ is estimated in the same manner as before. Note the similarity with class-based adaptation (7). The main uncertainty in this approach is the granularity required in the topic clustering procedure (Kneser and Peters, 1997). On the other hand, it has the advantage of avoiding fragmentation, as only one background model needs to be trained.

6. Semantic knowledge

Approaches taking advantage of semantic knowledge purport to exploit not just topic information as above, but the entire semantic fabric of the corpus A, so they usually involve a finer level of granularity and/or some sort of dimensionality reduction.

6.1. Triggers

With explicit topic modeling, since the background data is not fragmented, it is possible to increase the level of granularity without causing estimation difficulties. In the limit, one can envision each "topic" to be reduced to as little as a particularly judicious word pair. This would be the case, for example, of the pair ("interest", "rate") in the banking application mentioned in the Introduction. Each such word pair is known as a word trigger pair (Lau et al., 1993). These triggers are usually exploited in the framework of exponential models, each relevant pair forming a constraint to be satisfied by the adaptive SLM (Rosenfeld, 1996).

Obviously, depending on the recognition task, some word pairs are more likely than others to embody a concept relevant to the application at hand. In practice, word pairs with high mutual information are searched for inside a window of fixed duration. Unfortunately, trigger pair selection is a complex issue. The central difficulty is that if w_1 triggers w_2, and w_2 triggers w_3, then w_1 had better trigger w_3. But this is hard to enforce if (w_1, w_3) happens to be a low frequency trigger pair, because it will most likely be pruned away by the trigger pair selection algorithm. As a result,

different pairs display markedly different behavior, and the potential of low frequency word triggers is severely limited (Rosenfeld, 1996).

Still, self-triggers have been shown to be particularly powerful and robust (Lau et al., 1993), which underscores the desirability of exploiting correlations between the current word and features of the large-span history.

6.2. Latent semantic analysis

Recent work has sought to extend the word trigger concept by using a more systematic framework to handle the trigger pair selection (Bellegarda, 1998a,b, 2000a,b; Coccaro and Jurafsky, 1998). This is based on a paradigm originally formulated in the context of information retrieval, called latent semantic analysis (LSA) (Deerwester et al., 1990). This paradigm relies on the concept of a *document*, i.e., a "bag-of-words" entity forming a semantically homogeneous unit. LSA reveals meaningful associations in the language based on word-document co-occurrences, as observed in a document collection pertinent to the current task. The resulting semantic knowledge is encapsulated in a continuous vector space (LSA space) of comparatively low dimension, where all words and documents in the training data are mapped. This mapping is derived through a singular value decomposition (SVD) of the co-occurrence matrix between words and documents. Thereafter, any new word and/or document is itself mapped into a point in the LSA space, and then compared to other words/documents in the space using a simple similarity measure (Bellegarda, 2000b).

This framework is very effective at reducing the underlying dimensionality of the discourse, and thus offers promise in tracking semantic changes. Consider a shift in subject matter. Since pre- and post-shift sub-corpora are essentially disjoint, the probabilities of almost all the words in the vocabulary are likely to change. Capturing the shift at the unigram level would therefore require an inordinate number of parameters, essentially proportional to the size of the vocabulary. Yet our intuition is that only a small fraction of all probability changes is actually relevant, so the "true" dimension of the semantic shift is probably much lower. In that light, encapsulating semantic knowledge within a relatively small number of dimensions is quite appealing for adaptation.

In (Bellegarda, 1998b, 2000a), the LSA framework was embedded within the conventional n-gram formalism, so as to combine the local constraints provided by n-grams with the global constraints of LSA. The outcome is an integrated SLM probability of the form (Bellegarda, 2000b):

$$\Pr(w_q|h_q,\tilde{h}_q) = \frac{\Pr(w_q|h_q)\rho(w_q,\tilde{h}_q)}{Z(h_q,\tilde{h}_q)}, \qquad (20)$$

where $\tilde{h}_q$ denotes the global ("bag-of-words") document history, $\rho(w_q,\tilde{h}_q)$ is a measure of the correlation between the current word and this global LSA history (see Bellegarda, 2000b), and $Z(h_q,\tilde{h}_q)$ ensures appropriate normalization. The language model (20) represents, in effect, a modified n-gram SLM incorporating large-span semantic information derived through LSA. Note that in most cases $\tilde{h}_q$ will have a much larger span than the n-gram history h_q in (2). On the other hand, $\tilde{h}_q$ is unordered.

Taking advantage of (20), adaptation can proceed separately for the n-gram and the LSA coefficient $\rho(w_q,\tilde{h}_q)$. By analogy with topic-based adaptation, the latter could conceivably be obtained as

$$\rho(w_q,\tilde{h}_q) = (1-\lambda)\rho_A(w_q,\tilde{h}_q) + \lambda\rho_B(w_q,\tilde{h}_q) \qquad (21)$$

with λ estimated on A in a similar fashion. A more direct implementation might even be possible if a suitable framework can be derived for LSA adaptation. For recent progress on this front (see Bellegarda, 2001).

7. Syntactic infrastructure

Approaches leveraging syntactic knowledge make the implicit assumption that the background and recognition tasks share a common grammatical infrastructure, so that grammatical constraints are largely portable from corpus B to corpus A. The background SLM is then used for initial

syntactic modeling, and the corpus A to re-estimate the associated parameters.

7.1. Structured language models

One way to implement this strategy is to exploit the structured language model framework (Chelba and Jelinek, 2000; Jelinek and Chelba, 1999). Structured language modeling takes into account the hierarchical nature of natural language by using syntactic information specifically to determine equivalence classes on the n-gram history (Chelba et al., 1997). In this approach, the model assigns a probability $\Pr(\sigma, \pi)$ to every sentence $\sigma = w_1, \ldots, w_N$ and its every possible binary parse π. The terminals of π are the words of σ with part-of-speech tags, as obtained from a suitable parser, which is typically trained on a domain-specific treebank. Non-terminal nodes are then annotated with the headword of the phrase spanned by the associated parse sub-tree.

Recent efforts have made this framework to operate efficiently in a left-to-right manner (Jelinek and Chelba, 1999), through careful optimization of both chart parsing (Younger, 1967) and search modules (Chelba and Jelinek, 2000). This leads to the language model:

$$\Pr(w_q|\bar{h}_q) = \frac{1}{Z(\bar{h}_q)} \sum_{\{\pi_q\}} \Pr(w_q|\bar{h}_q, \pi_q) \Pr(\bar{h}_q, \pi_q), \tag{22}$$

where $\bar{h}_q$ represents the sentence history so far, $\{\pi_q\}$ denotes the set of all possible (partial) parses up to that point, and $Z(\bar{h}_q)$ ensures appropriate normalization. Note that in most cases, the span of $\bar{h}_q$ will slot between the n-gram history h_q and the LSA history $\tilde{h}_q$ defined in the previous section.

In practice, since each parse π_q gives rise to a unique headword history, it is expedient to simplify the model by conditioning only on the last $(n-1)$ headwords, say p_q, as opposed to the entire partial parse. This reduces the structured SLM to a standard n-gram with history p_q, as opposed to h_q. Of course, this also means that local information is now lost, so the resulting SLM has to be combined with the usual n-gram conditioned on h_q (cf. Chelba and Jelinek, 2000). The final model is given by

$$\Pr(w_q|\bar{h}_q, \pi_q) = \Pr(w_q|h_q, p_q), \tag{23}$$

where the dependency involves both short-span (words) and larger-span (headwords) histories.

The portability of syntactic structure, as captured by a structured background model $\Pr_B(w_q|h_q, p_q)$, was recently studied in (Chelba, 2001). It was found that using a well-trained structured SLM as background model, and re-estimating the parameters using the adaptation corpus A, produced better results than training from scratch a (more unreliable, due to the paucity of data) structured SLM on A. This bodes well for the eventual success of the structured SLM approach in new domains, where a treebank may not be available for training purposes. Of course, the main caveat in structured language modeling remains the reliance on the parser, and particularly the implicit assumption that the correct parse will in fact be assigned a high probability. This effectively requires the use of a high-performance, domain-independent parser.

7.2. Syntactic triggers

Structured language models such as (23) operate at the level of the current sentence. Recent work (Zhang et al., 1999) proposed to extend them by also exploiting the syntactic structure contained in previous sentences. In (Zhang et al., 1999), the effective history is a document, as in LSA, but it is ordered, as in structured SLMs. The extracted features are essentially the syntactic counterparts of the triggers of Section 6.1. Two kinds of triggering events are considered: those based on the knowledge of the full parse of previous sentences, and those based on the knowledge of the syntactic/semantic tags to the left of and in the same sentence as the word being predicted. These events are then integrated as constraints in an exponential model (cf. Section 4 and the discussion below).

Although not yet implemented in an adaptation context, this concept may ultimately provide the necessary framework to extend the benefits of

structured language modeling to a span greater than that of a sentence.

8. Multiple sources

In approaches exploiting multiple knowledge sources, the corpus A is used to extract information about different aspects of the mismatch between training and recognition conditions. It stands to reason that, if it is helpful to address a particular type of linguistic mismatch in isolation, performance should be even better with an integrated approach to SLM adaptation.

8.1. Combination models

A popular way to combine knowledge from multiple knowledge sources is to use exponential models (Chen et al., 1998), because the underlying maximum entropy criterion offers the theoretical advantage of incorporating an arbitrary number of features while avoiding fragmentation. In addition, the features considered in the constraint specification (13) are not restricted to functions of frequency count. As alluded to in Sections 6 and 7, they can include syntactic and semantic information as well.

For example, in (Rosenfeld, 1996), trigger pairs were encoded as features along with conventional n-grams and distance-2 n-grams. More recently, in (Wu and Khudanpur, 1999), the following combination was proposed, based on the structured SLM (23):

$$\Pr(w_q|\bar{h}_q,\pi_q) = \Pr(w_q|h_q,p_q,t_q), \tag{24}$$

where t_q corresponds to topic information extracted from the available structured SLM history, $\bar{h}_q$ (i.e., the current sentence so far). Suitable marginal constraints are then derived from the available data, and the resulting ME solution (14) is found using the GIS algorithm.

While this approach is elegant and general, it is not without weaknesses. Avoiding fragmentation is not the same as guaranteeing a robust estimation. There is a danger in throwing in too many features, as brute force constraint inclusion tends to deteriorate performance due to overtraining effects (Peters and Klakow, 1999). So the bottleneck essentially shifts to the selection of useful features to be included in the modeling. An automatic iterative procedure for selecting features from a given candidate set is described in (Della Pietra et al., 1997). However, eliciting appropriate candidate sets remains the subject of intensive research (Rosenfeld, 2000).

8.2. Whole sentence models

All SLM adaptation techniques mentioned so far exploit the chain rule (1) to focus on the adaptation of $\Pr(w_q|h_q)$, i.e., the probability distribution of a single word. It has been argued (Rosenfeld, 1997) that this may be a significant hindrance to modeling linguistic supra-structure such as person and number agreement, semantic coherence, and even length. After all, integrating external influences on the current sentence (from previous sentences, or the general topic) must be factored into the prediction of every word, which may cause small biases that could compound.

One way to address these issues is to adopt a "bag-of-features" approach to each sentence, where features are arbitrary computable properties of the entire sentence. This is the case of the whole-sentence exponential model recently proposed in (Rosenfeld, 1997; Rosenfeld et al., 2001). Using again the notation $\sigma = w_1,\ldots,w_N$, it can be written as

$$\Pr(\sigma) = \frac{\Pr_0(\sigma)}{Z}\prod_{k=1}^{K}\exp\{\lambda_k I_k(\sigma)\}, \tag{25}$$

where $\Pr_0(\sigma)$ is an initial model estimate, Z is a global constant, and the feature-selecting indicator functions $I_k(\sigma)$ are conceptually similar to those discussed in Section 4, but now capture arbitrary properties of word sequences at the sentence level. From the discussion of Section 4, it is clear that (25) can be trivially re-purposed as an adaptation technique by selecting an initial estimate trained on the corpus B and specifying feature constraints using the corpus A.

Note that in this approach, normalization is infeasible, since it involves summation over all

possible sentences. Similarly, it is impossible to explicitly compute true expectations of features, so training the model requires the use of sampling methods, e.g., based on the Monte Carlo Markov chain framework (Rosenfeld, 1997). On the other hand, (25) can effortlessly accommodate all kinds of features, whether parse-based or semantic (Rosenfeld et al., 2001; Zhu et al., 1999b).

9. Conclusion

9.1. Summary

Language model adaptation refers to the process of exploiting specific, albeit limited, knowledge about the recognition task to compensate for any mismatch between training and recognition. More generally, an adaptive language model seeks to maintain an adequate representation of the domain under changing conditions involving potential variations in vocabulary, syntax, content, and style. This involves gathering up-to-date information about the current recognition task, whether a priori or possibly during the recognition process itself, and dynamically modifying the language model statistics according to this information.

The various techniques that have been proposed to carry out the adaptation procedure can be broadly classified into three major categories. Where a particular technique falls depends on whether its underlying philosophy is based on: (i) model interpolation, (ii) constraint specification, or (iii) meta-information extraction, the latter referring to task knowledge which may not be explicitly observable in the word sequence itself.

Model interpolation approaches derive (sparse) frequency counts from the adaptation corpus and fold them into a well trained, but possibly less relevant, SLM from the background corpus. Constraint-based methods select promising marginal constraints and other properties of the domain that the background SLM should satisfy, typically within a maximum entropy framework. Techniques exploiting meta-information rely on a variety of knowledge sources to appropriately update the semantic and/or syntactic characteristics of the background SLM. These include discourse topic, semantic links, syntactic structure, and a combination thereof.

9.2. Perspectives

The present overview has attempted to show that a collection of seemingly disparate (and historically unrelated) approaches could be interpreted, in light of the general adaptation framework illustrated in Fig. 1, as addressing different (but ultimately deeply intertwined) aspects of the general adaptation problem. Given this rather eclectic inventory of techniques, however, it is legitimate to ask where the field is headed. Arguably, SLM adaptation has made substantial progress over the past decade, but the proverbial order of magnitude breakthrough remains elusive. What seems to be the most promising way to reach it?

This question opens up a new perspective on the directions discussed in Sections 5–7. While they each follow a different path, their common thread is to exploit information, be it syntactic structure or semantic fabric, which involves a fairly high degree of cognition. This is precisely the kind of knowledge that humans naturally bring to bear when processing natural language, so it can be reasonably conjectured to represent a key ingredient for success. In that light, combination techniques discussed in Section 8, whose ultimate goal is to integrate all available knowledge sources, appear most likely to harbor a potential breakthrough. It is hoped that on-going efforts to leverage such latent synergies will lead, in the not-too-distant future, to more polyvalent, multifaceted, and effective solutions for language model adaptation.

References

Adda, G., Jardino, M., Gauvain, J.L., 1999. Language modeling for broadcast news transcription. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1759–1762.

Bahl, L.R., Jelinek, F., Mercer, R.L., 1983. A maximum likelihood approach to continuous speech recognition.

IEEE Trans. Pattern Anal. Mach. Intel. PAMI-5 (2), 179–190.

Bellegarda, J.R., 1998a. Exploiting both local and global constraints for multi-span statistical language modeling. In: Proc. 1998 Internat. Conf. Acoust. Speech Signal Process., Vol. 2. Seattle, WA, May 1998, pp. 677–680.

Bellegarda, J.R., 1998b. A multi-span language modeling framework for large vocabulary speech recognition. IEEE Trans. Speech Audio Proc. 6 (5), 456–467.

Bellegarda, J.R., 2000a. Large vocabulary speech recognition with multi-span statistical language models. IEEE Trans. Speech Audio Proc. 8 (1), 76–84.

Bellegarda, J.R., 2000b. Exploiting latent semantic information in statistical language modeling. Proc. IEEE 88 (8), 1279–1296.

Bellegarda, J.R., 2001. A novel approach to the adaptation of latent semantic information. In: Proc. 2001 ISCA Workshop on Adaptation Methods, elsewhere in these Proceedings.

Bellegarda, J.R., Nahamoo, D., 1990. Tied mixture continuous parameter modeling for speech recognition. IEEE Trans. Acoust. Speech Signal Process. ASSP-38 (12), 2033–2045.

Berger, A., Miller, R., 1998. Just-in-time language modelling. In: Proc. 1998 Internat. Conf. Acoust. Speech Signal Process., Vol. 2. Seattle, WA, May 1998, pp. 705–709.

Bertoldi, N., Brugnara, F., Cettolo, M., Federico, M., Giuliani, D., 2001. From broadcast news to spontaneous dialogue transcription: portability issues. In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.

Besling, S., Meier, H.G., 1995. Language model speaker adaptation. In: Proc. 1995 Euro. Conf. Speech Comm. Technol., Madrid, Spain, September 1995, pp. 1755–1758.

Chelba, C., 2001. Portability of syntactic structure for language modeling. In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.

Chelba, C., Jelinek, F., 2000. Structured language modeling. Computer, Speech, and Language 14 (4), 283–332.

Chelba, C., Engle, D., Jelinek, F., Jimenez, V., Khudanpur, S., Mangu, L., Printz, H., Ristad, E.S., Rosenfeld, R., Stolcke, A., Wu, D., 1997. Structure and performance of a dependency language model. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 5. Rhodes, Greece, September 1997, pp. 2775–2778.

Chen, L., Huang, T., 1999. An improved MAP method for language model adaptation. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 5. Budapest, Hungary, September 1999, pp. 1923–1926.

Chen, S.F., Rosenfeld, R., 2000. A survey of smoothing techniques for ME models. IEEE Trans. Speech Audio Proc. 8 (1), 37–50.

Chen, S.F., Seymore, K., Rosenfeld, R., 1998. Topic adaptation for language modeling using unnormalized exponential models. In: Proc. 1998 Internat. Conf. Acoust. Speech Signal Process., Vol. 2. Seattle, WA, May 1998, pp. 681–684.

Church, K.W., 1987. Phonological Parsing in Speech Recognition. Kluwer Academic Publishers, New York.

Clarkson, P.R., Robinson, A.J., 1997. Language model adaptation using mixtures and an exponentially decaying cache. In: Proc. 1997 Internat. Conf. Acoust. Speech Signal Proc., Munich, Germany, May 1997, pp. 799–802.

Coccaro, N., Jurafsky, D., 1998. Towards better integration of semantic predictors in statistical language modeling. In: Proc. Internat. Conf. Spoken Language Process., Sydney, Australia, December 1998, pp. 2403–2406.

Darroch, J.N., Ratcliff, D., 1972. Generalized iterative scaling for log-linear models. Ann. Math. Statist. 43 (5), 1470–1480.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. 41, 391–407.

Della Pietra, S., Della Pietra, V., Lafferty, J., 1997. Inducing features of random fields. IEEE Trans. Pattern Anal. Mach. Intel. PAMI-19 (1), 1–13.

Della Pietra, S., Della Pietra, V., Mercer, R., Roukos, S., 1992. Adaptive language model estimation using minimum discrimination estimation. In: Proc. 1992 Internat. Confer. Acoust. Speech Signal Process., Vol. I. San Francisco, CA, April 1992, pp. 633–636.

Donnelly, P.G., Smith, F.J., Sicilia, E., Ming, J., 1999. Language modelling with hierarchical domains. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Budapest, Hungary, Vol. 4, September 1999, pp. 1575–1578.

Federico, M., 1996. Bayesian estimation methods for N-gram language model adaptation. In: Proc. 1996 Internat. Conf. Spoken Language Process., Philadelphia, PA, October 1996, pp. 240–243.

Federico, M., 1999. Efficient language model adaptation through MDI estimation. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1583–1586.

Federico, M., de Mori, R., 1999. In: Ponting, K. (Ed.), Language Model Adaptation. Springer-Verlag, New York.

Galescu, L., Allen, J., 2000. Hierarchical statistical language models: experiments on in-domain adaptation. In: Proc. 2000 Internat. Conf. Spoken Language Proc., Beijing, China, October 2000, pp. I186–I189.

Gildea, D., Hoffman, T., 1999. Topic-based language modeling using EM. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 5. Budapest, Hungary, September 1999, pp. 2167–2170.

Gorin, A.L., 1995. On automated language acquisition. J. Acoust. Soc. Amer. 97, 3441–3461.

Gretter, R., Riccardi, G., 2001. On-line learning of language models with word error probability distributions. In: Proc. 2001 Int. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.

Hofmann, T., 1999a. Probabilistic latent semantic analysis. In: Proc. Fifteenth Conf. Uncertainty in AI, Stockholm, Sweden, July 1999.

Hofmann, T., 1999b. Probabilistic topic maps: navigating through large text collections. In: Lecture Notes Comp.

Science, No. 1642. Springer-Verlag, Heidelberg, pp. 161–172.

Iyer, R., Ostendorf, M., 1999. Modeling long distance dependencies in language: Topic mixtures versus dynamic cache models. IEEE Trans. Speech Audio Process. 7 (1).

Iyer, R., Ostendorf, M., Rohlicek, J.R., 1994. Language modeling with sentence-level mixtures. In: Proc. ARPA Speech and Natural Language Workshop. Morgan Kaufmann Publishers, pp. 82–86.

Janiszek, D., de Mori, R., Bechet, F., 2001. Data augmentation and language model adaptation. In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.

Jardino, M., 1996. Multilingual stochastic n-gram class language models. In: Proc. 1996 Internat. Conf. Acoust. Speech Signal Proc., Atlanta, GA, May 1996, pp. I161–I163.

Jelinek, F., 1985. The development of an experimental discrete dictation recognizer. Proc. IEEE 73 (11), 1616–1624.

Jelinek, F., Chelba, C., 1999. Putting language into language modeling. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 1. Budapest, Hungary, September 1999, pp. KN1–KN5.

Kellner, A., 1998. Initial language models for spoken dialogue systems. In: Proc. 1998 Internat. Conf. Acoust. Speech Signal Process., Vol. 1. Seattle, WA, May 1998, pp. 185–188.

Kneser, R., Peters, J., 1997. Semantic clustering for adaptative language modeling. In: Proc. 1997 Internat. Conf. Acoust., Speech Signal Process., Vol. 2. Munich, Germany, May 1997, pp. 779–782.

Kneser, R., Steinbiss, V., 1993. On the dynamic adaptation of stochastic language models. In: Proc. 1993 Internat. Conf. Acoust. Speech Signal Process., Vol. II, Minneapolis, MN, May 1993, pp. 586–588.

Kneser, R., Peters, J., Klakow, D., 1997. Language model adaptation using dynamic marginals. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 4. Rhodes, Greece, September 1997, pp. 1971–1974.

Kuhn, R., de Mori, R., 1990. A Cache-based natural language method for speech recognition. IEEE Trans. Pattern Anal. Mach. Intel. PAMI-12 (6), 570–582.

Lafferty, J.D., Suhm, B., 1995. Cluster expansion and iterative scaling for maximum entropy language models. In: Hanson, K., Silver, R. (Eds.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, Norwell, MA.

Lau, R., Rosenfeld, R., Roukos, S., 1993. Trigger-based language models: a maximum entropy approach. In: Proc. 1993 Internat. Conf. Acoust. Speech Signal Process., Minneapolis, MN, May 1993, pp. II45–II48.

Lefevre, F., Gauvain, J.L., Lamel, L., 2001. Towards task-independent speech recognition. In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT, May 2001.

Martin, S.C., Liermann, J., Ney, H., 1997. Adaptive topic-dependent language modelling using word-based varigrams. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 3. Rhodes, Greece, September 1997, pp. 1447–1450.

Masataki, H., Sagisaka, Y., Tawahara, T., 1997. Task adaptation using MAP estimation in N-gram language model. In: Proc. 1997 Internat. Conf. Acoust. Speech Signal Process., Vol. 1. Munich, Germany, May 1997, pp. 783–786.

Mohri, M., 2000. Minimization algorithms for sequential transducers. Theor. Comp. Sci. 234 (1–2), 177–201.

Mood, A., Graybill, F., Boes, D., 1974. Introduction to the Theory of Statistics. McGraw-Hill, New York.

Nasr, A., Esteve, Y., Bechet, F., Spriet, T., de Mori, R., 1999. A language model combining N-grams and stochastic finite state automata. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 5. Budapest, Hungary, September 1999, pp. 2175–2178.

Niesler, T., Woodland, P., 1996. A Variable-length category-based N-gram language model. In: Proc. 1996 Internat. Conf. Acoust. Speech Signal Process., Atlanta, GA, May 1996, pp. I164–I167.

Ohtsuki, K., Furui, S., Sakurai, N., Iwasaki, A., Zhang, Z.-P., 1999. Recent advances in Japanese broadcast news transcription. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 2. Budapest, Hungary, September 1999, pp. 671–674.

Pereira, F.C., Riley, M., 1997. Speech recognition by composition of weighted finite automata. In: Roche, E., Schabes, Y. (Eds.), Finite-State Language Processing. MIT Press, Cambridge, MA, pp. 431–453.

Peters, J., Klakow, D., 1999. Compact maximum entropy language models. In: Proc. 1999 Autom. Speech Reco. Understanding Workshop, Keystone, CO, December 1999, pp. I253–I256.

Rabiner, L.R., Juang, B.H., Lee, C.-H., 1996. An overview of automatic speech recognition. In: Lee, C.-H., Soong, F.K., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition—Advanced Topics. Kluwer Academic Publishers, Boston, MA, pp. 1–30 (Chapter 1).

Rao, P.S., Dharanipragada, S., Roukos, S., 1997. MDI adaptation of language models across corpora. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 4. Rhodes, Greece, September 1997, pp. 1979–1982.

Reichl, W., 1999. Language model adaptation using maximum discrimination information. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1791–1794.

Riccardi, G., Gorin, A.L., 2000. Stochastic language adaptation over time and state in natural spoken dialog systems. IEEE Trans. Speech Audio Proc. 8 (1), 3–10.

Rosenfeld, R., 1995. Optimizing lexical and N-gram coverage via judicious use of linguistic data. In: Proc. 1995 Euro. Conf. on Speech Comm. Technol., Madrid, Spain, September 1995, pp. 1763–1766.

Rosenfeld, R., 1996. A maximum entropy approach to adaptive statistical language modeling. In: Computer Speech and Language, Vol. 10. Academic Press, London, pp. 187–228.

Rosenfeld, R., 1997. A whole sentence maximum entropy language model. In: Proc. 1997 Autom. Speech Reco. Understanding Workshop, Santa Barbara, CA, pp. 230–237.

Rosenfeld, R., 2000. Two decades of statistical language modeling: Where do we go from here. Proc. IEEE 88 (8), 1270–1278.

Rosenfeld, R., Chen, S.F., Zhu, X., 2001. Whole sentence exponential language models: a vehicle for linguistic–statistical integration. Computer Speech and Language 15 (1), Academic Press, London.

Schwartz, R., Imai, T., Kubala, F., Nguyen, L., Makhoul, J., 1997. A maximum likelihood model for topic classification of broadcast news. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 3. Rhodes, Greece, September 1997, pp. 1455–1458.

Seymore, K., Rosenfeld, R., 1997. Using story topics for language model adaptation. In: Proc. 1997 Euro. Conf. Speech Comm. Technol., Vol. 4. Rhodes, Greece, September 1997, pp. 1987–1990.

Souvignier, B., Kellner, A., Rueber, B., Schramm, H., Seide, F., 2000. The thoughtful elephant: Strategies for spoken dialog systems. IEEE Trans. Speech Audio Proc. 8 (1), 51–62.

Wu, J., Khudanpur, S., 1999. Combining nonlocal, syntactic and N-gram dependencies in language modeling. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 5. Budapest, Hungary, September 1999, pp. 2179–2182.

Younger, D.H., 1967. Recognition and parsing of context–free languages in time N^3. Inform. Control 10, 198–208.

Zhang, R., Black, E., Finch, A., 1999. Using detailed linguistic structure in language modeling. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1815–1818.

Zhu, X.J., Chen, S.F., Rosenfeld, R., 1999b. Linguistic features for whole sentence maximum entropy language models. In: Proc. 1999 Euro. Conf. Speech Comm. Technol., Vol. 4. Budapest, Hungary, September 1999, pp. 1807–1810.

Zhu, X., Rosenfeld, R., 2001. Improving trigram language modeling with the world wide web. In: Proc. 2001 Internat. Conf. Acoust., Speech, Signal Proc., Salt Lake City, UT, May 2001.

Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J., Hetherington, L., 2000. JUPITER: A telephone-based conversational interface for weather information. IEEE Trans. Speech Audio Proc. 8 (1), 85–96.