Style Adaptation of Statistical Language Models

Dong Hoon Van Uytsel, Patrick Wambacq and Dirk Van Compernolle

K.U.Leuven

Abstract

We present a computationally inexpensive technique to perform style adaptation: a general training text corpus is statically adapted to the style of a given target recognition task by weighted counting. The n-gram language model derived from this weighted background corpus is then used as a component of a mixture language model in a word lattice rescoring framework. We specify two types of weighted counting and evaluate their effectiveness in terms of word recognition error rate. Adapting broadcast news style to news talkshow style, these methods yield a close to insignificant improvement with respect to the unweighted case. However, adapting financial newspaper style to news talkshow style, we observed a 0.7% absolute reduction of word recognition error rate.

1 Motivation and strategy

1.1 Language and domain

In a statistical speech recognizer, the task of a language model (LM) is to associate each *sentence hypothesis* generated by the search engine with its *probability* to occur in the spoken input from the user, such that less likely partial hypotheses can be aborted as early as possible.

The term *language* refers here to the discourse in a specific situation. In the speech recognition community this is what is usually meant by *domain*. The domain determines style (e.g. formal or informal, spontaneous or well-formed), topic, channel (e.g. face-to-face or telephone, number of speakers), register and more. In this work the complex nature of discourse is simplified into two main characteristics: *style* and *topic*. Furthermore, we do not consider the problem of style and topic boundaries.[1] We define style and topic implicitly with the domain in which they occur.

It is crucial for a LM to be trained on text data that match the target domain as closely as possible in terms of style and topic. Unfortunately, in most situations, only a limited amount of training text (Ω_T) in the *target* domain (T) is available (1K–100K words). On the other hand, media such as the Internet and CD-ROMs offer enormous collections (1M–1G words) of other texts (Ω_B) in a number of different domains (B denotes the ensemble of other domains, the *background* domain).

[1] I.e. addressing the question: where does one style/topic start and where does another one begin?

1.2 Language model adaptation

The challenge is to make efficient use of both Ω_T and Ω_B. Simply extending Ω_T with Ω_B is not optimal. Deriving a language model p_B from Ω_B in the same way as p_T was derived from Ω_T and subsequent model interpolation already gives better results, since optimization of the interpolation parameters on a subset held out from Ω_T will make the interpolation weight of p_B small and will thereby effectively reduce the influence of the misfit between p_B and Ω_T.

Adaptation methods attempt to reduce this misfit by taking Ω_T and/or p_T into account when deriving or applying p_B.

Language model adaptation is a problem of (a) extracting the relevant (i.e. useful in the target domain) knowledge from different sources, and (b) combining them. There are at least two approaches to (a):

1. Corpus adaptation: adapt Ω_B to match the T domain better and extract knowledge from it in the same way it is done from Ω_T .

2. Model adaptation: train a model on Ω_B and adapt or transform it to the T domain, while making use of Ω_T or a model trained on Ω_T.

How to combine LMs (sub-problem (b)) is a study field on its own. Although it is strongly related to (a) and certainly very important, we have addressed (a) while keeping (b) fixed to context-independent linear interpolation (Jelinek and Mercer 1980).

1.3 Topic and style

A domain features a number of typical topics as well as a typical speaking style. One can talk about different things in the same style, or talk about the same thing in different styles. From recognition experiments (Yu, Clark, Malkin and Waibel 1998) it is known that a style mismatch between training and test set seriously degrades the recognition accuracy (e.g. recognition of spontaneous speech with a LM trained on 'clean' transcripts). On the other hand, recognition is reasonable as long as the topic of the test set was observed, among others, in the training set (Seymore, Chen and Rosenfeld 1998). This may be caused by the fact that style is related with patterns of function words, while topic is related with patterns of content words. The notion of function and content words was introduced by Isotani and Sagayama (1993) and Geutner (1996). Function words are closed classes (articles, prepositions, pronouns, auxiliary verbs, ...); content words are open classes (nouns, adjectives, ...). Since function words are generally very frequent, and since their typical usage heavily depends on the discourse context (style) (Iyer 1997), it is very important to model them accurately by selecting textual training data with care. Moreover, content words are typically longer and are therefore more easily acoustically disambiguated. Function words are shorter on average, are more difficult to disambiguate acoustically and rely more heavily on the LM for accurate recognition.

Most of the adaptation methods found in literature focus on topic rather than style (Iyer and Ostendorf 1996, Rao, Monkowski and Roukos 1995, Lafferty and Suhm 1995, Seymore et al. 1998), or do not make a distinction (Crespo, Tapias, Escalada and Alvarez 1997, Ries 1997, Federico, Bunnell and Idsardi 1996) by the nature of their adaptation schemes.

Iyer (1997) explicitly deals with style differences between corpora. The methods proposed in the following section can be considered as a generalization of Iyer's (1997, Ch. 4) relevance weighting scheme.

2 Weighted counting

In this section, we propose three different adaptation schemes—transformation and two variants of relevance weighting—which make use of a *weighted counting* approach. The computational cost is low and it can easily be applied to the class of n-gram models, which have a behavior that is very well known and which are most commonly used in speech recognition. An n-gram model makes the assumption that the occurrence of a word in a sentence $S = w_1 w_2 \dots w_N$ only depends on the $n-1$ preceding words:[2]

$$\begin{aligned} p(S) &= \prod_{i=1}^{N} p(w_i | w_1 \dots w_{i-1}) \\ &= \prod_{i=1}^{N} p(w_i | w_{i-n+1} \dots w_{i-1}) \end{aligned} \tag{1}$$

This assumption limits the number of parameters (the probabilities) to estimate. The maximum likelihood estimate is a relative frequency:

$$p(w_i | w_{i-n+1} \dots w_{i-1}) \simeq \frac{C(w_{i-n+1} \dots w_i)}{C(w_{i-n+1} \dots w_{i-1})} \tag{2}$$

where $C(A)$ denotes the number of times A was observed in the training text. Weighted counting implies the weighting of these n-gram counts. The n-gram is then further built up and evaluated in the conventional way.

As Iyer (1997) we assume equivalence between style and part-of-speech patterns. A "style model" in this text is nothing more than an n-gram LM of parts-of-speech. Unlike Brown, Della Pietra, de Souza, Lai and Mercer (1992), the set of parts-of-speech is augmented with the M most frequent word/tag pairs (with $M = 50..200$), in order to preserve more style information. From our experiments the choice of M appeared not to be critical.

2.1 Probability transformation

Given a background model $p_B(S)$, a small amount of training text Ω_T in the T domain, a considerable amount of training text Ω_B in the B domain, how can we transform $p_B(S)$ to match the T domain better in terms of style?

[2]Each sentence is padded with special markers in order to keep the formulas and its implementation simple: $w_{-n+2} \dots w_0$ are begin-of-sentence symbols and w_N is the end-of-sentence symbol.

Let $p_T(S)$ be the T model we would like to compute. Then

$$\begin{aligned} p_T(S) &= p_B(S)\frac{p_T(S)}{p_B(S)} \\ &= \kappa_S \cdot p_B(S) \end{aligned} \tag{3}$$

where $\kappa_S = p_T(S)/p_B(S)$ is a weighting factor; contrary to the language model weighting factor used in the recognition phase, κ_S explicitly depends on the sentence S being recognized. It is high when S matches T better than B and low in the opposite case.

κ_S can be pre-computed such that evaluation of the model during recognition is not made more complex. This is done in the following way: the maximum likelihood estimate of $p(S)$, given an observation sequence $\Omega = S_1 S_2 \ldots S_m$ is

$$p_{\mathrm{ML}}(S) = \frac{C_\Omega(S)}{m}, \tag{4}$$

in which $C_\Omega(S)$ is the number of times S occurs in Ω. This gives an approximation for $p_T(S)$ if Ω belongs to the T domain, or for $p_B(S)$ if Ω belongs to the B domain. In our problem statement, Ω is in the B domain. So we can only obtain $C_B(S)$. But by (3), $C_T(S)$ is approximated by

$$C_T(S) \simeq \kappa_S \cdot C_B(S) \tag{5}$$

This means, during training each sentence observation is weighted with κ_S, which is equivalent to browsing through the complete corpus Ω and adding κ_S to the count associated with S instead of 1 each time S is observed. Afterwards the counts are globally scaled in such a way that they sum up to m.

A similar derivation holds for documents instead of sentences. This is done by replacing the symbol S with the symbol D for document whereby the observation of all sentences within the same document D are scaled with the same factor κ_D.

Training an n-gram LM requires n-gram counts, not sentence or document counts. From (3) and (1) it follows that the n-gram counts in a sentence S or a document D consisting of N_S resp. N_D words should be weighted by a factor κ_S^{1/N_S} resp. κ_D^{1/N_D}. It is anticipated that estimating κ_D^{1/N_D} is more robust than estimating κ_S^{1/N_S}, which, in the same way, is more robust than $\kappa_{w_1^n} = p_T(w_1^n)/p_B(w_1^n)$. On the other hand, time resolution (considering the train corpus as a time series of discrete events) is enhanced by estimating κ from smaller units (n-grams instead of documents).[3]

2.2 Style relevance weighting of data

The adaptation problem can alternatively be formulated as follows: given $p_B(S)$, Ω_T and Ω_B, how can Ω_B be transformed to match the style of T closer—and yield a more T-like model?

[3] This is comparable with the time-position uncertainty principle.

A straightforward answer is to weight the observation of a sentence with its probability of being in the T domain:

$$\begin{aligned} p(T|S) &= \frac{p(S|T)p(T)}{p(S|B)p(B)+p(S|T)p(T)} \\ &= \left(\frac{p_B(S)}{p_T(S)}\frac{1-p(T)}{p(T)}+1\right)^{-1} \\ &= \left(\frac{1-p(T)}{\kappa_S p(T)}+1\right)^{-1} \end{aligned} \tag{6}$$

The a priori target style probability $p(T)$ was found not to be critical within a boundary of ± 0.2 for word accuracy.[4] Again, we could also weight the observation of a document with its probability of being in the T domain:

$$p(T|D) = \left(\frac{1-p(T)}{\kappa_D p(T)}+1\right)^{-1} \tag{7}$$

Geometric averages

The document relevance weighting scheme by Iyer (1997) is similar to (7), but replaces κ_D by κ_D^{1/N_D} where N_D is the number of words in the document. Although this formula is numerically more tractable and less sensitive to errors of the estimation of $p_T(D)$ and $p_B(D)$, it is difficult to defend theoretically. We evaluate (6), a variant thereof in which κ_S is replaced by κ_S^{1/N_S}, (7) and two variants in which κ_D^1 is resp. replaced by κ_D^{1/N_D} and κ_D^{1/M_D}, M_D being the number of sentences in D.

On the other hand, $p(T|S)$ can be considered as $\prod_i p(T|w_1 \ldots w_i)$ where $S = w_1 \ldots w_{N_S}$. Therefore the N_S-th root of (6), i.e.

$$\bar{p}(T|w_1 \ldots w_i) = \left(\frac{1-p(T)}{\kappa_S p(T)}+1\right)^{-1/N_S} \tag{8}$$

may alternatively serve as count increment for the n-grams contained within S. It is a more reliable weight than (6), but less powerful in adaptation. Once more, we can apply the same reasoning on the document level and take the N_D-th root (per n-ngram) or M_D-th root (per sentence) of (7).

2.3 Estimating κ_S and κ_D

For the computation of κ_S a model $p_T(S)$ is needed. Seemingly the solution itself is used in order to obtain it. It is however possible to compute smaller models for T and B that need a small training amount, but still yield a proper estimate for

[4] Alternatively the a priori target style probability may be recursively estimated as:

$$p(T) = \sum_{S \in \text{all domains}} p(T|S)p(S) \simeq \sum_{l=1}^{m} p(T|S_l)$$

where $\Omega_B = S_1 \ldots S_m$. This has not been implemented in this work though.

κ_S. Here we exploit this with *part-of-speech* based models, assuming that B and T contain similar topics, and only style has to be compensated for.

Parts-of-speech-based (POS) models are of the form

$$p(w_n|w_1 \dots w_{n-1}) = p(c_n|c_1 \dots c_{n-1}) \cdot p(w_n|c_n) \tag{9}$$

where $c_1 \dots c_{n-1}$ is the most likely POS annotation for the word sequence $w_1 \dots w_{n-1}$, $c_n = \arg\max_c p(c|c_1 \dots c_{n-1})$. They assume independence between the POS sequence (the POS n-gram model $p(c_n|c_1 \dots c_{n-1})$) and the marginal distribution $p(w_n|c_n)$. Iyer (1997) observes that the POS n-gram model reflects a rather intuitive concept of style, while the marginal distribution accounts for the topic. Therefore POS-based LM and style are considered more or less equivalent in the rest of this paper, although this may not be a valid assumption when more fine-grained style distinctions have to be made.

Now for a sentence $S = w_1 \dots w_N$:

$$\begin{aligned} \kappa_S &= \frac{p_T(w_1 \dots w_N)}{p_B(w_1 \dots w_N)} \\ &= \prod_{i=1}^{N} \frac{p_T(c_i|c_{i-n+1} \dots c_{i-1}) p_T(w_i|c_i)}{p_B(c_i|c_{i-n+1} \dots c_{i-1}) p_B(w_i|c_i)} \\ &= \prod_{i=1}^{N} \frac{p_T(c_i|c_{i-n+1} \dots c_{i-1})}{p_B(c_i|c_{i-n+1} \dots c_{i-1})} \end{aligned} \tag{10}$$

since $p_T(w|c) \simeq p_B(w|c)$ assuming topic similarity. POS-based n-gram LMs may be trained on small amounts of (pre-tagged) training text.

In fact, any model that would give a proper estimate for κ_S highlights the mismatch between T and B and can be trained on a minimal amount of training. For our purposes however, we limited ourselves to POS trigrams.

3 Experiment setup

3.1 Data selection and preprocessing

In this subsection the target domain (news talkshows) and two background domains (broadcast news and financial newspaper) are defined. In Table 1 we brought together a few sample sentences from each domain.

3.1.1 Target domain: the Newshour corpus (NH)

As the recognition task the automatic transcription of news talkshows was chosen. Transcriptions and audio waveforms from Newshour, a typical news talkshow, are made available through the web.

Transcriptions of 300 conversations were collected, consisting of 32,000 sentences. 20,000 sentences totalling 362,833 words were randomly selected for training, the rest was reserved for cross validation and testing. Three shows not included within the training set were selected for automatic recognition testing. The audio files were segmented by hand into speaker turns.

Table 1: A few sample sentences in the specified target and background domains.

Newshour: news talkshows

```
i think it's a very difficult moral question.
all right.
but it was to the most troubling part of the past that mister clinton returned today goree island
off senegal's coast where millions of africans began forced journeys to the new world that ended
either in their death or in slavery on plantations in america or islands in the caribbean.
```

Broadcast News Corpus: broadcast news

```
the black vote can be pivotal and this year more black voters are asking that they be taken
seriously.
this is also an election day in dade county florida nobody cares much no polling places set up by
city or county officials anywhere near this tent city.  perot trails with sixteen percent.
```

Wall Street Journal: financial newspaper

```
much smaller portions of the business were divvied up among other current r. j. r. agencies with
the bulk of the rest going to two units of interpublic group of companies which is also a major r.
j. r. agency.
the diversified energy concern said in houston that the loan is unsecured and is payable in full
on november seventh nineteen ninety five.
while mr. drucker names at least ten effective leaders not one is a woman.
```

Most of the conversations are held between two to four speakers, of which one is the moderator. The other speakers give background information and opinions on recent news events. Very often the speakers agree and there is no debate. The conversations are not real in the sense that the targeted audience is usually the t.v. watching public, not the other side of the table.

The sentence length varies greatly. Moderator turns are typically very short and are often incomplete clauses in the grammatical sense. The other speakers tend to compress all the information they want to share in their turn into one sentence by coordinating shorter clauses.

There is a moderate amount of discourse markers and fillers such as "kind of" and "uh" and repetitions, but the speech is much more controlled and less spontaneous than the speech of the Switchboard task (Godfrey, Holliman and McDaniel 1992).

3.1.2 Background domain: the Broadcast News corpus (BN)

As the first background domain, the Broadcast News corpus was selected. We only used the 1996 part of the language model training data for reasons of technical tractability. This part consists of 8,510 documents summing up to 543,579 sentences or 8,193,083 words.

BN approaches the NH well in terms of topic. The style differs slightly since most of the speakers are well-trained. The information is denser and sentences shorter. Although, judging by inspection, sentences are more complex than NH's. They are more well-formed and contain less discourse markers than NH. Conversations are only a part of the corpus, and most often these conversations are "fake" (e.g. a newsreader talking to an anchorman), very much like in NH.

3.1.3 Another background domain: the Wall Street Journal corpus (WSJ)

We have conducted the same set of experiments with the very different Wall Street Journal Corpus. We used the 1989 part of the language model training data, which count 11,870 documents (i.e. articles); this is 241,469 sentences or 5,222,189 words.

Topic as well as style differ greatly from the two other corpora. WSJ focuses primarily on business and economy, while BN and NH report on news of a more general kind.

Concerning style, WSJ language is usually very formal and well-formed. It does not contain discourse markers. The sentences are longer and more complex than the sentences of NH and BN. The information density is very high. Indirect speech and raising (" ... industry sources said", "it seems ... ") are very frequently used.

3.2 POS tagging and POS n-grams

All the training data was automatically annotated with part-of-speech tags using Brill's transformation-based tagger (Brill 1992). The tag set consisted of part-of-speech tags and tags for fillers and hesitations. The set was augmented with all possible word/tag pairs of the 200 most frequent words of BN, in order to keep valuable style-related information in the POS-based model (Fig. 1). The resulting tag set counted 540 tags.

A./NN	AFF	EX-AUX	NNPS	UH
A/DT	ANA	FW	NNS	VB
ABOUT/JJ	AUX	GW	PDT	VBA
AFTER/NNP	AUX-N	JJ	PREP	VBD
AGAIN/NNP	CC	JJA	PRP	VBG
AGAIN/RB	CCC	JJR	PRP$	VBN
...	CD	JJS	PRPA	VBP
YES/NNP	CV	NEG	RB	VBZ
YES/UH	DT	NN	RBA	WDT
YOU'RE/PRP	DT-AUX	NNA	RBR	WP
YOU/PRP	EOS	NNP	RBS	WRB
YOUR/PRP$	EX	NNP.S.	RP	XX

Figure 1: Augmented POS set.

POS n-grams are obtained by tagging the background resp. target corpus. In the resulting stream of word/tag pairs, the pairs for which the word is not among the 200 most frequent are replaced by the corresponding tag. A back-off *n*-gram model (Katz 1987) is derived from this modified stream and optimized as proposed by Kneser and Ney (1995).

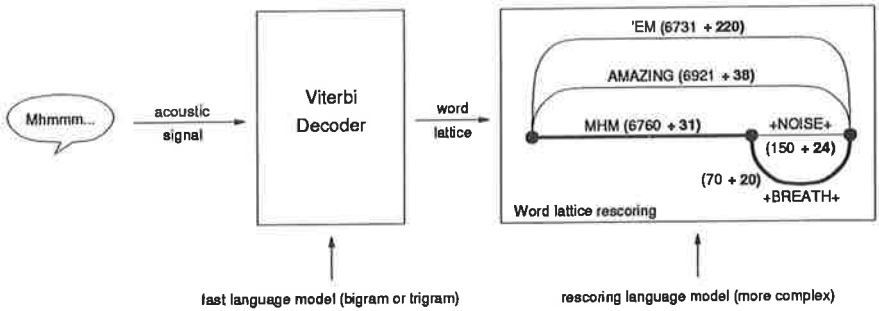

Figure 2: Word lattice rescoring paradigm.

3.3 Word lattice rescoring with a language model

For a varying language model, word lattice rescoring results show a better correlation with single-pass recognition accuracy than perplexity measurements on textual data, but still keeps computational cost manageable.

As shown in Fig. 2, it is possible to generate a word lattice from a viterbi decoding step (Ney and Aubert 1994). The language model used in the decoding has to be computationally simple and fast; usually it is a word bigram or a word trigram model. The word lattice is a compact representation of recognition hypotheses with a score close to the score of the best hypothesis (i.e. the recognition result of the single-pass decoding).

The rescoring process is a search of the path with the best (lowest) score. The total score is the sum of the scores of each segment. The score of the segment is a linear combination of the acoustic score (computed in the decoding step) and the rescoring language model score. The latter score depends on the preceding part of the path.

In our experiments, the lattices are kept fixed over all experiments, while the rescoring LM is varied. The obtained recognition accuracy after rescoring is used as an indicator of the quality of the rescoring language model.

3.4 Mixture language models

The LMs we apply here in rescoring are *mixture* models. They each consisted of:

1. a word trigram trained on NH (fixed);
2. a class bigram trained on NH with automatically acquired word classes (Ueberla 1997) (fixed);
3. a word trigram trained on *weighted* BN or WSJ data (variable).

These components are combined with linear context-independent interpolation (Jelinek and Mercer 1980). The interpolation weights are each time estimated by maximizing the likelihood of a cross-validation set held out from NH.

Table 2: Word recognition error rates on the NH task after word lattice rescoring with a mixture model (Section 3.4) containing a component trained on weighted data, either BN (column 2) or WSJ (column 3). Q denotes $(1-p(T))/p(T)$. The significance p(BN) and p(WSJ) of the best result w.r.t. the baseline result was computed with the Mann-Whitney-Wilcoxon test (Hogg and Tanis 1993, 625–628).

Weighting formula	BN (%)	WSJ (%)	p(BN)	p(WSJ)
1 (baseline)	37.2	38.4		
κ_D^{1/N_D}	37.1	38.2		
κ_S^{1/N_S}	37.0	38.8		
$(\kappa_D^{-1}Q+1)^{-1}$	37.3	38.0		
$(\kappa_S^{-1}Q+1)^{-1}$	36.9	37.9		
$(\kappa_D^{-1/M_D}Q+1)^{-1}$	37.2	38.2		
$(\kappa_D^{-1/N_D}Q+1)^{-1}$	37.1	37.9		
$(\kappa_S^{-1/N_S}Q+1)^{-1}$	36.8	37.7	15%	7%
$(\kappa_D^{-1}Q+1)^{-M_D}$	37.2	38.0		
$(\kappa_D^{-1}Q+1)^{-N_D}$	37.2	38.1		
$(\kappa_S^{-1}Q+1)^{-N_S}$	37.1	37.9		

4 Results and discussion

Word error rates (WER) after word lattice rescoring are summarized in Table 2. The LM score is $-\log p(w|h) \times A + B$. A is the LM factor and B is the word insertion penalty. The reported word error rates are obtained with optimal A and B.

Weighted counting is more effective on WSJ than on BN. All weighting schemes show an improvement of WER. One possible reason is that BN style differs only slightly from NH style.[5] The word error rate changes listed in the BN column are not significant. The best result was obtained with the $(\kappa_S^{-1/N_S}Q+1)^{-1}$ weighting scheme: a 0.3% WER drop with BN and 0.7% WER drop with WSJ, with respect to the baseline (no weighting). The significance of the BN result, 15%, is not low enough to draw strong conclusions.

For WSJ, computing the weights on sentences tends to be better than on documents. The benefit from a more fine-grained weighting seems to successfully counteract the loss of evidence per weight and of robustness.

All in all the gains from the proposed methods remain rather modest. We see at least two difficulties. First, n-gram LM probabilities need to be smoothed; the choice of the smoothing method may introduce unpredictable results when processing the output of two different LMs. A weighting method which depends

[5] BN also contains a small portion of news talkshows similar to NH, such that the assumption of disjunct styles is actually violated.

less on the outputs of more than one n-gram LM seems to be preferrable for that reason.

The second criticism is that POS n-grams are probably not sufficient to fully characterize style.[6] We intend to use more linguistic features similar to Biber, Conrad and Reppen (1998, Ch. 5, 6, 8), such that the outputs of the POS n-grams become less critical and a more fine-grained style distinction can be made.

Acknowledgements

The authors are thankful to Klaus Ries who spent a lot of time to discuss and share ideas. Some experimentation scripts were borrowed from him and from Hua Yu.

The first author is supported by IWT scholarship SB971170 and currently works as a visiting scholar at the Interactive Systems Laboratories (Carnegie Mellon University, USA) headed by Dr Alex Waibel.

References

Biber, D., Conrad, S. and Reppen, R.(1998), *Corpus Linguistics—Investigating language structure and use*, Cambridge University Press.

Brill, E.(1992), A simple rule-based part of speech tagger, *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, Trento, Italy.

Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C. and Mercer, R. L.(1992), Class-based n-gram models of natural language, *Computational Linguistics*.

Crespo, C., Tapias, D., Escalada, G. and Alvarez, J.(1997), Language model adaptation for conversational speech recognition using automatically tagged pseudo-morphological classes, *Proc. International Conference on Acoustics, Speech and Signal Processing 97*, Vol. II, Muenich, pp. 823–826.

Federico, M., Bunnell, H. and Idsardi, W.(1996), Bayesian estimation methods for n-gram language model adaptation, *Proc. International Conference on Spoken Language Processing 96*, Vol. I, pp. 240–243.

Geutner, P.(1996), Introducing linguistic constraints into statistical language modeling, *Proc. International Conference on Spoken Language Processing 96*, Vol. I, Philadelphia, PA, USA, pp. 402–405.

Godfrey, J. J., Holliman, E. and McDaniel, J.(1992), Switchboard: Telephone speech corpus for research and development, *Proc. International Conference on Acoustics, Speech and Signal Processing 92*, Vol. I, pp. 517–520.

Hogg, R. V. and Tanis, E. A. (eds)(1993), *Probability and Statistical Inference*, fourth edn, Macmillan Publishing Company.

Isotani, R. and Sagayama, S.(1993), Speech recognition using particle n-grams and content-word n-grams, *Proc. EUROSPEECH 93*, Berlin, Germany, pp. 1955–1958.

Iyer, R. and Ostendorf, M.(1996), Modeling long distance dependence in lan-

[6] Biber, D., personal correspondence.

guage, topic mixtures vs. dynamic cache models, *Proc. International Conference on Spoken Language Processing 96*, Vol. I, pp. 236–239.

Iyer, R. M.(1997), *Improving and predicting performance of statistical language models in sparse domains*, PhD thesis, Boston University, College of Engineering.

Jelinek, F. and Mercer, R. L.(1980), Interpolated estimation of Markov source parameters from sparse data, *in* E. S. Geltsema and L. N. Kanal (eds), *Pattern Recognition in Practice*, North Holland, Amsterdam.

Katz, S. M.(1987), Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**, 400–401.

Kneser, R. and Ney, H.(1995), Improved backing-off for m-gram language modeling, *Proc. International Conference on Acoustics, Speech and Signal Processing 95*, Vol. I, pp. 181–184.

Lafferty, J. and Suhm, B.(1995), Efficient iterative scaling of a class of maximum entropy models, *XV Workshop on Maximum Entropy and Bayesian Methods*, Los Alamos, USA.

Ney, H. and Aubert, X.(1994), A word graph algorithm for large vocabulary, continuous speech recognition, *Proc. International Conference on Spoken Language Processing 94*, Vol. III, Yokohama, Japan, pp. 1355–1358.

Rao, S. P., Monkowski, M. D. and Roukos, S.(1995), Language model adaptation via minimum discrimination information, *Proc. International Conference on Acoustics, Speech and Signal Processing 95*, Vol. I, pp. 161–164.

Ries, K.(1997), A class-based approach to domain adaptation and constraint integration for empirical m-gram models, *Proc. EUROSPEECH 97*, Vol. IV, Rhodes, Greece, pp. 1983–1986.

Seymore, K., Chen, S. and Rosenfeld, R.(1998), Nonlinear interpolation of topic models for language model adaptation, *Proc. International Conference on Spoken Language Processing 98*, Vol. VI, pp. 2503–2506.

Ueberla, J. P.(1997), Domain adaptation with clustered language models, *Proc. International Conference on Acoustics, Speech and Signal Processing 97*, Vol. II, Munich, Germany, pp. 807–810.

Yu, H., Clark, C., Malkin, R. and Waibel, A.(1998), Experiments in automatic meeting transcription using JRTK, *Proc. International Conference on Acoustics, Speech and Signal Processing 98*, Vol. II, Seattle, WA, USA, pp. 921–924.