TRANSFORMING OUT-OF-DOMAIN ESTIMATES TO IMPROVE IN-DOMAIN LANGUAGE MODELS

Rukmini Iyer *Mari Ostendorf*

Electrical and Computer Engineering Department
Boston University, Boston, MA 02215

ABSTRACT

Standard statistical language modeling techniques suffer from sparse-data problems when applied to real tasks in speech recognition, where large amounts of domain-dependent text are not available. In this work, we introduce a modified representation of the standard word n-gram model using part-of-speech (POS) labels that compensates for word and POS usage differences across domains. Two different approaches are explored: (i) imposing an explicit *transformation* of the out-of-domain n-gram distributions before combining with an in-domain model, and (ii) POS *smoothing* of multi-domain n-gram components. Results are presented on a spontaneous speech recognition task (Switchboard), showing that the POS smoothing framework reduces word error rate and perplexity over a standard word n-gram model on in-domain data, with increased gains using multi-domain models.

1. INTRODUCTION

Statistical language models, which characterize the probability of different word sequences, play an important role in state-of-the-art speech recognizers. The most commonly used statistical language modeling technique, referred to as n-gram language modeling, treats the word sequence $w_1, w_2, \dots, w_T$ as a Markov process with probability

$$P(w_1, w_2, \dots, w_T) = \prod_{i=1}^{T+1} P(w_i | w_{i-1}, \dots, w_{i-n+1}), \quad (1)$$

where w_0 and w_{T+1} are sentence boundary markers and n is typically restricted to 2 or 3, a bigram or trigram language model respectively. For notational simplicity, we use the bigram representation; however, all experiments in this paper use trigram models.

Although quite powerful given their simplicity, n-gram models suffer severe sparse-data problems in many real tasks where little domain-specific data is available for training language models. Class n-gram models [1] and models that exploit domain-specific knowledge have been previously used to deal with this problem; however, class models have not yet yielded performance gains and domain-dependent techniques are not easily portable to new domains. A different approach involves the use of text from other domains, motivated in particular by the increasing amount of data available in the form of newspaper text, transcribed television speech, and web hyper-text documents. Simplistic approaches to combining data from multiple domains include brute-force adding of raw counts or linear interpolation of word n-gram distributions. More successful techniques have proposed an adaptation of out-of-domain n-gram models [2] or discriminating out-of-domain data for relevance to the target domain [3].

In this paper, we use part-of-speech (POS) class conditioning to exploit differing word vs. POS usage across domains in two approaches: (i) imposing an explicit *transformation* of the out-of-domain n-gram distribution prior to combining with an in-domain model, and (ii) POS *smoothing* of multi-domain n-gram components. We investigate variations of these techniques that lead to effective use of new information from out-of-domain data, taking advantage of similarities to the target domain. Results, presented for the Switchboard corpus of spontaneous speech [4], show improvements in recognition performance when compared to using an in-domain word n-gram model.

POS conditioning in an n-gram model, described in Section 2, forms the basis for both the POS transformation and smoothing approaches. These two approaches are discussed in Section 3, which proposes alternatives for combining data and models from multiple domains. In Section 4, we outline the training and recognition paradigm, and present perplexity and word recognition results. Finally, Section 5 concludes with a discussion of the results and possible extensions of this work.

2. POS CONDITIONING OF N-GRAMS

This work is based on the hypothesis that the similarities between two domains can be characterized both in terms of style, roughly represented by POS sequences[1],

[1] We use linguistically motivated POS classes for representing style, as opposed to statistically derived classes, due to problems with robust estimation of such classes in sparse domains and since statistically-derived classes may not generalize across domains.

and content, as represented in the particular choice of vocabulary items. Separately accounting for these differences should lead to a more effective use of out-of-domain data in sparse domains.

POS class grammars, sometimes used to deal with sparse training problems, have been traditionally formulated as HMMs [1]. However, recent extensions obtain better results by including both word and class contexts as conditioning events [5]. We consider a simplified implementation of class grammars, where

$$P(w_i|w_{i-1}) = \sum_j P_X(c_j|w_{i-1}) P_Y(w_i|c_j, w_{i-1}), \quad (2)$$

where word w_i maps to part-of-speech classes c_j, w_{i-1} refers to the word n-gram history for w_i, and $X, Y \in \{I, O, I+O\}$ where I and O are henceforth used to subscript in-domain and out-of-domain models respectively. This representation makes it possible to use different aspects of out-of-domain data (style vs. content) with different weights.

The introduction of POS conditioning into the models raises the possibility of using a variety of back-off algorithms. In this work, we consider two formulations of the Witten-Bell smoothing scheme [6, 7]. A simple approach involves smoothing with lower order distributions:

$$\begin{aligned} P(c_i|w_{i-1}) &= \gamma(w_{i-1}) P_{ML}(c_i|w_{i-1}) \\ &\quad + (1-\gamma(w_{i-1})) P(c_i), \quad (3) \end{aligned}$$

where $P_{ML}()$ is the maximum likelihood estimate, and $\gamma(w_{i-1})$ is the smoothing parameter estimated as in [7]. Alternately, we explore a more complex back-off scheme, where less detailed, but not necessarily lower order, distributions are used for smoothing, for e.g.

$$\begin{aligned} P(c_i|w_{i-1}) &= \gamma(w_{i-1}) P_{ML}(c_i|w_{i-1}) + (1-\gamma(w_{i-1})) \\ &\quad \cdot \sum_k P(c_i|c_k) P(c_k|w_{i-1}), \quad (4) \end{aligned}$$

where

$$P(c_i|c_k) = \chi(c_k) P_{ML}(c_i|c_k) + (1-\chi(c_k)) P(c_i),$$

and c_k represents the k^{th} POS label for word w_{i-1}.

3. USING MULTI-DOMAIN DATA

The POS conditioning framework, described in Section 2 can be exploited to use multi-domain data while alleviating content and style differences. In this section, we introduce two approaches, one based on a POS *transformation* of out-of-domain distributions prior to estimating a multi-domain n-gram model, and the second based on a POS *smoothing* of multi-domain n-gram components.

3.1. POS Transformation of Out-of-Domain Data

In the first approach, we use Equation 2 to *transform* the out-of-domain n-gram distribution prior to combining with an in-domain distribution. Two such transformation schemes are explored here: transformation A ($X = I, Y = O$), under the hypothesis that class n-gram (style) differences across the two domains should be compensated for, and transformation B ($X = O, Y = I$), that assumes we should adjust for vocabulary differences. We use recognition performance as our criterion for selecting the appropriate transformation.

Once transformed, the POS-smoothed out-of-domain n-gram distribution, $\hat{P}_O()$, can be combined with an in-domain model ($X = I, Y = I$), $P_I()$, using linear interpolation,

$$P(w_i|w_{i-1}) = \lambda P_I(w_i|w_{i-1}) + (1-\lambda) \hat{P}_O(w_i|w_{i-1}), \quad (5)$$

where the interpolation weight λ can be estimated using the EM algorithm on a held-out, in-domain data set. We compared recognition performance of the interpolated transformed and un-transformed out-of-domain models. Transformation A outperformed the un-transformed model which outperformed transformation B, suggesting that it is useful to add new words, but more valuable if we can also compensate for style.

3.2. POS Smoothing of Multi-Domain Models

The transformation approach allows increasing *either* word or POS n-gram coverage, but it may be useful to increase *both* while discriminating for similarity to the target domain. Therefore, we investigated the POS smoothing framework Equation 2 with multi-domain components:

$$P(w_i|w_{i-1}) = \sum_j P_{I+O}(c_j|w_{i-1}) P_{I+O}(w_i|c_j, w_{i-1}). \quad (6)$$

No information from either domain is discarded in this scheme. The component distributions $P_{I+O}()$ can be estimated using different techniques as described below.

3.2.1. Linear Interpolation

A simple $2(n-1)$-parameter interpolation scheme for estimating parameters for data from n domains uses

$$\begin{aligned} P(w_i|w_{i-1}) &= \textstyle\sum_j (\sum_l \lambda_l P_l(c_j|w_{i-1})) \\ &\quad \cdot (\textstyle\sum_k \theta_k P_k(w_i|c_j, w_{i-1})), \quad (7) \end{aligned}$$

where $\{\lambda\}$ and $\{\theta\}$ are estimated iteratively using the EM algorithm. If $\lambda_l^{(p)}$ and $\theta_k^{(p)}$ represent the interpolation weights at iteration p, the corresponding weights at iteration $p+1$ are given by

$$\begin{aligned}
\lambda_l^{(p+1)} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_j \lambda_l^{(p)} P_l(c_j|w_{i-1}) P_{I+O}^{(p)}(w_i|c_j,w_{i-1})}{P_{I+O}^{(p)}(w_i|w_{i-1})} \\
\theta_k^{(p+1)} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_j \theta_k^{(p)} P_k(w_i|c_j,w_{i-1}) P_{I+O}^{(p)}(c_j|w_{i-1})}{P_{I+O}^{(p)}(w_i|w_{i-1})}
\end{aligned}$$

respectively, where $P_{I+O}^{(p)}()$ are parameters estimated using $\lambda_l^{(p)}$ and $\theta_k^{(p)}$.

The $2(n-1)$-parameter interpolation scheme can be further extended to incorporate context-dependent interpolation parameters as in Stolcke's Bayesian interpolation scheme [8]. Assuming some prior probability for the different domains, $\lambda_l(w_{i-1})$ and $\theta_k(c_j, w_{i-1})$ are estimated as

$$\begin{aligned}
\lambda_l(w_{i-1}) &= \frac{P(l)P_l(w_{i-1})}{\sum_m P(m)P_m(w_{i-1})} \\
\theta_k(c_j, w_{i-1}) &= \frac{P(k)P_k(c_j, w_{i-1})}{\sum_m P(m)P_m(c_j, w_{i-1})}.
\end{aligned}$$

3.2.2. Relevance Weighting

Instead of combining models, one might estimate the component distributions of Equation 6 based on the combined counts of different domains [3]. Relevance weighting involves estimating weights for the out-of-domain data to reflect in-domain similarity, and then estimating the n-gram model parameters using weighted counts. The relevance weights for each document D^i from the out-of-domain corpora are estimated as

$$P(I|D^i) = \frac{P_I(D^i)P(I)}{P_I(D^i)P(I) + P_O(D^i)P(O)}, \qquad (8)$$

where $P_I(D^i)$ and $P_O(D^i)$ can be computed using using word or POS n-gram models from the domains I and O respectively. The resulting relevance weights can be used for weighting out-of-domain counts. Parameter estimation with such weighted counts is covered in [3].

4. EXPERIMENTS

4.1. Paradigm

Perplexity and recognition experiments are reported on the Switchboard (SW) task, transcribing spontaneous, telephone conversational speech. The training data includes 2.1 million words of in-domain text from SW and 141 million words of broadcast news data (BN). The BN data includes spontaneous conversational speech from talk shows, as well as read speech in the form of news and voice-overs. Both SW and BN are POS-tagged using state-of-the-art POS taggers [11, 12], with a total of 80 POS tokens in the class vocabulary.

Recognition results are obtained using the N-best rescoring formalism [9] with the N-best hypotheses generated by the BBN Byblos System [10], a speaker-independent HMM system. More specifically, the top N sentence hypotheses ($N = 100$) are rescored by the language model, and a weighted combination of the HMM score and new language model scores is used to re-rank the hypotheses. The top ranking hypothesis is used as the recognized output.

4.2. Results

Results are first presented on the POS smoothed n-gram framework in Table 1. Replacing the standard in-domain word n-gram model with a corresponding POS-smoothed n-gram improves recognition performance from 41.1% to 40.5% with a small decrease in perplexity. Both back-off schemes result in improved recognition; however, the simple approach outlined in Equation 3 outperforms the more complex scheme from Equation 4. Hence, we use the simple back-off scheme in further experiments with the transformation model.

Table 1: Trigram perplexity and WER (%) results on dev96 for the SW model, comparing the new framework to a traditional word n-gram using different back-off schemes.

n-gram model	Back-off	Perplexity	WER (%)
traditional	simple	105.7	41.1
pos-smoothed	simple	95.9	40.5
pos-smoothed	complex	111.2	40.7

Table 2 looks at the different interpolation and relevance weighting strategies outlined in Section 3. All approaches to using multi-domain data gain over an in-domain model. The 2-parameter interpolation scheme outperforms the 1-parameter scheme, consistent with the recognition experiments in Section 3.1; new information is more important than similarity. The Bayesian multi-parameter scheme suffers from an explosion of parameters to estimate. The number of parameters can potentially be reduced, by conditioning the interpolation weights on a smaller context (unigram instead of bigram) [8], or by tying the interpolation weights as in [13]. The relevance weighting techniques for estimating multi-domain components of a POS-smoothed model perform almost as well as the 2-parameter interpolation scheme although the weighting schemes con-

Table 2: Trigram perplexity and WER (%) on dev96 using linear interpolation and relevance weighting for estimating the multi-domain (SW and BN) components of a POS smoothed n-gram model.

Interpolation	Perplexity	WER(%)
1-parameter	90.6	40.1
2-parameter	89.9	39.7
Bayesian multi-parameter	91.8	40.1

Relevance weighting	Perplexity	WER(%)
pos likelihood	136.7	39.8
word likelihood	127.5	39.8
pos, word likelihood	131.9	39.9

sistently associate perplexity increases with improved recognition performance. The best case offers a 1.4% absolute improvement in WER over the baseline trigram result.

5. CONCLUSIONS

The proposed POS-conditioning model exploits the potential of part-of-speech information for improving recognition performance over a standard word n-gram model. Results show that the new POS smoothing framework results in improved recognition performance over an in-domain word n-gram model, and further gains may be had with more complex interpolation strategies and/or improved weighting schemes for combining data/models across multiple domains.

Our approach to multi-domain modeling is both task-independent and easily portable to new domains, hence relevant for a broad range of speech research and commercial applications. We plan to study the effect of similar transformations with other languages such as Spanish as well as with other corpora such as the Wall Street Journal corpus, which differs significantly both in content and style from both domains used in the experiments reported here.

6. REFERENCES

[1] P. Brown *et al.*, "Class-Based N-gram Language Models of Natural Language", *Computational Linguistics,***18**, 467-479, 1992.

[2] S. Besling and H.-G. Meier, "Language Model Speaker Adaptation", *Proc. European Conference on Speech Comm. and Tech.,***3**, 1755-1759, 1995.

[3] R. Iyer, M. Ostendorf and H. Gish, "Using Out-of-Domain Data to Improve In-Domain Language Models", *IEEE Signal Processing Letters*, to appear.

[4] J. J. Godfrey, E. C. Holliman and J. McDaniel, "Switchboard: Telephone Speech Corpus for Research and Development", *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,***1**, 517-520, 1992.

[5] S. Roukos *et al.*, "LM95 Summer Workshop Project Report of the Phrase Structure Language Models Team", in *1995 Language Modeling Summer Research Workshop Technical Reports.*

[6] H. Witten and T. C. Bell, "The Zero Frequency Estimation of Probabilities of Novel Events in Adaptive Text Compression", *IEEE Transactions Inform. Theory*, **IT3-7** (4), 1085-1094, 1991.

[7] P. Placeway, R. Schwartz, P. Fung and L. Nguyen, "Estimation of Powerful LM from Small and Large Corpora", *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,***2**, 33-36, 1993.

[8] A. Stolcke, M. Weintraub *et al.*, "LM95 Summer Workshop Project Report of the Team for Portability of Language Models", in *1995 Language Modeling Summer Research Workshop Technical Reports.*

[9] M. Ostendorf *et al.*, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses", *Proc. ARPA Workshop on Speech and Natural Language,*83-87, 1991.

[10] L. Nguyen *et al.*, "The 1994 BBN/BYBLOS Speech Recognition System", *Proc. ARPA Spoken Language Technology Workshop,*77-81, 1994.

[11] M. Meteer, R. Weischedel and R. Schwartz, "POST: Using Probabilities in Language Processing", *Proceedings of IJCAI*, 1991.

[12] A. Ratnaparkhi, "A Maximum Entropy Part-Of-Speech Tagger", *Proceedings of the Empirical Methods in Natural Language Processing Conference*, May, 1996. University of Pennsylvania.

[13] F. Jelinek and R. L. Mercer, "Interpolation and Estimation of Markov Source Parameters from Sparse Data", *Pattern Recognition In Practice*, ed. E. S. Gelsema, L. N. Kanal, 1981.