INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Contextual prediction models for speech recognition

Yoni Halpern, Keith Hall, Vlad Schogol, Michael Riley,*
Brian Roark, Gleb Skobeltsyn, Martin Bäuml

Google Inc.
{yhalpern,kbhall,vlads,riley,roark,glebs,baeuml}@google.com

Abstract

We introduce an approach to biasing language models towards known contexts without requiring separate language models or explicit contextually-dependent conditioning contexts. We do so by presenting an alternative ASR objective, where we predict the acoustics and words given the contextual cue, such as the geographic location of the speaker. A simple factoring of the model results in an additional *biasing* term, which effectively indicates how correlated a hypothesis is with the contextual cue (e.g., given the hypothesized transcript, how likely is the user's known location). We demonstrate that this factorization allows us to train relatively small contextual models which are effective in speech recognition. An experimental analysis shows a perplexity reduction of up to 35% and a relative reduction in word error rate of 1.6% on a targeted voice search dataset when using the user's coarse location as a contextual cue.

Index Terms: speech recognition, language modeling

1. Introduction

Language model adaptation is critical for large scale speech recognition in a general setting. Adaptation is necessary when the particular use of the recognizer does not match the settings used to train the model. Language model adaptation has typically focused on techniques directly related to estimation: model interpolation, model mixtures, n–gram constraints in feature-based models, and trigger models [1].

In this paper we present a different modeling objective that leverages contextual cues that are available at the time of recognition. These contexts can include features such as the geolocation of the user, the time of day, or in cases where the particular user is known (e.g., for mobile voice search applications), information about the user or about clusters of similar users. Rather than modify the general language model or provide additional language models, we introduce a model which makes a prediction of the known contextual factor (e.g., the location of the device the user is speaking into). We then use this prediction to bias the recognizer towards contextually relevant hypotheses.

We present empirical results for our approach using a geographic contextual feature. Most importantly, our technique is able to reduce the error rate significantly on a set of geo-specific utterances without negatively impacting the error rate for general utterances.

2. Related Work

While our approach may appear similar in structure to topic-based language models [2], we do not modify the main language model, rather we add an extra term to our objective (see section 3). Previous topic modeling approaches typically modify the language model itself, either via something similar to traditional class-based language models [3] or via some kind of mixture or MaxEnt model [4, 5]. In these methods, the topic is predicted by the history and the current word is predicted by the history and the topic, suitably marginalized if more than one topic per history is allowed. In the current paper, we are investigating the use of features that are given, not predicted, so no marginalization is required.

Recently, [6] presented an exploration of interpolated models using geographic signals similar to what we will use in this paper. They present a technique for training and pruning a combined general and geo-specific language model. Each of the component geographical models are interpolated with the general model. The choice of component geographic model is determined by the location of the user's device on which the voice query is submitted. They show perplexity and word-error-rate reductions on data containing the geographic information about the user's device. Furthermore, they show that this model can be used for lattice rescoring as well as on-the-fly first-pass rescoring. Our work on geographic modeling is similar in that we train and evaluate our models on similar data. However, the advantage of our model over the interpolated language model is that we do not need to train per-geographic region language models, thus allowing us to experiment with very compact feature-based models.

3. Methodology

We introduce our approach as a modified form of the standard speech recognition noisy-channel formulation, whereby the word sequence W is predicted given the acoustic signal A and the known contextual feature C:

$$P(W|A,C) = \frac{P(A|W,C)}{P(A|C)}\frac{P(C|W)}{P(C)}P(W) \quad (1)$$

$$\approx \frac{P(A|W)}{P(A)}P(W)\frac{P(C|W)}{P(C)} \quad (2)$$

The right-hand side of Equation 1 is an exact factorization of the contextually dependent objective (the left-hand side). In Equation 2, we make a simplifying independence assumption, namely that the acoustic model is independent of the contextual cue. While this may not be the case, in this paper we do not attempt to model acoustic variation based on the contextual cue, limiting our focus to language variation.

Given these assumptions, Equation 2 contains the standard acoustic model and language model terms as well as an additional term $\frac{P(C|W)}{P(C)}$ which we refer to as the *contextual bias term*. This term can be interpreted as a measure of how well the words in the hypothesis are predictive of the known contextual

*This work was done while Yoni Halpern was an intern at Google. At that time, he was a Ph.D. student at New York University.

http://dx.doi.org/10.21437/Interspeech.2016-1358

feature. The ratio is 1.0 when the context is independent of the hypothesis, and hence does not change the score. Positive and negative correlation leads to a ratio greater than or less than 1.0, respectively.

3.1. Context Prediction Model

The contextual bias term we introduced in Equation 2 is made up of a prediction model $P(C|W)$ and a prior distribution over contextual features, $P(C)$. The prior is estimated using the relative frequency estimator over the training data. We train the prediction model using a Maximum Entropy criterion [7], allowing us to extract arbitrary features from a hypothesis W. The billions of examples in our training set necessitate training via efficient distributed training algorithms. We use the Iterative Parameter Mixtures SGD approach as presented in [8] and [9]. We do not use additional regularization, but do exclude features which occur fewer than 5 times in our training set.

For the experiments presented here, we use a simple set of features based largely on n-grams. For each word in a hypothesis, we collect the full n-gram and all lower order n-grams ending with that word. We include an additional skip-gram, which is a bigram covering the trigram context, but excluding the middle word. For each training instance, we have a bag-of-n-grams along with a bias feature (which models the class prior). While the form of the model does not require that the features be restricted to local n-grams, we found the bag-of-n-grams to be effective. Finally, we train models using a feature hashing approach [10]. We hash the feature identities and restrict the number of parameters using the modulo function, resulting in a simple vector of parameters.

3.2. Incremental Rescoring

The above models are suitable for an N-best rescoring setting, but are not straightforwardly applicable to incremental processing. The general formulation of our model allows us to predict the contextual feature just once per hypothesis. We make the assumption that our bag-of-n-gram models are suitable for assigning scores to prefixes of hypotheses as well as the entire hypothesis. Given such a model, for a string $W = \{w_1 \dots w_n\}$, we define the per-word factorization as:

$$\frac{P(C \mid W)}{P(C)} = \frac{P(C \mid w_1)}{P(C)} \prod_{k=2}^{n} \frac{P(C \mid w_1 \dots w_k)}{P(C \mid w_1 \dots w_{k-1})} \quad (3)$$

For each word, w_k, the incremental contribution of the contextual bias term is the ratio the probability of the context given the current prefix and the probability of the context given the previous prefix. Note that each of the numerators will cancel out the following denominator, leaving us with the original score. This allows us to construct an incremental scoring scheme that may be used during lattice rescoring. In the following empirical evaluation, we report results for N-best rescoring and do not evaluate lattice rescoring.

4. Empirical evaluation

All training and evaluation sets in this paper were English mobile search data from the United States.

4.1. Training Data

We collected 40B aggregated and anonymized training instances containing a typed user search query and a corresponding coarse geographic location for the device used to issue the query. We also held out 41M instances in order to measure each model's performance in terms of perplexity. Each geographic location was then mapped into coarse location clusters. We show results for three different clustering schemes: DMA [11] (210 clusters), ZIP2 (first two digits of the postal code, 100 clusters) and ZIP3 (first three digits of the postal code, 990 clusters).

4.2. Evaluation Data

As described above, we measure perplexity on a held-out set of instances from the same distribution as the training data. We evaluate word-error-rate (WER) on two datasets. The first set is a sample of general voice search utterances over a period of two months; we refer to this dataset as `VOICESEARCH` (12886 utterances, 62792 words). Each utterance is transcribed by three human raters; we keep those with agreement for the evaluation set.

The second set of data was also sampled from voice search utterances, but is focused on cases in which our current baseline system fails to provide a useful transcription. We identify *failed* utterances as those in which the user first speaks a search query and, within a short time period, revises the transcribed query by hand (using an on-screen keyboard or other typed input method). We limit this dataset to those in which the reference transcript contains an entity which is located within a fixed distance of the device location. This test, which we refer to as the `GEO-TARGETED` set, consists of 5,270 anonymized utterances (23,213 words). The baseline word error rates are much higher for this set, as expected. Note that the baseline sentence accuracy (SACC) is very low for this set by design. Because this test set is biased, we report results on both the `VOICESEARCH` dataset and the `GEO-TARGETED` dataset; the intention is to show that our approach can improve on the geographically informative *failed* queries without negatively impacting the general voice search utterances.

4.3. Metrics

We first show the impact of each model on test-set perplexity. Since our model can be applied in addition to any language model (n-gram-based, MaxEnt-based, mixutre models, etc.), we report the percentage of perplexity reduction instead of explicit perplexity values. Our model makes per-utterance predictions and therefore, we show the per-sentence perplexity reduction rather than a per-word quantity.

$$PPL = \exp\left(\frac{1}{N}\sum_{i}^{N} -\log\left(P_{LM}(W_i)\frac{P(C|W_i)}{P(C)}\right)\right) \quad (4)$$

$$= \exp\left(\frac{1}{N}\sum_{i}^{N} -\log P_{LM}(W_i) + \frac{1}{N}\sum_{i}^{N} -\log\frac{P(C|W_i)}{P(C)}\right)$$

$$= PPL_{LM} \cdot \exp\left(\frac{1}{N}\sum_{i}^{N} -\log\frac{P(C|W_i)}{P(C)}\right) \quad (5)$$

Equation 4 is the per-sentence perplexity of the combined language model and context prediction model. In Equation 5, we factor this term to be the per-sentence perplexity of the baseline LM, PPL_{LM}, and the relative perplexity contribution from the context prediction model (the second factor). We refer to this

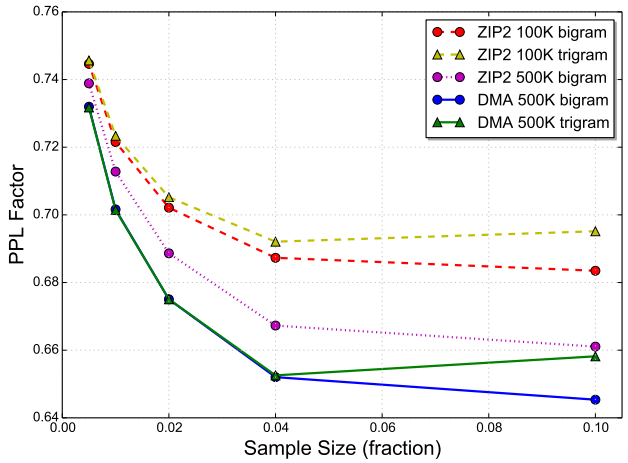

Figure 1: Perplexity reduction learning curves.

latter value as the PPL factor when reporting experimental results, i.e.,

$$\frac{PPL}{PPL_{LM}} = \exp\left(\frac{1}{N}\sum_{i}^{N} -\log\frac{P(C|W_i)}{P(C)}\right)$$

Additionally, we evaluate our models when used for N-best rescoring (where $N = 100$) on top of our baseline system. Our speech recognition system is based on a long short-term memory neural network acoustic model [12, 13] with a vocabulary of approximately 4 million words. The baseline language model is a Katz [14] smoothed 5-gram model pruned to 100M n-grams, trained using Bayesian interpolation to balance multiple sources [15]. Our second-pass rescoring LM is a distributed model trained on the same data fully concatenated using Katz backoff and pruned to 15 billion n-grams [16].

During N-best rescoring, we compute the contextual bias term for each hypothesis by predicting the known location cluster of the device from which the utterance was captured. In cases where we were unable to determine the location of the device, we do not apply our model. In the evaluation below we found that approximately 70% of the data contained usable postal-code information, which was required for the geographic clustering we performed.

We observed a small reduction in word-error-rate (less than a 0.1 % absolute reduction) on the VOICESEARCH test set. In order to get a more granular understanding of how our model performs for geographic contexts, we provide experimental results on the GEO-TARGETED test set.

4.4. Learning Curve

To establish how much of our data is needed to get optimal performance given a particular model size (hashed size), we evaluate the model over different sized training sets. We do this by randomly sampling (without replacement) from the full training set.

Table 1 shows that WER reductions and sentence accuracies (SACC – the number of sentences where the hypothesis matched the reference exactly), improve with more data. In Figure 1 we can see that perplexity improvements plateau as we sample at a rate of more than $1/10$ (using 1/10th the available training data).

Model	Sample	PPL Factor	WER	SACC
Baseline	-	1.000	25.0	12.56
DMA	0.005	0.732	24.8	13.20
DMA	0.01	0.702	24.7	13.41
DMA	0.02	0.675	24.7	13.58
DMA	0.04	0.652	24.6	13.79
DMA	0.01	0.645	24.6	13.82

Table 1: DMA models with hashed feature dimension of 500K. N-best oracle WER for the GEO-TARGETED test set is 15.6.

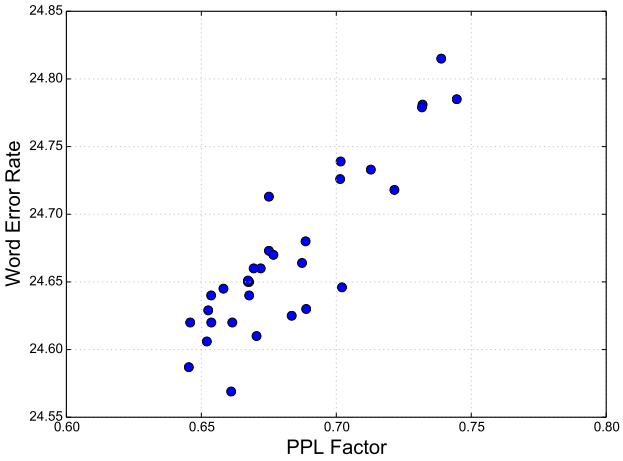

Figure 2: WER vs PPL factor. Lower is better on both axes.

4.5. Effect of Varying Feature Dimension

Model	Dim (K)	PPL Factor	WER	SACC
Baseline	-	1.000	25.0	12.56
DMA	100	0.677	24.7	13.52
DMA	500	0.653	24.6	13.67
ZIP2	100	0.687	24.7	13.49
ZIP2	500	0.667	24.7	13.59

Table 2: DMA and ZIP2 models for two different hash sizes. The dimension is the number of features per class (e.g., a ZIP2 model with hash dimension of 100K results in 10M parameters). The sample rate is 0.04.

Our models are built using hashed features, allowing us the ability to choose the exact size of the hashing dimension. Ideally, we use the smallest model which achieves a usable WER reduction. In Table 2 we see that a larger feature dimension improves perplexity, but leads to only very modest WER improvements.

4.6. Word-Error-Rate Results

In empirical analysis of speech recognition systems, the relationship between perplexity improvements and WER reductions is not always clear. In Figure 2, we plot the correlation between perplexity and WER reduction for the context prediction models under a number of different conditions where we varied the sampling rate, the size of the hashed feature dimensions, the order of the n-gram features and the clustering of the postal codes. In this case, it appears that perplexity reduction is a good indi-

ZIP2 Clusters			
Cluster	% of Train	Baseline	Test
60??? (Chicago Metro)	3.75	25.7	24.9
92??? (San Diego Area)	3.43	23.8	23.8
10??? (Manhattan & North)	3.29	23.7	22.8
75??? (Dallas Area)	3.08	24.2	24.2
90??? (Los Angeles)	2.69	25.5	24.7
ZIP3 Clusters			
Cluster	% of Train	Baseline	Test
100?? (in Manhattan)	2.26	21.5	21.5
606?? (in Chicago)	1.89	26.6	24.7
752?? (in Dallas)	1.57	23.6	23.8
900?? (in Los Angeles)	1.50	21.5	21.5
770?? (in Houston)	1.46	26.4	24.7
DMA Clusters			
Cluster	% of Train	Baseline	Test
New York, NY	8.22	24.1	23.8
Los Angeles, CA	6.84	24.5	24.5
Chicago, IL	3.99	25.8	24.8
Dallas-Ft. Worth, TX	3.81	26.2	26.2
San Francisco, CA	3.06	23.1	23.2

Table 3: The top 5 clusters in each of the three clustering schemes according to their frequency in the training data, the baseline WER, and the WER for the best model trained for that scheme.

cator of WER reduction. In Table 3 we show the WER changes for the five most frequent geographic clusters for the three different clustering approaches.

4.7. Examples

Hypothesis	Bias	Final
1. Rhode Island		
princeton city hall phone number	0.3	42.1
cranston city hall phone number	-3.0	40.7
2. Illinois		
metro train south shore to south bend indiana	-2.0	91.3
metra train south shore to south bend indiana	-3.2	90.5
3. Nebraska		
i was states	-0.0	51.4
iowa state's	-2.6	50.3
4. Nevada		
jeff hardy songs	0.1	58.9
fat freddy songs	-0.5	58.5

Table 4: Win examples: For each pair, the first example is preferred by the baseline system and the second example is preferred by the context prediction model (DMA Clusters with 500K feature dimensions and a sampling rate of 0.04.

To illustrate our model's ability to address geographically-related ASR errors, we isolated a few such examples from the manually transcribed `VOICESEARCH` test set. Table 4 shows four examples where the contextual bias term causes an incorrect 1-best transcript hypothesis to be replaced with a correct one, according to the manual transcriptions and our own verification. The table also displays the contextual bias term for each of the two hypotheses and the final cost of each hypothesis (lower costs are better).

A brief description of the improvement for each of the examples in Table 4:

1. The user is located in Rhode Island and our model correctly reranks the second hypothesis which contains the name of a city in Rhode Island.
2. The user is in the Chicago area and is asking for transit directions. Our model prefers the correct name for the transit authority there.
3. The user is in Nebraska and asks for a university in a neighboring state.
4. A user in Nevada is searching for a musician, though it is not clear why our model prefers the correct transcript in this case.

Hypothesis	Bias	Final
1. Massachusetts		
the first-ever shopkin made	0.2	84.7
the first ever shot in maine	-1.1	84.4
2. Minnesota		
chisholm ice racing kc pro	0.1	98.4
chisholm ice racing casey pro	-0.5	98.3
3. Alabama		
what is a power morcellator	0.0	113.5
what year is a power morcellator	-0.1	113.4

Table 5: Loss examples: For each pair, the first example is preferred by the baseline system and the second example is preferred by the context prediction model (DMA Clusters with 500K feature dimensions and a sampling rate of 0.04.

Similar to the improvements, we show some of the recognition errors introduced by our model in Table 5. A brief description of each:

1. The model prefers a transcript containing the word "maine" which is the name of a nearby state.
2. Our system prefers the name "casey" over the name "KC" for an ice racing league. There happens to be a place called Casey in Minnesota.
3. There is a slight preference for a transcript that contains an extra word which was not actually spoken.

5. Conclusion

We present a new modeling approach to allow for adaptation towards known contextual features during recognition. Our modeling framework allows for any multi-class classifier which produces probabilistic estimates for the classes. In our empirical analysis we employ a MaxEnt classifier under various training conditions allowing for compact models. We show that this model reduces perplexity by 35% on a heldout dataset and achieves a 1.6% relative reduction in WER on geographic voice search data without hurting performance on general voice search data.

In future work, we will explore both lattice rescoring and on-the-fly first-pass rescoring using the incremental formulation from Equation 3. We will also experiment with other contextual cues, such as temporal clusters (time-of-day, day-of-week, time-of-year).

6. References

[1] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, pp. 93–108, 2004.

[2] D. Gildea and T. Hofmann, "Topic-based language models using em," in *In Proceedings of EUROSPEECH*, 1999, pp. 2167–2170.

[3] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[4] R. M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 1, pp. 30–39, 1999.

[5] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *Computer Speech & Language*, vol. 14, no. 4, pp. 355–372, 2000.

[6] C. Chelba, X. Zhang, and K. Hall, "Geo-location for voice search language modeling," in *in Proc. Interspeech*, 2015.

[7] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–72, 1996.

[8] R. T. McDonald, K. B. Hall, and G. Mann, "Distributed training strategies for the structured perceptron," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, 2010, pp. 456–464.

[9] K. B. Hall, S. Gilpin, and G. Mann, "Mapreduce/bigtable for distributed optimization," in *NIPS LCCC Workshop*, 2010.

[10] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, "Feature hashing for large scale multitask learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1113–1120.

[11] Wikipedia, "Wikipedia: Media market," 2016. [Online]. Available: http://en.wikipedia.org/wiki/Media_market

[12] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 338–342.

[13] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of Long Short-Term Memory recurrent neural networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 1209–1213.

[14] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1987, pp. 400–401.

[15] C. Allauzen and M. Riley, "Bayesian language model interpolation for mobile speech input," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, 2011, pp. 1429–1432.

[16] P. Jyothi, L. Johnson, C. Chelba, and B. Strope, "Distributed discriminative language models for google voice search," in *Proceedings of ICASSP 2012*, 2012, pp. 5017–5021.