Cache-based Language Model Adaptation using Visual Attention for ASR in Meeting Scenarios

Neil Cooke
University of Birmingham
Edgbaston
Birmingham, United Kingdom
n.j.cooke@bham.ac.uk

Martin Russell
University of Birmingham
Edgbaston
Birmingham, United Kingdom
m.j.russell@bham.ac.uk

ABSTRACT

In a typical group meeting involving discussion and collaboration, people look at one another, at shared information resources such as presentation material, and also at nothing in particular. In this work we investigate whether the knowledge of what a person is looking at may improve the performance of Automatic Speech Recognition (ASR). A framework for cache Language Model (LM) adaptation is proposed with the cache based on a person's Visual Attention (VA) sequence. The framework attempts to measure the appropriateness of adaptation from VA sequence characteristics. Evaluation on the AMI Meeting corpus data shows reduced LM perplexity. This work demonstrates the potential for cache-based LM adaptation using VA information in large vocabulary ASR deployed in meeting scenarios.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language models*

General Terms

Algorithms, Experimentation, Measurement, Performance

1. INTRODUCTION

A key challenge in multimodal research is combining information from modalities to improve Automatic Speech Recognition (ASR) performance. ASR systems hypothesise the most probable word sequence given some speech, typically using a statistical acoustic model and Language Model (LM). LM adaptation is the process of modifying a generic LM to better model the target speech. Exploiting the fact that a speaker may say something again, cache-based LM Adaptation is shown to improve ASR performance [7].

In the ASR function of a multimodal system, LM adaptation can use the information from other modalities to provide additional contextual information; e.g., the topic being talked about may be indicated by what a person is looking at or pointing to; the LM adaptation process can increase the probabilities of words associated with that topic.

This work investigates cache-based LM adaptation. However, rather than using a cache containing words, the cache contains information from a person's Visual Attention (VA) sequence; i.e., the visual foci (e.g. person) or focus types (e.g. people) looked at and the corresponding viewing durations.

A key problem when using information from a modality to improve the recognition of speech is whether it is always appropriate; in the case of VA somebody may, at times, be talking about something completely unrelated to what they are looking at [6]. Appropriateness can be encouraged (although not guaranteed) by constraining the task, e.g. by asking people to talk directly about the landmarks on a map [1] or using a software application. However, it is more valuable to consider less constrained real-world scenarios that have practical applications, such as business meetings.

In a business meeting, people may look at one another, at shared information resources such as presentation material and at their notes. Crucially, they also look at nothing in particular; e.g., people avert their gaze to concentrate [5]. A question this study answers is, does people's speech differ depending on what they are looking at? e.g., in a meeting, one may expect to hear the word 'you' more often when someone is looking at one of their colleagues, as opposed to looking at a presentation screen. However, if someone says 'you' whilst looking at a presentation screen, then one may reasonably expect that the unspecified person was addressed by being looked at or called by name in the recent past.

From this requirement to understand what is being looked at over time to best utilise VA information, we propose and evaluate LM adaptation where the VA sequence information is considered as a cache. We investigate methods for its appropriate use and in doing so determine how best to utilise the VA data; Which VA sequence characteristics should the LM adaptation use? How can VA information be used to adapt LMs to improve ASR performance?

This paper proceeds as follows. Section 2 details previous work. Section 3 describes the LM Adaptation framework and section 4 its implementation. In section 5 we evaluate our models. Section 6 concludes with a discussion and outlines future work.

2. PREVIOUS WORK

In a previous study, we adapted a generic bigram language model so that words, primarily nouns associated with a specific object, were given higher probabilities when a person

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11 ...$10.00.

was looking at the object. We evaluated this approach on a map task showing a marginal improvement in ASR word error rates [4], in line with other similar studies [11] [10].

This research extends our earlier work in two ways. Firstly, we consider the sequence of VA prior to, and during, an entire utterance, rather than considering a single instance of VA that co-occurs with the word onset; this VA information forms the cache. We also make no a priori assumption that particular words are associated with particular visual focus, instead estimating LMs using AMI Multimodal meeting data [8] containing matched VA and speech transcriptions.

3. FRAMEWORK

Let W=$\{W_1$,. . .,W_n,. . .,$W_N\}$ be a word sequence, $s(n)$ be the onset time of the word W_n, V = $\{V_1$, . . ., V_m, . . ., V_M $\}$ be a discrete-valued VA sequence, D= $\{$ D_1, . . ., D_m, . . ., D_M $\}$ be the associated VA durations. For each m, let $r(m)$ be the onset time of the VA; i.e. at time $r(m)$ a person looks at a visual focus of type V_m for the duration D_m.

The word sequence probability $P(W)$ is calculated as a product of individual N-gram probabilities of order k which, unlike the traditional N-gram model, are dependent on the word onset times $s(n)$, and the VA sequence:

$$\begin{aligned} P(W) = \prod_{n=1}^{N} P_{s(n)}(W_n | W_{n-1}, \dots, W_{n-(1+k)}, \\ V_{m-(l-1)}, \dots, V_m, \\ D_{m-(l-1)}, \dots, D_m) \end{aligned} \tag{1}$$

where l is the length of the VA cache and m satisfies the condition $s(n) = r(m)$. The LM probability $P_{s(n)}(\cdot)$ is determined from the weighted interpolation of a baseline LM and an LM derived from the VA sequence:

$$\begin{aligned} P_{s(n)}(W_n | W_{np}, V_{mp}, D_{mp}) = \\ (\lambda - 1) P_b(W_n | W_{np}) \\ + \lambda P_{V,s(n)}(W_n | W_{np}, V_{mp}, D_{mp}) \end{aligned} \tag{2}$$

where λ represents the appropriateness of VA-based LM adaptation; it is a estimated probability that someone is talking about something related to their VA sequence. W_{np}, V_{mp} and D_{mp} are shortened forms of the conditionals given in expression 1.

$P_{V,s(n)}(\cdot)$ is an LM dependent on the VA sequence which is estimated from the weighted interpolation of LMs $P_o(\cdot)$ associated with each visual focus type o:

$$P_{V,s(n)}(W_n | W_{np}, V_{mp}, D_{mp}) = \sum_{o=1}^{O} \sigma_o P_o(W_n | W_{np}) \tag{3}$$

where $P_o(\cdot)$ is estimated from speech which occurs whilst a person is looking at o. σ_o is the weight given to $P_o(\cdot)$. O is the number of focus types, satisfying $\sum_{o=1}^{O} \sigma_o = 1$. The set of weights σ_o for all O is a function of V, D and the word onset time $s(n)$. This function for the set of weights may be considered as a weight distribution function $\sigma(V,D,s(n))$.

4. IMPLEMENTATION

The VA-cache LM Adaptation implementations investigated in this study differ by the weight distribution function and appropriateness method.

4.1 Weight Distribution Function (σ)

The weight distribution function $\sigma(V, D, s(n))$ determines the relative weights of LMs $P_o(.)$ associated with each visual focus type. As stated in Section 1, this study investigates which VA sequence characteristics the LM adaptation should use. Four functions are proposed, σ^{maxc},σ^{maxd},σ^{mixc} and σ^{mixd}.

For σ^{maxc}, the weight σ_o is 1 for the focus o that has the highest number of occurrences in V. All other weights are zero:

$$\sigma_o^{maxc} = \begin{cases} 1 & \text{if } \arg\max_\theta \frac{N_{\theta \in V}}{N} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $N_{\theta \in V}$ is the number of times the specific visual focus type θ is looked at and N is the length of the VA sequence V.

For σ^{maxd}, the weight σ_o for the visual focus o is 1 for the focus that is looked at for the longest duration. All other weights are zero:

$$\sigma_o^{maxd} = \begin{cases} 1 & \text{if } o = \arg\max_\theta \sum_{m=1, v_m=\theta}^{m=M} D_m \in D \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For σ^{mixc}, the weight σ_o for the focus type o is based on the number of occurrences in the VA sequence relative to all other focus types:

$$\sigma_o^{mixc} = \frac{N_{\theta \in V}}{N} \tag{6}$$

For σ^{mixd}, the weight σ_o for the focus type o is based on the number of occurrences of o in the VA sequence relative to all other focus types:

$$\sigma_o^{mixd} = \frac{\sum_{v_m=0, m=1}^{m=M} D_m \in D}{\sum_{m=1}^{M} D_m \in D} \tag{7}$$

4.2 Appropriateness measure (λ)

As discussed in Section 1, what is being looked at and what is being said may be unconnected. Therefore can VA-based LM adaptation be curtailed in these instances? To address this, we consider the VA prior to the onset of an utterance (sentence) separately to that of the VA during the utterance. The inspiration for this comes from cognitive psychology where various roles for eye movement are hypothesised[5]; mediating communication, e.g. engaging/disengaging from conversation; concentrating, e.g. averting our gaze from distractions; sentence planning e.g. looking at something before describing it; supporting word production, e.g. looking at something to recall its name [9]. The sentence planning and word production hypothesis are of interest because VA can be classified as either one or the other based on whether it appears before the onset of an utterance or during it; we propose that comparing VA from these two classes provides an opportunity to measure the appropriateness of adaptation, i.e. determine a suitable function for λ.

To demonstrate this approach, two caches are proposed - a cache containing VA information that occurs during Sentence Planning (SP) and a cache containing VA information that occurs during Word Production (WP). Building on the framework set out in section 3, let W_f be the first word in a sentence and $s(f)$ be its onset time. Let W_n be the current word in a sentence and $s(n)$ its onset time. At time $s(n)$, the caches for VA information that occurs during sentence planning, V^{SP}, and word production, V^{WP}, are subsets sequence of V:

$$V^{SP} = V_{m-l+1}, \ldots, V_{e-1} \tag{8}$$

$$V^{WP} = V_e, \ldots, V_m \tag{9}$$

Where the VA, V_m, is occurring at word W_n's onset time so that $r(m)$ satisfies $s(n) = r(m)$. Likewise, V_e is at the time of word W_f so that $r(e)$ satisfies $s(f) = r(e)$. l is the total length of the VA caches. Similar expressions may be derived for the associated durations, D.

We hypothesise that there is an increased chance that the visual foci looked at during sentence planning and word production will be similar if what is being looked and what is being said are related, i.e. the V^{SP} and V^{WP} caches will contain similar VA sequence data. In such cases, we propose to increase adaptation, i.e. set λ nearer to 1 than 0. Likewise, if there is less relation between what is being looked at and said then the V^{SP} and V^{WP} caches will differ and λ set nearer to 0.

The weights, σ_o for LMs associated with focus types o are calculated for caches V^{SP} and V^{WP} using one of the methods given in Section 4.1. The difference between the set of weights in the cache gives the level of appropriateness. As an examplar, we propose a distance metric that sums the absolute difference between corresponding weights in the two cache satisfying $0 \leq \lambda \leq 1$:

$$\lambda = 1 - \frac{\sum_{o=1}^{O} |\sigma_o^{SP} - \sigma_o^{WP}|}{2} \tag{10}$$

5. EVALUATION

5.1 Method

To evaluate VA cache LM adaptation, models were implemented which varied in terms of the weight distribution function σ (section 4.1) and the appropriateness λ (section 4.2). The models were evaluated in terms of their LM perplexity on the 56 AMI meeting corpus sessions which had VA and speech transcriptions; 9742 utterances in total. Their perplexity was compared to that of a baseline LM.

The baseline LM $P_b(\cdot)$ (expression 2) was implemented as a trigram LM with modified Kneser-Ney smoothing [3] and estimated by mixing LMs estimated from the spoken part of the British National Corpus [2] ($6,105,876$ utterances) and the sessions of AMI Meeting Corpus which have no VA transcriptions ($59,304$ utterances). The baseline LM vocabulary was constrained to that in the AMI meeting corpus ($14,399$ words).

Weight Distribution Function	LM Perplexity	Δ Perplexity
σ^{maxc}	89.3	-22.7
σ^{maxd}	101.2	-10.8
σ^{mixc}	105.8	-6.2
σ^{mixd}	109.4	-2.6

Table 1: Performance of the VA cache LM Adaptation weight distribution functions. Δ Perplexity is the change in perplexity against the baseline LM of 112.0**.**

The focus-type LM $P_o(\cdot)$ (expression 3) was estimated from speech segments that corresponded to whether someone was looking at a focus of a particular type from the three types specified in the AMI meeting VA transcriptions ('person', 'place' and 'unspecified'). Due to data sparsity, focus-type LMs for each session were estimated from the speech from all other sessions with VA transcriptions - a 'leave one out' evaluation strategy.

The LM Adaptation was implemented using Python scripting and the SRI Language Modeling Toolkit (SRILM) [12].

5.2 Experiments

The results of two intial experiments are reported. *Experiment 1: Weight distribution function* investigates the relative performance of the four proposed weight distribution functions $\sigma(V, D, s(n))$ (section 4.1). The appropriateness measure λ is constant and determined empirically. The motivation for this experiment is to explore which VA sequence characteristics should be used in LM adaptation.

Experiment 2: Appropriateness function investigates using the appropriteness function for λ (section 4.2). The motivation for this experiment is to test whether the difference between the VA sequence before and during an utterance can guide whether VA-based LM adaptation is appropriate.

In both experiments performance is measured using perplexity. Various VA cache lengths were evaluated; the following results (section 5.3) use $l = 5$ which was determined empirically.

5.3 Results

5.3.1 Experiment 1: Weight Distribution Function

Which VA sequence characteristics should VA cache LM adaptation use to give the greatest benefits? The results in Table 1 show that using the weight distribution function σ^{maxc} (row 1) resulted in the largest reduction in perplexity (-22.7). Selecting a single focus-type LM to mix with the baseline (rows 1 and 2) performs better than mixing focus-type LMs (rows 3 and 4). Counting the instances of VA onto a particular focus type (rows 1 and 3) resulted in lower perplexity than considering VA duration (rows 2 and 4).

5.3.2 Experiment 2: Appropriateness Function

What is the utility in exploiting the differences in eye movement behaviour hypothesised in cognitive psychology to control the adaptation weight λ? The results in Table 2 are similar to experiment 1; selecting a single focus type LM to interpolate with the baseline (rows 1 and 2) performs better than mixing focus type LMs (rows 3 and 4). The results show that varying λ to account for appropriateness has

Weight Distribution Function	LM Perplexity	Δ Perplexity
σ^{maxc}	88.5	-23.5
σ^{maxd}	104.7	-7.3
σ^{mixc}	105.3	-6.7
σ^{mixd}	110.4	-1.6

Table 2: Performance of the VA cache LM Adaptation with varying appropriateness function λ. Δ Perplexity is the change in perplexity against the baseline LM of 112.0.

a some benefit when using the weight distribution function σ^{maxc}.

6. DISCUSSION

A general framework for VA cache LM Adaptation has been outlined which accounts for the appropriateness of using VA information in ASR, given that the relationship between speech and what someone is looking at may vary. The framework extends the conventional N-gram model to account for the VA sequence. Various adaptation schemes were implemented and the best performing ones reduced perplexity when trained and evaluated using AMI meeting data. In using this data, the relationship between VA sequence and N-grams has been learnt; this is in contrast to previous studies where N-grams associated with a particular visual focus are boosted.

The large vocabulary of the LMs used in the evaluation provides scope for practical application. The preliminary results indicate some potential in this approach, however further work is required; e.g. only one distance metric was proposed for the appropriateness function λ. Recognising that improvements in perplexity do not necessarily lead to better ASR performance, extending this evaluation to incorporate a full ASR system should follow further investigation into optimising this technique.

7. REFERENCES

[1] A. H. Anderson et al. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[2] L. Burnard. Users reference guide for the British National Corpus. Technical report, Technical report, Oxford University Computing Services, 2000.

[3] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics Morristown, NJ, USA, 1996.

[4] N. Cooke and M. Russell. Gaze-contingent asr for spontaneous, conversational speech: An evaluation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4433–4436, 2008.

[5] Z. Griffin. Why look? Reasons for eye movements related to language production. *In M. Henderson and F. Ferreira Eds., The interface of language, vision, and action: Eye movements and the visual world*, pages 213–247, 2004.

[6] R. Jacob. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary), 2003.

[7] R. Kuhn and R. De Mori. A cache-based natural language method for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–582, 1990.

[8] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, page 4, 2005.

[9] A. Meyer and C. Dobel. Application of eye tracking in speech production research. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, 2003.

[10] S. Qu and J. Chai. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. *Proc. of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 284–291, 2007.

[11] R. Sarukkai and C. Hunter. Integration of eye fixation information with speech recognition systems. *5th European Conf. on Speech Communication and Technology*, pages 1639–1643, 1997.

[12] A. Stolcke. SRILM-an Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing.* ISCA, 2002.