Towards the Use of Inferred Cognitive States in Language Modeling

Nigel G. Ward, Alejandro Vega

Department of Computer Science, University of Texas at El Paso
nigelward@acm.org
avega5@miners.utep.edu

***Abstract*—In spoken dialog, speakers are simultaneously engaged in various mental processes, and it seems likely that the word that will be said next depends, to some extent, on the states of these mental processes. Further, these states can be inferred, to some extent, from properties of the speaker's voice as they change from moment to moment. As a illustration of how to apply these ideas in language modeling, we examine volume and speaking rate as predictors of the upcoming word. Combining the information which these provide with a trigram model gave a 2.6% improvement in perplexity.**

I. Introduction

The field of language modeling, like much of linguistics, has historically had a "written language bias" [1]: a tendency to focus on language as sequences of symbols, rather than as human products used for human purposes. While recent work has made incremental progress towards overcoming this bias — developing practical ways to improve and extend language models to better handle various phenomena of spontaneous language and dialog — we propose an extreme, alternate view, focusing on language behavior in dialog as the product of the speaker's cognitive processes. This suggests the strategy of using information about the speaker's cognitive state, at each moment, to enable better prediction of the word he or she is likely to say next.

To infer the mental state of a speaker we can use prosodic features computed over his or her recent speech. Previous research has demonstrated that it is possible to infer, at the utterance level, what sort of dialog act the speaker is doing and whether her or she is engaged, frustrated, pleased, confident, and so on, from the prosody. Less work has been done at the sub-utterance level, although it is more important for language modeling, but [2] showed that it is possible to identify disfluent states in speech, and to use this information to reduce word perplexities, and [3] showed that characterizing the context in terms of superficial states, such as "being 3–10 seconds into an utterance," can also help prediction of the upcoming word.

The exploitation of cognitive state information for language modeling can be approached in various ways. An interesting long term research program would be to first categorize the relevant cognitive states of speech, then build classifiers to detect those states from prosodic features, and finally relate the words of English to the states in which they tend to occur. Doing so will be difficult, not least because the number of possible states is large, as a person participating in a conversation has to do many many things (including processing what the other person just said, deciding when to respond and what sort of contribution to make, retrieving information from memory, and monitoring his own speech, among others [4]), and each of these cognitive processes has its own set of states.

In this paper we use a simpler approach: we do without explicit representations of cognitive states, instead directly relating recent prosodic features to the upcoming words. We consider two prosodic features: speaking rate and volume.

Section 2 illustrates how the probabilities of words vary with the speaking rate and volume of the immediately preceding context. Section 3 explains how this information can be combined with n-gram probability estimates. Section 4 presents the improvement obtained by doing so, and Section 5 discusses what should to be done next.

II. Initial Observations

We used the ISIP transcriptions of the Switchboard corpus [5], [6], a collection of telephone conversation of spontaneous topics between unacquainted adults.

A number of previous studies have used prosody in language modeling and speech recognition, as surveyed in [7], [8] and [9]. Our choice of prosodic features is different from that of previous work, reflecting our goal, of exploiting information related to cognitive states. Thus we use speaking rate, which may indicate degree of preparation and confidence, and volume, which may indicate engagement or dominance, for example.

The details of how we computed such features also reflect our aim. Our features are direct ones, in Shriberg's sense, not hand-labeled nor inferred to match hand-labeled tags. In contrast with the use of prosody for language modeling to capture syntactic and lexical-accent regularities, our prosodic features are not syllable-aligned nor syllable-normalized [8]. In contrast with uses of prosodic information related to task structure or dialog acts, our features are computed over local contexts, not over entire utterances.

A. Speaking Rate

Each token in the corpus was characterized as fast, slow, or of middling speaking rate. Specifically, a token less than 0.89 of the average duration for that word was considered fast, more than 1.11 of the average duration slow, and otherwise

previous speaking rate	characteristic words … …uncharacteristic words
fast	sixteen, carolina, o'clock, kidding, forth, weights, familiar, half, science, process, careful, matter, grand, doubt … …hm, uh-huh, ah, huh
middling	direct, mistake, mcdonald's, likely, wound, repairs, purchased, immigration, director, troops, lawyer, wears … …uh-huh, hi, um-hum
slow	goodness, gosh, agree, bet, let's, uh, god, um, grew, huh-uh, although, neat, either, definitely, true, am … …experience, yourself, ago
(none)	um-hum, uh-huh, hum, hm, oh, yep, yeah, wow, huh, yes, ah, right, okay, well, exactly, no, sure, which … …guess, know, mean, lot

TABLE I
CHARACTERISTIC AND UNCHARACTERISTIC WORDS IN VARIOUS SPEAKING-RATE CONTEXTS

lead-in	characteristic words … …uncharacteristic words
silent	um-hum, hum, hm, uh-huh, yep, ooh, oh, yeah, hi, ah, yes, wow, huh, okay, right, well, gee, exactly, huh-uh, um … …wear, hand, percent, own, ago
quiet	bet, know, mean, y[ou]-, although, seems, gosh, mostly, well, bye-bye, true, what's, eighty, which, let's, tend, looks … …hand, working, credit, ago
moderate	forth, francisco, hampshire, extent, colors, corps, dakota, trend, bag, whatnot, underneath, penalty, seats, minutes … …ooh, hm, um-hum, okay, hum
loud	sudden, opinions, hills, box, hand, restrictions, reasons, union, scale, industry, unusual, favorite, sorts, hook, hole … …uh-huh hm, um-hum, lunches

TABLE II
CHARACTERISTIC AND UNCHARACTERISTIC WORDS AFTER REGIONS OF VARIOUS VOLUME LEVELS

middling. Each token then was classified as after-slow, after-middling, after-fast, or after-silence, depending on the duration of the previous word, if any (tokens immediately preceded by a silence of at least 1.2 seconds, the value which we have found best for defining utterances for language modeling purposes, were considered to be "after-silence"). These characterizations were done from the transcripts, without reference to the actual speech signal.

We then calculated which words tended to occur in which contexts: Table I shows the most characteristic and most uncharacteristic, that is, words which are most and least likely to appear in the specific rate context, relative to their general frequency, as determined by the S ratio [3]. Examining the words in each category suggests some general patterns. Common after fast regions (words of relatively short duration) are high-content words, especially place names and numbers. Common after slow regions (words of relatively long duration) are assessments, disfluency markers, social expressions (*bye-bye, thank [you]*), expressions of belief (*definitely, unless, well, yes, [of] course, but, consider, absolutely, okay, must, generally, certainly, totally*), and the word *I*.

B. Volume

Volume information was computed over fixed-width regions, rather than over words, of 50 milliseconds, since this width gave best performance on the tuning data. The volume was normalized for each track, more for the sake of adapting to the recording conditions than for adapting to the speaker. Specifically, for each dialog side we took at a large sample of regions and used EM to find the mean volume of silence regions and the mean volume of speech regions. Regions with an energy closer to the silence mean than to the speaking mean were considered "silent," those with an energy within one standard deviation of the mean as "moderate," those less as "quiet," and those more as "loud." Each word was then associated with the loudness label over the 50 ms immediately preceding the word onset.

Table II shows the words most and least common after each volume tag. Common after quiet regions are expressions of belief (*[I] bet, [I] know, y[ou know], true*), of types and degrees of belief (*although, mostly, definitely, might, usually, tend, looks, guess, mostly*), and clause connectives (*well, then*). Common after middling-volume lead-ins are the tail ends of multi-word expressions (*[and so] forth, [San] Francisco, [New] Hampshire, [to some] extent*). Common after loud lead-ins are general content words, and, to a lesser extent, words pronounced while laughing.

Thus, as expected, these prosodic features do seem to reveal cognitive states, as seen by the types of words occurring in each context. Of course, one could also interpret these patterns of occurrence as reflecting communicative situations: for example, the tendency for expressions of belief to come after low-volume regions may reflect a communicative strategy of preceding important words with a quiet lead-in to give them more impact. For practical language modeling purposes, the interpretation we ascribe doesn't matter; what matters is whether regularities exist, and whether the regularities are nonredundant to the regularities captured by standard language models, which is the topic of the next section.

III. Combination with an N-gram Model

Thus the local prosodic context does provide useful information. However it is clear that this information does not obviate the need to exploit local context dependencies. Thus we integrate this information into an existing language model. The methods described in this section could be used with any language model, but for concreteness, we describe them as being applied to a trigram model, specifically the SRILM implementation of a backoff model [10].

Using prosodic context information to improve an existing model is a special case of the problem of language model adaptation [11], [12], for which many techniques are known. As our aim here is not to determine the best adaptation technique, but merely to determine whether the local prosodic context has useful information at all, we used only the very simple methods we developed for conditioning on time since various reference events [3], [13]. Here we recapitulate briefly.

The basic idea is to use probabilities based on the local prosodic context merely to tweak the backoff probabilities. For example, for a word occurring in context x, if the data indicates that the word is more common in that context than at other times, then we multiply the backoff probability by a scaling factor to reflect this. This gives the estimates:

$$P_t(w_i|c,x) = S(w_i|x)P_{\mathit{backoff}}(w_i|c) \qquad (1)$$

where c is the local context (for trigrams specifically, just the preceding two words) and S is the scaling factor.

The computation of S is somewhat *ad hoc* and complex, but it is based on the ratio R of the frequency of w_i in context x to the overall frequency of w_i. The complications include smoothing and the application of a weight k to adjust the strength of the contribution of the prosodic information: specifically S is derived from R raised to the power k, as described in [3].

Tweaking is omitted in two special cases: first, if the speed-context is "none", that is, if a word appears after no previous word, since in that case the bigram <s> w_i is perfectly adequate to represent this information, so tweaking can add no useful information; and second, if the speed-context is middling, since we found that tweaking in such cases hurts performance on average. Thus tokens in such contexts are excluded when computing the S ratios from the training data; similarly no tweaking is done for such tokens in the test data.

Finally there is a normalization step, so that the probability estimates for all words in the vocabulary in fact add to 1.

IV. Perplexity Results

The training, tuning, and test data were all subsets of Switchboard. The training data was 1000 tracks, consisting of about 652K words. Tuning of k was done using a separate set of tuning data consisting of 34K words. The best values for k were .99 for the rate information and .49 for the volume information. The test set consisted of 16 tracks from Switchboard, containing 10441 words and representing about 75 minutes of speech. The evaluation ignored end of sentence tags and out-of-vocabulary words.

	perplexity
baseline	107.8
speaking rate	105.0
volume	105.1

TABLE III
Perplexity on Switchboard

Conditioning on speaking rate gave a 2.8 point decrease in perplexity, which is a 2.6% improvement. Overall, the words which gave the maximum benefit were *um-hum, yeah, uh, oh, I, uh-huh*, and *you*, all of which are more common in slow contexts. We also tried a measure of speaking rate that was adapted to the current speaker, but this slightly hurt performance, to our surprise.

(Out of curiosity, we also examined whether conditioning on the speaking rate of the *current* word would also give an advantage (although including middling-rate tokens this time). This gave a perplexity of 107.6, a small improvement. The main source of the benefit is the fact that the model distinguishes words which tend to vary greatly in duration from those which don't. Of course, here we are trespassing on the domain of the acoustic model: since information about the likely durations of word may be represented, implicitly or explicitly, in the acoustic models, this perplexity improvement may not be indicative of better recognition in this case.)

Conditioning on volume gave a 2.7 point decrease in perplexity. Overall, the words which gave maximum benefit were *yeah, oh, um-hum, uh-huh, well, and, to*, and *of*. The first five of these words were more common in the after-silence condition; the last two were less common in this condition. Predictions were improved for words in all contexts, but most of the contribution was due to words in the after-silence context, notably those mentioned above. Words in the after-loud condition also contributing strongly, mostly due to words like *to, a, it, have*, and *of*, which tend to be more common after loud regions.

These perplexity benefits are larger than those seen in previous work [2], [3], probably because of the use of features that relate well to cognitive states (c.f. [3]), and because of the direct predictions of words from the local prosodic context, without reliance on hidden variables representing inferred cognitive state (c.f. [2]).

V. Summary and Future Work

This paper has shown that word probabilities vary with prosodic features of the local context, and that incorporating this information in a language model can decrease the perplexity on casual dialogs. Despite the use of features and techniques which are simple and *ad hoc* in many ways, a reasonable perplexity improvement was obtained. From this we conclude that a cognitive perspective on language modeling does indeed have promise.

The potential for further perplexity improvements is large. We know from experiments with people that the most promising direction for improving language models is the use of information beyond just the words said [14], but in this work we have only scratched the surface. Other predictive factors to consider include pitch height, pitch range, creakiness, and voicing fraction [13]. We also plan to use prosodic features of the interlocutor's recent behavior, as the cognitive processes and actions of the interactants in spoken dialog are often tightly coupled. We would like to develop a way to compute the scaling factors without discretizing speaking rate, etc. into categories. We may explore clustering of words or other dimensionality reduction techniques to combat sparseness.

We also need to demonstrate the value of these features for speech recognition. Technically this will be easy: the volume over the previous 50ms is easy to compute, the duration of the previous word hypothesis is available in the recognizer (although possibly with some inaccuracy), and augmenting a recognizer to include tweaking based on such features can be done trivially [15]. We expect that the perplexity improvements we found will be matched by reductions in word error rate, since the types of knowledge used are different from those captured by standard language models or acoustic models.

It is also worth noting that modeling the relation between local prosody and words may also be useful for applications other than speech recognition. For example, language generation and speech synthesis may be improved by better language modeling [16], and we expect that speech will be more natural and intelligible if the prosody is locally adapted to suit the upcoming words.

Creating a truly cognition-based language model remains a long-term project. Although cognitive-state modeling was our inspiration, and the features chosen do appear to be reflecting cognitive states, we still have a lot of work to do to create a cognitive model of human language production. The next step will be to identify the relevant cognitive states and use them as hidden variables. This could be done by boot-strapping from hand-labeled cognitive states, as suggested in the introduction, or perhaps by doing bottom-up clustering or factor analysis on the contexts predictive of words from various kinds. A subsequent step would be to augment this with mechanisms for modeling the transitions among states over time, perhaps initially by conditioning superficially on elapsed time [3], to create an actual cognitive process model.

Building such models should have benefits beyond the technical: many researchers have pointed out that dialog behaviors offer a unique window into human cognition and human social interactions [4], [17], [18], and predictive models of dialog behaviors and cognition over time may be of great value for improving our understanding of the dynamics of spoken dialog.

Acknowledgment

We thank David Novick and Olac Fuentes for comments. This work was supported in part by NSF Grant IIS-0415150 and by the US Army Research, Development and Engineering Command (RDECOM), via a subcontract to the USC Institute for Creative Technologies (ICT).

References

[1] P. Linell, *The Written Language Bias in Linguistics: Its nature, origins and transformations*. Routledge, 2005.

[2] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *ICASSP*, 1996, pp. 405–408.

[3] N. G. Ward and A. Vega, "Modeling the effects on time-into-utterance on word probabilities," in *Interspeech*, 2008, pp. 1606–1609.

[4] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[6] ISIP, "Manually corrected Switchboard word alignments," 2003, Mississippi State University. retrieved 2007 from http://www.ece.msstate.edu/research/isip/projects/switchboard/.

[7] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications, Vol. 138*. Springer-Verlag, 2004, pp. 105–114.

[8] S. Huang and S. Renals, "Modeling prosodic features in language models for meetings," in *Machine Learning for Multimodal Interaction IV (LNCS 4892)*, A. Popescu-Belis, S. Renals, and H. Bourlard, Eds. Springer, 2007, pp. 191–202.

[9] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework," in *ICASSP*, 2007.

[10] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Intl. Conf. on Spoken Language Processing*, 2002.

[11] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. Eurospeech*, 1997, pp. 1971–1974.

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

[13] N. G. Ward and A. Vega, "Using non-lexical context to improve a language model for dialog," *Speech Communication, submitted*, 2009.

[14] N. G. Ward and B. H. Walker, "Estimating the potential of signal and interlocutor-track information for language modeling," in *Interspeech*, 2009.

[15] N. Kiran and N. G. Ward, "Testing the value of a time-based language model for speech recognition," University of Texas at El Paso, Department of Computer Science, Tech. Rep. UTEP-CS-08-29, 2008.

[16] C. Brockmann, A. Isard, J. Oberlander, and M. White, "Modelling alignment for affective dialogue," in *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on User Modeling (UM-05)*, 2005.

[17] V. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, 1970, pp. 567–577.

[18] D. C. O'Connell and S. Kowal, *Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse*. Springer, 2008.