INTERSPEECH 2009 BRIGHTON

Estimating the Potential of Signal and Interlocutor-Track Information for Language Modeling

Nigel G. Ward, Benjamin H. Walker

Department of Computer Science, University of Texas at El Paso, USA

nigelward@acm.org, bhwalker@miners.utep.edu

Abstract

Although today most language models treat language purely as word sequences, there is recurring interest in tapping new sources of information, such as disfluencies, prosody, the interlocutor's dialog act, and the interlocutor's recent words. In order to estimate the potential value of such sources of information, we extend Shannon's guessing-game method for estimating entropy to work for spoken dialog. Four teams of two subjects each predicted the next word in a dialog using various amounts of context: one word, two words, all the words spoken so far, or the full dialog audio so far. The entropy benefit in the full-audio condition over the full text condition was substantial, .64 bits per word, greater than the .54 bit benefit of full text context over trigrams. This suggests that language models may be improved by use of the prosody of the speaker and context from the interlocutor.

Index Terms: entropy, perplexity, Shannon's guessing game, prediction, context, prosody

1. The Need

Better language models are needed to improve speech recognition. Today most work focuses on the challenges of mining more information from the word sequence [1], but additional types of information may also be of use. There have been scattered explorations in improving language modeling for dialog by use of pitch and energy features, words produced by the interlocutor, and the timing of the words [2, 3, 4, 5], but the results have not lived up to expectations, and it seems that interest in using such additional information has waned. This is not surprising, given how labor-intensive it is to build and test a novel addition to the language modeling repertoire.

However the limited success of past attempts does not tell us whether the sources of information themselves have little value or whether the models built were inadequate. Thus we need a way to estimate the potential value of considering new kinds of information in language models. This paper reports the development of such a method, and the discovery that applying information beyond the simple lexical context has the potential to greatly improve language models

2. Humans as Language Models

People are far better at speech recognition than machines, and the psychological literature is rich in studies of how quickly and accurately people can recognize speech input, as a function of different kinds of noise and context. However most of this work pertains more to acoustic modeling than language modeling. Even work which does study the effects of context is generally done to test specific hypothesis about neural pathways, using stimuli very unlike spontaneous speech. One exception is a study of the contributions of dialog context, which found that scrambling the utterances of a dialog reduced subjects' ability to correctly recognize the words [6].

In order to determine what kinds of syntactic knowledge are most useful for language modeling, Brill and Florian presented subjects with lists of the n-best hypotheses output by a speech recognizer, had them pick the one they thought mostly likely, and then asked them to indicate what type of knowledge they had used, choosing from a list of syntactic constructions and constraints [7]. This experiment design directly matches the way sophisticated language knowledge might be applied during a recognizer's second pass.

We set out instead to examine performance in a task resembling that of a language model in the first pass: that of determining (or predicting) the likely next word given some previous context. This is of course Shannon's original guessing-game method for estimating the entopy of English [8]. This method has been used previously as a source of insight for language modeling: Jelinek briefly mentions a study in which "humans beat the trigram model by factors of 3 or more in perplexity . . . mostly based on their ability to use . . . information that considerably precedes the currently guessed letters" [9], although this work was apparently not followed up.

In this paper we present a way to extend Shannon's method to spoken dialog, and use this method in various conditions to estimate the potential for improving language models of signal and interlocutor-track information.

3. Measuring the Entropy of Speech

While there have been indirect estimates of the entropy of speech [10], it appears that direct measurement has not previously been attempted. Shannon's method of guessing letters is impractical for audio presentation (at least in languages which have co-articulation or where letters do not map directly to phonemes), tedious for subjects, and not directly relevant to speech recognition. So we had the subjects predict words. Thus the subjects' task was to use left-context information to predict the next word.

Following Shannon, we allowed the subjects multiple guesses. A person who guesses word x before word y is, in some sense, expressing a higher probability estimate for x than y. To a first approximation, the best probability estimates are obtained by taking the probability (over all guesses) that the subject gets it right after exactly g tries as the probability that the subject is implicitly assigning to the gth guesses for each word. For example, if 15% of the time the first guess is correct, and if the subject's first guess for the next word in a given context is *cat*, then we take this as assigning 15% of the probability distribution to the word *cat*. This way of estimating prob-

abilities is mathematically identical to Shannon's, although the justification is different.

In a case where the subject never guesses the word, we still need a probability estimate; for this, we back off to the (unigram) probability of the word in a disjoint subset of the Switchboard corpus [11]. Thus we have what we need to compute the entropy (cross-entropy), which is, of course, the average over all words i of

$$- p_i \log q_i$$

where p_i is the likelihood of word i in the test data and q_i is the probability estimate of word i according to the model.

4. The Method

Since our motivation is the suspicion that language modeling has more to gain from exploiting new types of information than from using better text-based models, our hypothesis is that human guessers would do much better with audio and interlocutor-track information than with just more textual context. Thus, our two key conditions were unlimited textual context and unlimited text + speech + interlocutor's speech (where "unlimited" means "all the way back to the start of the dialog"). It is worth stressing that the different conditions apply, not to the word to be recognized, but to the *previous* words. We also included a bigram condition and a trigram condition to enable a rough calibration of the method relative to existing entropy estimators.

The detailed design of the experimental method was constrained by the nature of dialog and by the need to motivate and not fatigue our subjects, as explained below.

4.1. Stimulus Ordering

At first we considered a within-subjects design, in which each subject would first have to guess word n given one word of context, then two words, then all the previous words, then the entire previous speech signal. However this style of presentation would have several problems. Subjects might lose motivation to guess thoughtfully in the bigram and trigram cases, knowing that they would soon be getting more information. We were also concerned that the inability to let them go on as soon as they guessed the correct word (due to the need to have guesses in all the conditions) would also reduce motivation. Finally, we thought this method would be unnecessarily time-consuming.

We therefore opted for a between-subjects design, where each subject made decisions in each condition, but using different tracks. Moving to a between-subjects design allowed incremental presentation. In a pilot study we had subjects guess every word, in order. Thus, for example, after the first 24 words of the track, they were asked to guess the 25th. They were then told what the 25th word actually was and were asked to guess the 26th, and so on, The ability to follow a conversation as it unfolded, word by word, made the task interesting. This was also more efficient than having unrelated guesses, as the overlapping of the contexts of the guesses meant that the overhead of presenting the context was amortized over the guesses.

However word-by-word incremental presentation is incompatible with evaluating bigram- and trigram-based prediction abilities: subjects would get too much context. We therefore settled on having subjects guess every 10th word, which allowed some economy of context presentation, but made the bigram- and trigram-based guesses (mostly) uncorrupted by additional context. To further reduce the leakage of context into the bigram and trigram conditions, the presentations were scrambled in these conditions.

4.2. Stimulus Selection and Preparation

We selected four dialog-sides from Switchboard, choosing four that appeared to be of roughly similar complexity (as measured by SRILM-estimated trigram perplexities), and were sufficiently long: tracks 2238B, 2241B, 2260B, and 2168B. From each we selected 30 words: every 10th word, starting with the 25th: this gave us the four datasets. We prepared the text-based versions from the ISIP transcriptions [12]. Thus pauses were indicated with the token "[silence]" and no punctuation was shown.

In the audio condition, we presented the context up to the word to be guessed, except that we clipped early enough to remove any co-articulation cues to the upcoming word. In some cases this meant that subjects had to guess word n without being able to hear the audio for word $n-1$ at all; however they simultaneously were presented the context as text, so at least they knew what $n-1$ was. (If we were seeking a strict upper bound on the entropy this simultaneous presentation of text would be a problem, since it provided the subject veridical information on words that were acoustically indistinct or ambiguous, but this was not a problem for the purpose of comparing performance across conditions.)

4.3. Stimulus Presentation and Response Scoring

We wrote custom software to display textual context on the screen and to present the audio context in stereo. The software was operated by the experimenter. Before each guess, subjects were allowed to review or re-listen to the provided context as many times as they liked. (Although not representative of normal human listening (and perhaps also reading), this was done to make the task more comfortable for the subjects, especially when they needed to make multiple guesses.) In the audio condition they were allowed to control how much context they heard on replays. Interestingly, in the text conditions many subjects repeated the words to themselves out loud; perhaps to engage some neural circuits accessible only by auditory presentation.

In order to make the task more fun, subjects were run in pairs, with the teammates alternating turns, guess by guess. All teams acted at times competitively and at times cooperatively. For each prediction, teams were allowed up to 5 guesses. A guess was counted as correct if it exactly matched the token in the ISIP transcription [12], including such non-lexical items as *uh-huh*, *um-hum*, *uh*, and *um*. Prediction of silence (the sentence-end symbol) was not done: subjects were informed where the silences were. The correct words did not include fragments, [laughter], [vocalized-noise], or similar tokens. The guesses were typed in by the experimenter and the software recorded the guesses and checked for correctness.

Since merely recording guesses loses some information, we considered allowing the subjects to provide more information, for example a higher probability to a guess they felt sure of, or equal probabilities if they had no strong preference among two or three possibilities. More generally, we considered moving to a betting paradigm [13]. However we felt that most subjects wouldn't want to make such sophisticated estimates, and that the prospects for an increase in accuracy were not good [13], and probably not worth the time cost.

4.4. Participants and Protocol

The subjects were 8 adult native speakers of English, 18 to 57 years old, six male and two female. Subjects were recruited

Condition, Dataset	g 1	2	3	4	5	no	Entropy
Bigram, 1	1	1	0	0	1	26	8.38
Trigram, 2	1	0	0	0	1	27	9.35
Unltd. Text, 3	4	4	4	0	0	17	8.13
" + Audio, 4	10	2	2	0	0	15	6.68

Table 1: Number of times Team A required exactly g guesses to get the right answer; "no" indicates that they didn't get it at all.

Team	condition-dataset: entropy				avg.
A	b1: 8.38	t2: 9.35	u3: 8.13	a4: 6.68	8.14
B	t1: 8.09	u2: 8.10	a3: 6.53	b4: 7.61	7.58
C	u1: 7.67	a2: 7.82	b3: 8.94	t4: 7.65	8.02
D	a1: 7.36	b2: 9.30	t3: 8.02	u4: 7.06	7.93
avg.	7.88	8.64	7.91	7.25	7.92

Table 2: Per-Word Entropy, before normalization. b=bigram, t=trigram, u=unlimited text, and a = unlimited text + audio.

from among family and friends and compensated with $20. Education levels varied: one was soon to graduate from high school and the rest had at least some college, several having Bachelors degrees and one a Masters. Subjects were familiarized with the Switchboard genre and conventions, including the ISIP spelling conventions, by hearing a short description of the corpus and viewing a sample dialog transcript.

To allow for corrections for variations in performance — due to differences in the predictability of the datasets, differences in the ability of the subjects, and differences due to a training effect over time — each team saw a different combination of presentation conditions and datasets. All teams saw the datasets in the same order, but the conditions varied.

To keep our subjects from getting too tired, we had them guess only 30 words in each condition and allowed for breaks between conditions. It took each team about 90 minutes to complete the four conditions. Including the instructions and debriefing period, each session lasted about 105 minutes.

5. Results

Table 1 illustrates the data gathered, showing, for team A, how many words were guessed correctly on the first guess, on the second guess, and so on, for each condition. (This is analogous to Shannon's Table I, but transposed). The totals add up to 29, not 30, as one item had to be excluded due to a software error.

From this we computed the entropy, in the usual way, taking the average over all test items of –log q_i, where q_i was computed as explained above. The overall probability of success on the first guess was 15%, on the second 5%, on the third 4%, on the fourth 3% and on the fifth 2%. The entropy computation was done for each dataset for each team, giving the results in Table 2. It is clear that the teams varied in skill level, and that the datasets varied in difficulty. There also seems to be a learning effect as the teams gained experience with the Switchboard genre and with seeing spontaneous spoken dialog presented textually, as seen by the fact that 3 of the 4 teams did best on the last dataset.

These results were then normalized by row and by column. First we computed how much the average performance for each team differed from the global average. We then subtracted this difference from all scores by that team, to correct

	entropy per condition			
	bigram	trigram	unlimited	" + audio
Team A	8.21	8.41	7.92	7.14
Team B	8.61	8.47	7.71	6.88
Team C	8.85	8.21	7.61	7.00
Team D	8.56	8.02	7.71	7.39
average	**8.56**	**8.28**	**7.74**	**7.10**

Table 3: Per-Word Entropy, Normalized

	human		SRILM	
	relative entropy	perplexity reduction	relative entropy	perplexity reduction
Bigram	0.00	0%		
Trigram	–0.28	18%	–0.18	12%
Unlimited Text	–0.82	53%	-	-
" + Audio	–1.46	63%	-	-

Table 4: Relative Entropies, with bigrams at 0; and perplexity reductions, relative to bigrams

for their intrinsic skill level. Second, we computed the average performance on each dataset, and how much this differed from the global average. We then subtracted this difference from all scores on that dataset to correct for its intrinsic difficulty. Since each dataset was always presented in the same position (2238B always first, 2241B always second . . .), this normalization step also corrected for order effects. Table 3 shows the normalized results, re-ordered to have each condition in its own column.

Table 4 shows the benefits seen in each condition, relative to bigrams. For comparison, the right column shows the benefit of trigrams over bigrams when measured with SRILM run with default parameters on a large set of data (not the 116 specific words in the experiments): these figures correspond to a bigram perplexity of 102.9 and a trigram perplexity of 90.7. The fact that the human-estimated entropies came out higher than the SRILM-based estimates shows that our method is clearly not tapping all of the subjects' ability. The main problem is of course that subjects only got credit if they hit on the right word within 5 guesses; any intuitions regarding the likelihood of other possible words were untapped and unmeasured. However the fact that the improvement from bigrams to trigrams is in the same ballpark for humans and for SRILM adds credence to our estimates of the improvements possible in the other conditions.

6. Further Analysis

Entropy and perplexity measures count all words equally, but for speech recognition it is better to get the important words right: the content words matter more than the *um*s and *uh*s. We worried that the benefit of the full audio condition might be disproportionately due to such words. To see whether this was the case, we compared two sets of words, those which were guessed by the team with access to the full audio context but by no other teams, and those which were guessed by the team with access to the full lexical context but not by the teams with access only to the bigram or trigram context. We then roughly split these words into three categories, content words (*north, mismanaged*), function words (*like, the, and*) and discourse markers (*so, yeah, well*). The full textual context condition enabled the guessing of 22 words, of which 14% were content, 63%

function, and 24% discourse; the audio condition enabled the guessing of 16 additional words, of which 25% were content, 44% function, and 31% discourse. Thus it does seem that the added benefit of the full audio condition is not limited to words that don't matter.

To better understand the nature of the benefit of the full context audio condition, we attempted to infer the specific types of information that enabled the additional correct guess in each case. After we came up with an explanation for each of the 16 such words, we grouped them into categories.

The first category was 8 cases where the guessers were apparently benefiting from knowledge of the interlocutor's words. In 7 of these cases the correct word was a repetition of a word said by the interlocutor shortly before: *yeah*, *well*, *right*, *oh*, *over*, *machine*, and *live*. This is compatible with Ji and Bilmes' finding that the previous word by the interlocutor is a useful predictor [4]. Of these 7 cases, 1 was in response to a question using that word, *live*, and in 5 the correct word was utterance-initial: *yeah*, *well*, *right*, *oh*, *over*. In the remaining case, *point*, there was useful semantic context.

The second category was of 6 cases where the guessers apparently used the speaker's own prosody. This included 2 cases of list intonation, enabling the prediction of *or* and *the*, 1 topic-comment contour enabling the prediction of *is*, 1 continuation contour enabling the prediction of *and*, 1 clause-boundary contour enabling the prediction of *that*, and 1 disfluency context enabling the prediction of *maybe*.

Finally there was 1 word, *mismanaged*, that appeared to be predicted thanks to a negative tone of voice on the preceding words, and 1 inexplicable case, *like*.

7. Discussion

Our hypothesis was confirmed: humans do much better at guessing the next word when provided with all the available information — the full audio of both interlocutors — rather than just more textual context. This finding is suggestive but not conclusive: we do not know how to actually build language models that use such information, nor whether such improved language models will actually give us better speech recognition results. (Perplexity measures the extent to which the language model gives a good probability estimate for the correct word, but not the extent to which it gives appropriately lower estimates for rival words among the acoustic model's top candidates, which is also important for speech recognition.)

Nevertheless, this suggests that the current mainstream modularization of the speech recognition — in which only the acoustic models use signal information, with the language models operating purely in the symbolic realm — may need to be loosened.

As a way to estimate the entropy of spoken English dialog, our method has serious limitations. For example, our guessing game does not always elicit a probability estimate for the correct word, but it could probably be extended to do so, perhaps by presenting the word-to-guess partly obscured by noise, or by having subjects try to pick it out of a set with distractors. However, as our aim here is only to compare performance in different conditions, and there is no reason to think that additional information of various kinds would only benefit estimates for the 5 most likely words, it seems likely that the results do provide a useful estimate of potential overall benefit.

In passing, it is worth noting that to fully adapt Shannon's method for entropy estimation to work for spoken dialog, subjects would need to predict all the information-bearing aspects of the upcoming word, including its prosody and precise phonetic realization; this would be another interesting extension to develop. Doing so might cast light on problems of audio compression, by helping quantify the specific factors that are currently preventing us from compressing speech signals down from the kilobyte per second range to the tens of bits per second needed for the lexical content alone.

Although this paper has focused on language modeling for speech recognition, the lessons may also apply to dialog systems: language models used in the generation of the system's next utterance may also benefit from using more than just lexical context.

The methods developed will also be useful for more fine-grained future evaluations. We would like to tease out the specific contributions of the other track and of the speaker's track, and within that of the contributions of the phonetic details and of the prosody, and within prosody, of pitch, timing, rate, pause duration, energy, etc. Applying the method developed here to conditions where only such features are present, or only such features are masked, would enable us to identify the features with greatest promise for future language models.

8. Acknowledgments

We thank Alejandro Vega for SRILM wrangling and David G. Novick for helpful comments. This work was supported in part by the NSF under Grant No. 0415150 and by RDECOM via USC ICT.

9. References

[1] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proc. IEEE*, vol. 88, pp. 1270–1278, 2000.

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

[3] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Language and Speech*, vol. 41, pp. 439–487, 1998.

[4] G. Ji and J. Bilmes, "Multi-speaker language modeling," in *HLT*, 2004.

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

[6] D. G. Novick, K. Ward, and B. Corliss, "The effect of context on the intelligibility of dialogue," in *Eurospeech*, 1995.

[7] E. Brill, R. Florian, J. C. Henderson, and L. Mangu, "Beyond n-grams: Can linguistic sophistication improve language modeling?," in *Proc. of COLING-ACL-98*, 1998.

[8] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, pp. 50–64, 1951.

[9] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, vol. 73, pp. 1616–1624, 1985.

[10] A. M. Yaglom and I. M. Yaglom, *Probablity and Information*. D. Reidel Publishing (Kluwer), 1983.

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

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

[13] T. M. Cover and R. C. King, "A convergent gambling estimate of the entropy of English," *IEEE Transactions on Information Theory*, vol. IT-24, pp. 413–421, 1978.