UTEP Technical Report UTEP-CS-10-22

Studies in the Use of Time Into Utterance as a Predictive Feature for Language Modeling

Nigel G. Ward, Alejandro Vega

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518

email: nigelward@acm.org, avega5@miners.utep.edu

August 12, 2010

In [Ward and Vega, 2008] we examined how how word probabilities vary with time into utterance, and proposed a method for using this information to improve a language model. In this report we examine some ancillary issues in the modeling and exploitation of these regularities.

1 Introduction

Using time-into-utterance as a predictive factor for language modeling brings a benefit, although this is tiny compared to that obtainable by other non-lexical features [Ward and Vega, 2009], let alone ngram information. As some of the details of the computation may be relevant more generally, this report explores various aspects of this feature, using perplexity as the metric of quality and utility.

2 Length of Pause

A meta-parameter important for many useful temporal features is the length-of-pause to use to determine when a new utterance starts — that is, the amount of time that a speaker needs to remain silent in order to end the previous utterance and start a new one. Defining utterances is known to be tricky in general [Traum and Heeman, 1997, Deshmukh et al., 1998]. The trade-off for language modeling is that a shorter pause length will, *a priori*, give a large number of short utterances (bad for trigrams) while pushing more words into the early buckets, while a longer pause length will give a smaller set of longer utterances (better for trigrams unless the pause is so long that the pre-pause words are not informative) and push more words into later buckets.

We found that the best pause value, for the baseline trigram model, for the model augmented with time-into-utterance information, and for the model augmented with time-since-other's end information was 1.2 seconds in each case. However small variations in this value only slightly affected performance.

It is worth noting, however, that there may not be a single best pause value, as pauses of different lengths may tell us different things [Goldman-Eisler, 1967]. For example, the initial propensity to use *I* is much weaker if utterances are chosen as talk spurts set off by 2 second pauses than 1 second pauses (perhaps because the longer planning time makes it easier to retrieve other words or think less egocentrically).

3 Bucket Widths

Another set of meta-parameters specifies the widths for the buckets. In general we used 24 buckets: 5, each 0.1 seconds in width, from 0 to 0.5; 18, each 0.5 seconds in width, from 0.5 to 9.5; and one from 9.5 seconds out to infinity. We also experimented with combining some buckets, reasoning that doing so would reduce data sparsity, The best of those tried was obtained by merging some buckets before 0.5s and some after 7.0s; this gave a tiny but real performance benefit, of 0.024 points beyond that of the 24-bucket model (107.388 vs 107.412, where the baseline was 107.766).

4 Using Word Classes

As seen in Table 2 of [Ward and Vega, 2008], some buckets seem to be rich in semantically similar words, suggesting that variations in probability with time-into-utterance may be properties of word classes more than of individual words.

To examine this we looked first at two obvious categories: positive and negative emotion words. Using the Affective Norms for English Words, which lists for over 1000 common words the valence, on a 9-point scale [Bradley and Lang, 1999], Nisha Kiran, working in our laboratory, found that words in the first half second of utterances have significantly higher affect on average than words occurring later. Shreyas Karkhedkar used LIWC's positive emotion and negative emotion categories [Pennebaker et al., 2007], and found that the positive emotion words were almost twice as frequent from 0.2 to 0.4 seconds in as elsewhere, whereas negative emotion words are relatively rare until after 0.3 seconds (Figure 1).

It seems likely that these tendencies reflect rhetorical strategies, interpersonal strategies, parameters of memory retrieval processes, and/or cognitive processing constraints.

This suggests that word classes may provide better probability estimates in some cases, especially for less common words. For example, a word like *eighteen* does not occur often enough in any bucket to provide useful probability estimates (the S-ratios are all 1.0), but when pooled with the counts of other number words, there should be enough data to improve the predictions; that is, using classes may reduce the sparseness problem.

There are many ways to classify words, but for convenience we simply tried out some categories from the LIWC dictionary, including a few that appeared to be relevant (money, time, number,

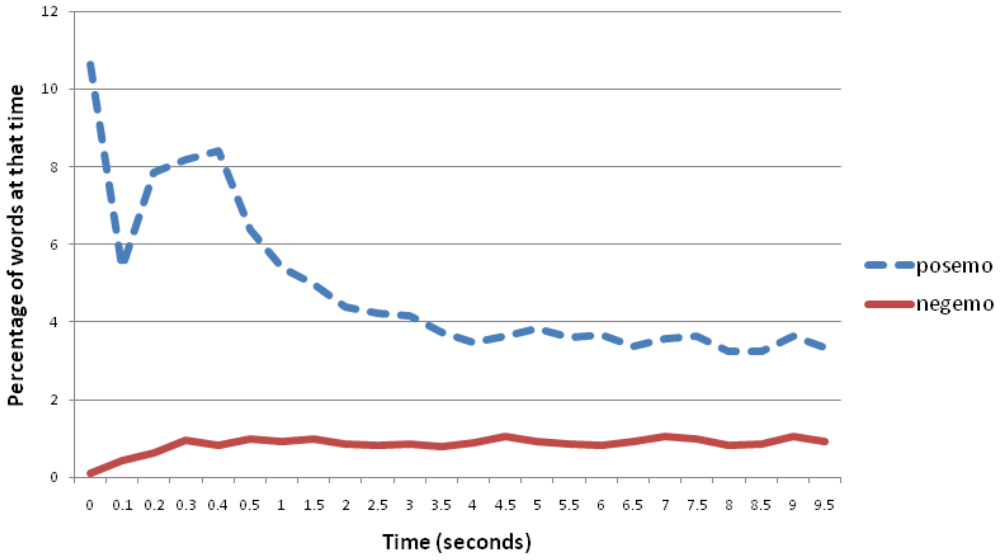

Figure 1: The probability of emotion words as a function of time into utterance.

LIWC class	examples	n	benefit
time words	*end, until, season*	239	0.006
number words	*second, thousand*	34	0.003
perceptual process words	*view, listen, feel*	273	0.003
articles	*a, an, the*	3	0.002
money words	*audit, cash, owe*	173	0.002

Table 1: Additional perplexity benefits obtained by using class-based S-ratios instead of word-based ratios, for certain classes. n is the number of words in the class in LIWC.

space), a few grammatical categories (verb, preposition, conjunction), the emotional categories (positive, negative), the cognitive mechanism category (*think, believe* ...) and the filler category. For each class we used only words of moderate frequency, excluding those among the most frequent 200 words overall, thinking that such words would almost certainly be modeled better by their own specific probabilities, and, as usual, excluding words not among in the top 5000 in the training. The classes that had positive contributions are shown in Table 1; the other classes tried gave nil or negative effects.

As the benefits were not large, we did not further use class-based estimates. However this is probably worth pursuing further: one could try more classes, soft classes, and classes found automatically by bottom-up clustering, derived perhaps from similarities in the patterns of temporal occurrence relative to various reference events in various corpora. Dimensionality reduction methods are another possible means to the same end [Bengio et al., 2003].

5 The Importance of *Temporal* Information

One could ask whether whether time into utterance, per se, is an informative feature. The fact that adding temporal information gives an advantage over a simple trigram model doesn't in itself prove that doing so would give an advantage over truly state-of-the-art models. In particular, one might ask whether the information given by time-into-utterance is truly new, or whether it is redundant to that handled by previous models, for example those which simply use more lexical context than trigrams, such as higher-order n-grams, models which use grammatical relations, trigger models, modes which support the application of contextual information across wider spans, and models which capture the evolution of cognitive state word-by-word (rather than second-by-second) [Gildea and Hofmann, 1999, Schwenk and Gauvain, 2004, Singh-Miller and Collins, 2007, Ji and Bilmes, 2006].

To determine whether the benefit could really be attributed to the novel idea here, the use of temporal information, we designed a simple experiment. If it were true that the time-of-occurrence of the words didn't actually matter, then we would see equal or better performance if we conditioned the probabilities not on time-into-utterance, but on word-into-utterance. For example, if the model were merely capturing syntactic regularities (for example, that verbs come in second position or later, *me* comes in third position or later), then a model using word-into-utterance should perform as well or better than a time-based one. Accordingly we built a version of the model where the buckets were based on ordinal position of the word in its utterance. (Computationally, this is equivalent to normalizing with respect to speaking rate, as measured by words per second.) Thus all words in second position fell into one bucket, all words in third position into another, and so on. Words after the 24th position were all cast into one bucket. The R and S values were then computed in the usual way, and the same experiments were run.

We also built a version conditioning probabilities on percent of time into utterance, using ten buckets. This would perform well if the probability variations depended on relative positions in utterances, for example, if there were a tendency that affected all words in the middle of utterances, regardless of whether the utterance is long or short.

Table 2 shows the results. Although conditioning on word-into-utterance also shows a benefit, it is less than that obtained by conditioning on time-into-utterance. Thus the temporal

	perplexity	benefit
baseline	107.766	-
time into utterance	107.412	0.354
word into utterance	107.449	0.317
percent into utterance	107.611	0.155

Table 2: Results of Conditioning on Various Measures of Distance into Utterance. The "benefit" is the perplexity decrease relative to the baseline model.

information itself is indeed providing useful extra information.

6 Testing in a Speech Recognizer

To explore how to use time-into-utterance information in a speech recognizer, Nisha Kiran modified the HTK system [Young et al., 2008]. In the lattice rescoring phase, we used the start time for each word hypothesis and looked up the time-based S-value, and then combined it with the (trigram) probability estimate based on the local context. This was done by a simple extension to HLRescore, done by modifying the function *TLatExpand()* in *HLat.c* to consult not only the standard language model but also the time-based values. These values were taken from simple look-up in an array that had been read in earlier. No normalization was done; that is, we used P_{bs} instead of P_n, to avoid unnecessary computation. Details of the modifications performed are given in [Kiran and Ward, 2008]. This would work also for the other contextual features described below.

In order to properly test whether the new language model actually improves speech recognition would require some additional work, including the creation of acoustic models suitable for Switchboard. Although we have not done this, two preliminary experiments in our laboratory, one with roughly trained acoustic models and one with off-the-shelf (VoxForge) acoustic models, obtained tiny but real decreases in error rates with time-into-utterance information [Kiran and Ward, 2008, Datta, 2009].

7 Patterns of Benefit and Harm

Table 3 of [Ward and Vega, 2008] illustrates how conditioning on time into utterance works. For that example the benefit of the model was strongest for the word *either*, which is common 0.2 to 2 seconds into utterances, and the word *said*, which is common 2 to 3 seconds into utterances. We examined such effects generally across the tuning data, looking for patterns in the ways that it helped and hurt prediction quality.

Although generally of value, adding this information had significant negative impact for one of the test dialogs. Upon listening, it turned out the speakers were familiar with each other, and in this respect the dialog was unlike the training data.

Later we did similar failure analysis on a model with additional predictive featues

[Ward et al., 2010]

Acknowledgment

This work was supported in part by NSF Grants IIS-0415150 and IIS-0914868 and REU supplements thereto, and by the US Army Research, Development and Engineering Command, via a subcontract to the USC Institute for Creative Technologies. We thank Nisha Kiran and Shreyas Karkhedkar for their contributions and discussions.

References

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[Bradley and Lang, 1999] Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

[Datta, 2009] Datta, S. (2009). Various test results. UTEP CS ISG internal memo, May 8, 2009.

[Deshmukh et al., 1998] Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. (1998). Resegmentation of Switchboard. In *ICSLP*, pages 1543–1546.

[Gildea and Hofmann, 1999] Gildea, D. and Hofmann, T. (1999). Topic-based language models using EM. In *Eurospeech*.

[Goldman-Eisler, 1967] Goldman-Eisler, F. (1967). Sequential temporal patterns and cognitive processes in speech. *Language and Speech*, 10:122–132.

[Ji and Bilmes, 2006] Ji, G. and Bilmes, J. (2006). Backoff model training using partially observed data: Application to dialog act tagging. In *HLT/NAACL*.

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

[Klakow and Peters, 2002] Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38:19–28.

[Pennebaker et al., 2007] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Technical report, LIWC.net (Linguistic Inquiry and Word Count).

[Schwenk and Gauvain, 2004] Schwenk, H. and Gauvain, J.-L. (2004). Neural network models for conversational speech recognition. In *Interspeech*.

[Singh-Miller and Collins, 2007] Singh-Miller, N. and Collins, M. (2007). Trigger-based language modeling using a loss-sensitive perceptron algorithm. In *IEEE ICASSP*.

[Traum and Heeman, 1997] Traum, D. and Heeman, P. (1997). Utterance units in spoken dialogue. In Maier, E., Mast, M., and LuperFoy, S., editors, *Processing in Spoken Language Systems*, pages 125–140. Springer-Verlag.

[Ward and Vega, 2008] Ward, N. G. and Vega, A. (2008). Modeling the effects on time-into-utterance on word probabilities. In *Interspeech*, pages 1606–1609.

[Ward and Vega, 2009] Ward, N. G. and Vega, A. (2009). Towards the use of inferred cognitive states in language modeling. In *11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 323–326.

[Ward et al., 2010] Ward, N. G., Vega, A., and Novick, D. G. (2010). Lexico-prosodic anomalies in dialog. In *Speech Prosody*.

[Young et al., 2008] Young, S. et al. (2008). The HTK book. from http://htk.eng.cam.ac.uk/docs/docs.shtml.