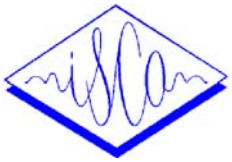

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

Modeling Disfluencies in Conversational Speech

Man-hung Siu † *Mari Ostendorf* ‡

† BBN Systems and Technologies, msiu@bbn.com
‡ Boston University, mo@bu.edu

ABSTRACT

Conversational speech is notably different from read speech in several ways, particularly in the presence of disfluencies but also in the frequent use of a small set of words that mark the flow of the discourse. Disfluencies are sometimes viewed as a "problem" in language modeling, where most previous work has focused on written text. In this paper, we take the view that disfluencies provide information themselves. In particular, we give evidence that filled pauses serve different functions, including marking linguistic unit and restart boundaries, and signaling hesitation where the speaker wants to hold the floor. The different functions can be connected to similar functions of other words common in spontaneous but not written speech, and the particular function affects the word conditioning choices in a variable ngram model. Thus, at least some of the idiosyncrasies of spontaneous speech can be viewed as a source of information for language modeling rather than an interruption in the linguistic structure.

1. INTRODUCTION

One difference between conversational speech and read speech is the presence of conversational speech markers such as *um, uh* and *you know*. In a conversational speech corpus such as Switchboard, these markers can account for a significant portion of the spoken word tokens, and their effect on language modeling is unclear. Previous work by Stolcke and Shriberg [1] indicates that filled pauses (*um* and *uh*) contain information for the prediction of neighboring words, and removing them from training and test actually increases the perplexity of neighboring words. They also argue that one function of *uh* and *um* is marking the beginning of a linguistic segment. In this paper, we investigated broader classes of conversational speech markers including conjunctions, discourse markers and filled pauses. For each of these markers, we investigated the possible functions they may serve and show that by modeling them appropriately, we can lower the perplexity of their neighboring words. Possible models include extending or reducing the ngram, using a class grammar in conjunction with a word ngram, and removing the disfluency markers from the word history.

The paper is organized as follows. First, we describe the data that we used and the conversational speech markers we looked at. Second, we list the possible functions these markers may serve. Third, we discuss the approach we took in separating the different functions of conversational speech markers and how to take advantage of these differences to improve the language model. Fourth, we present some experimental results. Finally, we conclude and discuss some possible future work.

2. DATA

All of the experiments reported in this paper are performed on the Switchboard corpus [3] which contains about 2.1M words of conversational dialogues over the telephone. In particular, we use a 1.6M word subset of the corpus which is annotated for conversational speech markers, linguistic segment boundaries and part-of-speech tags by LDC (http://www.cis.upenn.edu/~ldc). A 1.4M word portion is used for training and a 20K word portion is used for test. Within the annotated corpus, disfluencies and a small set of other conversational speech markers are labeled. These include coordinating conjunctions such as *and, but* or *so*; discourse markers such as *well, you know*; editing phrases such as *i mean*; and filled pauses such as *um, uh*, or *oh*. Repeat and repairs are also marked. For example,

> *{C But, } {F uh, } [[I, + I just,] + I] find it to be pretty offensive ...*

is a typical annotated conversational speech sentence starting with a coordinating conjunction (C) *but*, followed by a filled pause (F) *uh*, a repeat/repair of *I, I just, I* (+ indicates restart point) and then the sentence. In our work, a few common multi-word markers such as *you know* are treated as a single lexical item (*you_know*) to simplify the model. Just the conversational speech markers, without the repeat/repair, constituted more than 10% of the number of word tokens in the corpus.

Linguistic segment boundaries are also marked in the corpus. In conversational speech, sentences are rarely grammatically complete. For speech recognition in this domain, sentences are usually defined by "acoustic segment" boundaries which correspond to long stretches of silence or a change of turn. In contrast, "linguistic segment" boundaries mark a unit which the annotator interprets as complete but not necessarily as a grammatical sentence. In this paper, the term "sentence" will be used to denote a linguistic segment as marked in the annotated corpus. Experiments described in [4] suggest that language model perplexity can be reduced by working with linguistic segments (or "sentences") rather than acoustic segments.

CONJ	*and, but, so*
FILLER	*uh, um, oh*
DIS	*well, you_know*

Table 1: Conversational speech marker classes

In this work, we only look at the three most frequent different conversational speech marker classes. For each class, we model those words that are frequently observed. In Table 1, we show the conversational speech classes and the corresponding words that we model in this paper.

3. FUNCTIONS OF CONVERSATION MARKERS

The location of a conversational speech marker can be an indicator of the function of the marker. Stolcke and Shriberg [1] distinguished between sentence begin *uh*'s and sentence medial *uh*'s. In this work, we further partition markers in the middle of a sentence into repeat/repair markers and sentence medial depending on whether they are in the middle of a repeat/repair. We consider a conversational speech marker as "sentence begin" if it is the first word of a sentence or all its preceding words are also conversational speech markers. The reason is that in some instances, a sequence of conversational speech markers is at the beginning of a sentence, for example, *And uh he is ...* Although the *uh* is not the first word of a sentence, we conjecture that it serves the same function in terms of predicting the next word as in *Uh he is ...*

We hypothesize that markers can function either as an indicator of the beginning of a linguistic segment, the presence of a repeat/repair, or a filled pause (signaling hesitation). Our hypothesis that markers serve multiple functions raises the question of how to model the probability of their succeeding words. Choices we considered include extending or reducing the ngram, and skipping the markers in the word history.

4. APPROACH

To test the hypothesis that conversational speech markers may have different functions depending on their locations in the sentence, we initially looked at the perplexity of words following each marker class using a fixed trigram but comparing three locations: sentence begin, repeat/repair and sentence medial. Assuming that we know the locations of these markers in both training and test, we compared the perplexity for the same marker at different locations, as well as different ngram model variations. For each location, we then try to see whether we can improve the perplexity by either extending the ngram order or skipping the disfluent words. In other words, the probability of word w_i depends on the function of the history according to

$$F(w_{i-1}, w_{i-2}, \ldots) = \begin{cases} w_{i-1}, \ldots, w_{i-n} & \text{if } c_{i-1} \in D_1 \\ w_{i-1}, \ldots, w_{i-(n+1)} & \text{if } c_{i-1} \in D_2 \\ w_{i-2}, \ldots, w_{i-(n+1)} & \text{if } c_{i-1} \in D_3 \end{cases}$$

where D_j are different subsets of conversational speech marker classes and c_i is the class of word w_i.

The rationale here is that if the marker is an indicator of a new concept, we should not keep as much context information as for a regular word, i.e., reduce the ngram order. However, if the marker is blocking the context, we should remove it from the word history or extend the ngram to include more context to make up for the blockage.

Extending the ngram, however, inevitably increases the number of parameters to be estimated. In this work, we tried two ways to reduce the number of parameters needed when extending the ngram. First, we group conversational speech markers into disjoint classes (given marked data) and use a class ngram model, where

$$p(w_i|h_i) = p(w_i|c_i)p(c_i|F(c_{i-1}, c_{i-2}, \ldots)) \quad (1)$$

and h_i is the history of word w_i and c_i is the class of word w_i. All words are classes of themselves except the conversational speech markers. Instead of assuming that all information about succeeding words is captured by their classes, we can restrict the use of classes to the word history only, i.e.

$$p(w_i|h_i) = p(w_i|F(c_{i-1}, c_{i-2}, \ldots)). \quad (2)$$

For Equation 1 and 2, $F()$ is restricted to have $D_3 = \emptyset$. Finally, we can use a variable ngram on the words directly

$$p(w_i|h_i) = p(w_i|F(w_{i-1}, w_{i-2}, \ldots)). \quad (3)$$

If the functions of a marker at different locations are significantly different, then the conditional distribution of the succeeding words can be quite different. We can actually sharpen the conditional distributions by distinguishing conversational speech markers with different functions as different words, i.e. expand the dictionary. For example, we represent a sentence begin *uh* as *uh-B*, a sentence medial *uh* as *uh-M* and an *uh* in a repeat/repair as *uh-R*. Training and test can be modified this way by using the annotation. Since we have increased the size of our vocabulary on marked test data, we may hurt the overall test perplexity measure. When we do not have marking on test data, we can view the function-specific conversational speech markers as hidden states with the original words as seen observations. The perplexity of a test sentence can be computed through dynamic programming and summing of all possible paths through the sentence. Not only does this allow us to test on unmarked data, which is more readily available and realistic, the summing of different paths also smoothes the different conditional distributions and may alleviate the effect of data fragmentation. Furthermore, the test vocabulary is not increased and we can compare the perplexity result directly to the baseline ngram model.

5. EXPERIMENTS AND RESULTS

Our two approaches using word classes as shown in Equation 1 and Equation 2 degrade performance slightly. One possible explanation is that grouping words in the same conversational marker class may obscure the more important functional differences within and across words in the class. In addition, these words are well trained and there may be no advantage to grouping them into classes. We plan to investigate this more fully in the future, but relied on the last ngram model as shown in Equation 3 in the remaining experiments.

In our first experiment, we looked at the relationship between perplexity of words following different conversational speech markers and the position of these markers. These word perplexities are evaluated using a trigram trained on 1.4M words, where sentences are defined using the linguistic segmentation. Only 10k words are defined in the lexicon and out-of-vocabulary words in test are neglected in perplexity computation. Items which do not always function as conversational markers, such as *and*, *so* or *well*, are not modeled as having multiple functions in ngram training. For each word that follows each conversational speech marker, we compute four different perplexity numbers: the bigram perplexity, the trigram perplexity, the bigram and trigram perplexity with the conversational speech marker removed from the word history. Not all markers occur frequently in all three positions. In that case, only the frequent positions are evaluated.

As shown in Table 2, words following the same marker class at different positions have very different perplexity. Table 3 shows the perplexity of words following each marker word instead of the marker class, with the last column giving the number of times these markers occur in the 20K word test set. Most occur at the beginning of a sentence, except FILLER and *you_know*, which is consistent with distributional patterns of disfluencies observed by Shriberg [2] and the fact that discourse markers are usually sentence initial. In the case of FILLER, the perplexity of words following the three different positions are significantly different. The best ngram to use for these markers is also different, suggesting that the classes serve different functions. Trigrams are generally better than bigrams except for sentence medial *uh* or *you_know*. In that case, skipping the conversational marker helps, as also found in [1] for filled pauses.

In a second experiment, conversational speech markers that have different functions are distinguished as different words. To understand the effects of each marker, we train and test each marker independently. As in our first experiment, we evaluate the perplexity of the succeeding words in four different ways for all three functions, as shown in Table 4. Comparing Table 3 and Table 4, we notice that, relative to the case where only one function is represented, perplexity of words following all the markers except *so* is reduced for the the best ngram despite the effective increase in vocabulary size. In addition, the best ngram treatment changed and skipping is no longer useful. This is not compatible with the results in [1], a difference that may be explained by the more specific function dependence used here or by the fact that skips were accounted for in their training but not ours. For sentence begin markers, bigram and trigram perplexities are very similar. This is expected because in this detailed representation, the sentence begin information is encoded in the new word. For sentence medial markers, it is better to use bigrams instead of trigrams, which can be explained by the hypothesis that the conversational speech marker is used when people are uncertain about what to say and thus, previous context may not matter. In a repair, however, since people tended to repeat the same words before and after the marker, it is useful to have the extra context information.

In a third experiment, we ran the same models on unmarked test data. Similar to the previous experiments, we test each conversational speech marker word independently. In this case, the probability of a sentence is the sum of all paths through the model. Since we have to perform a dynamic programming over the sentence, we reported perplexity of the entire test set instead of only the words following the markers. Also, we use the best function-dependent ngram order we learned. Table 5 shows the perplexity of the test when we apply the new model on unmarked test data. As noted before, the vocabulary size in this case is the same as the baseline ngram, and thus the perplexity obtained is comparable with the baseline numbers. Since each marker constitutes a small portion of the overall data, the effect on the perplexity of the test is quite small. We notice that we obtain improvement in almost all the markers that we model. *Uh* and *you_know*, given that they occur more frequently, give the most gain. The perplexity of words following *so* degrade significantly when tested on marked data, but actually improve marginally on unmarked data. This may be explained by the fact that data fragmentation can hurt performance on marked test, but the smoothing effect of summing all paths compensates for some of the fragmentation effect.

In our last experiment, we built a model on all markers instead of one marker word at a time. One single ngram model is trained using the annotated data and tested on unmarked data. The combined model performs about 2.5% better than our baseline trigram model as shown in Table 6.

	Bigram	Trigram	Skip Bigram	Skip Trigram
CONJ-B	51	44	116	114
DIS-B	58	54	109	106
FILLER-B	55	47	99	98
FILLER-R	127	56	172	152
FILLER-M	1251	1745	1194	1703
YOU_KNOW-B	112	97	138	138
YOU_KNOW-R	85	21	60	53
YOU_KNOW-M	989	1431	790	1205

Table 2: Perplexity of words following conversational speech marker classes.

Word	Trigram	*and*	*but*	*so*
Perplexity	82.9	82.4	82.9	82.8
Word	*uh*	*oh*	*you_know*	*well*
Perplexity	81.8	82.9	82.5	82.8

Table 5: Perplexity of test after expanding the dictionary for each marker word.

6. CONCLUSIONS

In this paper, we described experiments in modeling conversational speech markers. We find that each marker can have multiple functions, including marking the beginning of a sentence, the presence of a restart, or a hesitation. The perplexity of words succeeding these markers can be very different and require different treatments, and it is useful

Word	Type	Bigram	Trigram	Skip Bigram	Skip Trigram	Count in Test
AND	CONJ-B	57	46	109	107	288
BUT	CONJ-B	37	38	104	103	147
SO	CONJ-B	53	44	148	148	80
WELL	DIS-B	57	56	114	112	108
OH	FILLER-B	44	45	78	76	75
UM	FILLER-B	19	15	268	269	24
UH	FILLER-B	81	62	93	92	193
UH	FILLER-R	128	55	148	162	55
UH	FILLER-M	1273	1776	1227	1746	272
YOU-KNOW	YOU-KNOW-B	112	97	138	138	48
YOU-KNOW	YOU-KNOW-R	85	21	60	53	8
YOU-KNOW	YOU-KNOW-M	989	1431	790	1205	96

Table 3: Perplexity of words following conversational speech markers using original dictionary.

Word	Type	Bigram	Trigram	Skip Bigram	Skip Trigram
AND	CONJ-B	44	46	169	165
BUT	CONJ-B	36	38	116	115
SO	CONJ-B	115	86	148	148
WELL	DIS-B	55	58	114	112
OH	FILLER-B	37	41	78	76
UM	FILLER-B	16	14	268	268
UH	FILLER-B	63	62	92	91
UH	FILLER-R	85	24	150	164
UH	FILLER-M	1076	1357	1247	1783
YOU-KNOW	YOU-KNOW-B	79	92	138	138
YOU-KNOW	YOU-KNOW-R	54	14	60	53
YOU-KNOW	YOU-KNOW-M	689	856	789	1202

Table 4: Perplexity of words following conversational speech markers after expanding the dictionary.

Grammar	Bigram	Trigram	New-model
perplexity	89.1	82.9	81.1

Table 6: Perplexity of test after expanding the dictionary for all marker word.

to model these functions separately with function-dependent word definitions and by treating the words differently in a variable ngram model. We also show that it is not necessary to have marked test data. By considering the marking as an unseen condition and summing over all possible functions, we can still improve perplexity of the test data.

In this work, we have looked at only a few classes of conversational speech markers, and therefore only obtained a small reduction in perplexity. To really take advantage of the extra information associated with some words in conversational speech, it is important to apply the general approaches of multi-function and variable ngram modeling to a broader set of words. In particular, the work of Stolcke and Shriberg [1] suggests that repeats can be handled by skipping words in the ngram history, and this class would fit well into our variable ngram model. Restarts are also relatively frequent in the switchboard corpus and could benefit from a variable ngram model. In addition, it may be useful to extend the approach of multi-function modeling to words in fluent regions that have different part-of-speech functions.

A problem that is raised by the multi-function word modeling is fragmentation of data. Here, we have avoided the problem by only working with word-function combinations that are frequently observed. Treating the function as hidden reduces the problem, but smoothing techniques are likely to be needed if the method is to be applied more broadly.

7. REFERENCES

1. A. Stolcke and E. Shriberg, "Statistical Language Modeling for Speech Disfluency", *Proc. 1996 IEEE ICASSP,* pp. 405-408.
2. E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D thesis, Department of Psychology, University of California, Berkeley, CA, 1994
3. J. Godfrey, E. Holliman and J. McDaniel. "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proc 1992 IEEE ICASSP,* pp 517-520.
4. M. Meteer and R. Iyer, "Modeling Conversational Speech for Speech Recognition", *Proc. 1996 Conf. on Empirical Methods in Natural Language Processing.*