LM Studies on Filled Pauses in Spontaneous Medical Dictation

Jochen Peters
Philips Research Laboratories, Weisshausstrasse 2, D-52066 Aachen
jochen.peters@philips.com

Abstract

We investigate the optimal LM treatment of abundant filled pauses (FP) in spontaneous *monologues* of a professional dictation task. Questions addressed here are (1) how to deal with FP in the LM history and (2) to which extent can the LM distinguish between positions with high and low FP likelihood. Our results differ partly from observations reported on *dialogues*. Discarding FP from all LM histories clearly improves the performance. Local perplexities, entropies and word rankings at positions following FP suggest that most FP indicate hesitations rather than restarts. Proper prediction of FP allows to distinguish FP from word positions by a doubled FP probability. Recognition experiments confirm the improvements found in our perplexity studies.

1 Introduction

Speech disfluencies are characteristic for spontaneous speech. Different disfluency types can be distinguished: Filled pauses (FP) such as ‘UH’ or ‘UM’, restarts or repairs, and repetitions. It is widely accepted that disfluencies considerably degrade the performance of speech recognition due to unexpected word sequences and due to the acoustic confusability of FP with short function words.

Most publications investigate different types of disfluencies in spontaneous *dialogues*. This paper, instead, reports analyses on spontaneous *dictation* of medical reports, i.e. on spontaneous *monologues*. Our studies focus on FP which are clearly dominant in our data (8% frequency) and which appear to be mainly associated with hesitations. As opposed to *dialogues*, FP are never used here to prevent interruptions by the dialogue partner as the speaker is searching for some formulation.

Central questions for language modeling are the optimal prediction of FP and its treatment in the LM history. Discarding FP from the history should be helpful if the sentence is continued after the interruption. For complete restarts, however, preceding words may be misleading and a conditioning on FP may be better. On *Switchboard*, (Stolcke and Shriberg, 1996) found that words following FP are better predicted if FP is *not* discarded from the history. This was attributed to the tendency of FP to appear at sentence boundaries where the word context from the preceding sentence appears to be harmful. Measurements after *sentence-internal* FP only, however, showed a local perplexity reduction for FP-cleaned histories by 20–30%. This was expected since most sentences are continued after the FP. These observations were confirmed by (Siu and Ostendorf, 1996) for sentence-internal FP but the local perplexity reduction due to skipping FP was much smaller. Interestingly, there, local trigram perplexities after FP are about 40% *worse* than bigram perplexities, no matter whether FP was discarded from the history or not. For a *How May I Help You* task, (Rose and Riccardi, 1999) report an improved LM prediction if FP is explicitly used for the conditioning of following words.

This paper is organized as follows: Section 2 describes the dictation task and our corpora. Section 3 lists three basic approaches to treat FP in trigram LMs. Section 4 discusses various perplexity comparisons, especially focussing on the question how to treat FP in the LM history. An extra study is concerned with LM uncertainties after FP. Finally, we analyze how well our LMs can discriminate FP from word positions. Section 5 summarizes our results and cites related speech recognition experiments.

2 Corpora

Our experiments are based on about 1.4 Mio. words of real-life dictated medical reports from various US hospitals which are partitioned into a *Train*, *Dev*, and *Eval* set (Table 1). The dictation style is fully spontaneous with repairs, repetitions, partial words, and – most frequent – filled pauses. Manual transciptions of these data include the annotation of FP. However, tags to distinguish between FP associated with hesitations, repairs, and restarts are missing. Here, as opposed to Switchboard, most FP are sentence-internal (ca. 70–80%).

A large background corpus provides formatted, i.e.

non-spontaneous reports which are mapped to the 60 k word list of our recognition system. To train LMs *including* FP this '*Report*' corpus was stochastically enriched with FP. Considering single or sequential FP/s as hidden events in the reports we randomly inserted them with their a-posteriori probabilities in the given word contexts. These probabilities are estimated using a bigram from the spontaneous training data. A similar approach was mentioned without details in (Gauvain et al., 1997). They report *increasing error rates* if too many FP are inserted by this method into the LM training data. This might be explained by the following observation: Adding FP in a context-dependent fashion diminishes the number of observed bi- and trigrams since words typically preceding or following FP "loose individual contexts" if many FP are inserted. For our *Report* corpus, the number of distinct uni- + bi- + trigrams drops from 107 M (without FP) to 98 M (after FP enrichment).

Corpus	Spont	# words	FP rate	OOV rate
Train	yes	1314 k	8.2 %	0.45 %
Dev	yes	81 k	6.3 %	0.23 %
Eval	yes	53 k	7.0 %	0.30 %
Report	no	1071 M	7.9 %	0.31 %

Table 1: *Characteristics of text corpora including FP. (The high OOV rate on* Train *is due to an extension of this data set after fixing our 60 k word list.)*

3 Language models

Mapping all filled pauses to a unique symbol FP we compare three LM approaches:

1. We treat FP as a regular word which is predicted by the LM and which conditions following words.
2. We use the LM for both words and FP but discard all FP from the conditioning histories.
3. We use a fixed, *context-independent* probability for FP of 0.08 (FP unigram). Here, words are predicted with a FP-free LM skipping FP in the history (as in approach 2.). Normalization is achieved by a scaling of word probabilities with $(1 - p_{\text{fix}}(\text{FP}))$. This simplistic approach relieves us from the need of FP-tagged corpora, but we clearly loose the discriminative prediction of FP.

Approaches 1. and 2. use count statistics with FP. As discussed above, the inclusion of FP "destroys" some possible word transitions. To exploit the knowledge about possible FP-cleaned transitions we successfully tested *merged* counts. Here, the sets of observed M-Grams in the corpus with and without FP are joined and counts of common M-Grams are added. (Doubled counts use modified discounting and the reduced FP-rate is compensated using marginal adaptation (Kneser et al., 1997).)

All reported results are obtained with linearly interpolated models from the spontaneous *Train* and the non-spontaneous *Report* corpus. (For trigrams, perplexities of these two component LMs are 95% and 19% above the perplexity of the interpolated LM.)

4 Experimental results

The three approaches are evaluated in terms of the overall perplexity (PP) and local values: PP_{FP} and PP_{word} are measured at FP and word positions only, and $\text{PP}_{\text{after }*}$ are measured immediately thereafter.

The results in Table 2 show that *discarding* FP from the history clearly *improves* the performance (2. versus 1.). The overall PP is reduced by 4–5%. Big reductions by 30–40% are found at positions immediately following FP. This, and the improvements as we go from bi- to trigrams (which are contrary to (Siu and Ostendorf, 1996)), indicates that sentences are – on average – continued after FP.

Using *merged* counts further improves our LMs. Gains are (almost) additive to those from FP-skipping. Especially, $\text{PP}_{\text{after FP}}$ decreases by another 10% for approach 2. which shows that the "recovered" FP-free M-Grams are indeed valuable if we use FP-free histories.

A comparison of $\text{PP}_{\text{after FP}}$ and $\text{PP}_{\text{after word}}$ confirms the common knowledge that word prediction after FP is pretty hard. Even the unigram perplexity is almost 50% higher for words following FP than for words following fluent contexts. This supports (Shriberg and Stolcke, 1996) where the reduced predictability after FP is partly attributed to the chosen words in those positions.

For trigrams, the discrepancy between $\text{PP}_{\text{after FP}}$ and $\text{PP}_{\text{after word}}$ is much larger. Asking "how unexpected is a word in a given context ?" we evaluated the entropy $H(h_i) = -\sum_w p_{\text{LM}}(w \mid h_i) \cdot \log p_{\text{LM}}(w \mid h_i)$ and the rank R_i of w_i following h_i in the distribution $p_{\text{LM}}(* \mid h_i)$. Both quantities were averaged over histories h_i ending on FP or on words.[1] Note that $e^{H_{\text{mean}}}$ represents a perplexity for the case that words following each history are distributed according to $p_{\text{LM}}(* \mid h)$. An actually measured PP above $e^{H_{\text{mean}}}$ indicates a bias in the corpus towards words with low $p_{\text{LM}}(w \mid h)$. The results from Table 3 show almost *no* such bias after words. After FP, however, following words are clearly biased to low probabilities within the trigram distributions. Also, the mean ranks are considerably higher after FP than after words.

Together, these findings support our impression that FP often represents a hesitation where the speaker is searching for a less common word or formulation.

[1] (Shriberg and Stolcke, 1996) report increasing entropies at FP versus word positions. Our studies confirm these results.

Table 2: *Perplexities and error bars (95% confidence) on the Dev set for linearly interpolated LMs.*

LM range	Appr.	Counts	$\mathrm{PP}_{\mathrm{overall}}$	$\mathrm{PP}_{\mathrm{FP}}$	$\mathrm{PP}_{\mathrm{word}}$	$\mathrm{PP}_{\mathrm{after\ FP}}$	$\mathrm{PP}_{\mathrm{after\ word}}$
			size: 81 k	5 k	76 k	5 k	76 k
Unigram	1. = 2.	with FP	786.5 ± 14.0	12.4 ± 0.0	1042.2 ± 17.9	1136.8 ± 85.7	767.1 ± 14.0
	3.	FP-free	786.4 ± 14.0	12.5 ± 0.0	1041.7 ± 17.9	1136.3 ± 85.6	767.0 ± 14.0
Bigram	1.	with FP	115.6 ± 2.4	11.0 ± 0.2	135.7 ± 3.0	957.5 ± 76.0	100.2 ± 2.2
	2.	with FP	112.0 ± 2.4	11.1 ± 0.2	131.0 ± 2.9	579.3 ± 50.6	100.2 ± 2.2
	3.	FP-free	110.9 ± 2.3	12.5 ± 0.0	128.6 ± 2.8	503.5 ± 42.6	100.1 ± 2.1
Trigram	1.	with FP	61.4 ± 1.4	10.4 ± 0.2	69.3 ± 1.6	605.9 ± 49.9	52.6 ± 1.2
	1.	**merged**	60.3 ± 1.3	9.8 ± 0.2	68.2 ± 1.6	646.3 ± 53.2	51.4 ± 1.1
	2.	with FP	59.2 ± 1.3	10.9 ± 0.2	66.4 ± 1.5	427.2 ± 39.8	51.8 ± 1.2
	2.	**merged**	57.5 ± 1.2	11.4 ± 0.2	64.2 ± 1.5	383.6 ± 34.5	50.5 ± 1.1
	3.	FP-free	57.9 ± 1.2	12.5 ± 0.0	64.3 ± 1.5	367.0 ± 33.0	51.1 ± 1.1

Table 3: *Measured PP versus* $e^{H_{\mathrm{mean}}}$ *and mean rank after histories ending on FP or on word (using pruned LMs).*

Range	Appr.	$\frac{\mathrm{PP}}{e^{H_{\mathrm{mean}}}}$ after		R_{mean} after	
		FP	word	FP	word
Uni	1. = 2.	1.6	1.1	1301	881
Tri	1.	2.6	1.2	1050	336
	2.	5.1	1.2	719	335

Recall that approach 3. cannot discriminate between positions with an increased or reduced FP probability. To evaluate the discrimination for approaches 1. and 2. we calculated $p(\mathrm{FP}|\ h)$ instead of $p(w \mid h)$ at each position in the corpus. The crucial result is that the mean FP probability is reduced by 48% and 45% (approach 1. and 2.) at *word* as compared to *FP* positions. This is an important feature of these LMs since small FP probabilities reduce confusions of proper words with FP.

5 Summary

Concerning the question how to best predict words next to FP we get the following results for our spontaneous dictation task: *Discarding* FP from the LM histories reduces $\mathrm{PP}_{\mathrm{overall}}$ by 4% and $\mathrm{PP}_{\mathrm{after\ FP}}$ by 30%. (The latter reduction is bigger than in (Stolcke and Shriberg, 1996). Note that our measurements *include* positions after sentence-initial FP which *suffer* from the FP-removal.) Count *merging* with FP-free M-Grams gives an additional reduction of $\mathrm{PP}_{\mathrm{overall}}$ by 3% and of $\mathrm{PP}_{\mathrm{after\ FP}}$ by 10%.

Comparisons of local perplexities and studies of entropies and word rankings indicate that FP often represents a hesitation as speakers are searching for a less common word or formulation which is hard to predict.

At positions following FP, trigrams outperform bigrams. This together with gains from discarded FP suggests that FP rarely represent sentence breaks or restarts.

We presented a new analysis of the LM's power to discriminate between FP and word positions. Predicting FP with a trigram allows to lower the FP probability at *word* positions by almost 50%. This is an important feature to reduce confusions of words with FP.

Speech recognition experiments are published in (Schramm et al., 2003). Using *merged* counts and *discarding* FP from the LM history reduces the error rate on *Eval* by 2.2% (relative) while PP is reduced by 7%.

Acknowledgements: I would like to thank Xavier Aubert and Hauke Schramm for their contribution of speech recognition experiments.

References

J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker. 1997. *Transcribing Broadcast News: The LIMSI Nov96 Hub4 System*. Proc. DARPA Speech Recognition Workshop.

R. Kneser, J. Peters, D. Klakow. 1997. *Language model adaptation using dynamic marginals*. Proc. EUROSPEECH, 4:1971–1974.

R.C. Rose, G. Riccardi. 1999. *Modeling disfluency and background events in ASR for a natural language understanding task*. Proc. ICASSP, 1:341–344.

H. Schramm, X. L. Aubert, C. Meyer, J. Peters. 2003. *Filled-pause modeling for medical transcriptions*. To appear at IEEE Workshop an Spontaneous Speech Processing and Recognition.

E. Shriberg, A. Stolcke. 1996. *Word predictability after hesitations: a corpus-based study*. Proc. ICSLP, 3:1868 -1871.

M. Siu, M. Ostendorf. 1996. *Modeling disfluencies in conversational speech*. Proc. ICSLP, 1:386–389.

A. Stolcke, E. Shriberg. 1996. *Statistical language modeling for speech disfluencies*. Proc. ICASSP, 1:405–408.