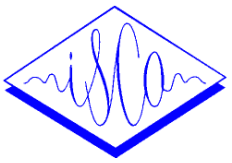

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

IMPROVING LANGUAGE MODEL PERPLEXITY AND RECOGNITION ACCURACY FOR MEDICAL DICTATIONS VIA WITHIN-DOMAIN INTERPOLATION WITH LITERAL AND SEMI-LITERAL CORPORA

Guergana Savova, Michael Schonwetter, Sergey Pakhomov,

Lernout and Hauspie, 5221 Edina Industrial Blvd, MN 55439, USA

emails: GSavova@lhsl.com, MSchonwetter@lhsl.com, SPakhomov@lhsl.com

ABSTRACT

We propose a technique for improving language modeling for automated speech recognition of medical dictations by interpolating finished text (25M words) with small human-generated literal or/and machine-generated semiliteral corpora. By building and testing interpolated (ILM) with literal (LILM), semiliteral (SILM) and partial (PILM) corpora, we show that both perplexity and recognition results improve significantly with LILM and SILM; the two yielding very close results.

1. INTRODUCTION

Language modeling for automatic speech recognition (ASR) for medical dictations has to deal with the characteristics of spoken language: disfluencies (filled pauses, restarts, repetitions), metalanguage, idiosyncrasies. The official transcriptions are fully edited texts that are readily available in large quantities and used for language modeling. However, this kind of corpus rarely represents n-grams that occur frequently in spontaneous speech. For example, a phrase like "um correction ah please go back to two and add the following" would not make it in the final edited text because it is considered extraneous material and is weeded out from the finished transcription. Language models (LM) trained on fully edited text are often not appropriate for the quasi-spontaneous speech of medical dictations.

One way to build LMs suitable for medical dictations is to train the LM on human-transcribed literal corpora that include all the speech actually uttered by the talker. However, this technique is extremely expensive, since millions of words are needed to build a reliable LM and the preparation of a suitable corpus would be unacceptably costly and time-consuming.

In this paper, we will show that *interpolation of a small literal and/or semiliteral corpus* can be used to train LMs that yield considerable improvements in LM perplexity and recognition accuracy when compared to LMs trained solely on fully edited text. In our approach, a small corpus (< 1 million words) of human-generated literal and/or machine-generated semiliteral transcriptions is interpolated with a larger fully edited corpus (> 20 million words). The interpolated LM (ILM) incorporates the variety and statistical reliability of the large corpus as well as spoken language characteristics; the latter are derived in a predictable way through an interpolation at a determined weight from the smaller literal or semiliteral corpus. Moreover, interpolation upon optimal weights yields better results than simply combining edited and literal texts into a single training set.

Modeling filled pauses (FPs) improves LMs for spontaneous speech (Shriberg, 1996; Pakhomov, 1999; Pakhomov and Savova, 1999). FP representation in the LMs significantly lowers the LM perplexity results. If FPs (ah, um) are not modelled, in our system they are usually recognised incorrectly as "thumb" or "arm". In our approach, we are going to model FPs following Pakhomov (1999) and Pakhomov and Savova (1999).

Our work makes use of the technique of linear interpolation. Rosenfeld (1994) offers an overview of combining different information sources for improving the LM performance. One way is to linearly interpolate two LMs with weights optimised for some particular data set. The combined model is, as Rosenfeld points out, an arithmetic average of the component models, while perplexity on the other hand, is a geometric average. Therefore, small linear contributions result in significant perplexity reductions (Rosenfeld, 1994).

We are concerned solely with within-domain adaptation. Across-domain adaptation combines sources from different discourse domains to produce a new LM, e.g. interpolating a Primary Care LM with Pediatrics LM. Within-domain adaptation improves the perplexity results as well (Besling and Meier, 1995).

Literal transcriptions of speech are used for acoustic modeling for continuous speech recognition systems. As the results below show, acoustic and language modeling have much more in common than one might think. The training corpus for the acoustic models can be "recycled" for language modeling. The technique described below has been implemented successfully in our system for automated telephony speech recognition of medical dictations for about a year.

2. MEDICAL DICTATION SPEECH

Our system is designed for speech recognition of medical dictations. The term quasi-spontaneous speech can be used when referring to medical dictations. Quasi-spontaneous speech has the characteristics of both prepared and unprepared speech. Physicians in the United States are required to document, usually through dictating, the results of each patient's examination. These dictations are then typed by medical transcriptionists who edit out the extraneous material and produce a neatly organised piece of written language. The final edited texts are kept for the records. Usually dictations are done over the phone, which poses acoustic signal quality issues that are beyond the scope of this paper.

3. TRAINING CORPORA

This section describes where each of the training corpora (headings A through H below) come from, how many words it is comprised of and whether artificial FP conditioning is applied to it. There are three types of corpora – partial, literal and semiliteral. We call a finished transcription "partial" because it represents only partially what was actually spoken by the talkers. Disfluencies such as repetitions, repairs and false starts, and extraneous material are edited out by the medical transcriptionists. These are the reports kept for the records. The language in the partial corpus is also edited for grammar. Literal transcriptions are those that include all the speech uttered by the talker. Semi-literal transcriptions (Pakhomov and Schonwetter, 2000) are generated by reusing partial transcriptions as training data for a finite state LM. The LM is smoothed with FP conditioning and a background LM constructed from utterances that are usually spoken but are not transcribed (e.g. "um delete that please"). The smoothed LM is used to re-recognize the dictation and a confidence measure is applied to the output to derive the semi-literal transcription.

All data come from our medical transcription operation. There is no overlap between the files comprising any of the corpora. The same talkers are represented in the training corpora. There is no file or talker overlap between the training and testing corpora. The naming convention for the corpora follows this format: type of corpus, word count, whether or not artificial FP conditioning was applied indicated by *fpcond* or *no_fpcond*. No artificial FP conditioning is applied to literal or semi-literal transcriptions and the *no_fpcond* extension in the name is left out.

A. partial_25M_no_fpcond: a corpus of 25,712,834 words from fully edited transcriptions, contains all available dictations from January'98 through July'99. The corpus represents over 200 talkers of mixed gender and professional medical status. A short example from a finished transcription file follows:

The patient was admitted, placed on heparin and Coumadin, bedrest with elevation of her left lower extremity. She was placed on Rocephin for her cellulitis and started on some albuterol nebs for her lungs. While in the hospital, she did quite well.

B. partial_25M_fpcond: filled pause conditioned

C.*partial_25M_no_fpcond*. The method for artificial FP conditioning follows Pakhomov (1999) and Pakhomov and Savova(1999).

D. partial_550K_no_fpcond: additional 550K words of fully edited finished transcriptions.

E. partial_550K_fpcond: artificial FP conditioning applied to partial_550K_no_fpcond

F. literal_550K: Hand-transcribed literal corpus of 550,000 words of all available word-by-word transcription data from approximately 20 talkers. As this is a literal, word-by-word transcription, all disfluencies, pronounced punctuation and metalanguage (e.g. "Thank you", "This is Doctor Smith dictating on patient John Smith", "End of dictation, thank you", "Correction, please go back to assessment and include there the following", etc.) are kept in the text. No final editing of any kind is applied to the corpus. Here is the same sample as above but as a literal file:

next line the patient was admitted comma placed on ah um heparin ah heparin and coumadin bedrest with elevation of her right correction left lower extremity period she was placed on rocephin for her cellulitis and ah started on ah some albuterol nebs for her lungs period ah while in the hospital she did quite well period this was doctor smith dictating on patient jane doe end of dictation thank you

G. *semiliteral_550K:* A training corpus of 550,000 words from December'99 – January'00 files from approximately 100 talkers. A sample of a semi-literal file is:

he should stop now and use of nicorette chewing gum of ah the ah rest of review of systems seems to be negative he's controlling pain with oxycontin and darvocet

H. *literal_275K_and_semiliteral_275K*: Randomly selected files from *literal_550K* of 275K words and *semiliteral_550K* of 275K words are combined in this corpus.

4. LANGUAGE MODELS (LMS)

Two groups of statistical trigram language models were built from the corpora or a combination of the corpora – non-interpolated (Table 1) and interpolated (Table 2) LMs. Bigram cutoffs were set to 0, trigram cutoffs were set to 1. The LMs were built using the Entropic Cambridge Research Lab's (ECRL) Transcriber language modeling tools (Valtchev et al, 1998). The same word list was used for building all LMs (25,000 words). The naming convention for the LMs is the names of the constituent corpora, e.g. literal_550K.trig was built from *literal_550K* corpus; part_25M_no_fpcnd_smltr_550K was built by interpolating *partial_25M_no_fpcond* LM with *semiliteral_550K* LM.

The Interpolated LMs for this research project followed the recommendations from the Interpolate option of the CMU SLM Toolkit (CMU SLM Toolkit, 1994), based on Rosenfeld (1994). For that purpose held-out data was compiled representing the same talkers from the test set with approximately 1000 words each. The word count for the held-out data is 13,422 words equally distributed among the test set talkers. There is no overlap between files in the held-out data and the test set, and between the held-out data and the training sets. The held-out files were used only to get the optimised weights.

Language Model	Training Corpus
Partial_25M_no_fpcond	Partial_25M_no_fpcond
Partial_25M_fpcond	Partial_25M_fpcond
Literal_550K	Literal_550K
Semiliteral_550K	Semiliteral_550K
Literal_275K_and_semi literal_275K	Literal_275K_and_semilite ral_275K

Table 1: Non-interpolated language models (NILMs)

As Rosenfeld (1994:79) describes "the program Interpolate

Language Model	weight for	weight for	weight for	weight for	weight for
	partial_25M_no_fpcond	partial_25M_fpcond	literal_550K	semiliteral_550K	literal_275K_and_semiliteral_275K
part_25M_no_fpcnd_lit_550K	0.495		0.505		
part_25M_no_fpcnd_smltr_550K	0.501			0.499	
part_25M_no_fpcnd_lit_smltr	0.458				0.542
part_25M_fpcnd_lit_550K		0.514	0.486		
part_25M_fpcnd_smltr_550K		0.539		0.461	
part_25M_fpcnd_lit_smltr		0.495			0.505

Language Model	weight for	weight for	weight for	weight for
	partial_25M_no_fpcond	partial_25M_fpcond	partial_550K_no_fpcond	partial_550K_fpcond
part_25M_no_fpcnd_part550K_no_fpcnd	0.779		0.221	
part_25M_no_fpcnd_part550K_fpcnd	0.71			0.29
part_25M_fpcnd_part550K_no_fpcnd		0.766	0.233	
part_25M_fpcnd_part550K_fpcnd		0.778		0.221

Table 2: Interpolated language models and weights for the constituent corpora (ILMs)

takes as input any number of probability streams. These are assumed to be the output of several language models on a common set of data [the held-out data, authors' note]. The program then runs the Estimation-Maximization (EM) algorithm, to find the set of weights that, when used for linearly interpolating the models, will result in the lowest perplexity on that data." It must be noted that the recommendation weights are tailored towards the held-out data.

5. TEST SET

The test set is distributed among 12 talkers, all Primary Care medical professionals – 6 male and 6 female. Each is represented by approximately 10 minutes of unedited speech. The total word count for the test set is 19,443 words. Training and testing talkers are completely different. The audio files are telephony speech with noise, background speech and, from time to time, static. We used our speaker independent (SI) telephony acoustic model for the recognition runs.

6. RESULTS AND DISCUSSION

Both perplexity and recognition results on the test files were obtained. Perplexity is a metric for evaluating language models. "Perplexity can be intuitively thought of as the weighted average number of choices a random variable has to make." (Jurafsky and Martin, 2000:225). Lowered perplexity indicates improvement in the LM. Perplexity alone is not a reliable indicator of accuracy. A complete recognition test has the advantage of showing the interaction between the acoustic and the language model and assessing the LM goodness on the overall search effort of the recognizer. Perplexity results, sorted from lowest to highest, are shown in Table 3. The shaded area represents the non-interpolated LMs (NILMs); the unshaded/clear area – the interpolated LMs (ILMs). Perplexity measures were calculated using the ECRL's LM Transcriber tools. Recognition results are shown in Table 4. Recognition tests were run on 4 LMs representative of the four groupings of the perplexity results.

Perplexity results fall naturally into 4 groups with clearly outlined ranges: group I with a perplexity result range 102 – 106; group II with perplexity result range 156 – 177; group III with range 216 – 241 and group IV with perplexity result range 272 – 328. The top four LMs (group I) are all LMs interpolated with a literal corpus, and an LM interpolated with the combination of literal and semiliteral corpora; the next best results are for group II which includes two LMs interpolated with semiliteral corpus and two non-interpolated LMs that include some representation of literal files in their training corpora. The worst perplexity results are with non-interpolated LMs, which did not have any literal representation in their training corpora. There is a 225-perplexity point difference between the top and the bottom LMs (Table 3). As it was pointed out before, the goal of this experiment was not to weigh interpolated vs. non-interpolated LM in general; rather it is to show that *witnin-domain interpolation with a small literal or semiliteral corpus improves both perplexity and recognition results as compared to interpolation with partial (that is fully edited) text or non-interpolated LM.*

language model	prplx	oov rate
1. part_25M_fpcnd_lit_550K	102.43	1.58%
2. part_25M_fpcnd_lit_smltr	103.66	1.58%
3. part_25M_no_fpcnd_lit_550K	104.05	1.58%
4. part_25M_no_fpcnd_lit_smltr	105.96	1.58%
5. literal_275K_and_semiliteral_275K	156.31	1.58%
6. part_25M_fpcnd_smltr_550K	159.86	1.58%
7. part_25M_no_fpcnd_smltr_550K	166.79	1.58%
8. literal_550K	176.48	1.58%
9. part_25M_fpcnd_part550K_no_fpcnd	216.34	1.58%
10. part_25M_fpcnd_part550K_fpcnd	220.80	1.58%
11. part_25M_no_fpcnd_part550K_fpcnd	222.40	1.58%
12. partial_25M_fpcond	241.16	1.58%
13. semiliteral_550K	272.08	1.58%
14. part_25M_no_fpcnd_part550K_no_fpcnd	292.14	1.58%
15. partial_25M_no_fpcond	327.10	1.58%

Table 3: Language model perplexity results, sorted by perplexity in ascending order (shaded area -> non-interpolated LMs; clear area -> interpolated LMs)

The results are consistent with Pakhomov, 1999 and Pakhomov and Savova, 1999 about the contribution of artificial filled pause conditioning of edited text (partial_25M_no_fpcond vs. partial_25M_fpcond). There is an improvement of 85.16 points in perplexity with the filled pause conditioned model as compared to the non filled paused conditioned one as Table 3 indicates. Results from semiliteral_550K are better than the non-filled pause conditioned LM probably because of the representation of patterned disfluencies modelled by the

semiliteral generation algorithm; however they are worse than partial_25M_fpcond probably because of the smaller corpus size for semiliteral_550K, therefore limited representation of trigrams and bigrams.

The results from this study also show that an LM from a small literal corpus of 550K words (literal_550K) decreases perplexity by 150 points as compared to the much bigger non-fp conditioned LM from a 25M word corpus (partial_25M_no_fpcond). However, a small corpus does not offer the variety of trigrams and bigrams that an interpolated LM built from literal and partial corpora has the potential to include. That is our explanation for the ordering of the top LMs.

A filled pause conditioned LM interpolated with a literal corpus yields the best results in perplexity and, as will be demonstrated below, recognition accuracy. The contribution of the literal corpus lies in the fact that it has representations of naturally occurring speech as well as the characteristics of semi-prepared speech. The top four LMs include interpolation with some amount of literal data – 550K words in LM 1 and 3, and 275K words in LM 2 and 4 (LM numbering refers to Table 3). The difference between the perplexity results of these four models is non-significant – within 4 points.

An interesting fact springs from the results in Group II – LMs 5, 6, 7 and 8 (LM numbering refers to Table 3). Interpolating with semiliteral files the size of the truth files does not yield as good results as the literal interpolation, however they are much better than both the non-filled pause conditioned and filled pause conditioned LMs from partial transcripts. This is truly remarkable since generating semiliteral files is done automatically and does not require any additional human effort besides starting the program.

Group III, LMs 9, 10, 11 and 12, includes interpolation with the additional 550K partial corpus. These LMs were built in order to get a truly controlled comparison when interpolating the same amount of words but different modes of transcribing – 550K words of partial, semi-literal and literal transcriptions.

The above discussed perplexity results fit nicely with recognition accuracy results (Table 4). We ran actual recognition on the test set with 4 language models, one from each perplexity result group. The baseline is *partial_25M_no_fpcond* LM, which seems to be the traditional method of building LMs for dictation applications. Again, we must point out that the test audio files were noisy telephony speech, which is not the most favourable condition. We believe that if the quality of the audio signal were better, the recognition results would be much better as well.

Interpolating the artificially FP conditioned corpus of partial transcriptions with the literal transcripts brings the best recognition and LM perplexity results. Interpolating the artificially FP conditioned corpus with semi-literal transcriptions of the same word count as the literal corpus brings results very close to the best. Applying just artificial FP conditioning to a partial corpus yields considerable improvements. Interpolating with a partial corpus and an FP conditioned LM brings a negligible improvement of 0.57 (the difference when subtracting the absolute values for LM12 and LM10 from Table 4). We conclude that an LM with an interpolated representation of literal or/and semiliteral transcriptions is superior because the literal and/or semiliteral corpora include the elements of spontaneous speech, which an edited text does not.

language model	% change	absolute
12. partial_25M_fpcond	19.53%	7.35
10. part_25M_fpcnd_part550K_fpcnd	21.05%	7.92
6. part_25M_fpcnd_smltr_550K	26.44%	9.95
1. part_25M_fpcnd_lit_550K	27.80%	10.46

Table 4: Recognition accuracy increase as compared to baseline partial_25M_no_fpcond (LM 15 from Table 3) (the LM numbering refers to Table 3)

7. CONCLUSIONS

In this experiment, it was shown that building an LM via within-domain interpolation of a larger FP conditioned corpus (> 20 million words) with a small human-generated literal corpus (< 1 million words) improves considerably both LM perplexity and recognition results for an automated speech recognition telephony system (ASRTS) for medical dictations. Interpolating a larger FP conditioned corpus (> 20 million words) with a small machine-generated semi-literal corpus (< 1 million words) yields respectable results. The baseline LM built from fully edited text performed poorly.

8. REFERENCES

1. Besling, S. and Meier, H. 1995. "Language Model Speaker Adaptation". In Proc. EUROSPEECH.
2. CMU SLM (Carnegie Mellon Statistical Language Modeling) Toolkit, Rev. 1.0, 1994.
3. Jurafsky, D. and Martin, J. 2000. Speech and Language Processing. Prentice Hall.
4. Pakhomov, S. and Schonwetter, M. 2000. A method and system for generating semi-literal transcriptions for speech recognition. Patent pending. Serial No.: 09/487398
5. Pakhomov, S. 1999. "Modeling Filled Pauses in Medical Dictations." Proc. ACL' 99.
6. Pakhomov, S. and Savova, G. 1999. "Filled Pause Distribution and Modeling in Quasi-Spontaneous Speech". In Proc. Disfluency Workshop. International Congress of Phonetic Sciences.
7. Rosenfeld, R. 1994. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. Ph.D. thesis, Carnegie Mellon University.
8. Shriberg, E. E. 1996. "Disfluencies in Switchboard", in Proc. ICSLP.
9. Shriberg, E. E. and Bates, R. and Stolcke, A. 1997. "A prosody-only decision tree model for disfluency detection", In Proc. EUROSPEECH.
10. Valtchev, V. Kershaw, D. and Odell, J. 1998. The Truetalk Transcriber book. Entropic Cambridge Research Laboratory, Cambridge, England.

ACKNOWLEDGEMENTS

We would like to thank Dr. Joan Bachenko for her professional guidance, encouragement, understanding and patience.