Optimizing Language Models for Speech Recognition

Suzanne Lynch Hruska* Maria Kiskowski† Jennifer Lefeaux‡ Kevin McCleary§
Dany Ngouyassa¶ Bryan Smith‖

Industry Mentor: Dr. Joan Bachenko **

July 28, 2000

Abstract

Speech recognition relies on language models trained to choose probable sequences of n or less words in speech data. We explore the effect of varying the value of n on the perplexity of the language model for different speakers. We find that for talkers who show high perplexity under low values of n, increasing the value of n will have a negative or null effect on performance (increasing perplexity), but that for talkers with low perplexity, increasing the size of n will further decrease perplexity. We also investigate the effects of the size of the training corpus on the performance of the language model for individual speakers, and find no correlation between the size of the corpus and the accuracy of the language model. Finally, by comparing accuracy and perplexity, we discover that there is no obvious correlation and suggest further study.

1 Introduction

1.1 Motivation: Medical Transcription Outsourcing

Linguistic Technologies Incorporated (LTI) uses speech recognition in medical transcription. A medical doctor speaks in dictation over the phone which is analyzed by a Speech Recognition System. The Speech Recognizer consists of an Acoustic Model that forms a hypothesis about the words spoken based on sounds, and a Language Model that finds probabilities for each of the given hypotheses (choosing one based on the highest probability). The Speech Recognizer outputs a transcription file which is edited by a medical transcriptionist (who has access to the original voice recording) to produce a final transcription. This final transciption is the product which is delivered back to the doctor to be put into patient files, sent to insurance companies, etc.

*Cornell University
†University of Notre Dame
‡North Carolina State University
§Kent State University
¶Indiana University
‖Tufts University
**Lernout and Hauspie - Linguistic Technologies, Inc.

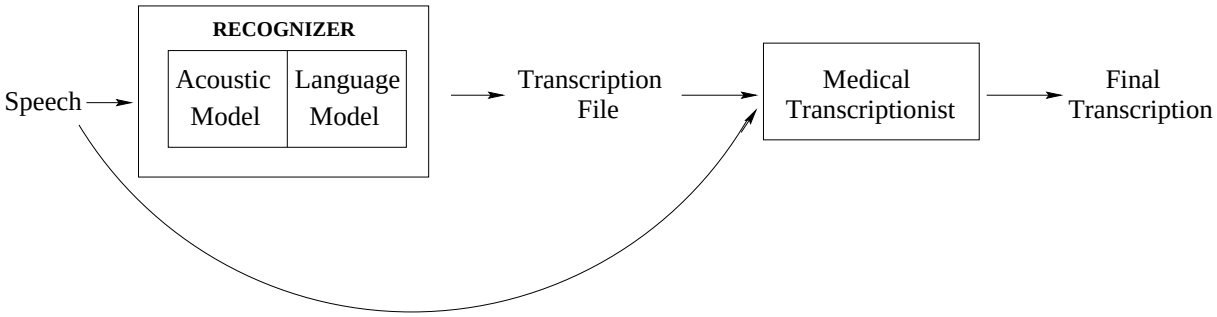

Figure 1: Speech Recognition in Medical Transcription Outsourcing.

The purpose of the Speech Recognition System in this application is to maximize the efficiency of the medical transcriptionist. Hence, LTI cares a great deal about the accuracy of the transcription file produced by the speech recognizer. If the recognizer does a poor job, it may take many times longer to edit this file than for the medical transcriptionist to have done all the work alone.

1.2 Language Models

A *language model* (LM) is a probabilistic model based on the Markov assumption that the probability of a word occurence depends on the words preceding it. An n-tuple of words is called an *n-gram*. When a language model computes the probability of a word occurence using the previous n-1 words it is called an n-gram LM. The Acoustic Model offers choices from which the Language Model chooses the word or word combination that produces an n-gram with the highest probability. In practice, language models are usually bigrams or trigrams. In this study, however, 4-gram, 5-gram and 6-gram models are also of interest.

The probability of an n-gram is computed from its frequency within a training text, or *corpus*. In most cases, corpora must be very large (generally greater than 10 million words) in order to have a sufficiently large sampling of n-grams and their relative frequencies for accurate probability assignments.

1.3 Quantitative Variables: Accuracy and Perplexity

Once a LM is built there are various ways to measure its performance. *Accuracy* and *correctness* measure how well the whole recognizer translated a speech. *Perplexity* measures the difficulty of the word choices made by the language model. *Out-of-vocabulary (OOV rate)* measures the number of words encountered by the LM that were not part of the training data.

A diagram illustrating the testing process is shown in Figure 2.

Accuracy and correctness use a perfect transcription of a speech, made by a person, to judge how well the recognizer interpreted the speech. These are called truth files. The formulas are

$$accuracy = \frac{H - I}{N} \times 100\%$$

and

$$correctness = \frac{H}{N} \times 100\%$$

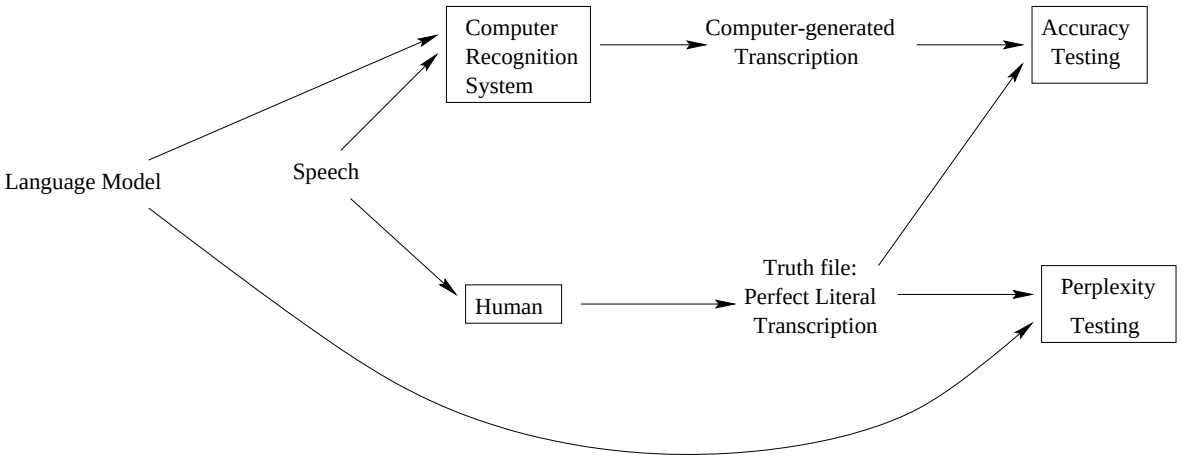

Figure 2: The Language Model Testing Process.

where H is the number of words correctly determined by the recognizer, N is the total number of words spoken, and I, the number of insertions, is the number of words not initially in the speech text but inserted by the computer. An alternative measurement is the Word Error Rate (WER), which is the complement of accuracy. All of these calculations are standard in the speech recognition community.

Perplexity, roughly speaking, is the average difficulty of the choices the language model has to make when interpreting the acoustic model's output of a new speech. Suppose $\{\mathbf{x}^i\}_{i=1}^M$ are all of the n-grams which have been determined, from the training text, to have non-zero probability p_i. Then $-\log p_i$ is a measure of the "surprise factor" associated with $\mathbf{x}^i$: If p_i is low, implying that $\mathbf{x}^i$ occurs infrequently, then $-\log p_i$ is large, indicating that its appearance would be surprising. Conversely, if $\mathbf{x}^i$ is a common n-gram, with probability close to 1, then $-\log p_i$ is close to 0, indicating that we are not surprised when it appears.

Entropy is the "average surprise rate": for a test corpus X of size N words in which the n-gram $\mathbf{x}^i$ appears N_i times, the entropy is defined as [6]

$$H(X) = -\sum_{i=1}^{M} \frac{N_i}{N} \log_2 p_i.$$

If N goes to infinity then $H(X) = -\sum_{i=1}^{M} \tilde{p}_i \log_2 p_i$, where $\tilde{p}_i$ is the "true" probability of the occurence of $\mathbf{x}^i$ (as opposed to the calculated approximate probability p_i). But for working purposes, we act as if the two are the same, so that

$$H(X) = -\sum_{i=1}^{M} p_i \log_2 p_i.$$

Perplexity is defined as [3]

$$perplexity = 2^{H(X)} = \prod_{i=1}^{M} \left(\frac{1}{p_i}\right)^{p_i}.$$

Thus perplexity is the geometric mean of the rates of occurence of each of the n-grams. It measures the average difficulty of the choices made by the language model: since perplexity

is maximized when entropy is maximized, the model is "most perplexed" when n-grams have similar rates of occurence — i.e., when choices are difficult to make. Similarly, the model is less perplexed when some n-grams are more likely to occur than others.

"Good" perplexity — usually in the 20 - 200 range — is a necessary condition for obtaining good accuracy in speech recognition. However, it should be noted that there are many other conditions that can cause poor accuracy even in the presence of good perplexity. In particular, disfluencies such as high filled pause rate or high false start rate can make it difficult for the recognizer to translate correctly extemporaneous speech. For this reason, we tested the perplexity of our LMs on the truth files of different talkers.

2 Problem Description

Two experiments were run to explore ways in which language model performance could be improved.

1. The possibility of a connection between speaker fluency and perplexity of various n-gram models was investigated. We tested the perplexity of five different language models on truth utterances from nine different speakers. The models varied in the value of n, with n ranging from 2 to 6.

2. The possibility of a connection between accuracy results and size of the training corpus was examined.

3 Methodology of Solution

The CMU Toolkit from Carnegie Mellon University has a program which creates a language model from a training text. We used this program to produce five language models: a bigram, trigram, 4-, 5-, and a 6-gram model.

The five language models were trained from the same corpus of 93 million words. Ninety-five percent of this corpus is composed of tokenized family medicine texts and 5% of tokenized truth data. Tokenization is a feature required by the LTI recognizer where special word sets such as hyphenations and contractions are replaced by formulas. Truth is the hand-transcribed literal texts, in our case taken from family medicine, and includes utterances such as "um" and "no, I mean..." that are spoken but not included in the final transcription. The copies of the truth text were attached end to end and then attached to the end of the larger collection of medical transcription text. The training data was collected between 1995 and 1999.

A disjoint collection of utterances and corresponding transcriptions were collected from nine family medicine doctors for our test data. The doctors were chosen based on a preliminary accuracy test provided by an LTI program to achieve the greatest range of fluency. The data for each talker was a collate of tokenized transcriptions collected by LTI from 1995-1997 of five to thirteen thousand words. There were 100-300 words per transcription.

We used the utterances for each talker to determine fluency of talkers both subjectively and by using an automated fluency meter. For our subjective test, three students listened to 2-3 utterances for each speaker (each of approximately 1 minute duration). Each utterance was assigned a score between 1 and 10, where 10 was reserved for talkers easiest to understand. The automated fluency meter used was developed by LTI and integrated a number of objective

quantities including accuracy based on a trigram language model, speaker pace and proportion of silence. We also found that the subjective student ratings had no correlation with the accuracy of the computer recognizer. Apparently, humans and computers evaluate speech quite differently!

The CMU Toolkit was used to measure the perplexity of the utterances for each of nine talkers (approx. 30 minutes of speech each) with respect to each of the five language models (2-,3-,4-,5-,6-gram). These twenty scores were then analyzed for interrelated trends in perplexity between fluency of talker and the type of language model. Analysis was performed to ensure that the test data size for each talker (which differed by more than 100% between speakers) did not confound our perplexity results.

4 Results

See Figures 3 through 7.

Table of Perplexity for 2-, 3-, 4-, 5- and 6-gram LMs for 9 Speakers

Talker	2-gram	3-gram	4-gram	5-gram	6-gram
	Perplexity	Perplexity	Perplexity	Perplexity	Perplexity
S1001	142.12	74.26	65.90	64.52	64.87
S1002	112.16	55.06	43.48	41.59	41.82
S100A	145.79	97.33	99.00	104.35	108.20
S5001	206.27	133.73	138.24	144.83	150.01
S5002	121.59	63.93	57.49	58.49	60.96
S7002	147.03	93.46	91.31	94.20	97.12
S7007	152.93	102.97	100.69	104.07	107.02
S7008	109.61	55.52	47.52	45.84	46.61
S9001	160.59	105.05	106.65	112.58	115.64

Figure 3: This data was used to create Figures 4-7.

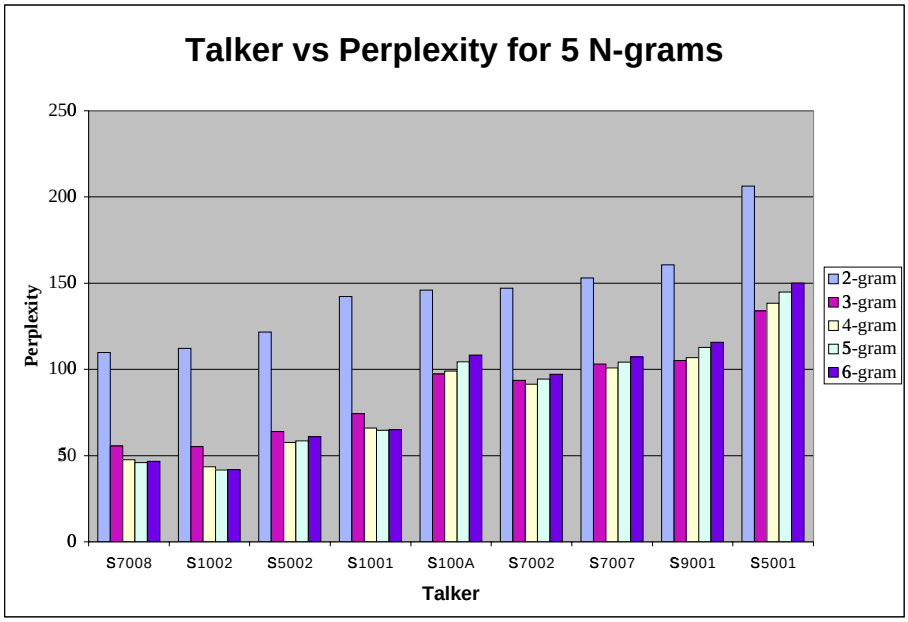

Figure 4: The baseline bigram model fares significantly poorer over the perplexity range than higher n-grams. Low bigram perplexities tend to suggest moving to the 5-, or 6-gram, while higher bigram perplexities tend to suggest using trigrams and 4-grams.

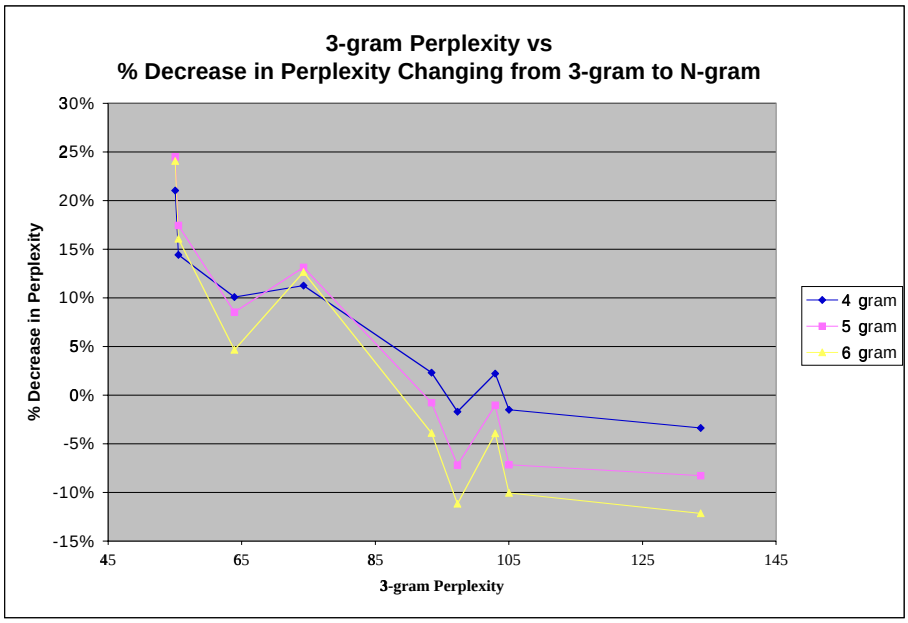

Figure 5: Compares a baseline trigram perplexity with decrease in perplexity associated with higher n-grams. The 4-,5-,6-grams optimize perplexity compared to the trigram model when trigram perplexity is around 50. But 4-,5-,6-grams increase the perplexity when trigram perplexity is at 90 or above.

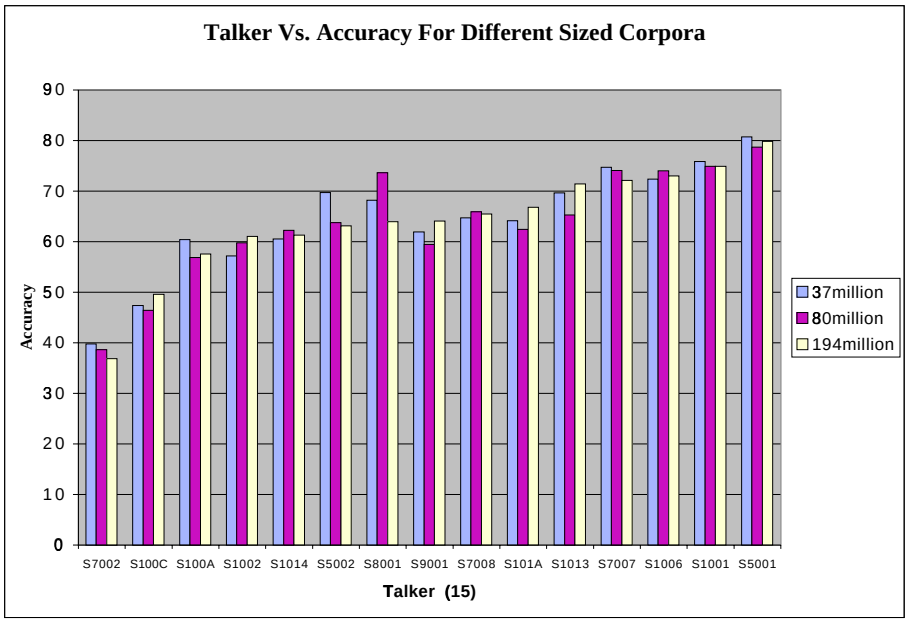

Figure 6: Each talker was tested on three different sized corpora. The results indicate no dependency of accuracy on corpus size in the range measured (37-194 million words).

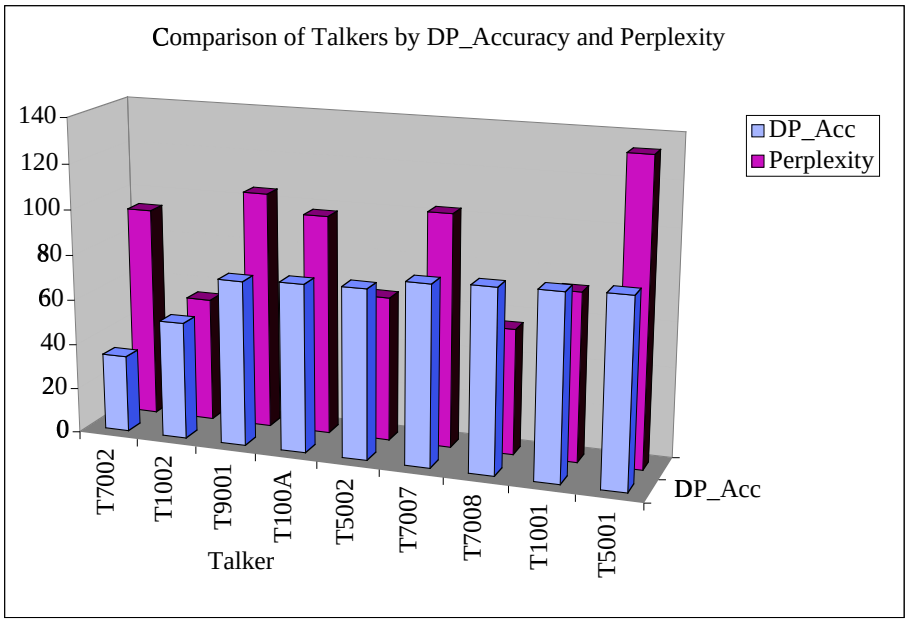

Figure 7: There is no obvious relationship or correlation between DP Accuracy and Perplexity based on the data. However, this graph raises new questions about perplexity and its potential for affecting accuracy results.[5]

5 Conclusions

It is important to note that the results for these experiments are specific to speakers of the English language, which has a relatively strict left-to-right word order with strong local dependence. Other languages may show different behavior. Also, our study was restricted to the sublanguage of Family Medicine.

5.1 Optimization of Perplexity

Perplexity score on a trigram language model may offer us insight on which n-gram will give the optimal perplexity for an individual speaker. If a talker has a high perplexity on the trigram model, then we expect that this person will do no better as the n-gram is increased. On the other hand, talkers who do well on the trigram perplexity test (e.g. below 75) usually benefit from a move towards a higher n-gram. See Figure 4.

5.2 Perplexity Gain vs. Cost

As the n-gram size for a language model is raised a major increase in computational intensity and storage requirements is experienced. As mentioned earlier, a slight decrease in perplexity may not lead to an increase in accuracy for a particular speaker (our ultimate goal).

By looking at Figure 4 it is observed that the substantial decrease in perplexity obtained by moving from bigram to trigram is unique. For any other increase in gram size, such as 4-gram to 5-gram, the decrease in perplexity (if any) is slight and in many cases an increase occurs.

When computational issues enter the picture it seems as though the gains in perplexity scores (i.e. decreased perplexity) may well be outweighed by the extra computational costs involved. Hence, we may not want to look past a 4-gram language model to improve the perplexity for any talker.

5.3 Effects of Corpus Size on Accuracy

We built three trigram models on corpora of varying size, and tested them on 50 speech files from fifteen different talkers. The three training corpora had sizes of 37 million words, 80 million words, and 194 million words. The variation of the corpus size appears to have no effect on accuracy results. Six of the speakers obtained the best accuracy from the model trained on the 37MW corpus, four experienced the best accuracy from the 80MW model, and five had best results from the 194MW corpus.

6 Acknowledgements

We thank the Institute for Mathematics and Its Applications and the University of Minnesota for their generous support. We also thank Dr. Joan Bachenko, especially for her encouragement and enthusiasm.

References

[1] Allen, James. 1995. Natural Language Understanding. Redwood City, CA: Benjamin/Cummings Publishing. Chapter 7 and Appendix C.

[2] The Carnegie-Mellon Statistical Language Modeling Toolkit, available on the Internet at: Speech at Carnegie-Mellon University, *http://www.speech.cs.cmu.edu/index.html*

[3] Jurafsky, Dan and James H. Martin. 2000. Speech and Language Processing. Englewood Cliffs, NJ: Prentice-Hall. Chapters 5,6,7.

[4] Rosenfeld, Ronald. *Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data.* Eurospeech 95.

[5] Savona, Guergana. 2000. *An Empirical Study of Language Model Adaptation.* LTI Internal Report.

[6] Schneider, Thomas. 2000. *Information Theory Primer. http://www.lecb.ncifcrf.gov/ toms/paper/primer/*