LARGE-VOCABULARY SPEECH RECOGNITION IN SPECIALIZED DOMAINS

Haakon Chevalier, Chuck Ingold, Carol Kunz, Chip Moore, Crispin Roven, Jon Yamron,
Bradley Baker, Paul Bamberg, Sarah Bridle, Tracy Bruce, Amy Weader

Dragon Systems, Inc., 320 Nevada St., Newton, Massachusetts 02160 USA

ABSTRACT

We report on research into the discrete-word speech recognition performance of several specialized language models optimized for four large domains of professional discourse. We describe the construction of these models and report perplexity and recognition results for each of the specialized domains. The data indicate that such specialization may significantly improve performance both before and after adaptation.

1. JUSTIFICATION

New users of Dragon's large-vocabulary discrete-utterance speech recognizer typically achieve an "out-of-the-box" performance of about 85% on their first 1000 words (including out-of-vocabulary words) on a wide variety of topics. With full adaptation the acoustic error rate soon falls typically to around 5%.

In the worst case, however, the perplexity of a dictated text can be so seriously misrepresented by an over-generalized language model as to extend the adaptation time considerably. Errors attributable to mismatches in word frequency are only one problem. Much more costly are the words that never appear in the language model at all. In some texts the new word error rate can exceed the acoustical error rate.

Of course, the vocabulary can be increased in size, but this only increases the chance of acoustical and perplexity related errors. How much improvement can be made instead by building a language model using a small amount of added information about the person dictating; what, more or less, is his or her field?

In this paper we will describe the building and testing of language models targeted towards specific fields. Section 2 describes our technique of combining n-gram statistics extracted from various sources into domain-specific language models. In section 3 we describe our testing methodology and in section 4 we present our results.

2. LANGUAGE MODELS

For the sake of this preliminary study, we used a bigram language model. In this set of experiments, the statistics for some of the less common words were combined to save memory.

The vocabulary size studied was 30,000 words, with a backup dictionary of 120,000 words. The domains used were:

1. General (not targeted).
2. Journalism (target: journalist for newspaper or newsmagazine).
3. General Legal (target: lawyer or paralegal).
4. General Medical (target: physician or transcriptionist).
5. General Business/Financial (target: financial analyst or executive).

The language is American English. For the General vocabulary, we used a word model which had previously been shown to have high performance on a wide variety of topics, and in fact was built using a significant amount of source appropriate for each of the other vocabularies. Note that, although domains 2 - 5 are more restricted than a completely general vocabulary, there is ample room for a large amount of variation.

2.1. Building the Vocabularies

We collected and parsed several gigabytes of text from over 2000 sources. A unigram and bigram frequency list were produced for each source. This source data varied widely in terms of quality and applicability to any particular specialized language model. In order to construct vocabularies from these sometimes less than ideal building blocks, we developed a technique called 'targeting'. Targeting enables a small unigram model to bootstrap a larger one.

Targeting makes use of the EM algorithm [1] to estimate which weighted mixture of the source unigram lists would most likely produce a given "target" list. The quantity minimized is the unigram perplexity of the target list given the mixture vocabulary. Unigram perplexity was used rather than bigram perplexity because it was easier to calculate and because the results were expected to be similar.

2.1.1. Target Model

We first built small domain-specific "target" unigram lists, using text judged by humans to be as close to our targets as possible. We attempted to attain a balance of varied texts, while keeping within the definition of the targets. The balancing was also done mostly by human judgement, in some cases supplemented by rudimentary usage statistics.

0-7803-2431-5/95 $4.00

2.1.2. Clustering

An initial test of the targeting algorithm revealed that the technique returned suspicious results if the source data were divided into too many separate unigram lists. In response, we added together the unigram and bigram frequency lists (extracted from our source texts) into a few dozen topic clusters.

The contents of these clusters were different for each vocabulary. Texts which were off-topic were combined into a smaller number of clusters than those related to the specialized domain. For instance when targeting the legal vocabulary, different kinds of legal documents were put into different clusters, while for other vocabularies, they were all combined into a single cluster.

2.1.3. Mixing the Clusters

Finally, the unigram and bigram clusters were combined with the weights suggested by the EM algorithm. The unigram frequencies were smoothed by adding a number between 0 and 1 such that the total number of words kept would be slightly more than the total number desired. [2]

3. Testing Methodology

3.1. Comparing the Domains

The distance between the unigram language models was measured. Assuming the models used are roughly correct for the domains, this should be an approximation of the distance between the domains themselves. The measure used was the symmetrical Kullback-Leibler distance between models Q and Q':

$$KL=\frac{1}{2}\sum_{w}[(P(Q_w)-P(Q'_w))(\log P(Q_w)-\log P(Q'_w))]$$

The results are summarized in Table I. From this data, the Business and Law unigram models are seen to be quite similar, while the Medical model is the most distinct.

3.2. Language Model against Test Texts.

3.2.1. Unigram tests

Two types of tests were used. First, the unigram language models were tested directly against sample texts. For this test, many test texts were used for each domain. A number of statistics were collected, including the number of words in each test text not contained in the first N words of the language model, as well as several measures of perplexity (entropy) of the resulting unigram language model [3], [4]. The measures of perplexity used differ in the treatment of out-of-vocabulary words:

a) *Known-word perplexity:* Out-of-vocabulary words are ignored in the calculation of perplexity.

b) *New-word perplexity:* Out-of-vocabulary words are given a value of one-half the smallest frequency for any word in the (possibly truncated and/or scaled) language model,

c) *Common perplexity:* Only words which are in all language models being tested contribute to the perplexity computation.

Each of the sample texts was tested against each of the five vocabularies.

Table I: Known-word Kullback-Leibler Distances Between Unigram Models

Jou	.47	0		
Bus	.66	.39	0	
Law	.61	.46	.22	0
Med	.89	.79	.81	.91
	Gen	Jou	Bus	Law

3.2.2. Recognition Tests

In a second series of tests, we compared the recognition performance on recorded speech using different language models. The starting acoustic models and all other parameters remained constant. Here again several types of tests were made:

1) In the first type of test, initial recognition performance on each recording was tested with both the acoustics and the language model completely untrained.

2) Adapted recognition was assessed using a "jackknife" test. In this case, the recognition results were reported after the acoustic and language models had adapted to other subject-related scripts recorded by a particular speaker.

We have made an effort to use predictive measures in stating performance. In particular, out-of-vocabulary words are considered errors, as well as words which have a different case from what was recognized. For instance, "an" (an article), "An" (a proper name) and "AN" (an acronym) are considered different words, even though at the beginning of a sentence the first two will appear identical. To this end, we quote an estimate of an additional statistic, which we term "Estimated Throughput." This represents the percent of utterances (or utterances plus keystrokes) which are devoted to dictation, as opposed to corrections, assuming all corrections are done by voice. Assuming a correction utterance takes about the same amount of time as a text utterance, this would be a measure of how fast the user can dictate.

To make an estimate of this figure, we make the approximation that all errors of the same type take the same amount of utterances or keystrokes to correct. The formula used is

$$ET = \frac{N}{N + (1 \cdot C) + (3 \cdot O) + (5 \cdot B) + (8 \cdot U)}$$

where N is the number of text utterances, C represents the number of errors in which the correct word ends up on a Choice-List, from which it can be selected by a single utterance or keystroke; O is the number of errors in which the correct word is considered but does not make the Choice-List; B represents the number of times the correct word is not considered but is available in the Backup Dictionary; U is the number of tokens which are completely missing, and typically have to be typed or spoken letter-by-letter. The weights on these factors represent the average number of keystrokes required to correct each type of error.

Whenever possible, test texts were drawn from different corpora from those used for targeting or source, and of course, the same text was never used for any two of the targeting, source, or test sets. Note that this differs from the traditional methodology in which the source and test texts are drawn from the same corpus. The reason is that we wanted to guard against making the specialized vocabularies too narrow, thereby artificially inflating performance results.

3.2.3. Correlation of Tests

Since the unigram testing is so much cheaper, it is of interest to try to determine what relevance it has to results of recognition tests. In preliminary tests, new word errors were found to be a good predictor of recognition performance, and change in perplexity between language models was found to be a good predictor of improvement in recognition performance.

4. RESULTS

A set of vocabularies was built, and the two types of tests mentioned above were run on some data. Both tests were run on data thought to be fairly close to the intended target. These include:

- General: previously tested against a large variety of articles on different subjects.
- Journalism: articles taken from different sections of 11 different newspapers, articles from news-wire services.
- Legal: circuit court opinions, software licenses, briefs
- Medical: radiologists' patient records, published articles, patient notes
- Business: agreements, financial reports, letters

For initial testing of the language models, a large number of test texts were used. However, for this article, results for both language-model and recognition tests are reported for only those texts for which scripts have been recorded. These consist of from 3 to 8 diverse texts in each domain, each containing approximately 500 to 1200 words.

We report in Table II the results of the common perplexity test on 30,000 word vocabularies. This test ignores the effect of words unknown in at least one language model. In Table III we report the new word error rate. In all cases, the lowest perplexity and new word error rate on a given test domain was achieved by the corresponding specialized language model. Similar results were obtained when the vocabularies were expanded to 120,000 words. On the domain-specific texts, the General language model has a mean new-word error rate of 2.29% at 120,000 words, while their corresponding Specialized language models achieve a mean rate of 1.04%.

Results for "out-of-the-box" recognition tests are shown in Table IV. Results of testing with partially adapted acoustics and word models are shown in Table V. Table V also shows the mean number of words on which adaptation took place before testing. In both tables, the results for the specialized text shown in the left column are described under the heading "Special." In general, the improvements resulting from the use of specialized vocabularies seem to continue, at least after moderate adaptation. In fact, for the Journalism and Legal test sets, the observed difference between the Specialized and General models is actually larger after adaptation. This may be due to acoustic effects on the unadapted models, rather than language-model effects.

Table II: Mean Common Perplexity for Unigram Models at 30,000 Words

		Language Models				
		Gen	Bus	Jou	Law	Med
Test Texts	Gen	1226	1836	1346	1698	1739
	Bus	1697	1315	1715	1635	2411
	Jou	1263	1684	1094	1565	2062
	Law	1558	1241	1534	1104	1998
	Med	1186	1589	1325	1498	876

Table III: Mean Percent Unknown Words for Unigram Models at 30,000 Words

		Language Models				
		Gen	Bus	Jou	Law	Med
Test Texts	Gen	2.24%	3.41%	2.56%	3.21%	3.77%
	Bus	3.25%	1.51%	2.62%	1.82%	3.83%
	Jou	4.20%	5.01%	3.44%	4.25%	7.02%
	Law	1.93%	1.53%	2.35%	1.17%	3.44%
	Med	6.67%	10.40%	10.25%	10.29%	2.30%

Table IV: Mean Word Percent Correct and Estimated Throughput for Unadapted Models

	Percent Correct		Est. Throughput	
	Special	General	Special	General
Bus	84.51%	83.61%	78.49%	72.83%
Jou	83.49%	83.44%	73.40%	71.55%
Law	88.37%	87.39%	82.41%	78.25%
Med	88.17%	82.91%	78.35%	65.63%

Table V: Mean Word Percent Correct for Partially Adapted Recognition

	Special	General	# of Words Adapted on
Bus	88.36%	87.93%	2020
Jou	87.31%	86.51%	2128
Law	92.37%	91.27%	2030
Med	92.64%	90.00%	1391

5. Conclusions

We have shown that the performance of large-vocabulary speech recognition systems can be markedly improved by the use of domain-specific language models in unadapted and partially adapted systems. It is our expectation that this improvement will also be maintained in systems which have been fully adapted acoustically (results for these further tests are not yet available). This means that users in these specific domains can immediately obtain better "out-of-the-box" performance and can reduce the time required to customize their vocabularies. The medical vocabulary, which is the most distinct, benefits the most from a language model which is very different from that of the other domains. The figures shown are based on a relatively small amount of data, but are very conservative. For the typical user, whose dictated texts are likely to be more uniform than our test texts, the fully adapted performance is likely to be considerably higher than quoted here. We also suggest the use of a new measure, Estimated Throughput, as a measure of recognition performance.

6. References

[1] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). "Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm", *J. Roy. Statist. Soc. B* **39** 1-38.

[2] Gale, W.A., and Church, K.W.,(1989) "What's Wrong with Adding One?", in *Corpus-Based Research Into Language*, in honor of Jan Aarts, Rodop, Amsterdam.

[3] Shannon, C.E. (1948) "Prediction and Entropy of Printed English", *The Bell System Technical Journal*, v.27.

[4] Cover, T.M. and King, R.C. (1978) "A Convergent Gambling Estimate of the Entropy of English", *IEEE Transactions on Information Theory*, Vol. IT-24, No.4.

[5] Dagan, I., Pereira, F., and Lee, L. (1994) "Similarity-Based Estimation of Word Cooccurrence Probabilities", *Proceedings of the 32nd Annual Meeting of the ACL.*