EXPERIMENTAL EVALUATION OF ITALIAN LANGUAGE MODELS FOR LARGE-DICTIONARY SPEECH RECOGNITION

M. Codognoto, L. Fissore, A. Martelli, G. Pirani, G. Volpi

ABSTRACT

This paper reports on experiments performed on the Italian language in order to assess the efficiency of probabilistic language models with reference to a task of large-dictionary speech recognition. Two different types of models, an M-gram and an Mg-gram one, have been investigated for comparison purposes.

The quality of the models trained on a corpus of 3.5 million words was measured in terms of perplexity and of the improvement achieved by integrating the language model in real speech recognition systems. Judging from this empirical measurement, the two language models exhibit equivalent performance for Italian, although perplexity measurements would suggest otherwise.

INTRODUCTION

Large-dictionary speech-recognition systems need some tool to integrate a-priori linguistic knowledge with acoustical evidence; in a probabilistic framework, this takes the form of a Language Model to estimate probabilities of words given the preceding part of the sentence ("predict the next word").

The experiments on the Italian language were performed with two classes of models (M-gram and Mg-gram models), which were described in (ref 1): an M-gram model predicts the next word on the basis of the last M-1 words in the sentence, while an Mg-gram model considers only the morphological-grammatical classes of the last M-1 words in the sentence.

MODEL AND EXPERIMENT DESCRIPTION

The language models that will be dealt with in the following take into account only the last two words to predict the current one, thereby being 3-gram and 3g-gram models (ref 1). The models were trained on a corpus of 3.5 million words, taken from the “Il Mondo” weekly magazine (economy, finance and politics). Training was performed on the basis of dictionaries of different sizes (about 1K, 3K, 8K, and 15K words) consisting of the most frequent words (word forms, not lemmata) from this same corpus. The training procedure is different depending on whether words or classes are to be considered as “atoms” of the model.

2.1 - 3-gram model: The objective of the training is the estimation of the probability $P(W_n | W_{n-2}, W_{n-1})$ by means of the approximate equation

$$
\frac{N(W_{n-2}, W_{n-1}, W_n)}{\sum_w N(W_{n-2}, W_{n-1}, w)}
$$

where $N(E)$ is the number of occurrences of trigram $E$ in the training corpus.

2.2 - 3g-gram model: In this case 205 classes have been defined in order to catch grammatical, morphological, and some surface semantical features of the words (e.g., verb in the first person singular, geographical proper noun, etc.). As each word can belong to more than one class, it is necessary to resort to a Hidden Markov Model for the language (ref 3). The goal of training is thus to estimate both the state transition probabilities $P(G_n | G_{n-2}, G_{n-1})$ and the emission probabilities $P(W_n | G_n)$. This is achieved through the forward-backward algorithm (ref 4), with a completely automatic procedure that does not require any manual labeling of the training data.
2.3 - Smoothing: In both cases, the impairment arising from the finite size of the training corpus and the consequent statistical insufficiency has been overcome through a non-linear backing-off approach using the First-time heuristic, allowing estimation of the probability of unseen events (ref 5). The identical approach seems to give satisfactory results for both 3-gram and 3g-gram models, despite the widely different degree of "sparseness" in the respective matrices.

PERPLEXITY MEASURES

In order to evaluate the quality of the different language models, an assessment of their predictive power is given by the perplexity, which is defined generally as (ref 2):

$$e^{-\sum_w P_t(w) \log(P_{model}(w))}$$

(2)

where the sum is over distinct events-in-context $w$.

However, the dictionaries for which the models are to be trained have different sizes, and in any case consist of many fewer distinct words than the corpus itself. As a consequence, many words both in the training and in the test data will be classified as "unknown", that is, not comprised in the dictionary; it would be a fictitious solution to restrict training and test data to a subset fully covered by the dictionary: indeed, it would just hide any problems with the modeling of "unknown" word behavior, and make it impossible to compare performance of models with different dictionaries on the same training and test data.

Moreover, models introduce a fictitious word called endsentence at the end of a sentence, to avoid doing predictions across sentence boundaries. Endsentence is usually very easy to predict, and we do not want to bias the perplexity measure in favor of tests with many short sentences; on the other hand, if a model should be weak in predicting endsentence, this would be a serious problem, and we would not want to mask it.

To take into account the problem of "unknown words" and "endsentence", while allowing consistent comparisons of the quality of different models, two practical perplexity measures have been introduced: the standard perplexity

$$-\sum_{w \neq \text{unknown}} \log(P_{model}(w | w \neq \text{unknown}))$$

(3)

c number of words $\neq$ unknown, endsentence

and the benchmark perplexity

$$-\sum_w \log(P_{model}(w))$$

(3')

c number of words $\neq$ unknown, endsentence

Note that "standard" perplexity tries to be "transparent" to the problem of unknown words, while the "benchmark" perplexity tries to penalize as much as possible any model which is weak in predicting the occurrence of unknown words or has too many unknown words - these could be serious drawbacks of a model in practical application. Standard perplexity seems to be closer to what has been actually reported in previous studies (ref 2); benchmark perplexity seems to be a more significant measure of model usefulness.

Standard and benchmark perplexities have been computed for both 3-gram and 3g-gram models, with reference to several tasks; the test data were taken from the same magazine as that used for training (chosen, of course, so as to be non overlapping with the training data; task MONDO), from a different source (agency news flashes; task ANSA) on the same subjects, and from texts on different subjects (scientific brochures; task BROCHURE).

The results are reported in Table I:
RECOGNITION EXPERIMENTS

In order to test the effectiveness of the language models in a more practical testbed, we have evaluated the improvement that such models introduce on the performance of two different speech recognition systems. The objective is to evaluate the language models, and not to compare the acoustical subsystems, that have been developed in completely different environments. One of them (hereafter denoted as SYS1) integrates the acoustic and linguistic knowledge during the recognition process, using the stack decoding algorithm (ref 6). The other (hereafter denoted as SYS2) uses the language model to perform a post-processing of the scored word hypotheses provided by the acoustic level (ref 7). The integrated approach allows faster (real-time) performance, while the post-processing approach is more flexible in allowing easy experimental comparisons of different language modeling approaches.

Both SYS1 and SYS2 have been tested on a set of 223 new sentences, specially written for the purpose, dealing with the same subjects as that of the training corpus, and made up of the most frequent 1000 words in the dictionary (task EXPER). The sentences were written to use as many different acoustic contexts as possible, and, as shown below, don't exhibit low perplexity. The sentences, amounting to a total of 2711 words, were uttered by one speaker inserting short silences between each pair of contiguous words. Perplexities and performance of SYS1 and SYS2 (in terms of percentage of correctly recognized words, both with and without the help of the language model) are reported in Table 2:

<table>
<thead>
<tr>
<th>Language model</th>
<th>Standard perplexity</th>
<th>Benchmark perplexity</th>
<th>Recognition with SYS1, speaker GV</th>
<th>Recognition with SYS2, speaker GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>n.a.</td>
<td>n.a.</td>
<td>88.0%</td>
<td>87.4%</td>
</tr>
<tr>
<td>3-gram</td>
<td>101</td>
<td>108</td>
<td>95.7%</td>
<td>94.6%</td>
</tr>
<tr>
<td>3g-gram</td>
<td>115</td>
<td>151</td>
<td>n.a.</td>
<td>95.4%</td>
</tr>
</tbody>
</table>

CONCLUSIONS AND REMARKS

The performance of the 3g-gram model in SYS2 can be surprising at first glance, particularly with reference to the perplexity measurements. This can be explained by the structure of the Italian language, which has many cases of strong grammatical differences between acoustically similar words. This statement can be supported by a modified perplexity measure, which accounts for acoustic similarity between words:
\[
\begin{align*}
- \sum_w P(w) \log(P_{\text{model}}(w|\text{acoustic cohort}))
\end{align*}
\] (4)

where the model is only applied to the cohort of words accepted by the acoustic subsystem SYS2 at each point. In this case, we have:

| Table 3 |
|-----------------|-----------------|
| **acoustically modified perplexity, 1K, EXPER** | **3-gram** | **3g-gram** |
| 2.2 | 1.9 |

which shows that the 3g-gram model cooperates better with the acoustics of Italian, and that its perplexity, higher than the 3-gram's when considering all of the dictionary words, goes actually somewhat lower when only applied to words acoustically similar to the correct one.

In conclusion, for the Italian language, the 3g-gram model seems to be very efficient, as it shows high performance even using a limited-size training corpus and is rather robust to the variability of the domain. It will of course continue to show the advantages outlined in (ref 1), but it now seems that these advantages shall not be payed for by a lower recognition rate.

Acknowledgment

The authors are grateful to R. Campo and L. Vacchetta, for their help in the development of the language model based on morpho-grammatical classes, and to M. Brandetti, M. Ferretti and S. Scarci for their support in recognition experiments.

REFERENCES


