Broadcast News LM Adaptation using Contemporary Texts

Marcello Federico & Nicola Bertoldi

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy
federico,bertoldi@itc.it

Abstract

This paper investigates the problem of dynamically updating the language model (LM) of a broadcast news speech recognition system, in order to cope with language and topic changes, typical of the news domain. Statistical adaptation methods are proposed that exploit written news sources which are daily available on the Internet, i.e. newswires and newspapers. Specifically, LM adaptation is performed by extending the basic lexicon, in order to minimize the out-of-vocabulary (OOV) rate, and by adapting the word probability distribution on the contemporary data. Experiments performed on 19 newscasts showed relative reductions of 58% on the OOV rate, 16% on the perplexity, and 4% on the word error rate.

1. Introduction

This paper investigates the issue of adapting a broadcast news (BN) language model (LM) by taking advantage of contemporary news that are daily available through the Internet, such as newswires and news agencies. Adaptation is performed on a day basis both at the level of lexicon and of n-gram probabilities. Lexicon adaptation is performed by introducing new words, partly observed in contemporary data, according to a strategy that tries to reduce the out-of-vocabulary rate. Language model adaptation is performed by selecting recent texts that go back in time up to one week. Such texts are then used to adapt a background LM through a mixture model, that naturally applies to interpolated n-gram LMs. More specifically, the proposed mixture model maintains the interpolation structure, which permits an efficient LM representation for speech decoding [1]. A similar idea was also proposed by [6] for back-off LMs.

LM adaptation was evaluated on 19 Italian television news shows spanning a period of one month. As a baseline, a 10xRT version of the ITC-irst BN transcription system was used [2].

2. Interpolated LMs

For the sake of clarity, let one first consider the case of a closed vocabulary LM. The probability of an n-gram $h_w$, where $h$ represents the history of word $w$, is computed by:

$$ Pr(w \mid h) = f^\star(w \mid h) + \lambda(h) Pr(w \mid \overline{h}). \quad (1) $$

where $f^\star$ is a discounted relative frequency that is smoothed with the zero frequency estimate

$$ \lambda(h) = 1.0 - \sum_{w \in V} f^\star(w \mid h) \quad (2) $$

weighted with the distribution of the lower order $n-1$-gram $\overline{h}, w$. Discounting is computed with a non linear smoothing method that is coupled with a frequency based pruning technique:

$$ f^\star(w \mid h) = \begin{cases} \frac{c(hw) - \beta}{c(h)} & \text{if } c(h) > \alpha \text{ and } c(hw) > \beta \\ 0 & \text{otherwise} \end{cases} \quad (3) $$

and, when $h = \emptyset$:

$$ f^\star(w) = \frac{c(w)}{N} \quad (4) $$

where $\beta$, $\alpha$, $c(\cdot)$, and $N$ are the discounting parameter, the pruning threshold, the training sample counting statistics, and the sample size, respectively. Different parameter settings are adopted according to the LM complexity.
For small to medium size LMs, no pruning threshold is used, i.e. $\alpha = 0$ for all levels, and the discounting parameter is estimated according to [4]. For large LMs, the pruning threshold $\alpha = 50$ is used and $\beta$ is set to 1. Experimentally, the estimate of $\beta$ by [4] results more accurate, while the value $\beta = 1$, coupled with trigram pruning, provided the best trade-off between performance and memory requirements.

### 2.1. Open Vocabulary

The extension to the open-vocabulary LM case is obtained by assuming a special out-of-vocabulary word class $\text{oov}$ and a “universal” vocabulary $U$, i.e. the expected vocabulary as the training corpus size grows to infinity. The $\text{oov}$ word class corresponds to all words not included into the observed vocabulary $V$. Typically, $n$-gram statistics including $\text{oov}$ are available when $V$ is a sub-dictionary of the words in the training corpus. Otherwise, any zero-frequency estimation method [5] can be applied to get a probability estimate of $\text{oov}$. The size of universal vocabulary $U$ can be empirically estimated [5] or set to a conventional (large) value. In the following experiments, the value $10^6$ was used. The aim of the vocabulary $U$ is to permit combining LMs with different vocabularies and to evaluate LMs with different out-of-vocabulary rates. A variant of the discounting method (3) is derived that exploits available statistics about the $\text{oov}$ class and uses the following constraints:

(i) the occurrence of a novel word resets the history

$$\Pr(w \mid h_1 x h_2) = \Pr(w \mid h_2) \quad \text{if } x \in U - V$$

(ii) the probability of a novel word does not depend on $h$

$$\Pr(w \mid h) = \Pr(w) \quad \text{if } w \in U - V$$

As a consequence of the two constraints, the presence of $\text{oov}$ forces backing-off to unigrams, and the zero-frequency probability $\lambda(h)$ must incorporate the discounted relative frequency (3) of the $\text{oov}$ class. Finally, the unigram distribution (4) is replaced with:

$$f^*(w) = \begin{cases} \frac{c(w)}{N} & \text{if } w \in V \\ \frac{c(\text{oov})}{N} \frac{1}{|U - V|} & \text{if } w \in U - V \end{cases}$$

(5)

Hence, the probability of a novel word is obtained by smearing the $\text{oov}$ probability over the estimated size of the $\text{oov}$ class.

### 2.2. Mixture of Language Models

Given $k$ interpolated LMs, we define the following mixture discounted relative frequency:

$$f^*_\text{mix}(w \mid h) = \sum_{i=1}^{k} \mu_i f_i^*(w \mid h)$$

(6)

where $\mu_i$ are weights of a convex combination. From the definition of the zero frequency probability, it follows that:

$$\lambda_{\text{mix}}(h) = \sum_{i=1}^{k} \mu_i \lambda_i(h)$$

(7)

which leads to the interpolation mixture model:

$$\Pr_{\text{mix}}(w \mid h) = f^*_\text{mix}(w \mid h) + \lambda_{\text{mix}}(h) \Pr_{\text{mix}}(w \mid \tilde{h})$$

(8)

An advantage of the proposed mixture model is that it preserves the basic interpolation scheme (1) and hence allows the efficient LM representation described in [1]. Moreover, the mixture weights $\mu_i$ can be estimated by applying the EM algorithm on the following log-likelihood:

$$\sum_{t=1}^{N} \log \sum_{i=1}^{k} \mu_i (f_i^*(w_t \mid h_t) + \lambda_i(h_t) \Pr_{\text{mix}}(w_t \mid \tilde{h}_t))$$

(9)

where the index $t$ scans all $n$-grams of a training sample. Improvements on the mixture model can be achieved by letting the interpolation weights $\mu_i$ depend on the most recent word of the history $h$. The following parameter tying was applied to most-recent context words:

- all words with frequency less or equal than 3 are tied;
- words with equal frequency between 3 and 10 are tied.

The mixture model can be used to combine one or more general (background) LMs with a (foreground) LM representing new features of the language we want to include. In this case, the mixture weights can be estimated on the training data of the foreground LM by applying a cross-validation scheme that simulates the occurrence of new $n$-grams.

### 3. Broadcast News Adaptation Experiments

For the experiments, the ITC-irst Italian broadcast news transcription system was used [2]. Acoustic models were trained over 47 hours of transcribed speech, taken from an extended version of the Italian Broadcast News Corpus (IBNC) [3]. A single step decoding was used to keep recognition response time below 10xRT on a 600MHz Pentium PC with 1Gb of RAM.

LM adaptation was experimented over a test set of 19 television news shows recorded in August-September 2000, for a total length of 6 hours. With respect to the acoustic conditions, the test set only includes wide-band speech.

A baseline 62K-word trigram Shift-1 LM, with pruning threshold set to 50, was estimated on a 215M-word corpus of Italian newspapers issued between January 1992
and July 2000. The newspaper LM, referred to by NP, provides, on the 19 test files, an average perplexity (PP) of 313, an out-of-vocabulary (OOV) rate of 1.58% and a word error rate (WER) of 25.31%.

To adapt the LM, different text sources were taken into account:

- manual transcripts of BNs between 1992 and 1999;
- newspaper articles of the test set period;
- news agency reports of the test set period;
- In particular, the latter two sources jointly produce about 60K words each day. Hence, the investigated idea is to update the LM on a day basis by adapting both the lexicon and the word probability distribution to contemporary news topics.

3.1. Domain Adaptation

As a first adaptation, the background LM NP was combined with a Shift-β LM estimated on the small corpus of BN transcripts, i.e. about 500K words. The resulting LM, named +BN, gave an OOV rate relative improvement of 0.2%, a PP relative reduction of 4.2%, and a WER relative reduction of 1.1%. The +BN LM will be taken as a reference from now.

3.2. Lexicon Adaptation

Different methods were investigated to extend the basic 62K-word lexicon in order to minimize the OOV rate. In particular, we associate the oov class with a unigram LM estimated over a list of new words, which is adapted day by day. Criteria used to evaluate lexicon adaptation are the achieved OOV rate, PP and WER. In fact, the last method implicitly takes into account the quality of the phonetic transcription used for the new words and the increase of word confusability due to the augmented lexicon. In our experiments, phonetic transcriptions of new words were always generated automatically. Two different word selection policies were considered:

(A) taking the most recently used new words;
(B) taking the most frequently used new words.

Strategy (A) picks words from the contemporary data, while strategy (B) from all the available texts. Concerning the notion of “contemporary” data, we also consider available texts of the same day of the newscast. In fact, this method exploits news sources which are indeed used by news-writers, too. Best OOV rates were obtained by combining both strategies, as is shown in Table 1.

The combined strategy (A)+(B) incrementally selects new words as follows: first, words are introduced according to time, i.e. from the broadcast day backward to at most three days, then, according to the training corpus frequency. In order to avoid mistyped words, contemporary words are selected that occur at least twice.

A new unigram frequency (5) is computed which smooths the oov class frequency over a piece-wise distribution defined on four disjoint sets of words: (i) new words from news of the same day, (ii) new words occurring in the preceding three days, (iii) new words taken from the training data, according to their frequency, (iv) and the oov class of the extended lexicon. Distributions of sets (i-iii) are estimated from different samples and weighted according to the estimated prior probability of each set.

Experiments were performed by extending the lexicon by 60K words. The resulting 122K-word LM was evaluated on the 19 test files and is referred to as LM +A.LX. Results are summarized in Table 2 and detailed in Figures 1-3, respectively. Observed OOV rate reductions range between 30% and 80%, with an average around 58%, while the mean absolute PP and WER reductions are 10 and 0.54%, respectively. As a comparison, an absolute WER reduction of 1.05% is observed if V is extended by exactly the OOV words in the test (≈ 800). Remarkably, the ratio between WER reduction and OOV reduction in both cases results very close, i.e 1.05/1.58 = 0.66 vs. 0.54/0.91 = 0.59. Hence, the potential confusability introduced by the vocabulary extension seems almost marginal.

3.3. N-gram Model Adaptation

Language model mixtures were estimated which combine trigrams from NP texts, BN transcripts, and contemporary texts from the last 7 days. The +BN lexicon of 62K

![Figure 1: OOV rates of the evaluated LMs.](image-url)
words was expanded up to 64K words, by including at most 2K words from the sets (i)-(ii), i.e. those occurring at least twice in the texts of the considered four days. The remaining, at most 60K new words, were associated to the OOV class as explained before. Mixture weights were estimated by cross-validation on the contemporary texts. Perplexity and WER results for all the single test files are reported in Figures 2-3, under the name +A.NG. From the perplexity point of view, adaptation outperforms the reference +BN, with an average reduction of 12%. With respect to only lexicon adaptation, i.e. +A.LX, the relative improvement is 9%.

By looking at the WER figures, smaller relative improvements are observed. Unfortunately, in 2 of 19 files, performance of the adapted LM is worse than of the +BN LM. Things are even worse if one considers the A.LX LM: WERs increase in 5 of 19 files. This contrasts with the homogeneous behavior showed by the A.LX LM, which improves the WER over all the test files. However, on the average, the adapted LM achieves WER relative reductions of 3.4% and 1.3% versus the +BN and A.LX LMs, respectively.

4. Conclusions

Two methods have been presented to adapt a BN LM to contemporary texts. The first method extends the lexicon with a strategy aimed at reducing the OOV rate, the second combines n-gram statistics of the current LM with those found in more recent texts. For n-gram adaptation, a mixture model was discussed that naturally applies to the class of interpolated LMs, allows to apply EM estimation, and allows an efficient search space representation for speech decoding. Consistent improvements have been achieved, on the average, by both methods in terms of OOV rate, perplexity, and WER. No significant computational overhead on speech decoding was observed. Recognition accuracy of the mixture model showed to be less stable as would be expected from the PP behavior. An explanation could be that PP is reliable only for sufficiently long text samples. Current work is devoted to find robust n-gram adaptation techniques that make use of automatically generated transcripts.

5. References


