Selecting relevant text subsets from web-data for building topic specific language models

Abhinav Sethy, Panayiotis G. Georgiou, Shrikanth Narayanan
Speech Analysis and Interpretation Lab
Integrated Media Systems Center
Viterbi School of Engineering
Department of Electrical Engineering-Systems
University of Southern California

Abstract

In this paper we present a scheme to select *relevant subsets* of sentences from a large generic corpus such as text acquired from the web. A relative entropy (R.E) based criterion is used to incrementally select sentences whose distribution matches the domain of interest. Experimental results show that by using the proposed subset selection scheme we can get significant performance improvement in both Word Error Rate (WER) and Perplexity (PPL) over the models built from the entire web-corpus by using just 10% of the data. In addition incremental data selection enables us to achieve significant reduction in the vocabulary size as well as number of n-grams in the adapted language model. To demonstrate the gains from our method we provide a comparative analysis with a number of methods proposed in recent language modeling literature for cleaning up text.

1 Introduction

One of the main challenges in the rapid deployment of NLP applications is the lack of in-domain data required for training statistical models. Language models, especially n-gram based, are key components of most NLP applications, such as speech recognition and machine translation, where they serve as priors in the decoding process. To estimate a n-gram language model we require examples of in-domain transcribed utterances, which in absence of readily available relevant corpora have to be collected manually. This poses severe constraints in terms of both system turnaround time and cost.

This led to a growing interest in using the World Wide Web (WWW) as a corpus for NLP (Lapata, 2005; Resnik and Smith, 2003). The web can serve as a good resource for automatically gathering data for building task-specific language models. Webpages of interest can be identified by generating query terms either manually or automatically from an initial set of in-domain sentences by measures such as TFIDF or Relative Entropy (R.E). These webpages can then be converted to a text corpus (which we will refer to as *web-data*) by appropriate preprocessing. However text gathered from the web will rarely fit the demands or the nature of the domain of interest completely. Even with the best queries and web crawling schemes, both the style and content of the web-data will usually differ significantly from the specific needs. For example, a speech recognition system requires conversational style text whereas most of the data on the web is literary.

The mismatch between in-domain data and web-data can be seen as a semi-supervised learning problem. We can model the web-data as a mix of sentences from two classes: in-domain (**I**) and noise (**N**) (or out-of-domain). The labels **I** and **N** are latent and unknown for the sentences in web-data but we usually have a small number of examples of in-domain examples **I**. Selecting the right labels for the unlabeled set is important for benefiting from it.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 145–148,
New York, June 2006.

Recent research on semi-supervised learning shows that in many cases (Nigam et al., 2000; Zhu, 2005) poor preprocessing of unlabeled data might actually lower the performance of classifiers. We found similar results in our language modeling experiments where the presence of a large set of noisy **N** examples in training actually lowers the performance slightly in both perplexity and WER terms. Recent literature on building language models from text acquired from the web addresses this issue partly by using various rank-and-select schemes for identifying the set **I** (Ostendorf et al., 2005; Sethy, 2005; Sarikaya, 2005). However we believe that similar to the question of balance (Zhu, 2005) in semi-supervised learning for classification, we need to address the question of distributional similarity while selecting the appropriate utterances for building a language model from noisy data. The subset of sentences from web-data which are selected to build the adaptation language should have a distribution similar to the in-domain data model.

To address the issue of distributional similarity we present an incremental algorithm which compares the distribution of the selected set and the in-domain examples by using a relative entropy (R.E) criterion. We will review in section 2 some of the ranking schemes which provide baselines for performance comparison and in section 3 we describe the proposed algorithm. Experimental results are provided in section 4, before we conclude with a summary of this work and directions for the future.

2 Rank and select methods for text cleaning

The central idea behind text cleanup schemes in recent literature, on using web-data for language modeling, is to use a scoring function that measures the similarity of each observed sentence in the web-data to the in-domain set and assigns an appropriate score. The subsequent step is to set a threshold in terms of either the minimum score or the number of top scoring sentences. The threshold can usually be fixed using a heldout set. Ostendorf (2005) use perplexity from an in-domain n-gram language model as a scoring function. More recently, a modified version of the BLEU metric which measures sentence similarity in machine translation has been proposed by Sarikaya (2005) as a scoring function. Instead of explicit ranking and thresholding it is also possible to design a classifier in a learning from positive and unlabeled examples framework (LPU) (Liu et al., 2003). In this system, a subset of the unlabeled set is selected as the negative or the noise set **N**. A two class classifier is then trained using the in-domain set and the negative set. The classifier is then used to label the sentences in the web-data. The classifier can then be iteratively refined by using a better and larger subset of the **I/N** sentences selected in each iteration.

Rank ordering schemes do not address the issue of distributional similarity and select many sentences which already have a high probability in the in-domain text. Adapting models on such data has the tendency to skew the distribution even further towards the center. For example, in our doctor-patient interaction task short sentences containing the word 'okay' such as 'okay','yes okay', 'okay okay' were very frequent in the in-domain data. Perplexity or other similarity measures give a high score to all such examples in the web-data boosting the probability of these words even further while other pertinent sentences unseen in the in-domain data such as 'Can you stand up please?' are ranked low and get rejected.

3 Incremental Selection

To address the issue of distributional similarity we developed an incremental greedy selection scheme based on relative entropy which selects a sentence if adding it to the already selected set of sentences reduces the relative entropy with respect to the in-domain data distribution.

Let us denote the language model built from in-domain data as P and let P_{init} be a language model for initialization purposes which we estimate by bagging samples from the same in-domain data. To describe our algorithm we will employ the paradigm of unigram probabilities though the method generalizes to higher n-grams also.

Let $W(i)$ be a initial set of counts for the words i in the vocabulary V initialized using P_{init}. We denote the count of word i in the j^{th} sentence s_j of web-data with m_{ij}. Let $n_j = \sum_i m_{ij}$ be the number of words in the sentence and $N = \sum_i W(i)$ be

the total number of words already selected. The relative entropy of the maximum likelihood estimate of the language model of the selected sentences to the initial model P is given by

$$H(j-1) = -\sum_i P(i) \ln \frac{P(i)}{W(i)/N}$$

If we select the sentence s_j , the updated R.E

$$H(j) = -\sum_i P(i) \ln \frac{P(i)}{(W(i)+m_{ij})/(N+n_j)}$$

Direct computation of R.E using the above expressions for every sentence in the web-data will have a very high computational cost since $O(V)$ computations per sentence in the web-data are required. However given the fact that m_{ij} is sparse, we can split the summation $H(j)$ into

$$\begin{aligned} H(j) &= -\sum_i P(i) \ln P(i) + \\ &\quad + \sum_i P(i) \ln \frac{W(i)+m_{ij}}{N+n_j} \\ &= H(j-1) + \underbrace{\ln \frac{N+n_j}{N}}_{T1} \\ &\quad \underbrace{- \sum_{i, m_{ij} \neq 0} P(i) \ln \frac{(W(i)+m_{ij})}{W(i)}}_{T2} \end{aligned}$$

Intuitively, the term $T1$ measures the decrease in probability mass because of adding n_j words more to the corpus and the term $T2$ measures the in-domain distribution P weighted improvement in probability for words with non-zero m_{ij}.

For the R.E to decrease with selection of sentence s_j we require $T1 < T2$. To make the selection more refined we can impose a condition $T1 + thr(j) < T2$ where $thr(j)$ is a function of j. A good choice for $thr(j)$ based on empirical study is a function that declines at the same rate as the ratio $\ln \frac{(N+n_j)}{N} \approx n_j/N \approx 1/kj$ where k is the average number of words for every sentence.

The proposed algorithm is sequential and greedy in nature and can benefit from randomization of the order in which it scans the corpus. We generate permutes of the corpus by scanning through the corpus and randomly swapping sentences. Next we do sequential selection on each permutation and merge the selected sets.

The choice of using maximum likelihood estimation for estimating the intermediate language models for $W(j)$ is motivated by the simplification in the entropy calculation which reduces the order from $O(V)$ to $O(k)$. However, maximum likelihood estimation of language models is poor when compared to smoothing based estimation. To balance the computation cost and estimation accuracy, we modify the counts $W(j)$ using Kneser-Ney smoothing periodically after fixed number of sentences.

4 Experiments

Our experiments were conducted on medical domain data collected for building the English ASR of our English-Persian Speech to Speech translation project (Georgiou et al., 2003). We have 50K in-domain sentences for this task available. We downloaded around 60GB data from the web using automatically generated queries which after filtering and normalization amount to 150M words. The test set for perplexity evaluations consists of 5000 sentences(35K words) and the heldout set had 2000 sentences (12K words). The test set for word error rate evaluation consisted of 520 utterances. A generic conversational speech language model was built from the WSJ, Fisher and SWB corpora interpolated with the CMU LM. All language models built from web-data and in-domain data were interpolated with this language model with the interpolation weight determined on the heldout set.

We first compare our proposed algorithm against baselines based on perplexity(PPL), BLEU and LPU classification in terms of test set perplexity. As the comparison shows the proposed algorithm outperforms the rank-and-select schemes with just 1/10th of data. Table 1 shows the test set perplexity with different amounts of initial in-domain data. Table 2 shows the number of sentences selected for the best perplexity on the heldout set by the above schemes. The average relative perplexity reduction is around 6%. In addition to the PPL and WER improvements we were able to acheive a factor of 5 reduction in the number of estimated language model parameters (bigram+trigram) and a 30% reduction in the vocab-

	10K	20K	30K	40K
No Web	60	49.6	42.2	39.7
AllWeb	57.1	48.1	41.8	38.2
PPL	56.1	48.1	41.8	38.2
BLEU	56.3	48.2	42.0	38.3
LPU	56.3	48.2	42.0	38.3
Proposed	**54.8**	**46.8**	**40.7**	**38.1**

Table 1: Perplexity of testdata with the web adapted model for different number of initial sentences.

ulary size. *No Web* refers to the language model built from just in-domain data with no web-data. *All-Web* refers to the case where the entire web-data was used.

The WER results in Table 3 show that adding data from the web without proper filtering can actually harm the performance of the speech recognition system when the initial in-domain data size increases. This can be attributed to the large increase in vocabulary size which increases the acoustic decoder perplexity. The average reduction in WER using the proposed scheme is close to 3% relative. It is interesting to note that for our data selection scheme the perplexity improvments correlate surprisingly well with WER improvments. A plausible explanation is that the perplexity improvments are accompanied by a significant reduction in the number of language model parameters.

5 Conclusion and Future Work

In this paper we have presented a computationally efficient scheme for selecting a subset of data from an unclean generic corpus such as data acquired from the web. Our results indicate that with this scheme, we can identify small subsets of sentences (about 1/10th of the original corpus), with which we can build language models which are substantially smaller in size and yet have better performance in both perplexity and WER terms compared to models built using the entire corpus. Although our focus in the paper was on web-data, we believe the proposed method can be used for adaptation of topic specific models from large generic corpora.

We are currently exploring ways to use multiple bagged in-domain language models for the selection process. Instead of sequential scan of the corpus, we are exploring the use of rank-and-select methods to give a better search sequence.

	10K	20K	30K	40K
PPL	93	92	91	91
BLEU	91	90	89	89
LPU	90	88	87	87
Proposed	**12**	**11**	**11**	**12**

Table 2: Percentage of web-data selected for different number of initial sentences.

	10K	20K	30K	40K
No Web	19.8	18.9	18.3	17.9
AllWeb	19.5	19.1	18.7	17.9
PPL	19.2	18.8	18.5	17.9
BLEU	19.3	18.8	18.5	17.9
LPU	19.2	18.8	18.5	17.8
Proposed	**18.3**	**18.2**	**18.2**	**17.3**

Table 3: Word Error Rate (WER) with web adapted models for different number of initial sentences.

References

Abhinav Sethy and Panayiotis Georgiou et al.. Building topic specific language models from web-data using competitive models. Proceedings of Eurospeech. 2005

Bing Liu and Yang Dai et al.. Building Text Classifiers Using Positive and Unlabeled Examples. Proceedings of ICDM. 2003

Kamal Nigam and Andrew Kachites McCallum et al.. Text Classification from Labeled and Unlabeled Documents using EM. Journal of Machine Learning. 39(2:3)103–134. 2000

Mirella Lapata and Frank Keller. Web-based models for natural language processing. ACM Transactions on Speech and Language Processing. 2(1),2005.

Philip Resnik and Noah A. Smith. The Web as a parallel corpus. Computational Linguistics. 29(3),2003.

P.G. Georgiou and S.Narayanan et al.. Transonics: A speech to speech system for English-Persian Interactions. Proceedings of IEEE ASRU. 2003

Ruhi Sarikaya and Agustin Gravano et al. Rapid Language Model Development Using External Resources For New Spoken Dialog Domains Proceedings of ICASSP. 2005

Tim Ng and Mari Ostendorf et al.. Web-data Augmented Language Model for Mandarin Speech Recognition. Proceedings of ICASSP. 2005

Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Computer Science, University of Wisconsin-Madison.