STRATEGIES FOR LANGUAGE MODEL WEB-DATA COLLECTION

Vincent Wan , Thomas Hain

Department of Computer Science
University of Sheffield, UK
{v.wan,t.hain}@dcs.shef.ac.uk

ABSTRACT

This paper presents an analysis of the use of textual information collected from the internet via a search engine for the purpose of building domain specific language models. A framework to analyse the effect of search query formulation on the resulting web-data language model performance in an evaluation is developed. The framework gives rise to improved methods of selecting n-gram search engine queries, which return documents that make better domain specific language models.

1. INTRODUCTION

The construction of a competitive automatic speech recognition (ASR) system requires considerable amounts of data for both acoustic and language modelling. It is a well known disadvantage of such systems that, for optimal performance, the data has to originate from the specific task and domain. For acoustic modelling the recording of sufficient data can be costly and time consuming. In the case of language modelling collection of sufficient data is overwhelming, especially in the case of tasks covering inter-human interaction. The use of *background language models* trained on large amounts of spoken and written text helps but the overall system performance is still considerably poorer without in-domain data.

Recently it was shown that data collected from the world-wide-web via a search engine could aid in the collection of in-domain data [1, 2, 3, 4]. Search engine queries are formed from n-grams obtained from a small sample of in-domain data. The text retrieved from these queries is then normalised, filtered and used to train a standard n-gram language model. While that model could be used directly it was found to be beneficial to interpolate it with a generic background model. This approach is in particular appealing both for tasks concerning conversational speech as well as speech from highly specialised areas because the world-wide-web holds many transcripts of speech as well as a wealth of specialised material. Wide-spread use of the technique was made for the transcription of conversational telephone speech in recent U.S. NIST evaluations [5] and for meeting room transcriptions [6]).

Experimental evidence suggests that the selection of the search queries has a considerable impact on the performance of the resulting language models, both in terms of perplexity and word error rates. This fact was also noted in recent work by Sethy et al. [7], who proposed multiple changes to the original techniques: Firstly, the search for query terms is based on the relative entropy between an in-domain *topic model* and a background model; and secondly, both the topic and the background language are updated according to relevance estimates based on log-probabilities of the prior language models then data selection was performed on an utterance level.

This paper develops a framework to analyse the effect of search query formulation on the resulting performance in an evaluation. In contrast to the work by [7] the formulation operates on a "per n-gram" basis. In order to retain robustness with in-domain data we derive simple measures for the selection of query terms.

The rest of the paper is organised as follows: section 2 describes web-data collection mathematically and motivates the use of *search models* in section 3. Section 4 provides an analysis and supporting experimental results. Section 5 concludes the paper.

2. COLLECTING WEB-DATA

Let B denote the background text, for example, a corpus of generic conversational speech that is topic independent. Let T be a small corpus that indicates the topic of interest and serves as the seed for the collection of a larger corpus C from the internet. Let E be the evaluation corpus, which may be identical to T but in reality should be different.

Assume that the language models are unsmoothed n-grams of arbitrary history depth, so the probability of an n-gram given a model derived from B is denoted

$$P(w|h,B) = \frac{N(w,h,B)}{N(h,B)} \tag{1}$$

where h is the history of word w, $N(w,h,B)$ is the count of the n-gram (w,h) in corpus B and $N(h,B)$ is similarly defined as the count of h in B.

The log likelihood of the corpus E given the model derived from B is

$$\log P(E|B) = \sum_w \sum_h N(w,h,E) log P(w|h,B) \tag{2}$$

When collecting web-data C, the aim is to ensure that the language model BC derived from an interpolation of B and C is more likely to generate E than the model B alone.

$$\log P(E|BC) > \log P(E|B) \tag{3}$$

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-145)

Substituting both sides with (2) and rearranging gives

$$\sum_w \sum_h N(w,h,E) \left[\log P(w|h,BC) - \log P(w|h,B)\right] > 0 \tag{4}$$

It is clear from the $N(w,h,E)$ term in (4) that only the n-grams present in E need to be investigated and only those n-grams with sufficient frequency are likely to provide significant contributions. Furthermore, to ensure, with certainty, that the data collection is working we can require that

$$\left[\log P(w|h,BC) - \log P(w|h,B)\right] > 0 \qquad \forall\,(w,h) \tag{5}$$

Note that this is a sufficient but not necessary condition. Assuming that BC is a linear interpolation of B and C,

$$P(w|h,BC) = \lambda P(w|h,C) + (1-\lambda)\,P(w|h,B) \tag{6}$$

When (6) is substituted into (5) the inequality becomes

$$\frac{P(w|h,C)}{P(w|h,B)} > 1 \tag{7}$$

which says that the language models will be improved, irrespective of the interpolation weights and excluding degenerate cases, when the likelihoods of the n-grams computed using a model derived from C are greater than the corresponding likelihoods computed using a model of B.

3. SEARCH MODELS

A search model is a model that predicts the outcome of a web-data collection. With these models, the n-gram probabilities of the collected corpus resulting from a given set of queries may be estimated, as outlined above.

3.1. Probability estimate assumption model

In the simplest case we assume that the probability distribution of the collected data is a linear combination of T and B. In other words it is assumed that the probability distributions of T and B are well estimated, but with incorrect proportions. So the search will yield similar distributions:

$$P\,(w|h,C) = \alpha P(w|h,T) + (1-\alpha)\,P(w|h,B) \tag{8}$$

assuming that B covers T fully. Substituting into (7) gives

$$\frac{P(w|h,T)}{P(w|h,B)} > 1 \tag{9}$$

Hence, composing queries from high likelihood ratio n-grams will improve the result. However, the probability estimation from T will be poor and so this method of choosing queries may not be totally reliable.

3.2. Count based model

Instead of placing assumptions on the probability distributions, another search model assumes that the n-gram count histograms

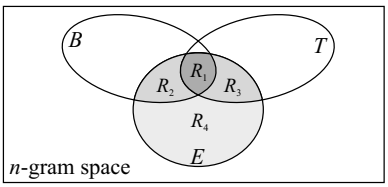

Fig. 1. Partitioning the n-gram space

obtained in the collection C are scaled versions of the counts in T and B,

$$N(w,h,C) = \alpha N(w,h,T) + \beta N(w,h,B) \tag{10}$$

α and β relate to the size of the collected corpus (N_C). Substituting (10) into (1) then (7) gives,

$$\frac{\frac{\alpha N(w,h,T)+\beta N(w,h,B)}{\alpha N(h,T)+\beta N(h,B)}}{\frac{N(w,h,B)}{N(h,B)}} = \frac{\alpha\frac{N(w,h,T)}{N(w,h,B)}+\beta}{\alpha\frac{N(h,T)}{N(h,B)}+\beta} > 1 \tag{11}$$

and hence

$$\frac{N(w,h,T)}{N(w,h,B)} > \frac{N(h,T)}{N(h,B)} \tag{12}$$

which is the same result as (9) even though the probability estimate is different. Since the aim is to satisfy the inequality (4) by ensuring that the summation is over predominantly positive terms

$$N(w,h,E)\log\frac{\alpha\frac{N(w,h,T)}{N(w,h,B)}+\beta}{\alpha\frac{N(h,T)}{N(h,B)}+\beta} \gg 0 \tag{13}$$

It is evident from the numerator of (13) that the larger the ratio $\frac{N(w,h,T)}{N(w,h,B)}$ the greater the value. The contribution will be very large if $N(w,h,B) \to 0$. Hence it is beneficial to search for n-grams that have $N(w,h,T) \gg 0$ and $N(w,h,B) = 0$. More interestingly, the denominator suggests that it is beneficial to boost those n-grams that are more frequent in T than in B *and* have corresponding histories that have a comparable or a lesser count in T than in B. For example, one might use as a query an n-gram that occurs in T but not in B if the corresponding $(n-1)$-gram history occurred frequently in B.

3.3. Partitioned n-gram space model

The count based model may be refined by writing $N(w,h,C)$ in terms of various n-gram subsets. On the whole, only the n-grams that are in set E need to be considered as they form the evaluation base. Splitting the n-gram space defined by E enables each n-gram to be treated differently depending upon whether it lies in a part of E that intersects with T, B or both: with the aid of figure 1,

$$N(w,h,C) = \tag{14}$$
$$\begin{cases} \nu_1 N(w,h,T) + \nu_2 N(w,h,B) & (w,h) \in R_1 \\ \nu_3 N(w,h,B) & (w,h) \in R_2 \\ \nu_4 N(w,h,T) & (w,h) \in R_3 \\ \nu_5 & (w,h) \in R_4 \end{cases}$$

Language model	PPL on E	Interpolation weights
B model	140.4	
T model	146.8	
B and T interp.	95.6	B=0.46; T=0.54
C_{freq} model	234.4	
B, C_{freq} interp.	119.1	B=0.70; C=0.30
B, T, C_{freq} interp.	91.3	B=0.35; T=0.51; C=0.14
$C_{\notin B}$ model	237.9	
B, $C_{\notin B}$ interp.	114.9	B=0.68; C=0.32
B, T, $C_{\notin B}$ interp.	90.4	B=0.35; T=0.50; C=0.15

Table 1. Perplexities on E of baseline language models and models derived from count based approaches.

Further assuming that E and T are identical then

$$N(w,h,C) = \begin{cases} \nu_1 N(w,h,T) + (\nu_2+\nu_3)N(w,h,B) & (w,h) \in R_1 \\ \nu_4 N(w,h,T) + \nu_5 & (w,h) \in R_4 \end{cases} \quad (15)$$

Since all of these cases are subsumed by the count based models above then the results are identical, however the overall sum (4) is split into the associated parts.

4. ANALYSIS

Experiments were performed on the AMI meeting corpus [8]. The background text B consisted of 15M words from Switchboard, Fisher and ICSI meetings [9] corpora. The T set consisted of 118 thousand words taken from a subset of the ES*a and ES*b recordings of the AMI corpus. The evaluation set E consisted of 90 thousand words from the corresponding ES*c recordings. Web-data collections (C) were obtained using the tools from the University of Washington [3] with some additional text normalisation to further improve the quality of the data.

Table 1 lists perplexity results in three sections. The first section gives the baseline perplexity on E using models constructed from B and T. The second section shows results of a count based web-data collection obtained by searching for the 448 most frequent 4-grams of T (the least frequent having a count of 4), using one 4-gram per search. 5M words were collected, which, after normalisation, resulted in 3.9M words in C_{freq}. In the third section, 3.6M normalised words of web-data $C_{\notin B}$ were similarly collected using the 432 most frequent 4-grams of T that were not found in B (the least frequent 4-gram query had a count of 2). The improvement is evident and, supporting the result of (13), shows that simply choosing the most frequent 4-grams is not the best approach. Language models created from web-data collected using the improved count based approach were tested successfully in the RT05s meeting transcription evaluations [6].

Figure 2 is a histogram of the number of 4-grams that have a particular occurrence in T and a log likelihood ratio in a particular range. It shows how collecting C_{freq} is not optimal. The least frequent n-gram queried had a count of 4. It can be seen from figure 2b that some of the queried 4-grams have negative log-likelihood ratios. This violates the sufficient condition (5) and, therefore, may affect the resulting web-data language model negatively. Furthermore, a significant proportion of these 4-grams have a low likelihood ratio and so (4) is reduced.

Table 2 shows the results of using the log likelihood ratio (9) for choosing search queries. All n-grams in T were grouped by their log likelihood ratio according to the ranges shown in column 1 of the table and separate collections where made for each range using single 4-gram queries. For comparison purposes, the sizes of the (normalised) collections were limited to a maximum of about 4M words. 4-grams with ratios greater than 12 were also queried but they returned very little data. The distribution of the queries across each log likelihood range can be inferred from figure 2.

Table 2, column 3 shows the change in perplexity after each web-data model is interpolated with the background model. It indicates that, for AMI data, there is a "sweet spot" in the log likelihood ratio ranges between 2 and 6, each of which have lower perplexities than the C_{freq} collection. The simple probability estimate model (9) indicated that queries based on 4-grams with higher likelihood ratios should yield better models than lower ratio queries. However, this is true only up to a point as the term $N(w,h,E)$ in (4) also has an influence: higher likelihood ratio n-grams generally occur less frequently so the gain achieved by boosting them is lessened. This is clearly shown in figure 2. The result may also be partly related to the relatively small number of words returned by searches for high likelihood ratio queries: it is not uncommon for high ratio queries to return no results.

The mass distribution columns of table 2 show the proportions of the web-data intersecting with E and B. The regions R_1 to R_4 are defined with the aid of figure 1 but replacing E with C and T with E. The results of $(R_1 + R_3)$ show that only a small proportion of the collected web-data actually intersects with E: approximately 1% of the 4-gram mass and between 7% and 9% of the 3-gram mass. This small percentage overlap is unsurprising as E is only small but it may also indicate that some additional filtering of the web-data may be necessary. Interestingly, there is a substantial difference between the percentage overlap of 3-grams and of 4-grams in R_1 and R_2. It suggests that querying for 4-grams actually returns many more topic relevant 3-grams. The 3-gram overlap in R_2, which shows the overlap between the web-data and B, is enormous and indicates that a general search will just retrieve a lot of B material. The factor seven difference (between 3 and 4-grams) in the amount of overlap in R_1 is especially interesting as it relates to the likelihood ratio directly and suggests that it may actually be easier find in-domain material by searching for 3-grams.

Note that in many cases it is possible to achieve more significant perplexity reductions by collecting more web-data. For example, it is possible to achieve a perplexity of 115 by collecting 7.5M words for the log likelihood range 2 – 3. The final perplexity obtained by interpolating all the models of table 2 using the weights in column 4 was 109.7. Curiously, the interpolation weights tell a different story from the perplexity numbers: the weights decrease steadily as the likelihood ratio increases giving no indication of the "sweet spot" mentioned above.

LLR range	Number of words collected	ppl on E after interp. with B	Overall interp. weight	% 4-gram mass distribution of collected data				% 3-gram mass distribution of collected data			
				R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
B	–	140.4	0.610								
1 – 2	3,763,221	122.0	0.103	1.37	13.26	0.06	85.31	8.85	30.59	0.12	60.44
2 – 3	3,988,426	117.7	0.081	1.28	12.26	0.08	86.38	8.47	29.23	0.19	62.11
3 – 4	4,004,998	118.3	0.049	1.14	11.62	0.08	87.16	7.86	28.69	0.18	63.26
4 – 5	3,785,525	117.4	0.047	1.14	11.12	0.08	87.66	7.73	27.81	0.21	64.25
5 – 6	2,638,330	118.2	0.042	1.04	10.38	0.09	88.49	7.33	26.97	0.24	65.45
6 – 7	1,512,204	121.0	0.022	1.03	10.52	0.09	88.36	7.26	27.05	0.25	65.44
7 – 8	887,684	123.3	0.014	1.23	11.03	0.10	87.64	7.85	27.39	0.23	64.53
8 – 9	410,533	124.2	0.021	1.07	10.59	0.09	88.25	7.36	26.85	0.27	65.52
9 – 10	303,744	131.5	0.001	0.94	10.51	0.07	88.48	6.96	27.42	0.19	65.43
10 – 11	133,526	136.4	0.000	0.89	10.40	0.07	88.64	6.90	27.64	0.26	65.20
11 – 12	71,767	130.0	0.010	0.92	8.87	0.10	90.11	6.66	24.78	0.31	68.25

Table 2. Results of collecting web-data according to log likelihood ratio value

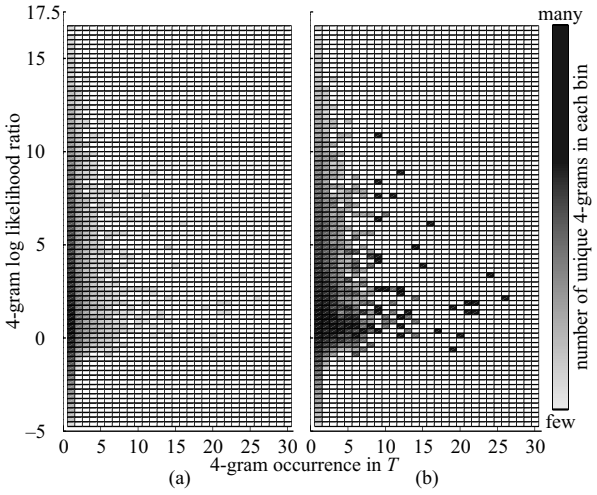

Fig. 2. A 3D histogram of the number of unique 4-grams with a certain count in T and have a log likelihood ratio within a certain range: (a) is the unnormalised histogram showing that the vast majority of 4-grams occur once; (b) has the histogram peaks in each column normalised to the same height to make the more frequent 4-grams visible.

5. CONCLUSION

This paper introduced the concept of a search model of web-data collection for improving n-gram language models. The simplest of search models have shown better ways of gathering web-data by choosing the search queries more carefully. Evidence seems to indicate that queries should be chosen carefully according to both n-gram likelihood ratio and n-gram counts, with some trade-off between the two. The theoretical analysis also suggests that searching for relatively more frequent n-grams of T that have histories well represented in B will also be beneficial.

6. REFERENCES

[1] M. Mahajan, D. Beeferman, and X. D. Huang, "Improved topic-dependent language modelling using information retrieval techniques," in *Proc. ICASSP*, 1999.

[2] X. Zhu and R. Rosenfield, "Improving trigram language modelling with the world-wide-web," in *Proc. ICASSP*, 2001.

[3] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc HLT*, 2003.

[4] R. Sarikaya, A. Gravano, and Y. Gao, "Rapid language model development using external resources for new spoken dialog domains," in *Proc. ICASSP*, 2005, vol. 1, pp. 573–576.

[5] A. Lee, "2004 fall rich transcription speech-to-text evaluation," in *Proc. NIST RT04F Workshop*, 2004.

[6] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005.

[7] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competitive models," in *Proc. Interspeech*, 2005.

[8] J. Carletta, S. Ashby, S. Bourban, M. Guillemot, M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma, "The AMI meeting corpus," in *Proc MLMI*, 2005.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, 2003.