EFFICIENT CLASS-BASED LANGUAGE MODELLING FOR VERY LARGE VOCABULARIES

E. W. D. Whittaker and P. C. Woodland

Cambridge University Engineering
Department, Trumpington Street,
Cambridge CB2 1PZ, UK.

ABSTRACT

This paper investigates the perplexity and word error rate performance of two different forms of class model and the respective data-driven algorithms for obtaining automatic word classifications. The computational complexity of the algorithm for the 'conventional' two-sided class model is found to be unsuitable for very large vocabularies (>100k) or large numbers of classes (>2000). A one-sided class model is therefore investigated and the complexity of its algorithm is found to be substantially less in such situations. Perplexity results are reported on both English and Russian data. For the latter both 65k and 430k vocabularies are used. Lattice rescoring experiments are also performed on an English language broadcast news task. These experimental results show that both models, when interpolated with a word model, perform similarly well. Moreover, classifications are obtained for the one-sided model in a fraction of the time required by the two-sided model, especially for very large vocabularies.

1. INTRODUCTION

Class-based language models have frequently been shown to improve the performance of speech recognition systems when combined with conventional word-based language models even when a large amount of training data is available [1, 2, 3]. However, the vocabulary size required for a highly-inflected language like Russian is almost seven times greater (430k words) than that for a language like English (65k words) if the same vocabulary coverage is desired (1.1% OOV-rate) [4]. This presents a problem for the automatic clustering algorithm for the conventional class model since the scaling properties of the algorithm make clustering extremely time consuming. In this paper we present a comparison of the conventional class model, referred to here as the two-sided model, against an alternative formulation referred to as a one-sided model. Both these models and the greedy automatic clustering algorithms, which produce classifications by maximising the training set likelihood, are described in Section 2. It is then demonstrated that the clustering operation for the one-sided model can be performed significantly faster than that for the two-sided model with little or no loss in performance from the word/class model combination. The comparison is made in terms of perplexity on Russian and English data and the results are presented in Section 3. In particular the time taken to classify words for the two different types of class model is examined and these results are presented in Section 4. In Section 5 a comparison is made of the performance of the interpolated word and class models in terms of word error rate from lattice rescoring experiments on an English language broadcast news task.

E.W.D.Whittaker was supported in this research by an EPSRC studentship. His current affiliation is with Compaq Computer Corporation's Cambridge Research Laboratory (edw@crl.dec.com).

2. CLASS-BASED LANGUAGE MODELLING

The effects of sparsity in a corpus can be reduced to some extent by mapping each of the N_V vocabulary words w into N_C classes—where $N_C < N_V$—and collecting N-gram statistics for the mapped corpus. A deterministic word-to-class mapping

$$C : w \rightarrow c = C(w), \tag{1}$$

in which a word may only belong to one class, may be obtained using an automatic clustering algorithm. In this work, words are clustered into classes automatically using the training set likelihood of a class bigram model as the optimisation criterion.

An interpolated word and class model is then built to combine the specificity of the word model with the generalisation ability of the class model. Varying the number of classes changes the ability of the class model to generalise to unobserved word sequences. The optimal number of classes required to complement the performance of the word model is generally dependent on the amount of training data available.

The two types of class model which have been investigated are described in the next two sections.

2.1. Two-sided class model

The two-sided class model is represented by the following two component probabilities:

$$P_0(w_i \mid C(w_i)) \cdot P_1(C(w_i) \mid C(w_{i-N+1}), \ldots, C(w_{i-1})) \tag{2}$$

i.e. a unigram class membership component and a class N-gram component. The model is two-sided (symmetric) since the same classification function C is used to map words in the history and also the current word. This particular model has received most attention in the literature on language modelling since once C has been determined, the N-gram class component can be built in an identical manner to that for word N-gram models. This makes it particularly easy to implement.

The clustering algorithm used in these experiments is the exchange algorithm described in [5] in which each word in the vocabulary is moved in turn to all available classes and left in the

class for which the increase in the class bigram training set likelihood was greatest. The operation is greedy since no consideration is made of subsequent configurations. A naive implementation of the clustering algorithm scales quadratically in the number of classes since each time a word is moved to one class, all class bigram counts are potentially affected. However, by only considering those counts that actually change, the algorithm can be made to scale somewhere between linearly and quadratically in the number of classes [1]:

$$\mathcal{O}(I \cdot (2 \cdot B + N_V \cdot N_C \cdot (N_C^{pre} + N_C^{suc}))), \quad (3)$$

where N_C^{pre} and N_C^{suc} are the average number of predecessor and successor classes respectively, for which values must be updated each time a word is moved. Irrespective of these improvements the algorithm still scales approximately linearly in the size of the vocabulary. The factor B is the number of unique bigrams in the corpus which is a function of the corpus and vocabulary sizes.

2.2. One-sided class model

The probability component which represents the one-sided model[1] used in this work is given by the following[6]:

$$P(w_i \mid C(w_{i-N+1}), \ldots, C(w_{i-1})). \quad (4)$$

The current word is now conditioned directly on the preceding words which are mapped into classes. The same classification function is used for all word positions however this is not obligatory (as it is not for the two-sided model) and further improvements have been obtained when an independent classification function is determined for each position in the word history.

The action of the clustering algorithm for the one-sided model is essentially identical to that for the two-sided model. Each word is moved among the available classes and the configuration which maximises the class bigram training set likelihood is chosen. The fundamental difference is that each time a word is moved to a new class, the counts involving other classes are not affected. Consequently, the algorithm scales linearly in the number of classes. This is illustrated by the order of the algorithm as follows:

$$\mathcal{O}\left(I \cdot N_V \cdot N_C \cdot \left(\frac{B}{N_V} + 1\right)\right). \quad (5)$$

In addition, factoring N_V back in to the above expression results in $B + N_V$ inside the brackets. Since for most corpora B dominates and generally scales substantially less than linearly with the size of the vocabulary, this will be shown to be a significant advantage of clustering for the one-sided class model when large vocabularies are involved.

3. PERPLEXITY EXPERIMENTS

The British English BNC corpus and the Russian corpus used for the perplexity experiments each comprise around 100 million words and are partitioned into `training`, `dev-test` and `eval-test` sets in the ratio 98:1:1. More details can be found in [4]. The baseline performance of the back-off word trigram models built for each corpus are given in Table 1 together with the number of N-gram events stored in each model. The models presented in this paper have all had singleton bigrams and trigrams removed unless stated otherwise.

Language	Word trigram perplexity on `eval-test`	Model size (parameters)
Russian (65k)	413.3	10,896,660
Russian (430k)	677.0	12,177,700
English (65k)	216.1	12,431,060

Table 1. Perplexities of word trigram models on `eval-test` portion of corpus for 65k and 430k vocabularies on Russian and 65k vocabulary on BNC together with the number of parameters in each trigram model.

Automatically derived classifications were produced for a range of different numbers of classes (204, 504, 1004, 2004, 3004, 4004 and 5004)[2] for both the two-sided and one-sided algorithms. For each set of classifications a back-off class trigram model was built and all singleton bigrams and trigrams were discarded. Consequently each model contained a different number of parameters. However, it was considered more important that the class models did not contain singleton N-gram events that had been discarded from the word model since these were considered more likely to contribute to any differences observed in model performance. Results are reported for English and Russian with a 65k vocabulary and also for Russian with a 430k vocabulary. Two-sided classifications for the latter were too time consuming to generate and so are not given.

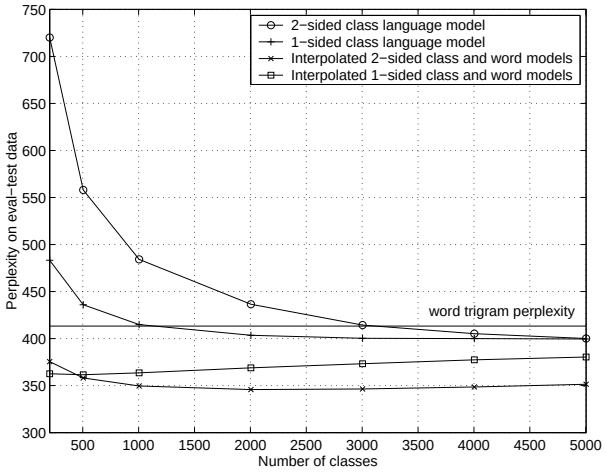

Fig. 1. Russian (65k): perplexity results for stand-alone class and interpolated word/class models on `eval-test` data.

In ASR applications, the performance improvements using class models have generally been obtained by combining them with a word model using linear interpolation. Perplexity results for both the stand-alone model and the interpolated word and class model are reported where the optimal interpolation weights were determined on the held-out `dev-test` data and all perplexity results

[1]The one-sided model has been mentioned occasionally in the literature but the authors are unaware of any other experimental results obtained using the model.

[2]The extra four classes in each model contain symbols that were not considered for clustering: the unknown, number and word boundary symbols.

are computed on the appropriate corpus's `eval-test` data. A plot of the perplexities of each of the seven different class models both alone and combined with the word model is shown in Figure 1 for Russian and in Figure 2 for English.

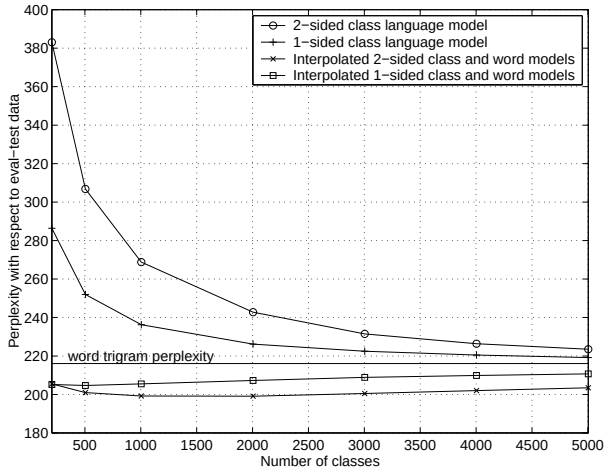

Fig. 2. English (65k): perplexity results for stand-alone class and interpolated word/class models on `eval-test` data.

	Perplexity		% improvement
N_C	PP_{class}	PP_{interp}	over word trigram
204	773.9	567.1	16.2
504	701.9	566.8	16.2
1004	670.7	572.4	15.5
2004	653.0	582.1	14.0
3004	648.6	590.5	12.8
4004	648.2	597.9	11.7
5004	647.7	602.7	11.0

Table 2. Russian corpus (430k): perplexity on `eval-test` data of stand-alone 1-sided class trigram models and interpolated word and class trigram models.

The improvements of the interpolated models over the baseline word trigram that were obtained with Russian are greater (up to 16.3%) than those obtained for English (up to 7.9%). This was attributed mainly to the greater sparsity of the Russian corpus. For the interpolated models, the optimal number of classes for the 65k vocabularies on both corpora was 2004 for the two-sided model and 504 for the one-sided model. This reflects the similar quantity of training data used for each language. Also, the one-sided model captures less general dependencies than the two-sided model hence the optimal number of classes will generally be lower so as to best complement the specific dependencies captured by the word model.

It is interesting that for both Russian vocabulary sizes a number of different stand-alone class models had a lower perplexity than the word model and also contained fewer parameters. It is particularly interesting to note that for the Russian 430k vocabulary the stand-alone 5004 class trigram model also outperforms a word 4-gram model (with all singleton N-grams discarded) which has a perplexity of 659.6 on the same data. Moreover, the class model is 32% smaller than the word 4-gram model. The poor performance of the word 4-gram model on Russian is partly due to the difficulty of generating reliable probability estimates with insufficient data. However, the problem is also partly explained by the fact that higher order word N-grams become less useful for modelling highly-inflected languages like Russian in which a more flexible word ordering is permitted.

4. ALGORITHM PERFORMANCE

In Figure 3 the clustering time per iteration through the vocabulary of the two algorithms on a 300MHz Sparc Ultra2 processor are presented for different numbers of classes on the Russian corpus for both the 65k and 430k vocabularies. Although not shown here, this plot is similar to that obtained for the BNC corpus. Clustering the 430k vocabulary for the two-sided model was not performed for more than 2004 classes due to the excessive computation time required.

The class/word bigram counts used in the one-sided clustering algorithm were stored using a sparse matrix implementation so as to be memory efficient when a large vocabulary and large numbers of classes were used. However, for a very large vocabulary it was found that the clustering speed was reduced adversely by the relatively larger proportion of words that appear in each class context when a small number of classes is used. An implementation using arrays (essentially equivalent to that used in the two-sided clustering algorithm) was also implemented and found to be much faster but it could only be used for 204 and 504 classes due to memory limitations. The timing points for this version of the one-sided algorithm are shown using circles in Figure 3.

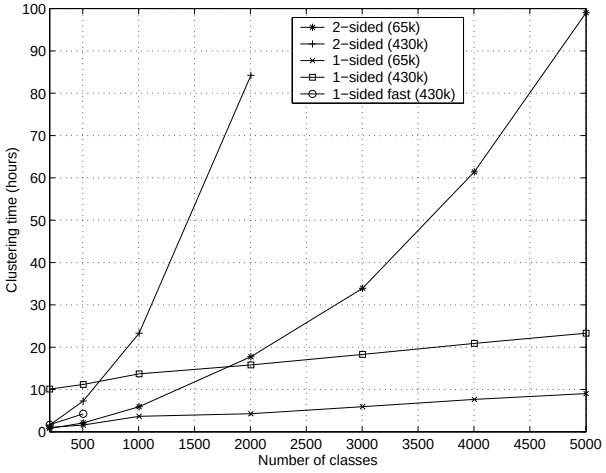

Fig. 3. Clustering times per iteration for 1-sided and 2-sided models with both a 65k and 430k Russian vocabulary (fast implementation of one-sided clustering algorithm is marked using circles).

The graph in Figure 3 clearly shows the speed advantage of the clustering algorithm for the one-sided model as it was predicted by the order of the algorithm in Equation 5. Moreover, for the much larger 430k vocabulary there is simply an offset in the overall clustering time. The scaling in the number of classes is otherwise similar to that for the 65k vocabulary.

5. RECOGNITION EXPERIMENTS

This section shows that the improvement in clustering speed using the one-sided class model formulation is not obtained at the expense of recognition performance.

The 1997 DARPA HUB4 broadcast news evaluation was chosen for the experiments and we perform lattice rescoring experiments on lattices generated using the 1997 HTK broadcast news transcription system [7]. The language model training data comprised 132 million words of the LDC broadcast news texts, the transcriptions of the 1997 broadcast news training data (added twice) and the 1995 Marketplace transcriptions. A word trigram model was built using the same vocabulary that was used to generate the original lattices. This baseline word trigram employed Katz back-off with Good-Turing discounting and had singleton bigrams and trigrams removed to produce a model containing around 16.5 million parameters. The optimal number of classes and the interpolation weights between the word and class models were optimised on the development lattices. The number of classes was varied in increments of 100 between 100 and 1500 classes and the interpolation weights evaluated in increments of 0.1. The optimal number of classes for the two-sided model was found to be 1000 with interpolation weights of 0.7 (word) and 0.3 (class). For the one-sided model the optimal number of classes was 400 with interpolation weights 0.6 (word) and 0.4 (class). The perplexity of the models on the reference transcription and the word error rate results are given in Table 3.

Model	PP_{ref}	$\%WER$	% rel. imp.
Interp. 2-sided 1000	161.6	17.8	2.2
Interp. 1-sided 400	162.4	17.8	2.2
Baseline word trigram	171.4	18.2	—

Table 3. Perplexity and word error rate on evaluation data of the optimised, interpolated word and class models and the baseline word trigram model.

Both model combinations give an improvement in performance over the baseline word trigram model, and each improvement is statistically significant at the 99% level using the NIST Matched Pair Sentence Segment test. Also, interpolating a word 4-gram model (with bigram, trigram and 4-gram cutoffs of 1,3,3) with a class 4-gram model (with the same cutoffs) reduced the word 4-gram baseline result of 17.6% to 17.1% for the interpolated word and two-sided model, and to 17.2% for the interpolated word and one-sided model. Both these improvements were also found to be statistically significant at the 99% confidence level using the same test.

6. DISCUSSION

It is clear from the perplexity results that have been presented for Russian that combined class and word based language modelling can produce significant improvements in performance. For a language like Russian where higher order word N-gram models do not significantly improve performance, combinations of word and class models appear to offer a very appealing solution. Although the improvements in perplexity were shown to be generally less for English, they still translated into significant reductions in word error rate on an English language broadcast news task.

The advantage of using the one-sided class model has been clearly demonstrated for a situation where a very large vocabulary is necessary. Automatic classifications can be obtained in significantly less time than for the two-sided class model with little or no loss in performance. Also, although it has not been explicitly investigated here, clustering can be used as a pruning technique for obtaining smaller and/or more robust stand-alone language models. This was demonstrated with the perplexity results presented in Section 3. For such situations a larger number of classes is generally required hence the scaling properties of the algorithm for the one-sided class model would recommend its use.

7. CONCLUSION

In this paper we have presented a comparison of two-sided and one-sided class models and have shown the performance of the two models in terms of perplexity and word error rate to be comparable. We also showed that the computational complexity of the algorithm for obtaining classifications for the one-sided model scaled much more favourably than that for the two-sided model particularly when large vocabularies and/or large numbers of classes were involved. In addition, results were also obtained which showed that combinations of word and class models offer a good solution to modelling highly inflected languages especially when higher order word N-gram models give little improvement.

8. REFERENCES

[1] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," *Speech Communication*, vol. 24, pp. 19–37, 1998.

[2] T.R. Niesler, E.W.D. Whittaker, and P.C. Woodland, "Comparison of Part-of-speech and Automatically Derived Category-based Language Models for Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998, pp. 177–180.

[3] J.T. Goodman, "Putting it all together: Language Model Combination," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000, pp. 1647–1650.

[4] E.W.D. Whittaker and P.C. Woodland, "Comparison of Language Modelling Techniques for Russian and English," in *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia., 1998.

[5] R. Kneser and H. Ney, "Improved Clustering Techniques for Class-based Statistical Language Modelling," in *Proceedings of the European Conference on Speech Communication and Technology*, Berlin, Germany, 1993, pp. 973–976.

[6] E.W.D. Whittaker, *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*, Ph.D. thesis, Cambridge University, 2000.

[7] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker, and S.J. Young, "The 1997 HTK Broadcast News Transcription System," in *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, 1998, pp. 41–48.