International Conference on Artificial Neural Networks, ICANN02. To appear.

An Efficiently Focusing Large Vocabulary Language Model

Mikko Kurimo and Krista Lagus

Helsinki University of Technology, Neural Networks Research Centre
P.O.Box 5400, FIN-02015 HUT, Finland
Mikko.Kurimo@hut.fi, Krista.Lagus@hut.fi

Abstract. Accurate statistical language models are needed, for example, for large vocabulary speech recognition. The construction of models that are computationally efficient and able to utilize long-term dependencies in the data is a challenging task. In this article we describe how a topical clustering obtained by ordered maps of document collections can be utilized for the construction of efficiently focusing statistical language models. Experiments on Finnish and English texts demonstrate that considerable improvements are obtained in perplexity compared to a general n-gram model and to manually classified topic categories. In the speech recognition task the recognition history and the current hypothesis can be utilized to focus the model towards the current discourse or topic, and then apply the focused model to re-rank the hypothesis.

1 Introduction

The estimation of complex statistical language models has recently become possible due to the large data sets now available. A statistical language model provides estimates of probabilities of word sequences. The estimates can be employed, e.g., in speech recognition for selecting the most likely word or sequence of words among candidates provided by an acoustic speech recognizer.

Bi- and trigram models, or more generally, n-gram models, have long been the standard method in statistical language modeling [1]. However, the model has several well-known drawbacks: (1) an observation of a word sequence does not affect the prediction of the same words in a different order, (2) long-term dependencys between words do not affect predictions, and (3) very large vocabularies pose a computational challenge. In languages with syntactically less strict word order and a rich inflectional morphology, such as Finnish, these problems are particularly severe.

Information regarding long-term dependencies in language can be incorporated into language models in several ways. For example, in word caches [1] the probabilities of words seen recently are increased. In word trigger models [2] probabilities of word pairs are modeled regardless of their exact relative positions.

[1] n-gram models estimate $P(w_t|w_{t-n+1}w_{t-n+2}\dots w_{t-1})$, the probability of nth word given the sequence of the previous $n-1$ words. The probability of a word sequence is then the product of probabilities of each word.

Mixtures of sentence-level topic-specific models have been applied together with dynamic n-gram cache models with some perplexity reductions [3]. In [4] and [5] EM and SVD algorithms are employed to define topic mixtures, but there the topic models only provide good estimates for the content word unigrams which are not very powerful language models as such. Nevertheless, perplexity improvements have been achieved when these methods are applied together with the general trigram models.

The modeling approach we propose is founded on the following notions. Regardless of language, the size of the active vocabulary of a speaker in a context is rather small. Instead of modeling all possible uses of language in a general, monolithic language model, it may be fruitful to *focus* the language model to smaller, topically or stylistically coherent subsets of language. In the absence of prior knowledge of topics, such subsets can be computed based on content words that identify a specific discourse with its own topics, active vocabulary, and even favored sentence structures.

Our objective was to create a language model suitable for large vocabulary continuous speech recognition in Finnish, which has not yet been extensively studied. In this paper a focusing language model is proposed that is efficient enough to be interesting for the speech recognition task and that alleviates some of the problems discussed above.

2 A Topically Focusing Language Model

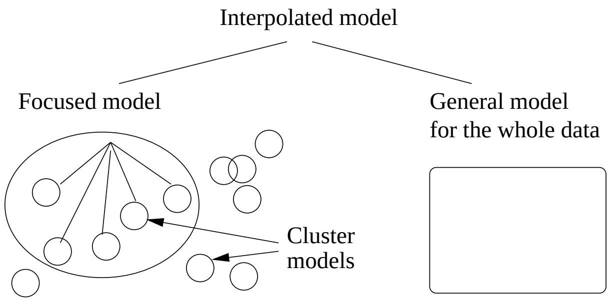

Fig. 1. *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

The model is created as follows:

1. Divide the text collection into topically coherent text 'documents', such as paragraphs or short articles.
2. Cluster the passages topically.
3. For each cluster, calculate a small n-gram model.

For the efficient calculation of topically coherent clusters we apply methods developed in the WEBSOM project for exploration of very large document collections [6][2]. The method utilizes the Self-Organizing Map (SOM) algorithm [7] for clustering document vectors onto topically organized document maps. The document vectors, in turn, are weighted word histograms where the weighting is based on idf or entropy to emphasize content words. Stopwords (e.g., function words), and very rare words are excluded, inflected words are returned to base forms. Sparse random coding is applied to the vectors for efficiency. In addition to the success of the method in text exploration, an improvement in information retrieval when compared to the standard tf.idf retrieval has been obtained by utilizing a subset of the best map units [8].

The utilization of the model in text prediction comprises the following steps:

1. Represent recent history as a document vector, and select the clusters most similar to it.
2. Combine the cluster-specific language models of the selected clusters to obtain the focused model.
3. Calculate the probability of the predicted sequence using the model and interpolate the probability with the corresponding one given by a general n-gram language model.

For the structure of the combined model, see Fig. 1. When regarded as a generative model for text, the present model is different from the topical mixture models proposed by others (e.g. [4]) in that here a text passage is generated by a very sparse mixture of clusters that are known to correspond to discourse- or topic-specific sub-languages.

Computational efficiency. Compared to the conventional n-grams or mixtures of such, the most demanding new task is the selection of the best clusters, i.e. the best map units. With random coding using sparse vectors [6] the encoding as a document vector takes $\mathcal{O}(w)$, where w is the average number of words per document. The winner search in SOM is generally of $\mathcal{O}(md)$, where m is the number of map units and d the dimension of the vectors. Due to sparse documents the search for the best map units is reduced to $\mathcal{O}(mw)$. In our experiments ($m = 2560$, $w = 100$, see Section 3.) running on a 250 MHz SGI Origin a single full search among the units took about 0.028 seconds and with additional speedup approximations that benefit from the ordering of the map, only 0.004 seconds. Moreover, when applied to rescoring the n best hypotheses or the lattice output in two-pass recognition, the topic selection need not be performed very often. Even in single-pass recognition, augmenting the partial hypothesis (and thus the document vectors) with new words requires only a local search on the map. The speed of the n-gram models depends mainly on n and the vocabulary size; a reduction in both results in a considerably faster model. The combining, essentially a weighted sum, is likewise very fast for small models. Also preliminary experiments on offline speech recognition indicate that the relative increase

[2] The WEBSOM project kindly provided the means for creating document maps.

of the recognition time due to the focusing language model and its use in lattice rescoring is negligible.

3 Experiments and Results

Experiments on two languages, Finnish and English, were conducted to evaluate the proposed unsupervised focusing language model. The corpora were selected so that each contained a prior (manual) categorization for each article. The categorization provided a supervised topic model against which the unsupervised focusing cluster model was compared. For comparison we implemented also another topical model where full mixtures of topics are used, calculated with the EM-algorithm [4]. Furthermore, as a clustering method in the proposed focusing model we examined the use of K-means instead of the SOM.

The models were evaluated using perplexity[3] on independent test data averaged over documents. Each test document was split into two parts, the first of which was used to focus the model and the second to compute the perplexity.

To reduce the vocabulary (especially for Finnish) all inflected word forms were transformed into base forms. Probabilities for the inflected forms can then be re-generated e.g. as in [9]. Moreover, even when base forms are used for focusing the model, the cluster-specific n-gram models can, naturally, be estimated on inflected forms. To estimate probabilities of unseen words, standard discounting and back-off methods were applied, as implemented in the CMU/Cambridge Toolkit [10].

Finnish corpus. The Finnish data[4] consisted of 63 000 articles of average length 200 words from the following categories: Domestic, foreign, sport, politics, economics, foreign economics, culture, and entertainment. The number of different base forms was 373 000. For general trigram model a frequency cutoff of 10 was utilized (i.e. words occurring fewer than ten times were excluded), resulting in a vocabulary of 40 000 words. For the category and cluster specific bigram models, a cutoff of two was utilized (the vocabulary naturally varies according to topic). For the focused model, the size of the document map was 192 units and only the best cluster (map unit) was included in the focus. The results on a test data of 400 articles are presented in Fig. 2.

English corpus. The English data consisted of patent abstracts from eight subcategories of the EPO collection: A01–Agriculture; A21–Foodstuffs, tobacco; A41–Personal or domestic articles; A61–Health, amusement; B01–Separating, mixing; B21–Shaping; B41–Printing; B60–Transporting. Experiments were carried out using two data sets: pat1 including 80 000 and pat2 with 648 000 abstracts, with an average length of 100 words. The total vocabulary for pat1 was nearly 120 000 base forms, the frequency cutoff for the general trigram model 3

[3] Perplexity is the inverse predictive probability for all the words in the test document.

[4] The Finnish corpus was provided by the Finnish News Agency STT.

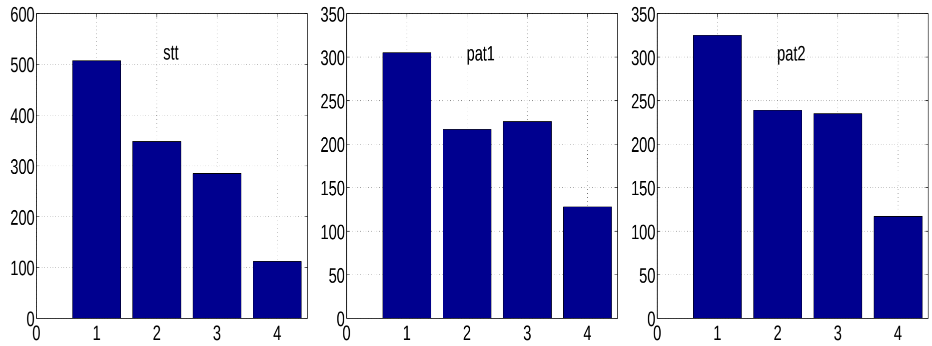

Fig. 2. The perplexities of test data using each language model for the Finnish news corpus (stt) on the left, for the smaller English patent abstract corpus (pat1) in the middle, and for the larger English patent abstract corpus (pat2) on the right. The language models in each graph from left fo right are: 1. General 3-gram model for the whole corpus, 2. Topic factor model using mixtures trained by EM, 3. Category-specific model using prior text categories, and 4. Focusing model using unsupervised text clustering. The models 2–4 were here all interpolated with the baseline model 1. The best results are obtained with the focusing model (4).

words resulting in vocabulary size 16 000. For pat2 these figures were 810 000, 5, and 38 000, respectively. For the category and cluster specific bigram models a cutoff of two was applied. The size of the document map was 2560 units in both experiments. For pat2 only the best cluster was employed for the focused model, but for pat1, with significantly fewer documents per cluster, the amount of best map units chosen was 10. The results on the independent test data of 800 abstracts (500 for pat2) are presented in Fig. 2.

Results. The experiments on both corpora indicate that when combined with the focusing model the perplexity of the general 'monolithic' trigram model improves considerably. This result is, as well, significantly better than the combination of the general model and topic category specific models where the correct topic model was chosen based on manual class label on the data. When K-means was utilized for clustering the training data instead of SOM, the perplexity did not differ significantly. However, the clustering was considerably slower (for an explanation, see Sec.2 or [6]). When applying the topic factor model suggested by Gildea and Hofmann [4] with each corpus we used 50 normal EM iterations and 50 topic factors. The first part of a test article was used to determine the mixing proportions of the factors and the second part to compute the perplexity (see results in Fig. 2).

Discussion. The results for both corpora and both languages show similar trends, although for Finnish the advantage of a topic-specific model seems more pronounced. One advantage of unsupervised topic modeling over a topic model based on fixed categories is that the unsupervised model can achieve an arbitrary granularity and a combination of several sub-topics.

The obtained clear improvement in language modeling accuracy can benefit many kinds of language applications. In speech recognition, however, it is central to discriminate between the acoustically confusable word candidates, and the average perplexity is not an ideal measure for this [11, 4]. Therefore, a topic for future research (as soon as a speech data and a text corpus of related kind can be obtained for Finnish), is to examine how well the improvements in modeling translate to advancing speech recognition accuracy.

4 Conclusions

We have proposed a topically focusing language model that utilizes document maps to focus on a topically and stylistically coherent sub-language. The longer-term dependencies are embedded in the vector space representation of the word sequences, and the local dependencies of the active vocabulary within the sub-language can then be modeled using n-gram models of small n. Initially, we aimed at improving statistical language modeling in Finnish, where the vocabulary growth and flexible word order offer severe problems for the conventional n-grams. However, the experiments indicate improvements for modeling English, as well.

References

1. P. Clarkson and A. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," In *Proc. ICASSP*, pp. 799–802, 1997.
2. R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," In *Proc. ICASSP*, pp. 45–48, 1993.
3. R.M. Iyer and M. Ostendorf, "Modelling long distance dependencies in language: Topic mixtures versus dynamic cache model," *IEEE Trans. Speech and Audio Processing*, 7, 1999.
4. D. Gildea and T. Hofmann, "Topic-based language modeling using EM," In *Proc. Eurospeech*, pp. 2167–2170, 1999.
5. J. Bellegarda. "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, 88(8):1279–1296, 2000.
6. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. "Organization of a massive document collection," *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000.
7. T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 2001. 3rd ed.
8. K. Lagus, "Text retrieval using self-organized document maps," *Neural Processing Letters*, 2002. In press.
9. V. Siivola, M. Kurimo, and K. Lagus. "Large vocabulary statistical language modeling for continuous speech recognition," In *Proc. Eurospeech*, 2001.
10. P. Clarkson and R. Rosenfeld, "Statistical language modeling using CMU-Cambridge toolkit," in *Proc. Eurospeech*, pp. 2707–2710, 1997.
11. P. Clarkson and T. Robinson. "Improved language modelling through better language model evaluation measures," *Computer Speech and Language*, 15(1):39–53, 2001.