A Dynamic Language Model

Based on Individual Word Domains

E. I. Sicilia-Garcia, Ji Ming, F. J. Smith
School Computer Science
Queen's University of Belfast
Belfast BT7 1NN, Northern Ireland
e.sicilia@qub.ac.uk

Abstract

We present a new statistical language model based on a combination of individual word language models. Each word model is built from an individual corpus which is formed by extracting those subsets of the entire training corpus which contain that significant word. We also present a novel way of combining language models called the "union model", based on a logical union of intersections, and use this to combine the language models obtained for the significant words from a cache. The initial results with the new model provide a 20% reduction in language model perplexity over the standard 3-gram approach.

Introduction

Statistical language models are based on information obtained from the analysis of large samples of a target language. Such models estimate the conditional probability of a word given a sequence of preceding words. The conditional probability can be further used to determine the likelihood of a sentence through the product of the individual word probabilities. A popular type of statistical language model is the dynamic language model, which dynamically modifies conditional probabilities depending on the recent word history. For example the cached-based natural language models (Kuhn R. & De Mori R., 1990) incorporates a cache component into the model, which estimates the probability of a word depending upon its recent usage. Trigger based models go a step further by triggering associated words to each content word in a cache giving each associated word a higher probability (Lau et al., 1993).

Our statistical language model, based upon individual word domains, extends these ideas by creating a new language model for each significant word in the cache. A significant word is hard to define; it is any word that significantly contributes to the content of the text. We define it as any word which is not a stop word, i. e. articles, prepositions and some of the most frequently used words in the language such as "will", "now", "very", etc. Our model combines individual word language models with a standard global n-gram language model.

A training corpus for each significant word is formed from the amalgamation of the text fragments taken from the global training corpus in which that word appears. As such these corpora are smaller and closely constrained; hence the individual language models are more precise than the global language model and thereby should offer performance gains. One aspect of the performance of this joint model is how the global language model is to be combined with the individual word language models. This is explored later.

This paper is organised as follows. Section 1 explains the basis for this model. The mathematical background and how the models are combined are explained in section 2. In the third section, a novel method of combining the word models, the probabilistic-union model is explained. Finally, results and conclusion are drawn.

1 Dynamic Language Model based on Word Language Models

Our dynamic language model builds a language model for each individual word. In order to do this we need to select which words are to be classified as significant and furthermore create a language model for them. We excluded all the stop words (‘is’, ’to’, ‘an’, ‘some’) due to their high frequency within the text and their limited contribution to the thematic of the text. A list of stop words was obtained by merging together the lists used by various www search engines, for example Altavista.

Secondly we need to create a dictionary that contains the frequency of each word in the corpus. This is needed because we want to exclude those non-stop words which appear too often in the training corpus, for example words like ‘dollars’, ‘point’, etc. A hash file is constructed to store large amounts of information so that it can be retrieved quickly.

The next step is to create the global language model by obtaining the text phrases and their probabilities. Frequencies of words and phrases are derived from a large text corpus and the conditional probability of a word given a sequence of preceding words is estimated. These conditional probabilities are combined to produce an overall language model probability for any given word sequence. The probability of a sequence of words is:

$$\begin{aligned} P(w_1 \cdots w_n) &= P(w_1^n) = \\ &= P(w_1) \times P(w_2 \mid w_1) \times \cdots P(w_n \mid w_1^{n-1}) = \\ &= \prod_{i=1}^{n} P(w_i \mid w_1^{i-1}) \end{aligned} \quad (1)$$

where $w_1^n = \{w_1, w_2, w_3, \cdots, w_n\}$ is a sentence or sequence of words. The individual conditional probabilities are approximated by the maximum likelihoods:

$$P_{ML}(w_i \mid w_1^{i-1}) = \frac{freq(w_1^i)}{freq(w_1^{i-1})} = \frac{freq(w_1 \cdots w_{i-1} w_i)}{freq(w_1 \cdots w_{i-1})} \quad (2)$$

where $freq(X)$ is the frequency of the phrase X in the text.

In equation (2), there are often unknown sequences of words i.e. phrases which are not in the dictionary. The maximum likelihood probability is then zero. In order to improve this prediction of an unseen event, and hence the language model, a number of techniques have been explored, for example, the Good-Turing estimate (Good I. J., 1953), the backing-off method (Katz S. M., 1987), deleted interpolation (Jelinek F. and Mercer R. L., 1984) or the weighted average n-gram model (O’Boyle P., Owens M. and Smith F. J., 1994). We use the weighted average n-gram technique (WA), which combines n-gram[1] phrase distributions of several orders using a series of weighting functions. The WA n-gram model has been shown to exhibit similar predictive powers to other n-gram techniques whilst enjoying several benefits. Firstly an algorithm for a WA model is relatively straightforward to implement in computer software, secondly it is a variable n-gram model with the length depending on the context and finally it facilitates easy model extension[2]. The weighted average probability of a word given the preceding words is

$$P_{WA}(w \mid w_1 \cdots w_m) = \frac{\lambda_0 P_{ML}(w) + \sum_{i=1}^{m} \lambda_i P_{ML}(w \mid w_{m+1-i} \cdots w_m)}{\sum_{i=0}^{m} \lambda_i} \quad (3)$$

where the weighted functions are:

$$\begin{aligned} \lambda_0 &= Ln(N), \\ \lambda_i &= Ln\left(freq\left(w_{m+1-i} \cdots w_m\right)\right) \cdot 2^i \end{aligned} \quad (4)$$

N is the number of tokens in the corpus and $freq\left(w_{m+1-i} \cdots w_m\right)$ is the frequency of the sentence $w_{m+1-i} \cdots w_m$ in the text.

The maximum likelihood probability of a word is:

---

[1] A n-gram model contains the conditional probability of a word dependant on the previous *n* words. (Jelinek F., Mercer R.L. and Bahl L. R., 1983)

[2] The *“ease of extension”* applies to the fact that additional training data can be incorporated into an existing WA model without the need to re-estimate smoothing parameters.

$$P_{ML}(w) = \frac{freq(w)}{N} \quad (5)$$

$freq(w)$ is the frequency of the word *w* in the text. This language model (defined by equation (3) and (5)) is what we term a standard n-gram language model or global language model.

Finally the last step is the creation of a language model for each significant word, which is formed in the same manner as the global language model. The word language-training corpus to be used is the amalgamation of the text fragments taken from the global training corpus in which the significant word appears. A number of choices can be made as to how the word-training corpus for each significant word can be selected. We initially construct what we termed the "paragraph context model", entailing that the global training corpus is scanned for a particular word and each time the word is found the paragraph containing that word is extracted. The paragraphs of text extracted for a particular word are joined together to form an individual word-training corpus, from which an individual word language model is built. Alternative methods include storing only the sentences where the word appears or extracting a piece of the text M^- words before and M^+ words after the search word.

Additionally some restrictions on the number of words were imposed. This was done due to the high frequency of certain words. Such words were omitted since the additional information that they provide is minimal (conversely language models for "rare" words are desirable as they provide significant additional information to that contained within the global language model). Once individual language models have been formed for each significant word (trained using the standard n-gram approach as used for the global model), there remains the problem of how the individual word language models will be combined together with the global language model.

2 Combining the Models

We need to combine the probabilities obtained from each word language model and from the global language model, in order to obtain a conditional probability for a word given a sequence of words. The first model to be tested is an arithmetic combination of the global language model and the word language models. All the word language models and the global language model are weighted equally. We believe that words, which appear far away in the previous word history, do not have as much importance as the ones closest to the word. Therefore we need to make a restriction in the number of language models. First, the conditional probabilities obtained from the word language models and the global language model can be combined in a linear interpolated model as follows:

$$P(w \mid w_1^n) = \lambda_G P_{Global}(w \mid w_1^n) + \sum_{i=1}^{m} \lambda_i P_i(w \mid w_1^n) \quad (6)$$

$$\text{where} \quad \lambda_G + \sum_{i=1}^{m} \lambda_i = 1 \quad (7)$$

and $P_i(w \mid w_1^n)$ is the conditional probability in the word language model for the significant word w_i, λ_i are the correspondent weights and m is the maximum number of word models that we are including.

If the same weight is given to all the word language models but not to the global language model and if a restriction on the number of word language models to be included is enforced, the weighted model is defined as:

$$P(w \mid w_1^n) = \alpha \cdot P_{Global}(w \mid w_1^n) + \frac{(1-\alpha)}{m}\left[\sum_{i=1}^{m} P_i(w \mid w_1^n)\right] \quad (8)$$

and α is a parameter which is chosen to optimise the model.

Furthermore, a method was used based on an exponential decay of the word model probabilities with distance. This stands to reason, as a word appearing several words previously will generally be less relevant than more recent

words. Given a sequence of words, for example, "We had happy times in America…"

We	Had	**Happy**	**Times**	In	**America**
5	4	**3**	**2**	1	

where 5, 4, 3, 2, 1 represent the distance of the word from the word *America*, *Happy* and *Times* are significant words for which we have an individual word language models. The exponential decay model for the word *w*, where in this case *w* represents the significant word *America*, is as follows:

$$P(w \mid w_1^n) = \frac{\left(\begin{array}{c} P_{Global}(w \mid w_1^n) + P_{Happy}(w \mid w_1^n) \cdot \exp(-3/d) \\ + P_{Times}(w \mid w_1^n) \cdot \exp(-2/d) \end{array} \right)}{1 + \exp(-3/d) + \exp(-2/d)} \quad (9)$$

where $P_{Global}(w \mid w_1^n)$ is the conditional probability of the word w following a phrase $w_1 \cdots w_n$ in the global language model. $P_{Happy}(w \mid w_1^n)$ is the conditional probability of the word w following a phrase $w_1 \cdots w_n$ word language model for the significant word *Happy*. The same definition applies for the word model *Times*. *d* is the exponential decay distance with *d*=5, 10, 15,etc. The decaying factor *exp(-l/d)* introduces a cut off:

$$\text{if} \quad l \geq d \quad \Rightarrow \quad \exp(-l/d) = 0$$

where l is the word model to word distance

d is the decay distance

Presently the combination methods outlined above have been experimentally explored. However, they offer a reasonably simplistic means of combining the individual and global language models. More sophisticated models are likely to offer improved performance gains.

3 The Probabilistic-Union Model

The next method is the Probabilistic-Union model. This model is based on the logical concept of a disjunction of conjunction which is implemented as a sum of products. The union model has been previously applied in problems of noisy speech recognition, (Ming J. et al., 1999). Noisy conditions during speech recognition can have a serious effect on the likelihood of some features which are normally combined using the geometric mean. This noise has a zeroing effect upon the overall likelihood produced for that particular speech frame. The use of the probabilistic-union reduces the overall effect that each feature has in the combination, therefore loosening any zeroing effect.

For the word language model, some of the conditional probabilities are zero or very small due to the small size of some of the word model corpora. For these word models, many of the words in the global training corpus are not in the word-model training-corpus dictionary. And so, the conditional probability will be in many cases zero or near zero reducing the overall probability. As in noisy speech recognition we wish to reduce the effect of this zeroing in the combined model. The probabilistic-union model is one of the possible solutions for the zeroing problem when combining language models.

The union model is best illustrated with an example when the number of word models to be included is m=4 and if they are assumed to be independent probabilities.

$$P_{Union}^{(1)}(w) = \Psi_1(P_1 \cdot P_2 \cdot P_3 \cdot P_4) \quad (10)$$

$$P_{Union}^{(2)}(w) = \Psi_2(P_1P_2P_3 \oplus P_1P_2P_4 \oplus \cdots) \quad (11)$$

$$P_{Union}^{(3)}(w) = \Psi_3(P_1P_2 \oplus P_1P_3 \oplus P_1P_4 \oplus \cdots) \quad (12)$$

$$P_{Union}^{(4)}(w) = \Psi_4(P_1 \oplus P_2 \oplus P_3 \oplus P_4) \quad (13)$$

where $P_{Union}^{k}(w) = P_{Union}^{k}(w \mid w_1^n)$ is the union model of order k. $P_i = P_i(w \mid w_1^n)$ is the conditional probability for the significant word w_i and Ψ_k is a normalizing constant. The symbol '$\oplus$' is a probabilistic sum, i.e. its equivalent for 1 and 2 is:

$$P_{1\text{and}\,2} = P_1 \oplus P_2 = P_1 + P_2 - P_1P_2 \quad (14)$$

The combination of the global language model with the probabilistic-union model is

defined as follows:

$$P(w \mid w_1^n) = \alpha P_{Global}(w \mid w_1^n) + (1-\alpha) P_{Union}(w \mid w_1^n) \quad (15)$$

Results

To evaluate the behaviour of one language model with respect to others we use perplexity. It measures the average branching factor (per word) of the sequence at every new word, with respect to some source model. The lower the branching factor, the lower the model errors rate. Therefore, the lower the branching (perplexity) the better the model. Let w_i be a word in the language model and $w_1^m = \{w_1, w_2, w_3, \cdots, w_m\}$ a sentence or sequence of words. The perplexity of this sequence of words is:

$$Perplexity(w_1 w_2 \cdots w_n) = PP(w_1^n) = \\ = \exp\left(-\frac{1}{n} \times \sum_{i=1}^{n} Ln\left(P_{WA}\left(w_i \mid w_1^{i-1} \right) \right) \right) \quad (16)$$

The Wall Street Journal (version WSJ0[3]) contains about 38 million words, and a dictionary of approximately 65,000 words. We select one quarter of the articles in the global training corpus as our training corpus (since the global training corpus is large and the normalisation process takes time). To test the new language model we use a subset of the test file given by WSJ0, selected at random. The training corpus that we are using contains 172,796 paragraphs, 376,589 sentences, 9,526,187 tokens. The test file contains 150 paragraphs, 486 sentences, 8824 tokens and 1908 words types. Although the size of this test file is small, further experiments with bigger training corpora and test files are planned.

Although in our first experiments we use 5-grams in the calculation of the word models, the size of the n-gram has been reduced to 3-grams because the process of normalisation is slow in these experiments.

[3] CSR-I(WSJ0) Sennheiser, published by LDC , ISBN:1-58563-007-1

The model based on a simple weighted combination offers improved results, up to 10% when α=0.6 in Eq. (8) and a combination of a maximum of 10 word models. Better results were found when the word models were weighted depending on their distance from the current word, that is, for the exponential decay model in Eq. (9) where d=7 and the number of word models is selected by the exponential cut off (Table 1). For this model improvements of over 17% have been found.

	Exponential Decay d			
cut	5	6	7	8
4d	15.53%	16.31%	16.46%	**16.44%**
5d	15.90%	16.42%	16.52%	16.43%
6d	15.92%	16.45%	**16.53%**	16.41%
7d	**16.02%**	**16.46%**	16.51%	16.40%
8d	**16.02%**	**16.46%**	16.51%	16.39%
9d	15.97%	16.45%	16.51%	16.39%

Table 1. Improvement in perplexity for the exponetial decay models with respect to the Global Language Model over the basic 3-gram model.

For the probabilistic-union model, we have as many models as numbers of word language models. For example, if we wish to include *m=4* word language models, the four union models are those with orders 1 to 4 (equation (13) to (15)). The results for the probabilistic union model when the number of words models is m=5 and m=6 are shown in the tables below.

	Union Model Order				
alpha	5	4	3	2	1
0.3	13%	15%	-2%	-15%	-25%
0.4	**13%**	18%	6%	-3%	-10%
0.5	12%	19%	11%	4%	-1%
0.6	12%	**19%**	13%	9%	5%
0.7	11%	18%	**14%**	11%	8%
0.8	9%	15%	13%	**11%**	**9%**
0.9	6%	10%	9%	8%	8%

	Union Model Order					
alpha	6	5	4	3	2	1
0.3	13%	15%	-2%	-13%	-22%	-30%
0.4	**13%**	18%	6%	-2%	-8%	-13%
0.5	13%	20%	11%	5%	1%	-3%
0.6	12%	**20%**	14%	9%	6%	3%
0.7	11%	18%	**14%**	**11%**	9%	7%
0.8	9%	16%	13%	11%	**10%**	**9%**
0.9	6%	11%	10%	9%	8%	7%

Table 2. Improvement in perplexity of the Probabilistic-Union Model with respect to the Global Language Model over the basic 3-gram model.

The best result obtained so far, is an improvement of 20% when a maximum of 6 word models and the order is 5, i.e. sums of the products of pairs (Table 2).The value of alpha is 0.6.

Conclusion

In this paper we have introduced the concept of individual word language models to improve language model performance. Individual word language models permit an accurate capture of the domains in which significant words occur and hence improve the language model performance. We also describe a new method of combining models called the probabilistic union model, which has yet to be fully explored but the first results show good performance. Even though the results are preliminary, they indicate that individual word models combined with the union model offer a promising means of reducing the perplexity.

Weighted Eq. (8)	10%
Exponential Decay Eq. (9)	17%
Union Model 5 words	19%
Union Model 6 words	20%
Union Model 7 words	19%

Table 3. Improvement in perplexity for different combinations of word models.

Acknowledgements

Our thanks go to Dr. Phil Hanna for his collaboration in this research.

References

Good I. J. (1953) "The Population Frequencies of Species and the Estimation of Population Parameters". Biometrika, Vol. 40, pp.237-254.

Jelinek F., Mercer R. L. and Bahl L. R. (1983) "*A Maximum Likelihood Approach to Continuous Speech Recognition*". IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 5, pp.179-190.

Jelinek F. and Mercer R. L. (1984) "*Interpolated estimation of Markov Source Parameters from Sparse Data*". Pattern Recognition in Practice. Gelsema E., Kanal L. eds. Amsterdam: North-Holland Publishing Co.

Katz S. M. (1987) "*Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser*". IEEE Transactions On Acoustic Speech and Signal Processing. Vol. 35(3), pp. 400-401.

Kuhn R. and De Mori R. (1990) "*A Cache-Based Natural Language Model for Speech Recognition*". IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 12 (6), pp. 570-583.

Lau R., Rosenfeld R., Roukos S. (1993). "*Trigger-based Language models: A Maximum entropy approach*". IEEE ICASSP 93 Vol2, pp 45-48, Minneapolis, MN, U.S.A., April.

Ming J., Stewart D., Hanna P. and Smith F. J. (1999) "*A probabilistic Union Model for Partial and temporal Corruption of Speech*". Automatic Speech Recognition and Understanding Workshop. Keystone, Colorado, U. S. A., December.

O'Boyle P., Owens M. and Smith F. J. (1994) "*Average n-gram Model of Natural Language*". Computer Speech and Language. Vol. 8 pp 337-349.