MODELLING WORD-PAIR RELATIONS IN A CATEGORY-BASED LANGUAGE MODEL

T.R. Niesler and P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, England

ABSTRACT

A new technique for modelling word occurrence correlations within a word-category based language model is presented. Empirical observations indicate that the conditional probability of a word given its category, rather than maintaining the constant value normally assumed, exhibits an exponential decay towards a constant as a function of an appropriately defined measure of separation between the correlated words. Consequently a functional dependence of the probability upon this separation is postulated, and methods for determining both the related word pairs as well as the function parameters are developed. Experiments using the LOB, Switchboard and Wall Street Journal corpora indicate that this formulation captures the transient nature of the conditional probability effectively, and leads to reductions in perplexity of between 8 and 22%, where the largest improvements are delivered by correlations of words with themselves (self-triggers), and the reductions increase with the size of the training corpus.

1. INTRODUCTION

Language models based on *n*-grams of word-categories are intrinsically more compact than their word-based counterparts, and are truly able to generalise to unseen word sequences [4]. However their inability to model relationships between particular words limits their performance and prevents them from exploiting large training sets.

The category-based models in question employ variable-length word-category *n*-grams[1] [4], and in this work the categories correspond to part-of-speech classifications as defined in the LOB corpus [1]. Words may belong to multiple categories, and consequently the model bases its probability estimates on a set of possible classifications of the word history into category sequences. Each such classification has an associated probability, and is updated recursively for each successive word in a sentence during operation of the model. An underlying assumption is that the probability of a word depends only upon the category to which it belongs, and therefore its occurrence is equally likely at any point in a corpus at which this category occurs. Factors such as the topic and style of the text cause certain words to occur in groups, however, thereby violating this assumption. This paper presents a technique by means of which this is taken into account by explicit modelling of the transient nature displayed by the probabilities of correlated words as a function of the separation between them.

2. TERMINOLOGY

Let $w(i)$ and $v(i)$ denote the i_{th} word in the corpus and its category respectively, while w_j and v_k denote a particular word and category from the lexicon[2], where $j \in 0 \ldots N_w - 1$ and $k \in 0 \ldots N_v - 1$, (w_j, v_k) is a valid word-category pair from the lexicon, and N_w and N_v are the number of different words and categories respectively.

[1]Referred to as "**varigram**" models hereafter

[2]The possible category assignments for each word in the vocabulary.

Now consider the effect which the occurrence of a **trigger** word (w_{trig}, v_{trig}) has on the subsequent probability of occurrence of a **target** word (w_{targ}, v_{targ}). Define the distance d between this trigger-target pair to be the number of times a word belonging to category v_{targ} is seen after witnessing the trigger and before the first sighting of the target itself, so that $d \in \{0, 1, 2, 3, \ldots, \infty\}$. This definition of distance between the trigger and the target has been employed as a way of minimising syntactic effects on word co-occurrences, notably the phenomenon that certain categories very rarely follow certain others. Syntactic effects should be reduced as much as possible since they are already modelled by the category n-gram component of the language model.

In the following a distinction will be drawn between the case where trigger and target are the same word (termed **self-triggers**) and the case where they differ (referred to as **trigger-target pairs**).

Word-pairs have been combined with word n-gram language models both within a maximum-entropy framework [6] and by linear interpolation [3]. The development here differs by taking explicit account of the distance between word occurrences, and by taking specific advantage of the category-based model.

3. PROBABILISTIC FRAMEWORK

Let the assumption that the probability of a word $w(i)$ depends only upon $v(i)$ be referred to as the *independence assumption*, and for a particular word let this fixed probability be $p(w_j|v_k) = P_w$. Empirical investigation of $p(w_j|v_k)$ as a function of the distance d reveals an exponential decay towards a constant for words between which a recency relationship exists. Figure 1 illustrates this for the case where the trigger is the titular noun "*president*" and the target the proper noun "*congress*". Data is drawn from the WSJ0 corpus (refer to section 5) and the probability P_w is shown.

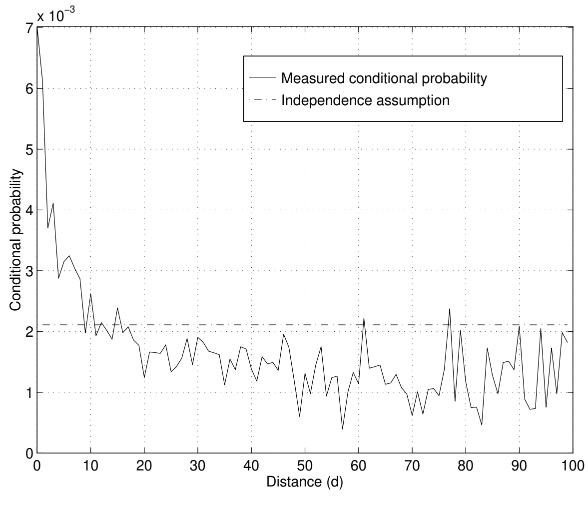

Figure 1: Measured $P(w_j|v_k, d)$

This transient behaviour displayed in this graph is typical, and has motivated the following postulate:

$$p(w_j|v_k, d) = P_b + \gamma \cdot e^{-\rho \cdot d} \tag{1}$$

which is an exponential decay towards a constant probability P_b, γ and ρ defining the strength and rate of decay respectively. Assuming that the triggers occur independently with probability P_a, it follows that the probability mass function $p(d)$ for the target occurrence after sighting the trigger is given by :

$$p(d) = \kappa \cdot \left(\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) \right) \cdot \left(P_b + \gamma \cdot e^{-\rho \cdot d}\right) \tag{2}$$

The normalising constant κ accounts for the probability mass associated with cases in which a trigger follows another trigger before sighting the target.

The empirical estimates of figure 1 have been obtained by binning counts over the graphed distance range. From a storage point of view the potentially extremely large number of word-pair relations however make this approach infeasible for large-scale application, and hence it is not possible to estimate the parameters of equation (1) from a direct fit to the data. The estimation of P_b and then of γ and ρ is treated in the following two sections.

3.1. Estimating P_b

The probability P_b may be estimated from the tail of the distribution, where the transient effect of the exponential term in (1) is assumed to be insignificant. Were the trigger and target to occur independently, their separating distance would have a geometric distribution, and we use its mean μ_g as a rough estimate of the actual mean :

$$\mu_g = \frac{N_{tc} - N_{tt}}{N_{tt}}$$

where

$$N_{tt} = N(w_{trig}, v_{trig}) + N(w_{targ}, v_{targ})$$

and

$$N_{tc} = N(v_{targ}) \quad \text{or} \quad N_{tc} = N(v_{targ}) + N(w_{trig}, v_{trig})$$

when the trigger and target belong to the same or different categories respectively, and where $N(w_{trig}, v_{trig})$, $N(w_{targ}, v_{targ})$ and $N(v_{targ})$ are the number of times the trigger word, the target word, and the target category each appear in the training corpus.

P_b is estimated using counts of all trigger-pair occurrences with distances beyond this mean, i.e.:

$$P_b = \frac{N(w_j, v_k)|_{d>\mu_g}}{N(v_k)|_{d>\mu_g}} \tag{3}$$

where the numerator and denominator on the right hand side are the respective number of times the target word-category pair (w_{targ}, v_{targ}) and the target category v_{targ} have been seen at distances exceeding μ_g.

3.2. Estimating γ and ρ

Expressions allowing the determination of γ and ρ from the mean and mean-square distances separating trigger and target have been derived. Since mean and mean-square calculation requires little storage, this represents a memory-efficient alternative to a direct fit of the conditional probability function (1) to measured binned data. Two successive functional approximations of equation (2) are made, the first using $log(1+x) \approx x$ to find that [5]:

$$p(d) \approx \tilde{p}(d) = \kappa \cdot \left(P_b + \gamma \cdot e^{-\rho \cdot d}\right) \cdot (1 - P_a - P_b)^d \cdot \Phi^{1 - e^{-\rho \cdot d}} \tag{4}$$

where

$$\Phi = e^{\frac{-\gamma}{(1-P_a-P_b)\cdot(1-e^{-\rho})}} \tag{5}$$

This approximation is good while $\frac{\gamma}{1-P_a-P_b} \ll 1$, which is true when $P_a \ll 1$, $P_b \ll 1$ and $\gamma \ll 1$, as may be expected for content words. As a second step, (4) is approximated by:

$$\hat{p}(d) = \kappa \cdot \left[\epsilon_1 \cdot (1 - P_1)^d \cdot P_1 + \epsilon_0 \cdot (1 - P_0)^d \cdot P_0\right] \tag{6}$$

where

$$\kappa \left(\epsilon_1 + \epsilon_0\right) = 1 \tag{7}$$

The functional form of (6) has the following motivations:

- As the superposition of two geometric terms, it retains the overall geometric character exhibited empirically.
- The faster geometric component should model the initially more rapid decay of the observed distribution (which is in turn due to the higher conditional probability at small d).
- The slower geometric component should model the tail of the observed distribution.
- Closed form expressions for the mean and mean-square exist.

Now by imposing the constraint

$$\lim_{d \to \infty} \tilde{p}(d) = \lim_{d \to \infty} \hat{p}(d)$$

we find from (4) and (6) that :

$$P_0 = P_a + P_b \quad \text{and} \quad \epsilon_0 = \frac{P_b \cdot \Phi}{P_a + P_b} \tag{8}$$

Furthermore, requiring $\tilde{p}(0) = \hat{p}(0)$ we find from (4) and (6) that:

$$P_1 = \frac{P_b + \gamma - \epsilon_0 \cdot P_0}{\epsilon_1} \tag{9}$$

and finally, requiring $\tilde{p}(1) = \hat{p}(1)$ we find that:

$$P_1 = 1 - (1 - P_0) \cdot e^{\left[\frac{\rho \cdot [(P_b + \gamma) \cdot \ln(\Phi) - \gamma]}{P_b + \gamma - P_b \cdot \Phi}\right]} \tag{10}$$

The values of P_a and P_b are known, and in order to solve for ϵ_0, ϵ_1 γ and ρ from the above equations, the mean $\overline{d}$ and mean-square $\overline{d^2}$ distance of the distribution are employed. However, when estimated from data these quantities have been found to be particularly sensitive to outliers, in particular trigger and target words separated by large quantities of text and occurring in unrelated parts of the training corpus. Robustness is significantly improved by measuring the mean and mean-square within only a predetermined range of distances, $d \in \{0 \cdots N-1\}$. Refer to these as **truncated** mean and mean-square estimates, and denote them by $\overline{d}(N)$ and $\overline{d^2}(N)$ respectively. Since equation (6) is the superposition of two geometric terms, we may express the truncated mean and mean-square as a linear combination of the corresponding terms for truncated geometric distributions (refer to the appendix):

$$\overline{d}(N) = \kappa \cdot [\epsilon_1 \cdot \mu(P_1, N) + \epsilon_0 \cdot \mu(P_0, N)] \tag{11}$$

and

$$\overline{d^2}(N) = \kappa \cdot [\epsilon_1 \cdot v(P_1, N) + \epsilon_0 \cdot v(P_0, N)] \tag{12}$$

Equations (5), (7), (8), (9), (10), (11) and (12) relate P_a, P_b, γ and ρ to ϵ_0, ϵ_1, κ, $\overline{d}(N)$ and $\overline{d^2}(N)$, and may be used to determine γ and ρ given the measured values P_a, P_b, $\overline{d}(N)$ and $\overline{d^2}(N)$. However, since it is not possible to do this analytically, these values are determined numerically.

3.3. Typical estimates

Figure 2 repeats the curves of figure 1, and adds the plot of equation (1) using the parameters P_b, γ and ρ determined from the results of sections 3.1 and 3.2. The estimated conditional probability reflects the true nature of the data much more closely than the constant value P_w used under the independence assumption.

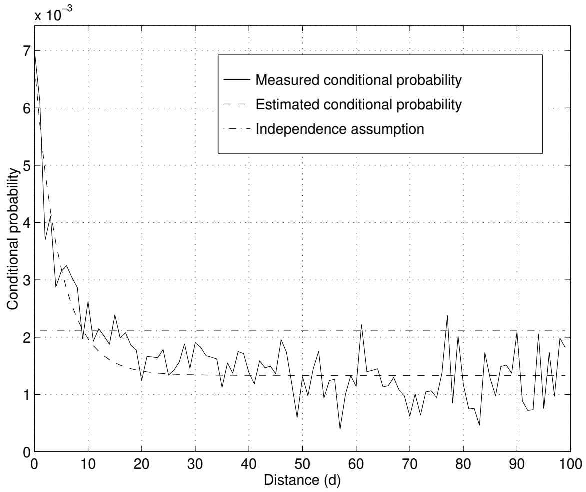

Figure 2: Measured and estimated $P(w_j|v_k, d)$

4. DETERMINING TRIGGER-TARGET PAIRS

While the number of possible self-triggers is bounded by the vocabulary size, the number of potential trigger-target pairs equals the the square of this number, and it is not possible to consider these relations exhaustively except for very small vocabularies. In order to identify suitable candidates in a feasible manner, an approach employing two passes through the training corpus has been developed [5]. The first pass identifies promising candidates by dynamically adding new contenders to a tentative list, and consequently updating their mean distances. At each update a t-test is performed, determining to a desired level of confidence whether the measured mean is lower than would be expected were the candidates to occur independently. These tests allow unpromising candidates to be pruned continually from the tentative list, and adjustment of the relevant pruning thresholds and confidence levels allows the rate of new acquisitions to be balanced against the rate at which unpromising relations are discarded, thereby avoiding the explosion in the number of considered word-pairs that would otherwise arise. During the second pass, each surviving pair is reconsidered by calculating its mean and mean-square distances over the entire training corpus, and executing a second t-test at completion to determine whether a the candidate relation should ultimately be retained or discarded. Finally, the values of the parameters P_b, γ and ρ are calculated for each remaining trigger-target pair.

The following table lists some examples of typical targets and their triggers as found by the described technique when applied to the LOB corpus. The bracketed designations are the grammatical categories of the words in question[3]. It is appealing to find such intuitive relationships in meaning between word pairs gathered according to purely statistical criteria.

Target	Triggers
discharged (JJ)	prison (NN), period (NN), supervision (NN), need (NN), prisoner (NN), voluntary (JJ), assistance (NN)
advocate (NN)	truth (NN), box (NN), defence (NN), honest (JJ), face (VB), case (NN), witness (NN), evidence (NN)
Cambridge (NP)	university (NN), educational (JJ), affected (VBN), Oxbridge (NP), tomorrow (NR), universities (NNS)
worked (VBN)	demand (NN), changes (NNS), cost (NN), strength (NN)

Table 1: Triggers and targets collected from the LOB corpus

5. PERPLEXITY RESULTS

The benefit of characterising trigger pairs as described in the previous sections was gauged by comparing the performance of a category-based language model employing the independence assumption with another using equation (1) but identical in all other respects. Experiments were carried out on the LOB, Switchbaord (SWBD) and Wall-street Journal (WSJ0) corpora, category-based language models having been constructed for each using a pruning threshold of 5e-6 during construction of the variable-length category *n*-grams [4]. The following table gives a brief description of each corpus, where N_{trn} and N_{tst} refer to the number of words in the training- and test-sets respectively.

Corpus	Source	N_{trn}	N_{tst}
LOB	Various (e.g. news, fiction etc.)	1.0 M	56K
SWBD	Telephone conversations	1.9 M	10K
WSJ0	Wall Street Journal (87-89)	37 M	92K

Table 2: Summary of the LOB, Switchboard and WSJ0 corpora

The details of the language models constructed for each of these corpora are summarised in table 3. Information for a standard trigram language model using the Katz backoff and Good-Turing discounting [2] is given in order to establish a baseline. The symbols N_v, N_{wng} and N_{cng} refer to the number of words

[3] JJ = adjective, NN = common noun, NNS = plural common noun, NP = proper noun, NR = singular adverbial noun, VB = verb base form, VBN = past participle.

in the vocabulary, the number of n-grams in the trigram, and the number of n-grams in the category language model respectively, while N_{st} N_{tt} are the number of self-triggers and trigger-target pairs for which parameters were estimated.

Corpus	N_v	N_{wng}	N_{cng}	N_{st}	N_{tt}
LOB	41K	1.14M	44,380	14,295	4,427
SWBD	23K	1.18M	54,547	8,615	4,262
WSJ0	65K	13.05M	174,261	56,928	133,608

Table 3: Language models for LOB, Switchboard and WSJ0.

Table 4 shows the trigram (TG) and varigram model perplexities, where the abbreviations "VG","VGST", "VGTT" and "VGSTTT" refer to the varigram by itself, with self-triggers, with trigger-target pairs and with both self-triggers and trigger-target pairs respectively.

Corpus	**TG**	**VG**	**VGST**	**VGTT**	**VGSTTT**
LOB	413.1	458.3	412.2	458.09	412.2
SWBD	96.6	145.3	134.1	143.70	133.4
WSJ0	132.2	469.4	381.0	441.40	366.6

Table 4: Perplexities for the LOB, Switchboard and WSJ0 corpora

6. DISCUSSION

The largest perplexity improvement is obtained for the WSJ0 corpus, which also has the largest number of self-trigger and trigger-target pairs. This stems from the much greater corpus size and consequent lower sparseness. For LOB and SWBD, on the other hand, many words occur too infrequently to make estimation of the conditional probability parameters possible, thus leading to a reduced number of trigger pairs.

For all three corpora, the addition of self-triggers has a more significant impact on the perplexity than does the introduction of trigger-target pairs. Self-triggers seem more reliable since the target, being its own trigger, is actually seen before being predicted to occur again. Trigger-target pairs, on the other hand, predict words that have either not yet been seen at all or have occurred in the distant past. Since such correlations are heavily dependent upon the topic of the passage, the effectiveness of a trigger-target association depends on how much the topics associated with a trigger coincide between the training- and test-set. For the LOB corpus, which is very diverse in the material it contains, there is a significant mismatch in this regard, leading to the observed very small impact of self-triggers on performance, while for the WSJ corpus the mismatch is smaller, leading to greater success.

The addition of the self-triggers increases the number of parameters in the model by $2 \cdot N_{st}$ (storage of γ and ρ). This increase is mild, and offers a favourable size versus performance tradeoff. For instance, the varigram with self-triggers for LOB uses 58,675 parameters and achieves a lower perplexity than the trigram with 1.1 million parameters. Furthermore, the effectiveness of both types of word-pair modelling improves with corpus size, and since the parameter determination and final implementation of the model has low memory requirements, the technique is suitable for use with large training sets. This complements the category-based model, whose performance does not improve in the same way.

Finally, inspection of the values of ρ assigned to trigger-target pairs, as well as cases in which a trigger successfully predicts a target show that correlations well beyond the range of conventional n-gram models are captured, and therefore the proposed technique is indeed able to model long-range dependencies.

7. CONCLUSION

A new technique for modelling the empirically observed transient character of the occurrence probability between related words in a body of text has been introduced. Procedures both for the identification of such word pairs as well as for the estimation of the three parameters required by the parametric model have been developed. Experiments demonstrate that meaningful relations are indeed identified, and that the transient behaviour (which often spans many words) is successfully captured by the proposed model. Perplexity reductions of between 8 and 22% were achieved, where the greatest improvement seen was for the largest and least-sparse corpus, and the most significant impact on performance was displayed by word correlations with themselves (self-triggers). The modelling technique is able to reduce the performance limit displayed by category-based models for large corpora, thereby improving their good performance versus size tradeoff.

8. APPENDIX

Assuming a limited range $d \in \{0 \cdots N-1\}$ with constant probability of success P_x at each d, the probability density function over this range is :

$$P(d) = \frac{(1-P_x)^d \cdot P_x}{1-(1-P_x)^N}$$

for which the moment generating function is:

$$M_d(t) = \frac{P_x \cdot (1-(e^t \cdot (1-P_x))^N)}{(1-(1-P_x)^N \cdot (1-e^t \cdot (1-P_x))}$$

and from which straightforward algebra leads us to expressions for the mean $\mu(P_x, N)$:

$$\mu(P_x, N) = \frac{(1-P_x)\left[1 + [(N-1)(1-P_x) - N] \cdot (1-P_x)^{(N-1)}\right]}{P_x \cdot (1-(1-P_x)^N)}$$

A similar but cumbersome development leads to an expression for the mean-square $v(P_x, N)$ [5].

9. REFERENCES

[1] Johansson, S; Atwell, R; Garside, R; Leech, G. *The tagged LOB corpus user's manual*; Norwegian Computing Centre for the Humanities, Bergen, 1986.

[2] Katz, S. *Estimation of probabilities from sparse data for the language model component of a speech recogniser*; IEEE Trans. ASSP, vol. 35, no. 3, March 87, pp. 400-1.

[3] Ney, H; Essen, U; Kneser, R; *On structuring probabilistic dependencies in stochastic language modelling*, Computer Speech and Language, vol. 8, pp. 1-38, 1994.

[4] Niesler, T.R; Woodland, P.C. *A variable-length category-based n-gram language model*, ICASSP 96, vol. 1, pp. 164-7.

[5] Niesler, T.R; Woodland, P.C. *Word-pair relations for category-based language models*, Tech. report CUED/F-INFENG/TR.281, Dept. Engineering, University of Cambridge, U.K., 1997.

[6] Rosenfeld, R; *Adaptive statistical language modelling : a maximum entropy approach*, PhD thesis, School of Computer Science, CMU, April 1994.