Word-pair relations for category-based language models

T.R. Niesler and P.C. Woodland

CUED/F-INFENG/TR.281

February 1997

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

trn@eng.cam.ac.uk
pcw@eng.cam.ac.uk

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.

Contents

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 report 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. Let $N_c(w_j)$ and $N_c(v_k)$ denote the total number of times w_j and v_k respectively occur in the corpus.

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}). Refer to the sequence consisting of all trigger occurrences as well as all words belonging to the target category as the **trigger-target stream**, and denote it by $S(w_{trig}, v_{targ})$. Let the total number of words in the stream $S(w_{trig}, v_{targ})$ be N_s, and the number of occurrences of the trigger and target words respectively be $N_s(w_{trig})$ and $N_s(w_{targ})$. It will be assumed henceforth that the stream has been taken from the training corpus, and under this condition we note that $N_s(w_{trig}) = N_c(w_{trig})$ and $N_s(w_{targ}) = N_c(w_{targ})$. Furthermore:

$$N_s = \begin{cases} N_c(v_{targ}) & \text{if} \quad v_{trig} = v_{targ} \\ N_c(v_{targ}) + N_c(w_{trig}) & \text{otherwise} \end{cases} \tag{1}$$

Now the overall probability of occurrence respectively of the trigger and target within $S(w_{trig}, v_{targ})$ are given by:

$$p_s(w_{trig}) = \frac{N_s(w_{trig})}{N_s} \tag{2}$$

and

$$p_s(w_{targ}) = \frac{N_s(w_{targ})}{N_s} \tag{3}$$

but since

$$p\left(w_{targ} | v_{targ}\right) = \frac{N_c(w_{targ})}{N_c(v_{targ})}$$

we see that the stream and category-conditional word probabilities are related by:

$$p\left(w_{targ} | v_{targ}\right) = K_s \cdot p_s(w_{targ}) \tag{4}$$

with

$$K_s = \frac{N_s}{N_c(v_{targ})} \tag{5}$$

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

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

Define the distance d between a trigger-target pair to be the number of times that a word belonging to category v_{targ} is seen after witnessing the trigger and before the first sighting of the target itself, so that $d \in \{0, 1, 2, 3, \ldots, \infty\}$ is the separating distance in the trigger-target stream. This definition of distance 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 [5] 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*. Empirical investigation of the category-conditional probability $p(w_j|v_k)$ 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*". [3]

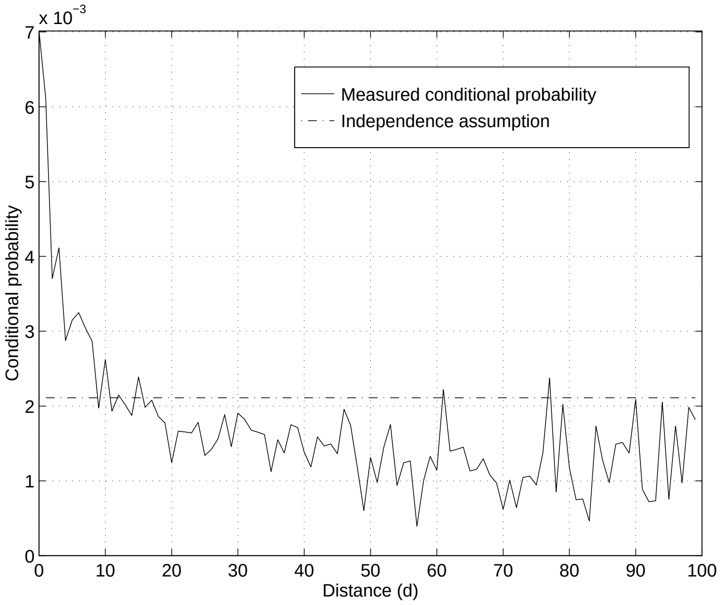

Figure 1: Measured $P(w_j|v_k, d)$

This transient behaviour displayed in this graph is typical, and has motivated the following postulated form of the category-conditional probability:

$$p(w_{targ}|v_{targ}, d) = P_v + \gamma_v \cdot e^{-\rho \cdot d} \tag{6}$$

which is an exponential decay towards a constant probability P_v in which γ_v and ρ define the strength and rate of decay respectively. The stream-probability is found by scaling according to equation (4) :

$$p_s(w_{targ}, d) = P_b + \gamma \cdot e^{-\rho \cdot d} \tag{7}$$

with $P_v = K_s \cdot P_b$ and $\gamma_v = K_s \cdot \gamma$. Assuming that the triggers occur independently in the stream with probability $P_a = p_s(w_{trig})$, it follows that the probability mass function $p(d)$ for the target occurrence after sighting the trigger is given by :

$$p_s(d) = \kappa \cdot \left(\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) \right) \cdot \left(P_b + \gamma \cdot e^{-\rho \cdot d}\right) \tag{8}$$

[3] Data is drawn from the WSJ0 corpus (refer to section 5).

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. However, from a storage point of view the potentially extremely large number of word-pair relations make this approach infeasible for large-scale application, and hence it is not possible to obtain the parameters of equation (7) 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 (7) 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 :

$$\mu_g = \frac{N_s - N_s(w_{trig}) - N_s(w_{targ})}{N_s(w_{trig}) + N_s(w_{targ})}$$

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

$$P_b = \frac{N_s(w_{trig})|_{d>\mu_g}}{N_s(v_{trig})|_{d>\mu_g}} \qquad (9)$$

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

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 (6) to measured binned data. In order to obtain closed-form expressions for the mean and mean-square, the exact distribution was approximated by one of the form:

$$\hat{p}(d) = \kappa \cdot \left[\epsilon_1 \cdot (1 - P_1)^d \cdot P_1 + \epsilon_0 \cdot (1 - P_0)^d \cdot P_0\right] \qquad (10)$$

where

$$\kappa\,(\epsilon_1 + \epsilon_0) = 1 \qquad (11)$$

The following equations relate the parameters of the exact and approximate distributions, details of their derivation are shown in appendix A.

$$\Phi = e^{\frac{-\gamma}{(1-P_a-P_b)\cdot(1-e^{-\rho})}} \qquad (12)$$

$$P_0 = P_a + P_b \qquad (13)$$

$$\epsilon_0 = \frac{P_b \cdot \Phi}{P_a + P_b} \qquad (14)$$

$$P_1 = \frac{P_b + \gamma - \epsilon_0 \cdot P_0}{\epsilon_1} \qquad (15)$$

$$P_1 = 1 - (1 - P_0) \cdot e^{\left[\frac{\rho\cdot[(P_b+\gamma)\cdot\ln(\Phi)-\gamma]}{P_b+\gamma-P_b\cdot\Phi}\right]} \qquad (16)$$

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 measured 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, $d \in \{0 \cdots N_T - 1\}$. Expressions for the mean-and mean-square expected for such **truncated** measurements under the independence assumption have been derived in appendix B. Since equation (10) is the superposition of two geometric terms, we may employ the results of this appendix and express the truncated mean $\overline{d}(N_T)$ and mean-square $\overline{d^2}(N_T)$ as a linear combination of the corresponding terms for truncated geometric distributions:

$$\overline{d}(N_T) = \kappa \cdot [\epsilon_1 \cdot \mu(P_1, N_T) + \epsilon_0 \cdot \mu(P_0, N_T)] \tag{17}$$

and

$$\overline{d^2}(N_T) = \kappa \cdot [\epsilon_1 \cdot \upsilon(P_1, N_T) + \epsilon_0 \cdot \upsilon(P_0, N_T)] \tag{18}$$

Equations (11), (12), (13), (14), (15), (16), (17) and (18) relate P_a, P_b, γ and ρ to ϵ_0, ϵ_1, κ, $\overline{d}(N_T)$ and $\overline{d^2}(N_T)$, and may be used to determine γ and ρ given the measured values P_a, P_b, $\overline{d}(N_T)$ and $\overline{d^2}(N_T)$. This is accomplished numerically by means of nested bisection searches, since it is not possible to find a closed-form solution to this system of equations.

3.3. Typical estimates

Figure 2 repeats the curves of figure 1, and adds the plot of equation (6) 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 used under the independence assumption.

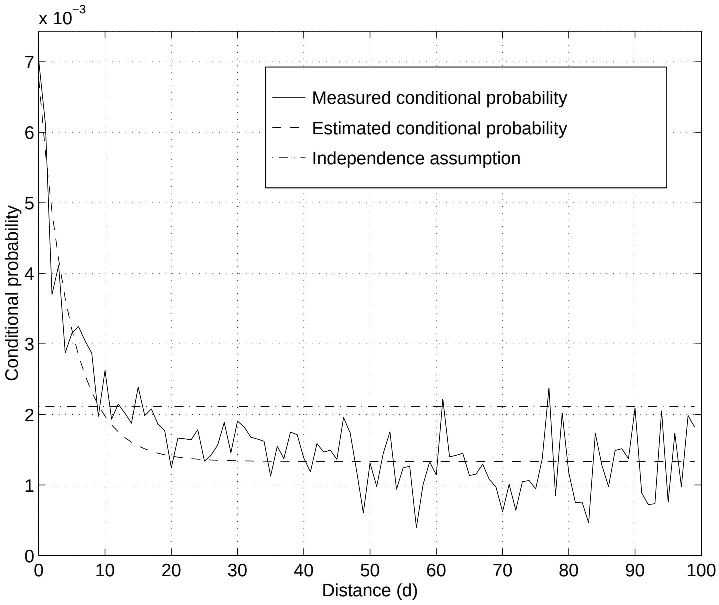

Figure 2: Measured and estimated $P(w_j|v_k, d)$

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.

4.1. First-pass

This first stage of processing provides each word in the lexicon that is to be considered as a potential target with a **tentative-list** and a **fixed-list** for trigger candidates. The latter holds words for which a reliable correlation with the target has been established, while the former lists those for which no decision could yet been reached regarding the presence or absence of such relationship, and includes storage for the cumulative totals required for distance-mean and -variance calculation. At a given point within the corpus, let the trigger-target pair have been seen N_m times at separations $\{d_0, d_1, \ldots, d_{N_m-1}\}$, each falling within the chosen truncation interval N_T, i.e.

$$d_k < N_T \quad \forall \;\; k \in \{0, 1, \ldots, N_m - 1\}$$

so that the measured truncated mean is given by:

$$\mu_{meas} = \frac{1}{N_m} \cdot \sum_{k=0}^{N_m-1} d_k$$

and the measured variance by:

$$\sigma^2_{meas} = \frac{1}{N_m - 1} \cdot \sum_{k=0}^{N_m-1} (d_k - \mu_{meas})^2$$

Finally, assuming that the trigger and target are not correlated but occur independently, we find from appendix B that the expected value of the truncated mean is given by:

$$\mu_{exp} = \frac{(1 - P_{tt}) \cdot \left[1 + (1 - P_{tt})^{N_T - 1} \left[(N_T - 1) \cdot (1 - P_{tt}) - N_T\right]\right]}{P_{tt} \cdot \left(1{-}(1{-}P_{tt})^{N_T}\right)}$$

where $P_{tt} = P_s(w_{trig}) + P_s(w_{targ})$ is the probability of occurrence of either trigger or target calculated uniformly over the stream. Truncated mean- and variance-measurements are once again used to reduce sensitivity to outliers. The statistics for members of the tentative list are updated on sighting the associated target during the sequential processing of the corpus, and after each such update two hypothesis tests, termed the **fix**- and **kill**-tests respectively, are used to decide upon the strength of the correlation. Since both the mean and the variance are measured from the data, and since empirical observations of the samples d_k have shown them to posses approximately normal distributions, the *t*-test has been employed for this purpose as follows:

- **The kill-test.**

  When the measured mean μ_{meas} is found to exceed the expected mean μ_{exp} by a specified margin δ_{kill} and to a confidence of $(1 - \alpha_{kill})100\%$, the kill-test succeeds and the trigger candidate is deleted from the tentative list. In particular, let:

  $$\mu_{kill} = \mu_{exp} \cdot (1 + \delta_{kill})$$

  The critical value μ_t of the mean is

  $$\mu_t = \mu_{meas} - t(\alpha_{kill}, N_m - 1) \cdot \frac{\sigma_{meas}}{\sqrt{N_m}}$$

  where $t(\alpha_{kill}, N_m - 1)$ is the value obtained from the t-distribution for confidence $(1 - \alpha_{kill})100\%$ and $N_m - 1$ degrees of freedom. The kill-test succeeds when $\mu_t > \mu_{kill}$, and the following figure illustrates these conditions.

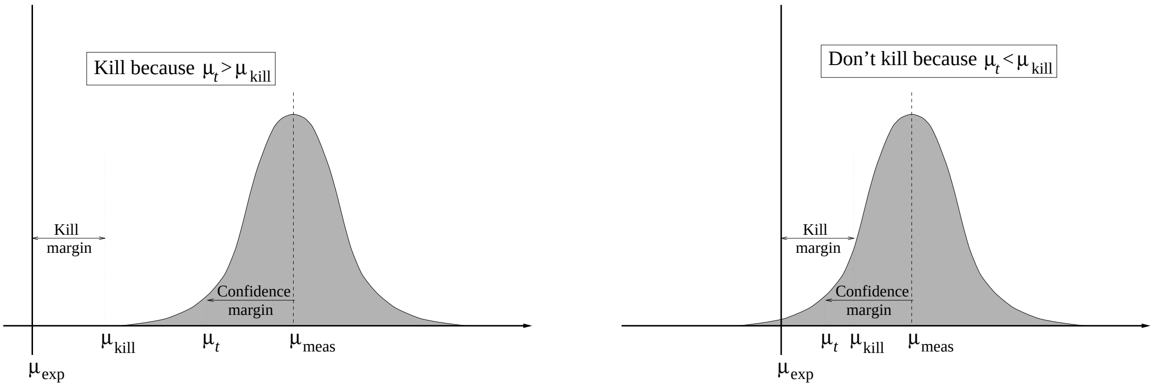

Figure 3: Kill test.

- **The fix-test**

  When the expected mean μ_{exp} is found to exceed the measured mean μ_{meas} by a specified margin δ_{fix} and to a confidence of $(1 - \alpha_{fix})100\%$, the fix-test succeeds and the trigger candidate is moved from the tentative- to the fixed-list.

  $$\mu_{fix} = \mu_{exp} \cdot (1 - \delta_{fix})$$

  The critical value μ_t of the mean is

  $$\mu_t = \mu_{meas} + t(\alpha_{fix}, N_m - 1) \cdot \frac{\sigma_{meas}}{\sqrt{N_m}}$$

  where $t(\alpha_{fix}, N_m - 1)$ is the value obtained from the t-distribution for confidence $(1 - \alpha_{fix})100\%$ and $N_m - 1$ degrees of freedom. The kill-test succeeds when $\mu_t < \mu_{fix}$, and the following figure illustrates these conditions.

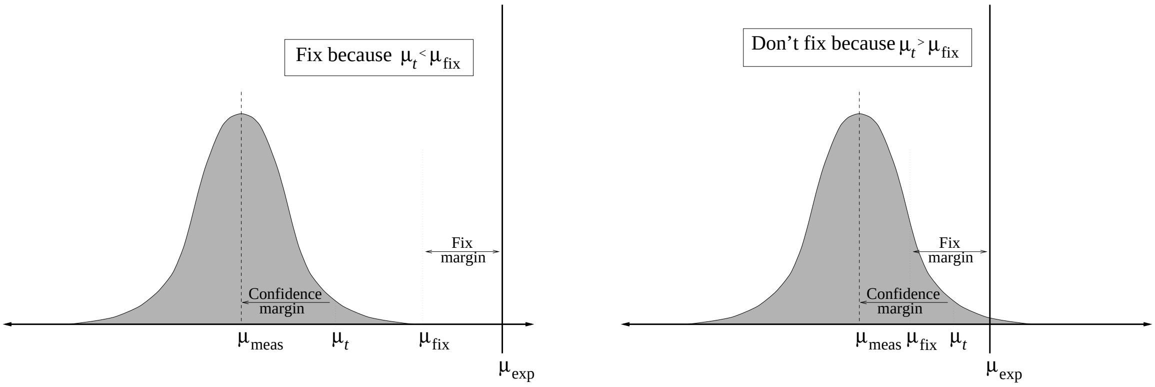

Figure 4: Fix test.

This mechanism allows unpromising candidates to be pruned continually from the tentative list, thereby counteracting the explosion in the number of considered word-pairs that would otherwise arise. The following figure illustrates this by showing the growth in the number of tentative triggers when the kill-test is disabled and when it is active.

Once a correlation has been established (and the fix-test succeeds), the trigger is moved from the tentative to the fixed list, and no further statistics need be gathered during this pass. Separate tentative- and fixed-lists are maintained since the latter can be made much more compact, not including any storage for mean or variance, and this is extremely important in view of the generally very large number of trigger-target candidates considered during the first pass.

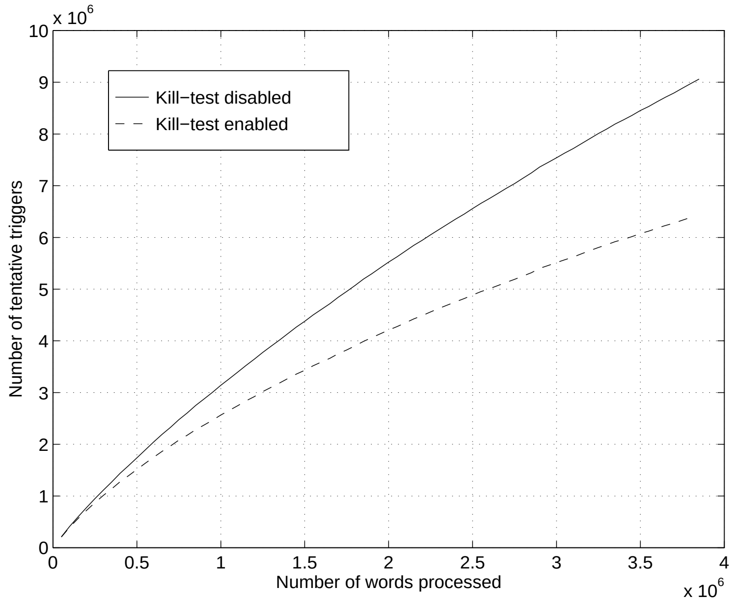

Figure 5: The effect of the kill-test on the total number of tentative triggers during processing

Initially all tentative- and fixed-lists are empty. Furthermore, a record of the H most recent unique words is maintained during processing. As each word in the corpus is processed sequentially, each member of this history is hypothesised as a possible trigger word, and for each of these candidates the following processing steps are performed:

- If this is not the first sighting of the target since the trigger, then END.
- If the trigger is already in the fixed-list, then END.
- If the trigger is not yet in the tentative-list, then:
  - Add it to the tentative-list.
  - Initialise the cumulative sum-of-distance and sum-of-squared-distance fields with the first measurement.
  - END.
- If the trigger is already in the tentative-list, then:
  - Update the sum-of-distance and sum-of-squared-distance fields with the new measurement.
  - Calculate the distance mean and variance.
  - Calculate the expected mean under the independence assumption.
  - Perform FIX and KILL *t*-tests :
    - ∗ Case A: The measured mean exceeds the independence-mean by a desired margin and to a desired level of confidence: conclude that there is no correlation.
      ⇒ **KILL**: remove the trigger from the tentative list.
    - ∗ Case B: The measured mean is lower than the independence-mean by a desired margin and to a desired level of confidence: conclude that there is a correlation.
      ⇒ **FIX**: remove the trigger from the tentative list and add it to the fixed-list.
    - ∗ Case C: neither of the above: conclude that there is as yet insufficient data to reach a conclusion, do nothing.
  - END.

The result of the first-pass is the set of all trigger-target relations in the fixed-lists on completion, those remaining in the tentative-lists being discarded.

4.2. Second-pass

Since the fix-test in the first-pass uses means and variances usually gathered only over a small portion of the training set, many of the detected correlations are local anomalies, and do not generalise to the corpus as a whole. Consequently the second-pass recalculates the means and variances for all candidates over the entire training set, and applies the fix-test again to each. Those failing are discarded, and those succeeding are retained and their measured means and mean-squares used to calculate the parameters P_b, γ and ρ of the postulated conditional probability function.

4.3. Regulating memory usage

Selection of the fix-test margin and confidence level allows the rate of transferrals from the tentative- to the fixed-list to be regulated, and thus gives control over the growth and the final number of fixed-triggers. The kill-test parameters, on the other hand, affects the rate of deletions from the tentative-list. Finally, the length of the history H determines the rate of addition of new tentative trigger candidates for the current target.

The size of the tentative-list is of prime practical importance during first-pass processing, since each entry requires significantly more storage than in the fixed-list. Despite the control over its size afforded by the choice of H and the kill-test parameters, it may still be difficult to limit the number of trigger-target pairs considered to practical levels. The following two refinements are employed as additional measures in this regard.

1. **Exclusion lists**

   Semantic correlations may be expected chiefly among content words, and since the grammatical functions of words are known, it is possible to exclude non-content words from consideration as triggers or targets during processing. Practically this is achieved by means of an **exclusion-list** containing all grammatical categories that should be disregarded in this way.

2. **Background culling**

   Observations during first-pass processing have shown a large number of tentative-list members to be predominantly idle. These are infrequent trigger-target candidates which, once added to the list, are neither fixed nor killed due to an insufficient number of measurements and long periods between updates. In order to reduce the number of these cases, a process termed **background culling** has been introduced. During processing the distance to the last update is monitored for members of the tentative list, and the decision boundary for the kill-threshold is moved gradually towards that of the fix-threshold as this time increases, thereby relaxing the kill-test and ultimately forcing a fix/kill decision. The rate at which this occurs is normalised with respect to the frequency of the trigger, so that a single global parameter may be used to set the severeness of pruning.

   Background culling is an approximation necessitated by practical considerations, and will generally introduce errors by eliminating valid but infrequent trigger-target relations. However it allows the size of the tentative list to be regulated to practical levels for large corpora and vocabularies, as illustrated in the following figure.

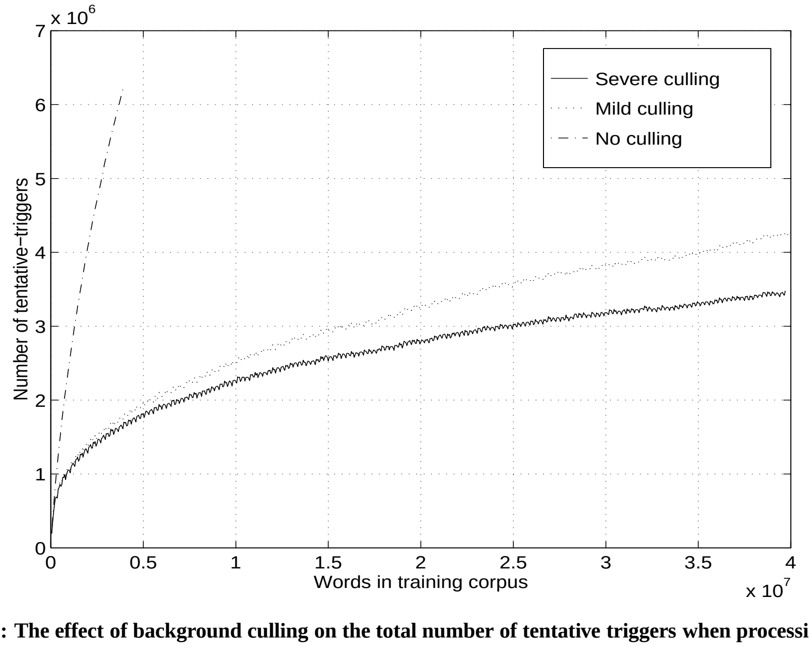

Figure 6: The effect of background culling on the total number of tentative triggers when processing WSJ0

4.4. Example pairs

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[4]. 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)
dry (JJ)	currants (NNS), suet (NN), teasp. (NNU), wines (NNS), raisins (NNS)
judicial (JJ)	legal (JJ), binding (JJ), rules (NNS)
semiindustrialised (JJ)	world (NN), substantial (JJ), fall (NN), trade (NN), demand (NN), supply (NN)
cinema (NN)	directors (NNS), viewing (NN), film (NN), festival (NN), tastes (NNS)
current (NN)	inductance (NN), constant (NN), capacitor (NN), voltage (NN),
drowning (NN)	respiration (NN), failure (NN), inhaled (VBN), body (NN), spasm (NN), sea (NN), salt (JJ), minutes (NNS), lethal (JJ), water (NN), resuscitation (NN), recovery (NN), asphyxia (NN), survival (NN)
rotor (NN)	r.p.m. (NNU), values (NNS), blade (NN), pitching (NN), speed (NN), flapping (NN), wind (NN), tunnel (NN), helicopter (NN), body (NN), rotors (NNS)
syntax (NN)	language (NN), categories (NNS), formal (JJ), syntactic (JJ), grammatical (JJ), morphology (NN)
transfusion (NN)	bleeding (NN), blood (NN), cells (NNS), ml (NNU), reaction (NN), haematoma (NN), transfusions (NNS), patient (NN), group (NN), treated (VBN)
increases (NNS)	salary (NN), agreement (NN), salaries (NNS)
raisins (NNS)	list (NN), lemon (NN), milk (NN), salt (NN), teasp. (NNU), brandy (NN), mixed (JJ), currants (NNS), suet (NN), sugar (NN), nutmeg (NN), oz (NNU), sultanas (NNS), eggs (NNS), peel (NN), apples (NNS)
Orpheus (NP)	Heurodis (NP), Orfeo (NP), tale (NN), fairy (NN), Eurydice (NP)
Verwoerd (NP)	policy (NN), Africa (NP), South (NP)

Table 1: Triggers and targets collected from the LOB corpus

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

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 (6) 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 topics (e.g. news, fiction, science etc.)	1,003,839	55,933
SWBD	Spontaneous telephone conversations	1,860,178	10,179
WSJ0	Reportage from 1987-9 (inclusive) issues of the Wall Street Journal	37,346,080	92,024

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 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	41,097	1,142,457	44,380	14,295	4,427
SWBD	22,643	1,183,880	54,547	8,615	4,262
WSJ0	65,000	13,047,678	174,261	56,928	133,608

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

Table 4 shows the trigram (TG) and varigram model perplexities, where the abbreviations "VG","VG+ST", "VG+TT" and "VG+ST+TT" 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**	**VG+ST**		**VG+TT**		**VG+ST+TT**	
			pp	%	pp	%	pp	%
LOB	413.14	458.34	412.23	10.1	458.09	0.1	412.17	10.1
SWBD	96.57	145.28	134.09	7.7	143.70	1.1	133.41	8.2
WSJ0	132.21	469.40	381.01	18.80	441.40	6.0	366.57	21.9

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

9. APPENDIX A

By virtue of the category-conditional probability function chosen to model occurrence correlations between trigger and target words, the distribution function is found to be :

$$p(d) = \kappa \cdot \left(\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) \right) \cdot \left(P_b + \gamma \cdot e^{-\rho \cdot d}\right) \tag{19}$$

where

- d is the distance separating trigger and target.
- P_a is the probability of occurrence of the trigger.
- P_b, γ and ρ are the parameters describing the transient occurrence probability of the target with respect to a trigger sighting.
- κ is a normalising constant.

This appendix describes the swo-stage algebraic approximation of (19) by a distribution of the form

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

9.1. First approximation

The objective here is to eliminate the product operator from equation (19), since its presence makes algebraic manipulation difficult. Consider the first term on the right-hand side of (19) :

$$\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) = (1 - P_a - P_b)^d \cdot \left(\prod_{i=0}^{d-1} \left(1 - \zeta \cdot e^{-\rho \cdot i}\right) \right) \tag{20}$$

where

$$\zeta = \frac{\gamma}{1 - P_a - P_b} \tag{21}$$

Take the logarithm and apply the first-order Taylor approximation $log(1 + x) \approx x$ to find:

$$\begin{aligned} log \left(\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) \right) &\approx d \cdot log(1 - P_a - P_b) - \sum_{i=0}^{d-1} \zeta \cdot e^{-\rho \cdot i} \\ &= d \cdot log(1 - P_a - P_b) - \frac{\zeta \left(1 - e^{-\rho \cdot d}\right)}{1 - e^{-\rho}} \end{aligned}$$

Now, taking the inverse logarithm and resubstitute (21) into the above we obtain

$$\prod_{i=0}^{d-1} \left(1 - P_a - P_b - \gamma \cdot e^{-\rho \cdot i}\right) \approx (1 - P_a - P_b)^d \cdot \left[e^{-\frac{\gamma \cdot \left(1 - e^{-\rho \cdot d}\right)}{(1 - P_a - P_b) \cdot (1 - e^{-\rho})}} \right] \tag{22}$$

For clarity now define:

$$\Phi = e^{-\frac{\gamma}{(1 - P_a - P_b) \cdot (1 - e^{-\rho})}} \tag{23}$$

and it follows from (19), (22) and (23) that

$$p(d) \approx \tilde{p}(d) = \kappa \cdot \left(P_b + \gamma \cdot e^{-\rho \cdot d}\right) \cdot (1 - P_a - P_b)^d \cdot \Phi^{1-e^{-\rho \cdot d}} \tag{24}$$

This approximation is good when $\frac{\gamma}{1-Pa-P_b} \ll 1$, which is true when $P_a \ll 1$, $P_b \ll 1$ and $\gamma \ll 1$, as may be expected for content words.

9.2. Second approximation

Since we would ultimately like to find closed-form expressions for the approximate mean and mean-square of the probability distribution (19), and this is not yet possible using (24), we will further approximate the latter by:

$$\hat{p}(d) = \kappa \cdot \left[\epsilon_1 \cdot (1 - P_1)^d \cdot P_1 + \epsilon_0 \cdot (1 - P_0)^d \cdot P_0\right] \tag{25}$$

where

$$\kappa\,(\epsilon_1 + \epsilon_0) = 1 \tag{26}$$

The functional form of (25) has the following motivations:

- As the superposition of two geometric terms, it retains the overall geometric character exhibited empirically by the distribution.
- 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.

Note firstly that

$$\sum_{d=0}^{\infty} \hat{p}(d) = 1$$

and that for $0 \leq P_0 \leq 1$ and $0 \leq P_1 \leq 1$:

$$0 \leq \hat{p}(d) \leq 1 \qquad \forall \;\; d \in (0, 1, 2, \ldots, \infty)$$

so that it represents a valid probability mass function. In order to solve for the parameters of (25) in terms of the parameters of (24), we impose the following three constraints :

1. Equality in the limit as $d \to \infty$: From (24) we find that

   $$\lim_{d\to\infty} \tilde{p}(d) = \kappa \cdot P_b \cdot (1 - P_a - P_b)^d \cdot \Phi$$

   and from (25), assuming $P_0 < P_1$

   $$\lim_{d\to\infty} \hat{p}(d) = \epsilon_0 \cdot (1 - P_0)^d \cdot P_0$$

   and so by requiring

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

   we may choose

   $$P_0 = P_a + P_b \tag{27}$$

   and

   $$\epsilon_0 = \frac{P_b \cdot \Phi}{P_a + P_b} \tag{28}$$

2. Equality at $d = 0$: From (24) we find that

$$\tilde{p}(0) = \kappa \cdot (P_b + \gamma)$$

and from (25)

$$\hat{p}(0) = \kappa \left[\epsilon_1 \cdot P_1 + \epsilon_0 \cdot P_0 \right]$$

so that, for $\tilde{p}(0) = \hat{p}(0)$ we find

$$P_1 = \frac{P_b + \gamma - \epsilon_0 \cdot P_0}{\epsilon_1} \tag{29}$$

3. Equality of the first derivative at $d = 0$: From (24) we find that

$$\frac{\partial}{\partial d}\tilde{p}(d) = \kappa \left[-\rho\gamma e^{-\rho d} + \ln\left(1 - P_a - P_b\right) \left(P_b + \gamma e^{-\rho d}\right) + \ln\left(\Phi\right) \rho e^{-\rho d} \cdot \left(P_b + \gamma e^{-\rho d}\right) \right] \cdot \left(1 - P_a - P_b\right)^d \cdot \Phi^{1 - e^{-\rho d}}$$

from which, taking $d = 0$, we obtain

$$\frac{\partial}{\partial d}\tilde{p}(d) \mid_{d=0} = \kappa \cdot \left[-\rho \cdot \gamma + (P_b + \gamma) \cdot \ln(1 - P_a - P_b) + \rho \cdot (P_b + \gamma) \cdot \ln(\Phi) \right] \tag{30}$$

Similarly, from (25)

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

from which, taking $d = 0$, we obtain

$$\frac{\partial}{\partial d}\hat{p}(d) \mid_{d=0} = \kappa \cdot \left[\epsilon_1 \cdot \ln(1 - P_1) \cdot P_1 + \epsilon_0 \cdot \ln(1 - P_0) \cdot P_0 \right] \tag{31}$$

Using (27) and (28) we may write

$$\begin{aligned} \epsilon_0 \cdot \ln(1 - P_0) \cdot P_0 &= \frac{P_b \cdot \Phi}{P_a + P_b} \cdot \ln(1 - P_0) \cdot (P_a + P_b) \\ &= P_b \cdot \Phi \cdot \ln(1 - P_0) \end{aligned} \tag{32}$$

Now, by requiring

$$\frac{\partial}{\partial d}\tilde{p}(d) \mid_{d=0} = \frac{\partial}{\partial d}\hat{p}(d) \mid_{d=0}$$

we find from (30), (31) and (32) that:

$$\epsilon_1 \cdot \ln(1 - P_1) \cdot P_1 = \ln(1 - P_0) \cdot \left[P_b + \gamma - P_b \cdot \Phi \right] + \rho \cdot \left[(P_b + \gamma) \cdot \ln(\Phi) - \gamma \right]$$

$$\Rightarrow (P_b + \gamma - P_b \cdot \Phi) \cdot \ln(1 - P_1) = \ln(1 - P_0) \cdot \left[P_b + \gamma - P_b \cdot \Phi \right] + \rho \cdot \left[(P_b + \gamma) \cdot \ln(\Phi) - \gamma \right]$$

$$\begin{aligned} \Rightarrow \ln(1 - P_1) &= \frac{\ln(1 - P_0) \cdot \left[P_b + \gamma - P_b \cdot \Phi \right] + \rho \cdot \left[(P_b + \gamma) \cdot \ln(\Phi) - \gamma \right]}{P_b + \gamma - P_b \cdot \Phi} \\ &= \ln(1 - P_0) + \frac{\rho \cdot \left[(P_b + \gamma) \cdot \ln(\Phi) \right]}{P_b + \gamma - P_b \cdot \Phi} \end{aligned}$$

so that, finally, we obtain

$$P_1 = 1 - (1 - P_0) \cdot e^{\left[\frac{\rho \cdot \left[(P_b + \gamma) \cdot \ln(\Phi) - \gamma \right]}{P_b + \gamma - P_b \cdot \Phi} \right]} \tag{33}$$

10. APPENDIX B

Consider an experiment consisting of Bernoulli trials with probability of success P_x. The number of repetitions before witnessing the first positive result is described by the binomial distribution:

$$P_{bin}(d) = (1 - P_x)^d \cdot P_x \tag{34}$$

Now select the subset of trials for which $d < N$, where N is an integer greater than zero. The probability distribution over this interval $P(d)$ may be determined by applying the normalisation requirement

$$\sum_{d=0}^{N-1} P(d) = 1$$

to equation (34) and obtaining

$$\begin{aligned} P(d) &= \frac{P_{bin}(d)}{\sum_{d=0}^{N-1} P_{bin}(d)} \\ &= \frac{(1-P_x)^d \cdot P_x}{1-(1-P_x)^N} \qquad \text{with} \quad d \in \{0, 1, \ldots, N-1\} \end{aligned} \tag{35}$$

In this appendix we find expressions for the mean and mean-square of this **truncated binomial distribution**, and begin by calculating the moment generating function:

$$\begin{aligned} M_d(t) &= \sum_{d=0}^{N-1} e^{t \cdot d} \cdot P(d) \\ &= \frac{P_x}{1-(1-P_x)^N} \cdot \sum_{d=0}^{N-1} (1 - P_x)^d \cdot e^{t \cdot d} \\ &= \frac{P_x \cdot (1-(e^t \cdot (1-P_x))^N)}{(1-(1-P_x)^N) \cdot (1-e^t \cdot (1-P_x))} \\ &= \Upsilon \cdot \frac{1-(e^t \cdot (1-P_x))^N}{(1-e^t \cdot (1-P_x))} \end{aligned} \tag{36}$$

where

$$\Upsilon = \frac{P_x}{\left(1-(1-P_x)^N\right)} \tag{37}$$

Taking the derivative of (36) with respect to t we find:

$$\begin{aligned} \frac{\partial}{\partial t} M_d(t) &= \Upsilon \cdot \frac{e^t(1 - P_x) \cdot \left(1 - e^{Nt}(1 - P_x)^N\right) - N \cdot e^{Nt} \cdot (1 - P_x)^N \cdot \left(1 - e^t \cdot (1 - P_x)\right)}{(1 - e^t \cdot (1 - P_x))^2} \\ &= \Upsilon \cdot \frac{e^t(1 - P_x) - N \cdot e^{Nt}(1 - P_x)^N + (N - 1) \cdot e^{(N+1)t} \cdot (1 - P_x)^{N+1}}{(1 - e^t \cdot (1 - P_x))^2} \end{aligned} \tag{38}$$

and to obtain the mean $\mu(P_x, N)$ we set $t = 0$:

$$\begin{aligned} \mu(P_x, N) &= \frac{\partial}{\partial t} M_d(t) \mid_{t=0} \\ &= \Upsilon \cdot \frac{(1 - P_x) \cdot \left[1 + (1 - P_x)^{N-1} \left[(N - 1) \cdot (1 - P_x) - N\right]\right]}{P_x^2} \end{aligned} \tag{39}$$

Taking the second derivative of (36) with respect to t we find:

$$\begin{aligned}\frac{\partial^2}{\partial t^2} M_d(t) &= \Upsilon \cdot \frac{\left(1-e^t\,(1-P_x)\right)^2 \cdot \left[e^t\,(1-P_x) - N^2 e^{Nt}\,(1-P_x)^N + (N-1)\,(N+1)\,e^{(N+1)t}\,(1-P_x)^{N+1}\right]}{(1-e^t\,(1-P_x))^4} \\ &+ 2 \cdot \Upsilon \cdot \frac{\left[e^t\,(1-P_x) - N \cdot e^{Nt}\,(1-P_x)^N + (N-1) \cdot e^{(N+1)t} \cdot (1-P_x)^{N+1}\right] \cdot \left(1-e^t\,(1-P_0)\right) \cdot e^t\,(1-P_x)}{(1-e^t\,(1-P_x))^4} \end{aligned} \tag{40}$$

and obtain the mean-square $\nu(P_x, N)$ of the distribution by again setting $t = 0$

$$\begin{aligned}\nu(P_x, N) &= \frac{\partial^2}{\partial^2 t} M_d(t) \mid_{t=0} \\ &= \Upsilon \cdot \frac{P_x^2 \cdot \left[(1-P_x) \cdot \left\{1 - N^2\,(1-P_x)^{N-1} + (N-1)\,(N+1)\,(1-P_x)^N\right\}\right]}{P_x^4} \\ &+ 2 \cdot \Upsilon \cdot \frac{\left[(1-P_x) \cdot \left\{1 - N\,(1-P_x)^{N-1} + (N-1)\,(1-P_x)^N\right\}\right] \cdot P_x \cdot (1-P_x)}{P_x^4} \end{aligned} \tag{41}$$