Comparative evaluation of word- and category-based language models

T.R. Niesler and P.C. Woodland

CUED/F-INFENG/TR.265

5 July 1996

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

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

Abstract

Conventional n-gram language models employ the occurrence counts of word n-tuples to calculate probabilities for word sequences. It has been demonstrated, however, that language models using n-tuples of word-categories rather than words exhibit certain advantages, such as the intrinsic ability to generalise to unseen word sequences, and attactive size versus performance tradeoffs. This document compares the behaviour of word- and category-based language models in detail, and among the significant findings are that the category-based model is less likely to deliver very small probability estimates, that it performs better in situations where the word-model backs-off, and that the category-based model is less sensitive to changes in the character of the test-text.

Contents

1. INTRODUCTION

This document describes an in-depth comparison of a conventional n-gram language model[1], and an n-gram language model based on syntactic word categories [3]. The purpose of the investigation is to determine the strengths and weaknesses of each approach as a means of identifying ways in which either may be improved.

2. THE CORPUS

All language models used in the experiments described in the following sections were constructed from the WSJ0 corpus, which consists of 37,346,118 words and 1,625,606 sentences[2] of newspaper text collected from the Wall Street Journal over the period 1987-89 inclusive. Being a considerably sized body of text, findings made using WSJ0 are expected to hold also for larger corpora such as NAB1, on which several state-of-the-art recognition systems have been based.

2.1. Preprocessing

The verbalised pronunciation processed version of the WSJ0 corpus (vp) was used to build all language models. The filters `vp2svp1` and `sgml2text`[3] were employed to obtain plain text output, after which the following steps of preprocessing were carried out prior to model construction :

1. Insert a separating space between the end of a word and a comma where this is missing, e.g. "million," ⇒ "million ,". This prevents words such as the former from appearing as separate entries in the vocabulary.
2. Remove all periods from the ends of words (e.g. "Mr." ⇒ "Mr"). This was done for compatibility across corpora, since the trailing period is absent in some text sources.
3. Correct common misspellings present in NAB corpus (e.g "milllion" ⇒ "million"). A total of 3,238 corrections of this type were made.
4. Map American to British spellings, e.g "color" ⇒ "colour". This was done to maintain compatibility with the LOB corpus [1], which consists of approximately 1 million words of British English, and was used to tag the WSJ0 corpus (more detail is given later). A total of 134,109 such corrections were made.

2.2. Test-set

The standard setaside dev-test text for WSJ0 was used as a test-set for all experiments in this work. Consisting of 91,896 sentences and 2,104,322 words, its size is approximately 6% of that of the training-set. The test-set was subjected to the same preprocessing steps as the training corpus.

2.3. Vocabulary

All tests employed a vocabulary consisting of the 65,000 most-frequent words in the preprocessed WSJ0 corpus.

2.4. Out-of-vocabulary (OOV) words.

All OOV words were mapped to the special symbol "UW", which is used as part of the context in the n-gram models to predict future words, but is excluded from the perplexity calculation. The OOV rate for the described vocabulary is 0.62% on the test-set.

[1] A model consisting of word n-tuples, and referred to as a **word-based model** hereafter.
[2] These figures have been calculated after subjecting the corpus to the preprocessing steps described in the following section.
[3] These filters are part of the CMU language modelling toolkit.

2.5. Summary of corpus statistics

The chief statistics given in the preceding sections are summarised below for convenience.

Training-set:

Number of sentences	1,625,606
Number of words	37,346,118
Number of different words	162,002

Test-set:

Number of sentences	91,896
Number of words	2,104,322

Vocabulary:

Size (words)	60,000
OOV rate on test-set	0.62%

3. THE CATEGORY-BASED MODEL

A language model based on variable-length word-category n-grams[4] was constructed for the WSJ0 corpus using the techniques described in [3]. The model allows words to belong to multiple categories, and consequently 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. The model construction procedure increases the length of individual n-grams according to the expected gain in performance based on a cross-validation criterion, thus maintaining compactness and avoiding overtraining. The particular category definitions used were the syntactic part-of-speech (POS) classifications present in the LOB corpus [1], with additional classes added for (i) plural and genitive forms of letters, and (ii) various contractions.

In order to supply the necessary POS classifications for each word in the untagged WSJ0 corpus, a language model constructed from the LOB corpus itself and employing an augmented lexicon for improved coverage was configured as a tagger [3] and used to tag the training-set. The explicit estimation of the probability of OOV words within each category allows unknown words to be tagged with fair accuracy.

The tagged WSJ0 corpus was consequently used to construct three varigram models, employing likelihood pruning thresholds of 1e-4, 1e-5 and 5e-6 respectively. Figure 1 shows the number of category n-grams as a function of tree depth[5] for the resulting models, and figure 2 the performance of the models when the n-grams are limited to various maximum lengths. The overall complexity of the final language models and their performance on the test-set is given in table 1.

Tree threshold	No. of n-grams	Test-set perplexity
1e-4	25,484	537.79
1e-5	107,979	495.71
5e-6	172,749	485.28

Table 1: Language models for the LOB corpus

[4] Referred to as a "**varigram**" model hereafter.

[5] The category n-grams are organised conceptually as trees [3]. The depth within the tree (i.e. the length of the path from a node to the root) corresponds to the length of the n-gram it signifies.

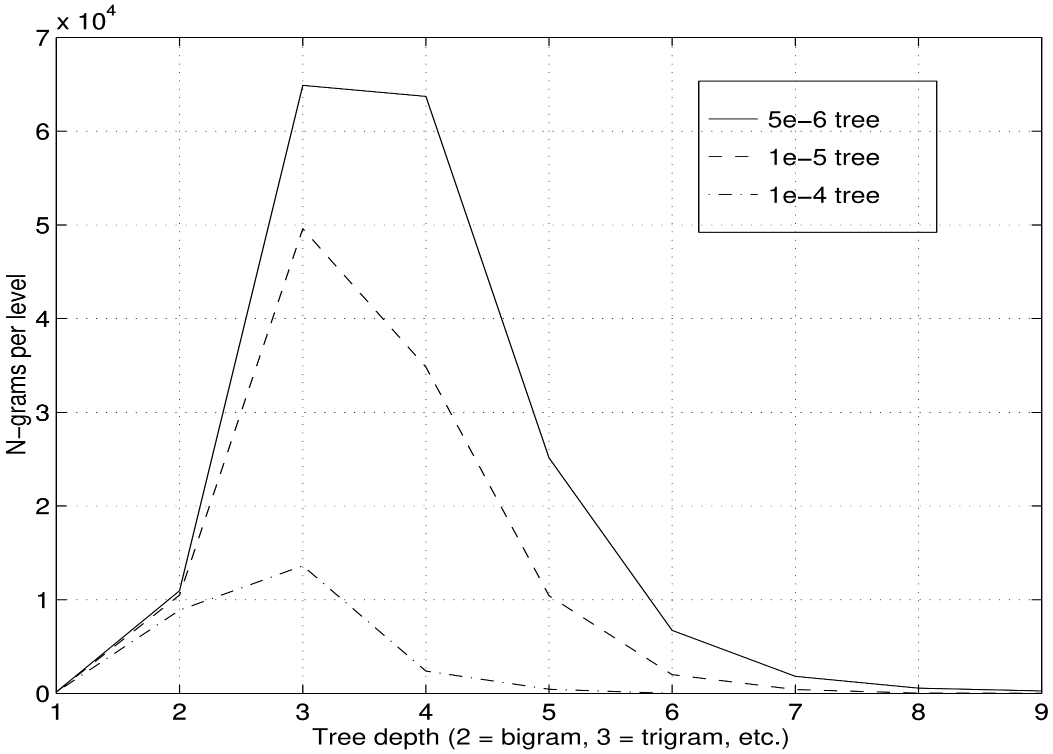

Figure 1: N-grams in each level for each of the category-based trees

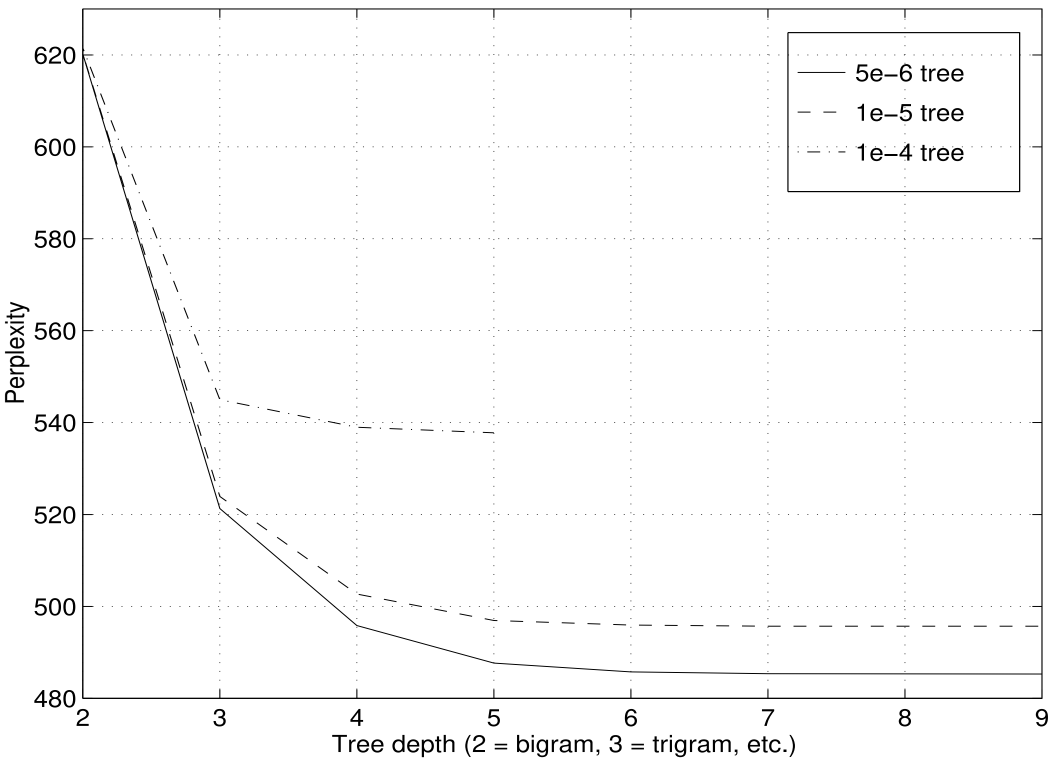

Figure 2: Performance of category-based models for various maximum tree-depths

4. THE WORD-BASED MODEL

Word-based n-gram language models employing the Katz back-off in conjunction with Good-Turing discounting [2] to provide probability estimates for unseen events were constructed from the WSJ0 corpus.

Since a likelihood pruning technique was seen to be effective in reducing the complexity of varigram models [3], its utility to word-based models was investigated here by employing the same cross-validated performance measure to decide whether to extend a word n-gram to a word (n+1)-gram or not. Both bigram and trigram models were constructed in this way, and from their performance (shown in the following two subsections) it is clear the the pruning technique is less successful for word-models. It was verified that the best word-based bigram and trigram matched the performance of corresponding models constructed under identical conditions with the CMU toolkit very closely.

4.1. Word-based bigram

The following word-based bigram models were constructed by varying the pruning threshold:

Threshold	No. of bigrams	Perplexity
No pruning	4,359,331	204.42
1e-7	4,051,165	205.63
1e-6	3,405,922	215.69

4.2. Word-based trigram

The word-based trigram model was constructed by extending the n-grams of the bigram model obtained with a pruning threshold of 1e-7. Due to storage limitations, only trigrams whose contexts occurred at least 6 times as bigrams were collected. However, since this affects only very sparsely-trained contexts, its effect is assumed to be insignificant.

Threshold	No. of trigrams	Perplexity
No pruning	9,959,732	132.08
0.0	9,603,396	132.06
1e-7	9,548,230	132.06
1e-6	9,294,003	132.11
1e-5	8,367,348	133.35
1e-4	6,554,027	140.26
1e-3	4,208,587	157.72

5. LANGUAGE MODEL COMPARISON

This section analyses the relative performance of the category based model (constructed with a pruning threshold of $5e-6$) and the word-based bigram and trigram (both constructed with a pruning threshold of $1e-7$).

5.1. Overall probability estimates

Aim : to investigate differences in the values of the probability estimates made by the category- and word-based models over the entire test-set.

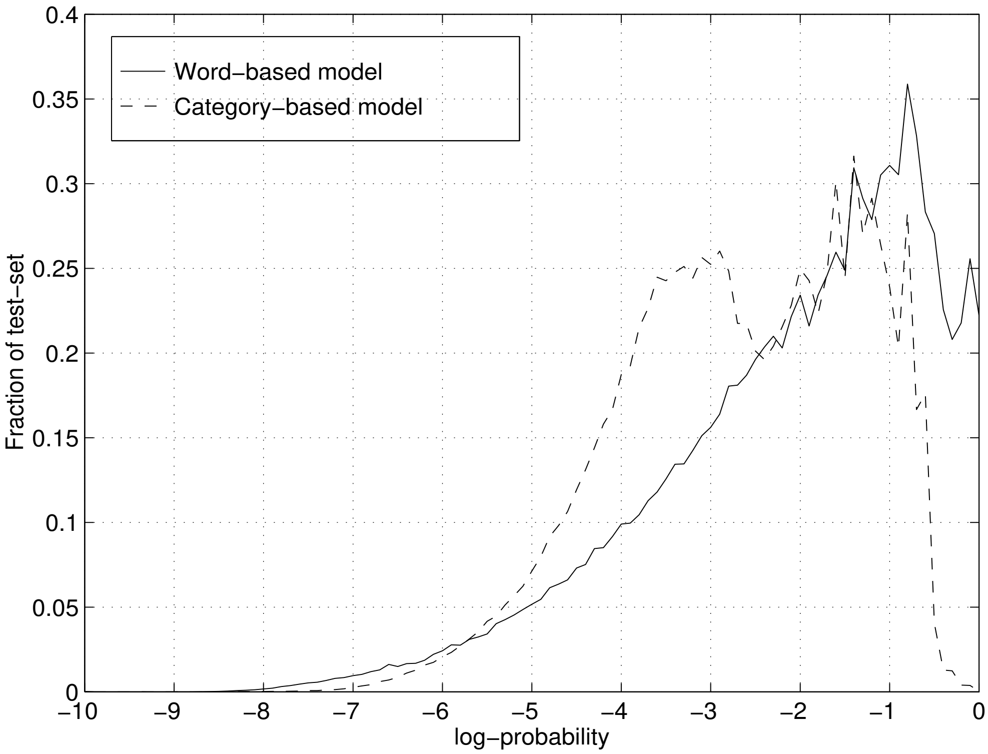

Figure 3: Overall distribution of log-probabilities produced by word- and category-based models

The histogram in Figure 3 shows the proportion of words in the test-set predicted with various log probabilities by both the category- and the word-based model. It is evident that :

- The category-based model assigns very low probabilities (smaller than 10^{-6}) to a smaller proportion of events.
- A much larger proportion of words are predicted with rather high probability (between 1.0 and 0.1) by the word-based than by the category-based model.

Conclusion : The ability of the category-model to generalise to unseen word sequences leads to the smaller number of very low probabilities, but this same characteristic does not allow it to capture word-specific relations, the strongest of which lead to the high-probability estimates produced by the word-based model.

5.2. The effect of backoffs

Aim : to investigate the probability estimates made by the category- and the word-based model in the cases where the latter does not back-off, or backs-off to various degrees.

Table 2 shows the perplexities calculated for words predicted with each type of n-gram request and type of backoff by the trigram language model. The class- and word-based models are denoted by "CBM" and "WBM" respectively.

Type of n-gram request	% of test-set	CBM perplexity	WBM perplexity
Unigram	1.07 %	564.58	1,027.61
Bigram (found)	19.33 %	321.63	95.25
Bigram (backed-off to unigram)	2.34 %	15,110.33	73,541.35
Trigram (found)	59.69 %	232.77	30.79
Trigram (backed-off to bigram)	13.36 %	3,418.93	3,530.23
Trigram (backed-off to unigram)	4.21 %	31,202.23	288,236.96

Table 2: Perplexities for various types of n-gram requests.

Figures 4 and 5 plot the difference in the log probabilities produced by the word- and the category-based models respectively for all cases in which the former did not need to back-off, against the log of the count of that particular n-gram in the training-set (indicated by "EC"). The two figures show the behaviour for the bigram and for the trigram word-models respectively[6]. From the graphs we see that the category-based model matches the word-based model most closely in performance for rare n-grams and for very frequent n-grams . The former is ascribed to the increasing unreliability of word n-gram counts as a consequence of data sparseness, and the latter to the most frequent n-grams consisting of words that belong to categories with a small number of members (e.g. articles, prepositions etc.).

Conclusion : When the word-based model does not back-off (a situation true for approximately 80% of the words in the test-set), it does significantly better on average than the category-based model. When backoffs do occur this is no longer true since the category-based model is intrinsically able to generalise to unseen word sequences and is therefore often able to deliver superior probability estimates. Furthermore, it is interesting to note that a significant proportion (approximately 22%) of the predictions made by the category-based model are as good or better than those of the word-model even when backing-off does nor occur.

5.3. Per-category analysis

Aim : To analyse the total contribution to the overall likelihood made by individual word-categories, as well as average likelihood associated with each for non backed-off word n-grams .

The category is taken to be that assigned to the word by the same tagger used to tag the training corpus. The following figures[7] illustrate the absolute fraction of the total log probability each category is accountable for (figures 6 and 7), as well as the per-category average likelihood (fig 8). The horizontal axis indicates the frequency with which each such category occurs in the test-set.

Conclusions :

- On average, the word-based model performs better than the category-based model. Since the results are for non-backoff estimates only, this result is to be expected given the findings of the previous section.

[6] A subset of 130,324 points drawn uniformly from the test-set is shown in the plot.

[7] The following most significant grammatical categories have been labelled in these figures: common noun (NN), plural common noun (NNS), adjective (JJ), proper noun (NP), verb base form (VB), past tense of verb (VBD), past participle (VBN), present participle (VBG), third-person singular verb (VBZ), adverb (RB), preposition (IN), cardinal (CD), singular or plural article (ATI), singular article (AT), coordinating conjunction (CC), subordinating conjunction (CS), letter of the alphabet (ZZ), end-of-sentence marker (SE), infinitival "to" (TO), unit of measurement (NNU).

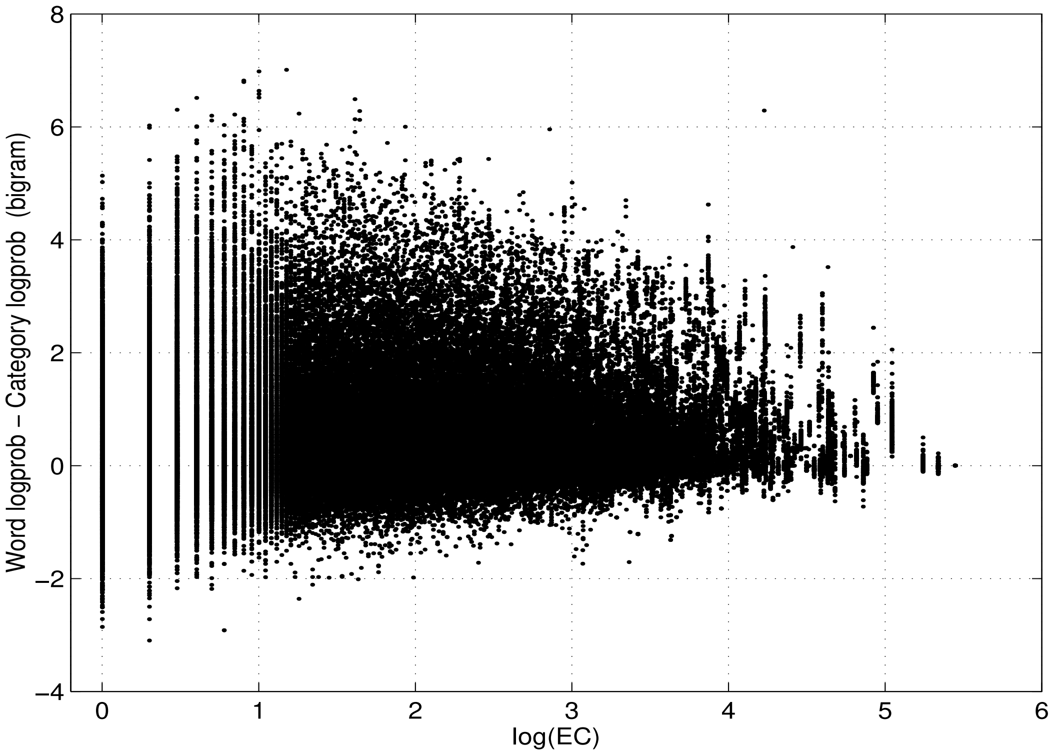

Figure 4: Difference in log probabilities as a function of event count for word-bigram

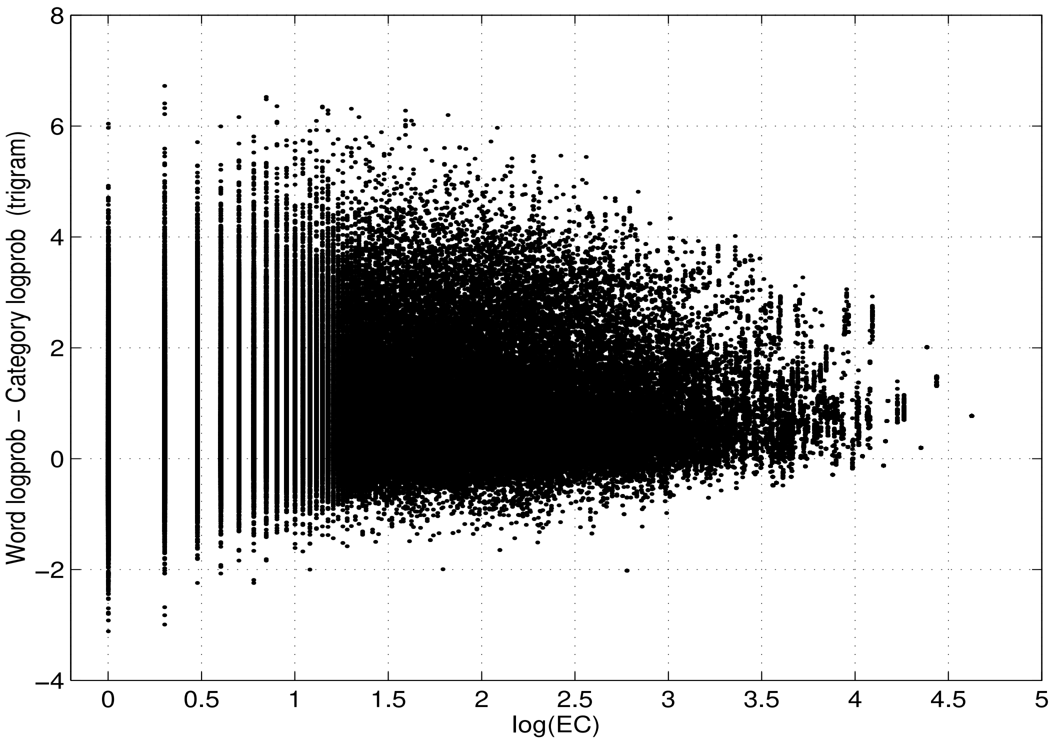

Figure 5: Difference in log probabilities as a function of event count for word-trigram

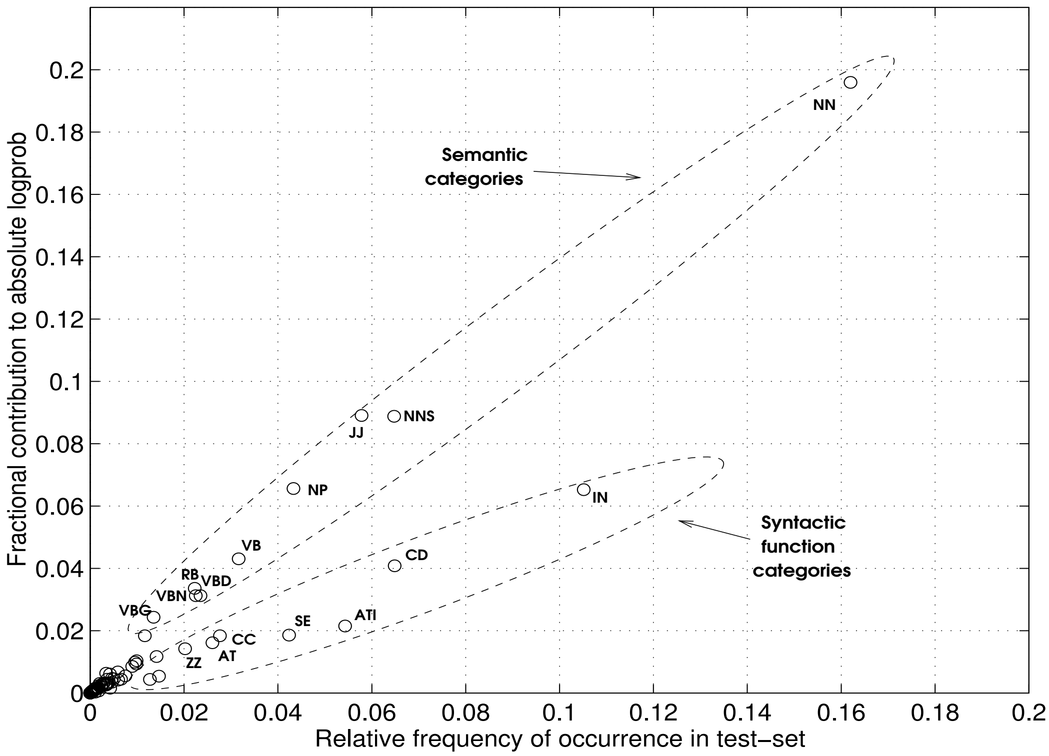

Figure 6: Contribution to total log probability of each category in the training-set (word model)

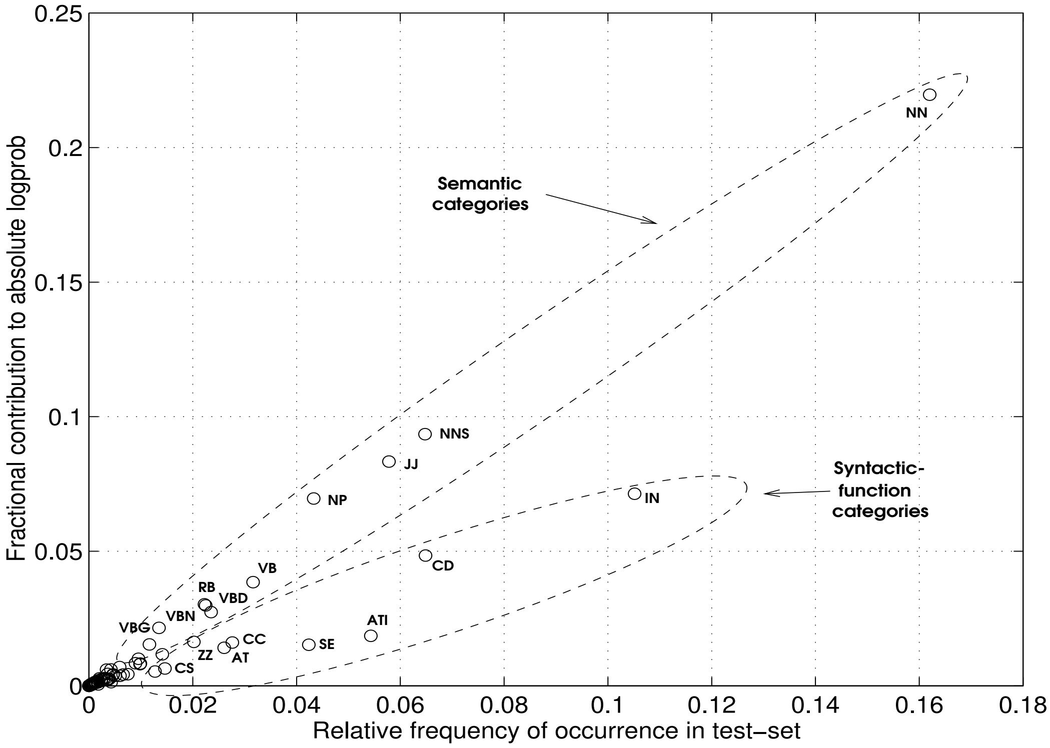

Figure 7: Contribution to total log probability of each category in the training-set (category model)

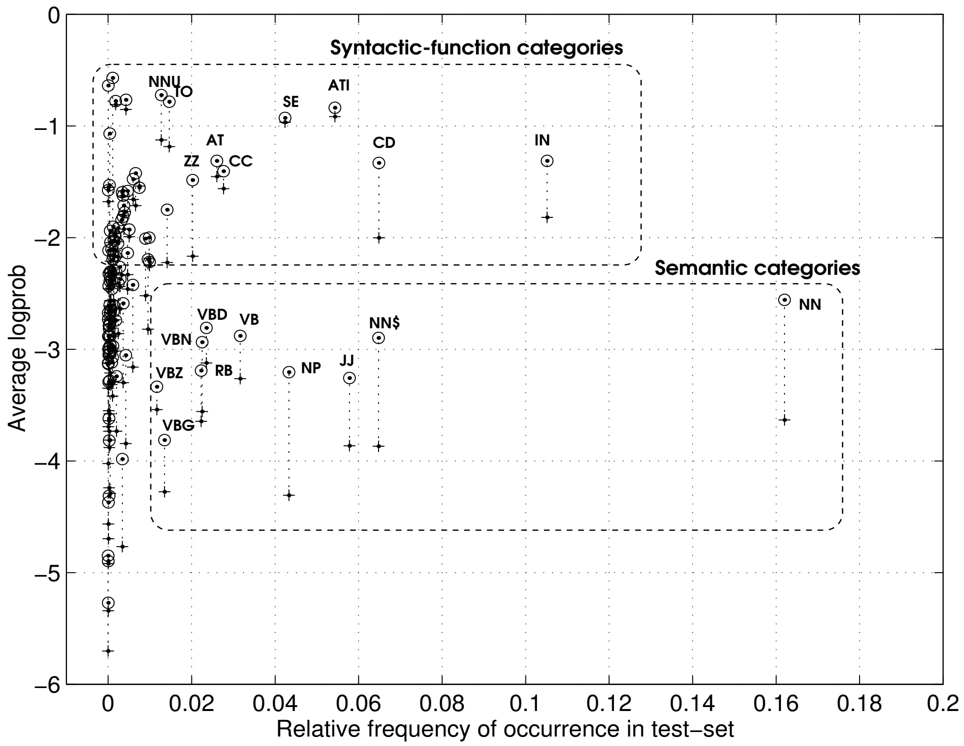

Figure 8: Average log probability within each category for both word- and category-based models

- There is no clear relationship between the frequency of occurrence of a category and its perplexity, except that the categories with lowest average likelihood also have been seen a very small number of times (this is not visible in figure 8, but becomes evident on expansion of the horizontal axis by several orders of magnitude). On further investigation it becomes apparent that these categories are sparsely trained as they have been seen only a small number of times in the training-set.
- Although some categories have a low average likelihoods, they are infrequent and therefore make a minor contribution to the overall likelihood.
- Common nouns (NN) are the most significant contribution to overall likelihood, followed by plural common nouns (NNS), adjectives (JJ) and proper nouns (NP).
- In general it seems that words with semantic content, such as nouns and adjectives, seem to be harder to predict (have lower probability) than syntactic function words (prepositions, articles, conjunctions and personal pronouns). This is true for both word- and category-based models. The two vertically separated groupings in figure 8 indicate this relation.
- There appears to be an approximately linear relationship between the frequency with which a a category occurs and its contribution to the overall likelihood. Furthermore, the constant of proportionality is different for words with significant semantic content and syntactic function words - this is emphasised by the two elongated groupings in figures 6 and 7.

5.4. Per-n-gram analysis

Aim : The findings of the preceding sections have made it clear that word-based n-grams carry a significant amount of information that cannot be captured by their category-based counterparts. The aim of this section is to investigate what proportion of word n-grams play a significant role in improving upon the category-based model.

In order to achieve this aim, the total contribution to the likelihood difference between word- and category-based models made by each distinct word-trigram in a non-backoff situation is calculated, and graphed in order of decreasing value. Figure 9 shows a normalised plot of these results. Note the portion of the curve above 1.0, which is due to those n-grams assigned higher probabilities by the category-based model.

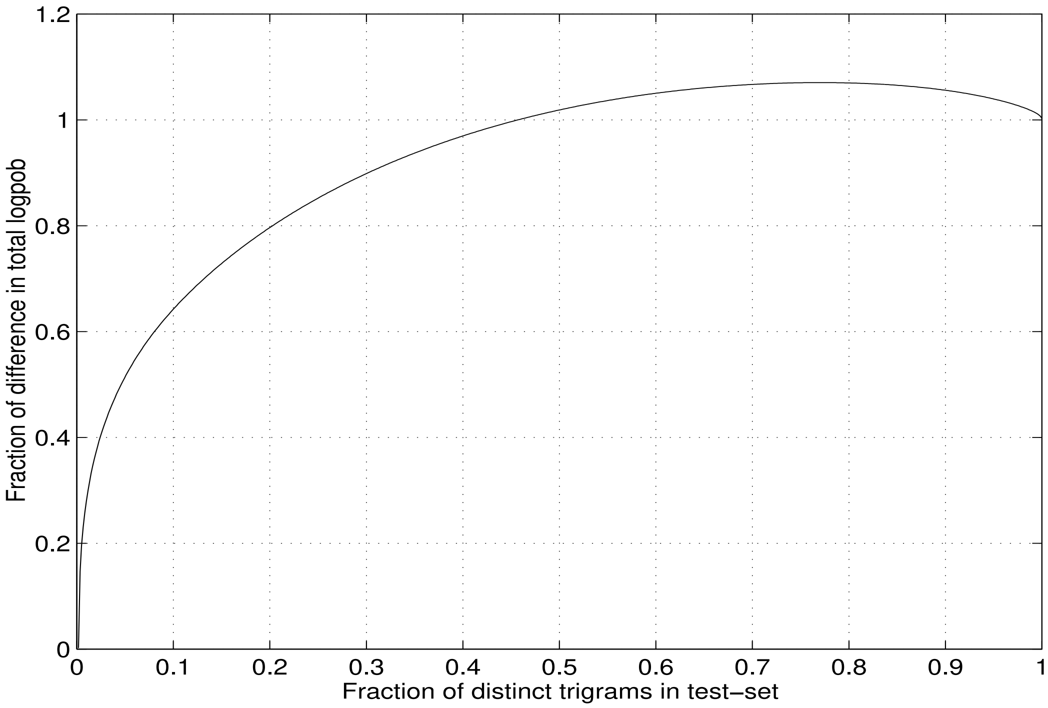

Figure 9: Contribution to difference in log probabilities generated by word- and category-based models

Conclusion : From figure 9 we may deduce that approximately half of the word trigrams contribute to the lead which the word-based model has over its category-based counterpart, the other half being predicted equally well or better by the latter. Furthermore, more than 50% of the improvement is contributed by only 5% of the trigrams. Thus, were category- and word-based models to be used in conjunction with one another, it should be possible to make the latter significantly more compact.

As a matter of interest, table 3 lists a few examples of trigrams for which the word- and category-models fare better respectively.

Word-model better	Category-model better
a border patrol	abortive rising of
a roman catholic	also compared him
accepted accounting principles	announcers or analysts
across national borders	as the point
adam and eve	be announced to
below zero fahrenheit	both declined that
caught by surprise	closely held to
declaration of independence	declined to quote
five year old	farmers now raise
former investment banker	five hundred percent
great barrier reef	grins and says
have grown accustomed	he agreed the
increase cash flow	Italian or French
lowest discount fares	last year shares
McDonnell Douglas Astronautics	minister or president
more than doubled	more than us
nobel peace prize	new orders of
open heart surgery	open not closed
possible business combination	previously reported of
president Francois Mitterrand	president and Mr
racist and sexist	rest of an
registered as unemployed	rose to say
satellite into orbit	seven million futures
sentiment remained bearish	so healthy that
soviet air defences	spokesman for an
television sports commentator	twenty thousand percent
Vancouver British Columbia	vision than chief

Table 3: Trigrams modelled better by word- and category-models respectively.

5.5. Robustness to domain-change

Preceding experiments have compared the two language models on a test-set whose character corresponds closely to that of the training-set, in that the text is derived from the same newspaper (WSJ) and from the same period. An important issue in language modelling is how well the model will fare on data from a different domain (e.g. from a different newspaper, or with an entirely different character and style). In order to investigate the performance of the two language models on domains other than WSJ, the following four additional test-sets were compiled from the LOB corpus [1].

- **Press reportage** (consist of categories A, B and C from the LOB corpus, which are made up of various newspaper articles, editorials and reviews not from WSJ).
- **Religion** (category D of the LOB corpus, which contains text concerning religious topics).
- **Scientific writing** (category J of the LOB corpus).
- **Fiction** (categories K, L, N, P and R from the LOB corpus, consisting of adventure, mystery, Western, romantic and humorous fiction).

Table 4 summarises the performance of both the word- and category-based language models on these test-sets. Performance on the standard WSJ0 test-set is also shown for comparison.

Test-set	No. of words	% OOV	Perplexity (WBM)	Perplexity (CBM)
WSJ0 dev-test	2,104,322	0.62	132.06	485.28
Press reportage	174,465	5.15	472.78	608.54
Religion	33,215	6.00	527.13	570.41
Scientific writing	154,755	7.24	589.71	649.38
Fiction	239,927	7.10	657.40	711.87

Table 4: Performance of the category- and word-based models on different test-sets.

Conclusion : While the word-based model outperforms the category-based model in all cases, it is interesting to see that, whereas the perplexity of the former increases by a factor of between 3.57 and 4.98, the perplexity of the latter does so only by a factor of between 1.17 and 1.47, indicating a reduced sensitivity to a change of domain.

6. REFERENCES

[1] Johansson, S; Atwell, R; Garside, R; Leech, G. *The Tagged LOB corpus user's manual*; Norwegian Computing Centre for the Humanities, Bergen, Norway 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 1987, pp. 400 - 401.

[3] Niesler, T.R; Woodland, P.C. *Variable-length category-based n-grams for language modelling*, Tech. report CUED/F-INFENG/TR.215, Dept. Engineering, University of Cambridge, U.K., April 1995.