6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A MISSING-WORD TEST COMPARISON OF HUMAN AND STATISTICAL LANGUAGE MODEL PERFORMANCE

Marie Owens, Anja Krüger, Paul Donnelly, F J Smith and Ji Ming
School of Computer Science
The Queen's University of Belfast
Belfast BT7 1NN
Northern Ireland
Email: fj.smith@qub.ac.uk

10.21437/Eurospeech.1999-40

ABSTRACT

A suite of missing-word tests based on text extracts selected randomly from two different text corpora provided a metric which was used in an evaluation of human performance, an evaluation of language model performance and a cross-comparison of the performances. The effects of providing different sizes of context for the missing word (ranging from two words to three sentences) were examined and two main patterns became clear from the results:

- surprisingly, for tests where the language model was able to take advantage of all the context information provided (i.e. where the context consisted of just a few words) it outperformed humans;
- conversely, humans outperformed the language model when the size of context given for the missing word exceeded the size, which the language model could usefully, employ in its probability calculations (typically more than six words).

INTRODUCTION

Missing-word tests [1] have a major advantage over the traditional perplexity metric so often used for language model evaluation: they facilitate direct comparison of language model performance with human language skills. A missing-word test is based on an extract of text from which a single word has been removed but its position flagged. The language model or human is then set the task of 'deciding' what word is missing. Tests requiring human subjects to guess which letter or word is missing from a text extract have been widely used as a means of investigating human language skills [2,3,4,5] but to our knowledge no other research group has employed this or any other metric for direct comparison of the performance of language models with human performance.

The experiment described in this paper used missing-word tests to compare the collective performance of 48 humans with an n-gram variable-order statistical language model. The experiment had the specific focus of trying to gain an insight into human use of context information. The effects of providing different sizes of context for the missing word were examined; the sizes of context ranged from two words to three sentences. The effects of supplying left context only, right context only or context on both sides were also examined. It was hoped that the study would provide guidelines on how to improve statistical language model performance so that language models could come closer to emulating human performance.

EXPERIMENT DETAILS

The experiment used 400 text extracts – 200 randomly selected from each of two corpora:

(i) the domain-specific VODIS corpus of train timetable enquiries and responses [6] containing 75K word tokens and 2.2K word types and

(ii) a multi-domain corpus consisting of the text from three months of the Times and Sunday Times newspapers containing 7.3 M word tokens and 75K types.

Each of the 400 extracts gave rise to three tests:

- a *left* context test where only preceding context was provided;
- a *right* context test where only succeeding context was provided;
- a *left-right* context test where preceding and succeeding context were provided.

The human subjects ranged in age from 18 to 58. There were 28 males and 20 females. They were native English speakers and all had at least a GCSE/O-level (i.e. school-leavers) qualification in English. They were paid for taking part.

In the experiments with humans each test was completed separately by two subjects. Thus, a total of 2400 (400x3x2) tests were completed by the human subjects. The tests were distributed randomly among the 48 subjects with each subject completing 50 tests.

An incremental style was used to present the missing-word tests to humans. Each incremental test consisted of up to six trials with increasing context shown if the missing word was not guessed in an earlier trial. Context was increased two words at a time in the first four trials (from two, to four, to six, then eight words) before moving to sentence level increments. In the fifth trial the remaining words of the sentence on the left side for *left* tests and right side for *right* tests were added while for *left-right* tests the remaining words on each side were added. In the sixth trial an additional complete sentence was added on the left, right or on both sides as appropriate.

As an example the context presented during the first three trials of a *left-right* test are shown below:

Example of a left-right-context test

Trial 1 2-word context:
demand _______ much

Trial 2 4-word context:
probably demand _______ much of

Trial 3 6-word context:
We probably demand _______ much of our

Solution: *too*

The tests were administered by computer using software developed for the experiment. Supervisors were in attendance throughout the experiment.

To conduct the language model tests a weighted-average variable-order n-gram language model [7] with n<=8 was trained for each corpus using the whole corpus less the test extracts. Software was developed to enable the language model to complete the same tests as were undertaken by the humans.

RESULTS

In this section we present our main results which address the following issues:

- which context type (*left*, *right* or *left-right*) is most beneficial;
- is there a limit to the size of context which brings a gain in performance;
- does the performance vary with the type of corpus.

Human Experiments

Results for the three types of context presentations (*left*, *right* and *left-right*) may be directly compared with each other for the first four trials only because in those trials the increments result in the same amount of context information being shown for all three types of test, e.g. a two-word *right* test shows two words after the missing word (? word1 word2) and a two-word *left-right* test shows one word on each side of the missing word (word1 ? word2). This is not the case for trials 5 and 6; e.g. when incrementing to the next sentence boundary for a *left* or *right* test only one sentence is completed whereas for a *left-right* test two sentences are completed (one on each side).

In Table 1 the results for the three context types are presented; the cumulative percentage success for the two corpora are shown for each trial, e.g. for the 800 *left* tests incrementing the context from 2 words to 4 words enabled subjects to guess an additional 8% of the 800 missing words correctly, bringing the cumulative percentage success up to 16%.

Table 1 shows that by the end of Trial 4 in the *left* and *right* tests around 24% of words were guessed correctly as against 42.5% in the *left-right* tests. Given that English flows from left to right it is surprising that the results for *right* tests are so similar to those for *left* tests. The finding that context on both sides is superior to left or right contexts points to the conclusion that the words closest to the missing word contain the most relevant cue information. However, the continued improvement in the later trials showed that humans still obtain helpful information some sentences away from the missing word.

Table 1: Human Performance
Cumulative percentage success for the two corpora combined with increasing context
for *left, right* and *left-right* tests

Trial	Context Size	Left (800 tests)	Right (800 tests)	Left-right (800 tests)
1	2 words	8	11	13
2	4 words	16	18	26
3	6 words	21	21	35
4	8 words	24.5	24	42.5
5	full sentence	28	25.5	50.5
6	next sentence	30	28	56

The results for *left*, *right* and *left-right* were examined for each corpus individually (see Table 2). In the VODIS tests more words were correctly guessed in all three types of context than in the Times tests. This may be due to the greater frequency of short, repetitive phrases which occur naturally in colloquial, domain-specific text like “British Rail can I help?” and “Thank you very much”.

Language Model Experiments

The corresponding results for the n-gram language model (see Table 3) highlight its restricted ability to use n-grams greater than six words in length, e.g. trials 5 and 6 brought no improvement in language model performance and so have been omitted from Table 3. Context on both sides was more beneficial than just left or just right, while differences between left versus right were negligible. Overall, the language model performed better for VODIS tests than Times tests which is not surprising given the restricted domain of the VODIS tests.

Table 2: Human Performance
Cumulative percentage success with increasing context for VODIS and Times *left*, *right* and *left-right* tests
(number of tests in brackets)

Trial	Context Size	VODIS left (400)	VODIS right (400)	VODIS left-right (400)	Times left (400)	Times right (400)	Times left-right (400)
1	2 words	10.5	12.5	18	5	9	8
2	4 words	20	21	33	13	15	20
3	6 words	26	25	42	16.5	17.5	27.5
4	8 words	30	28	51.5	19	20	33.5
5	full sentence	33	29	58.5	22.5	22	43
6	next sentence	35.5	31.5	63	25	24	49

Table 3: Language Model Performance
Cumulative percentage success (and number of correct guesses) with increasing context for VODIS and Times *left*, *right* and *left-right* tests
(number of tests in brackets)

Context Size	Vodis left (200)	Vodis right (200)	Vodis left-right (200)	Times left (200)	Times right (200)	Times left-right (200)
2 words	24	23.5	42	16	16	20
4 words	26	25	48	17	17	28.5
6 words	26	25.5	50	17	17	29.5
8 words	26	25.5	50	17	17	32

Comparison Human – Language Model Performance

In Figure 1 human and language model performances are compared for *left* tests. In *left*, *right* and *left-right* tests humans outperform the language model for a context greater than about six words – this is clearly related to the inability of the simple n-gram language model to improve its performance when supplied with wider context information. When the results for each corpus are examined individually the ‘crossing point’ of six words is found to be consistent over all *left* and *right* tests in both corpora whereas it is about seven words for the VODIS *left-right* tests and almost eight words for the Times *left-right* tests, reflecting the ability of the language model to make some use of the wider context in *left-right* tests.

The most unexpected finding uncovered from our data is that for both corpora the language model outperforms the humans for contexts smaller than six words. The finding is less surprising for the domain-specific VODIS corpus where it seems plausible that the language model could be at an advantage due to its lexicon being restricted to the training text which comes from the same narrow domain as the test text. However, the Times training corpus was selected to cover a wide range of subjects and domains and one might therefore expect humans to be at an advantage due to their larger lexicon and ability to apply general knowledge. A detailed study of all individual tests, which is planned for the near future, should provide insight into this finding and answer other questions such as whether the human improvement in performance as context increases is due to the presentation of key words which trigger human recognition of the precise domain of the test and whether the grammatical category of the missing word and/or its near neighbours has an influence on performance. Further experiments will show whether the improvement in human performance as context increases is sustained well beyond three sentences.

CONCLUSION

Our experiments have provided us with a valuable collection of data, which deserves more detailed analysis. The main findings to date can be summarised as follows:

(i) Context on both sides of a missing word is more beneficial to humans and language model than the same size of context just on the left side or just on the right side.

(ii) Context on the right side is as beneficial to humans and language models as context on the left side.

(iii) Humans still gain benefit from context as it increases up to a few sentences away from the missing word. This should be investigated further.

(iv) The n-gram language model outperforms the humans up to a context size of six words – not being able to take advantage of wider contexts. Cache-based language models [8, 9, 10, 11] should be evaluated using wider-context missing-word tests.

(v) Text type (i.e. its domain) has a similar effect on human and language model performance (based on the two types in our experiment). It would be interesting to explore other types of corpora e.g. multi-domain transcribed speech.

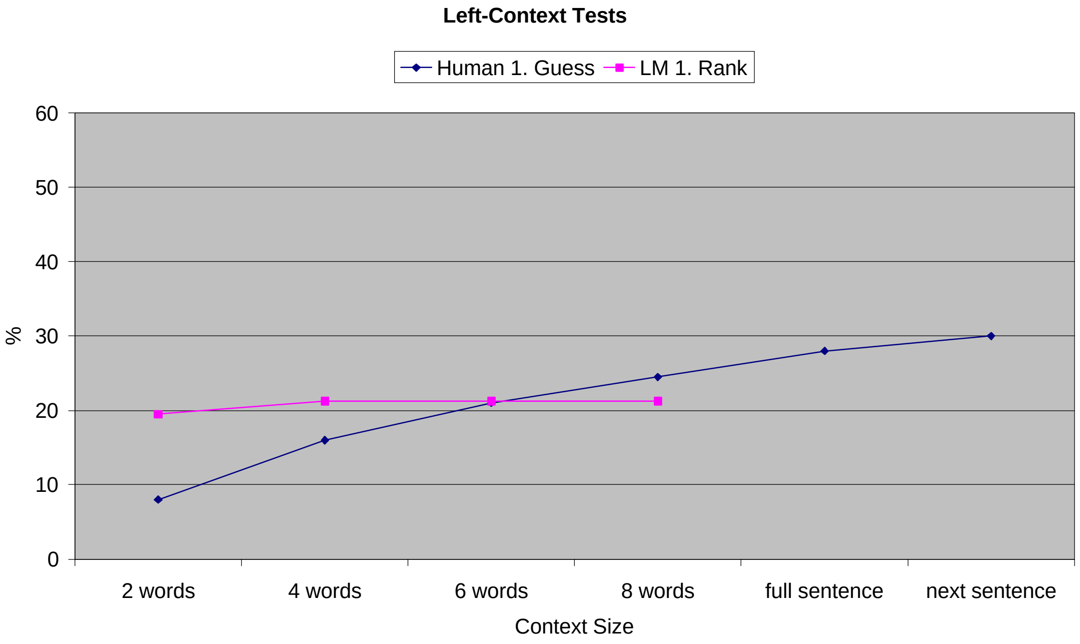

Figure 1: Comparison of Human and Language Model Performance in Left-Context Tests

It should encourage the language modelling community to learn that an n-gram language model is already capable of performance levels so similar to humans for small-context missing-word tests.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding provided by the UK Engineering and Physical Sciences Research Council (GR/L34419) and the provision of the VODIS corpus by BT Research Laboratories.

REFERENCES

[1] Owens, M., O'Boyle, P., McMahon, J., Ming, J., Smith, F. J. (1997). A comparison of human and statistical language model performance using missing-word tests. *Language and Speech*, *40* (4), 377-389.

[2] Shannon, C. E. (1951). Prediction and entropy of printed English. Bell System Technical Journal, 50-64.

[3] Aborn, M., Rubenstein, H., & Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology*, *57*, 171-180.

[4] Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*, 415-433.

[5] Fillenbaum, S., Jones, L. V., & Rapoport, A. (1963). The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. *Journal of Verbal Learning and Verbal Behaviour*, *2*, 186-194.

[6] Cookson, S. (1988). Final evaluation of VODIS. *Proceedings of Speech '88*, 7th Symposium. Edinburgh, pp. 1311-1320.

[7] O'Boyle, P., Owens, M. & Smith, F. J. (1994). A weighted average n-gram model of natural language. *Computer Speech and Language*, *8*, 337-349.

[8] Bahl, L. R., Brown, P. F., deSouza, P. V. & Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *37* (7), pp. 1001-1008.

[9] Lau, R., Rosenfeld, R. & Roukos, S. (1993). Trigger-based language models: a maximum entropy approach. Proceedings *International Conference on Acoustic, Speech and Signal Processing*, *II*, pp. 45-48.

[10] Kuhn, R. & de Mori, R. (1992). A cache based natural language model for speech recognition, *IEEE Transactions PAMI*, *14*, pp. 570-583.

[11] Wright, J. H. Carey, M. J. & Parris, E. S. (1995). Topic discrimination using higher-order statistical methods. *Computer Speech and Language*, *9*, *381-409*.