The Combination of Thesaurus and Word Form Vectors

B. Faith and J. Jensen

Abstract

In this study, the Thesaurus and the Word Form Dictionaries are merged, and the performance of this new dictionary is compared to that of its individual elements. The new dictionary yields better normalized precisions and recalls, but the improvement is only slight and the results are sometimes inconsistent.

1. Introduction

One of the major areas of research in information retrieval is the investigation of dictionaries. There have been numerous studies trying to determing the best type of dictionary to use. In ISR-13, Section VI [1], E. M. Keen compared the performances of stem dictionaries versus "suffix s" dictionaries and found, with one exception, that the stem is superior to the suffix s. In Section VII, Keen continued these studies by comparing the thesaurus against the stem dictionary, and determined that while the results were very close, the thesaurus is nearly always superior. Additional investigations added phrases and hierarchy dictionaries to the thesaurus. The phrases only slightly improved performance, but the addition of the hierarchy actually hindered it.

It appears that one other study is required — a comparison of the results of the thesaurus and the word form dictionaries separately and then combined. Since the thesaurus and word form vectors yield recall and precision graphs that are very close to each other, it would seem that the two dictionaries should complement each other. The major question is whether the

improvement is significant enough to justify the extra time and cost involved in the computer execution.

2. Procedure

The procedure for accomplishing this study requires the use of the CRAN 200 Thesaurus and the CRAN 200 Word Form dictionaries. Each is tested separately with the 42 available queries and 200 documents, and the two are then concatenated by use of an object module. The object module contains two subroutines, MASTER and CRDCON. CRDCON merges the CRN2S and CRN2TH dictionaries, adding a constant to the concept numbers of the Word Form dictionary in order to maintain the separate identities of the two dictionaries for both queries and documents. The constant is added by means of the addition of a "DO-LOOP" to CRDCON. The constant is introduced to the system through the subroutine MASTER.

Each of the three runs (Thesaurus only, Word Form only, and merged Thesaurus and Word Form) are searched in the usual manner. First, a search of all the documents is made (the 0 iteration). Second, two searches are made with feedback (iterations 1 and 2) holding the ranks of the relevant documents constant. Finally a run is made with feedback but with the ranks of the relevant documents no longer being "frozen" (iteration 3). The results of these are averaged and precision versus recall graphs are drawn for both the Document-level averages and Recall-level averages.

For ease of comparison, Tables 1 and 2 provide the normalized recall and precision for the three dictionaries and their four iterations. Graphs 1 and 2 are the precision versus recall plots for the three dictionaries based on the Document-level averages, while Graphs 3 and 4 use the Recall-level averages. Graphs 1 and 3 are for the third iteration and Graphs 2 and 4 are for the zero iteration. Since iterations 1 and 2 are intermediate steps,

their recall versus precision graphs are not included.

3. Results

Upon examination of Tables 1 and 2, it is seen that the merged Thesaurus and Word Form Dictionary yields slightly better normalized recalls and precisions than the Thesaurus alone, and significantly better figures than the Word Form alone. The graphs also indicate this with the exception of Graph 3 where a portion of the word form curve is higher than the combined, and of Graph 2 where the Thesaurus curve is higher for most of the values. A survey of the 42 queries indicates that the Thesaurus alone outperforms the merged dictionary for only two of the queries, while the Word From outperforms the merged vector in six of the instances. For the most part, the merged dictionary yields essentially the same results as the other two, except that it outperforms the Thesaurus five times and the Word Form seven times.

While in general, the combined dictionary seems to represent a compromise between the Thesaurus and Word Form dictionaries, a number of individual queries yield confusing results. The relevant document ranks for query 1 are as follows:

Relevant Document Number	Combined Rank	Thesaurus Rank	Word Form Rank
59	1	27	1
58	2	41	2
8	6	1	3
60	9	150	8
13	29	37	40

For documents 59, 58, and 60, the rank on the combined dictionaries are the same or close to that of the Word Form Dictionary, while the combined rank

Dictionary	Iteration 0	Iteration 1	Iteration 2	Iteration 3
Thesaurus and Word Form	.8788	.9184	.9144	.9321
Thesaurus	.8733	.9070	.9119	.9321
Word Form	.8430	.8917	.8915	.8594

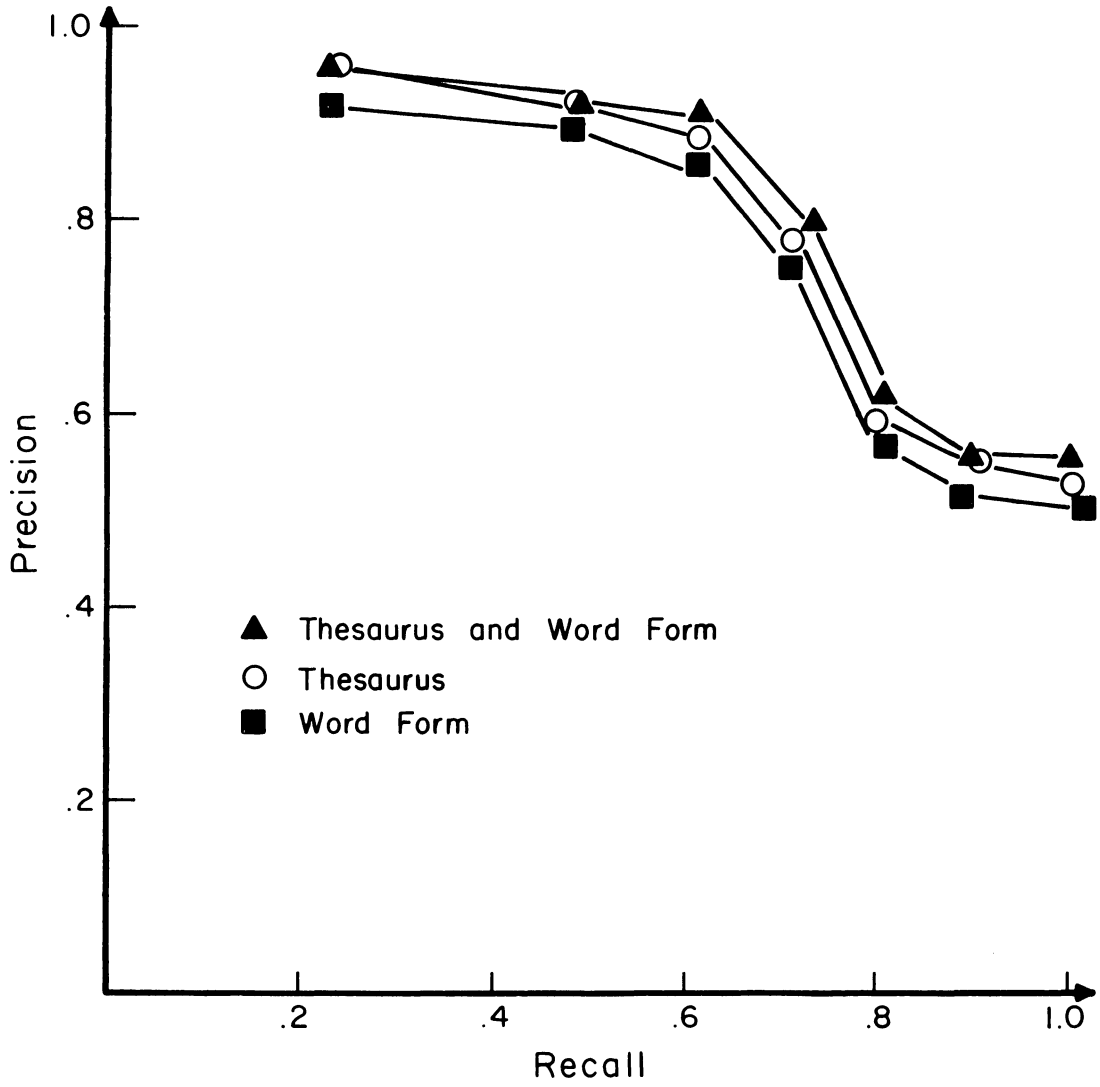
Normalized Recall

Table 1

Dictionary	Iteration 0	Iteration 1	Iteration 2	Iteration 3
Thesaurus and Word Form	.7035	.7448	.7431	.8747
Thesaurus	.6932	.7255	.7291	.8704
Word Form	.6659	.7039	.7079	.8594

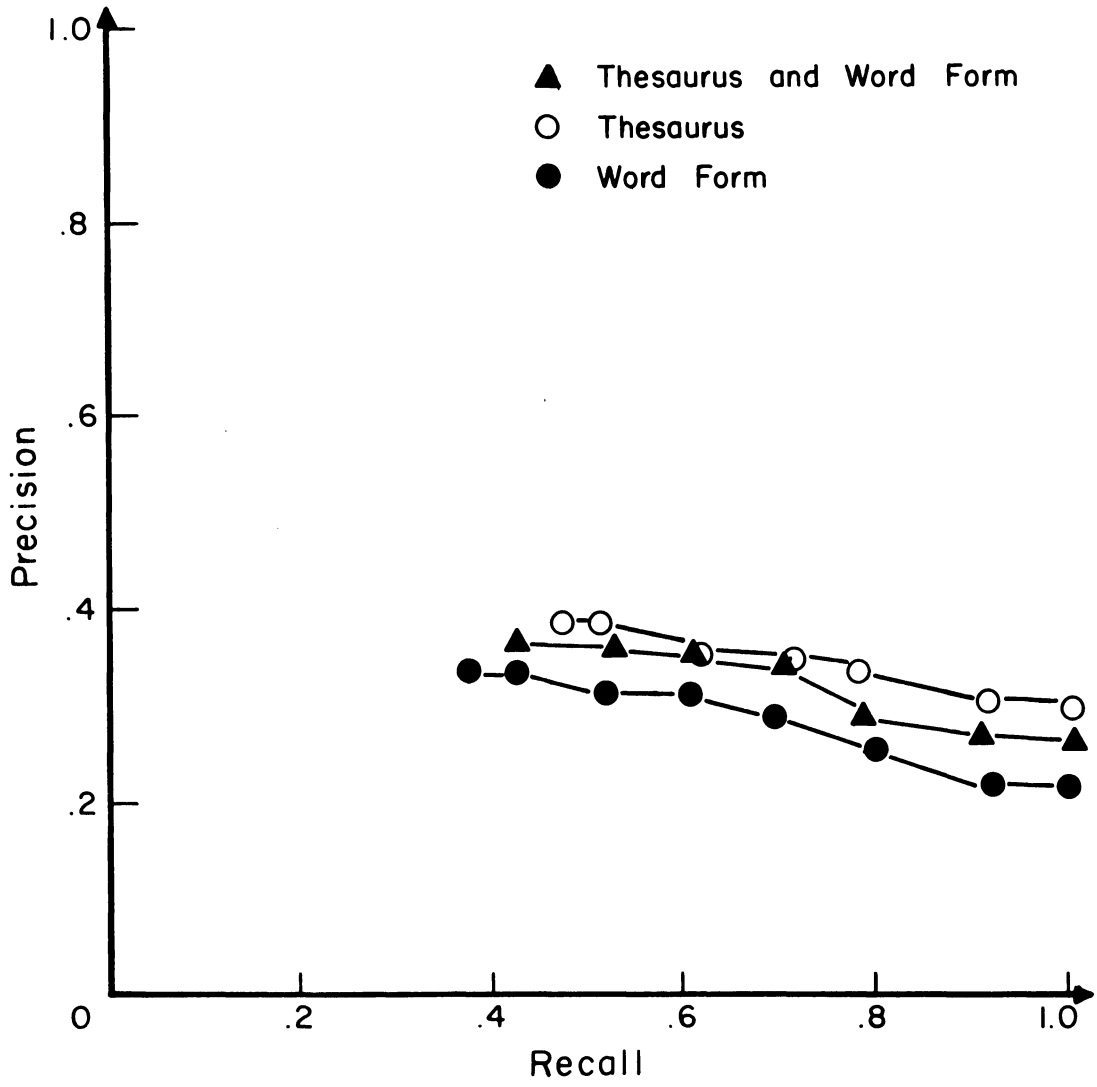
Normalized Precision

Table 2



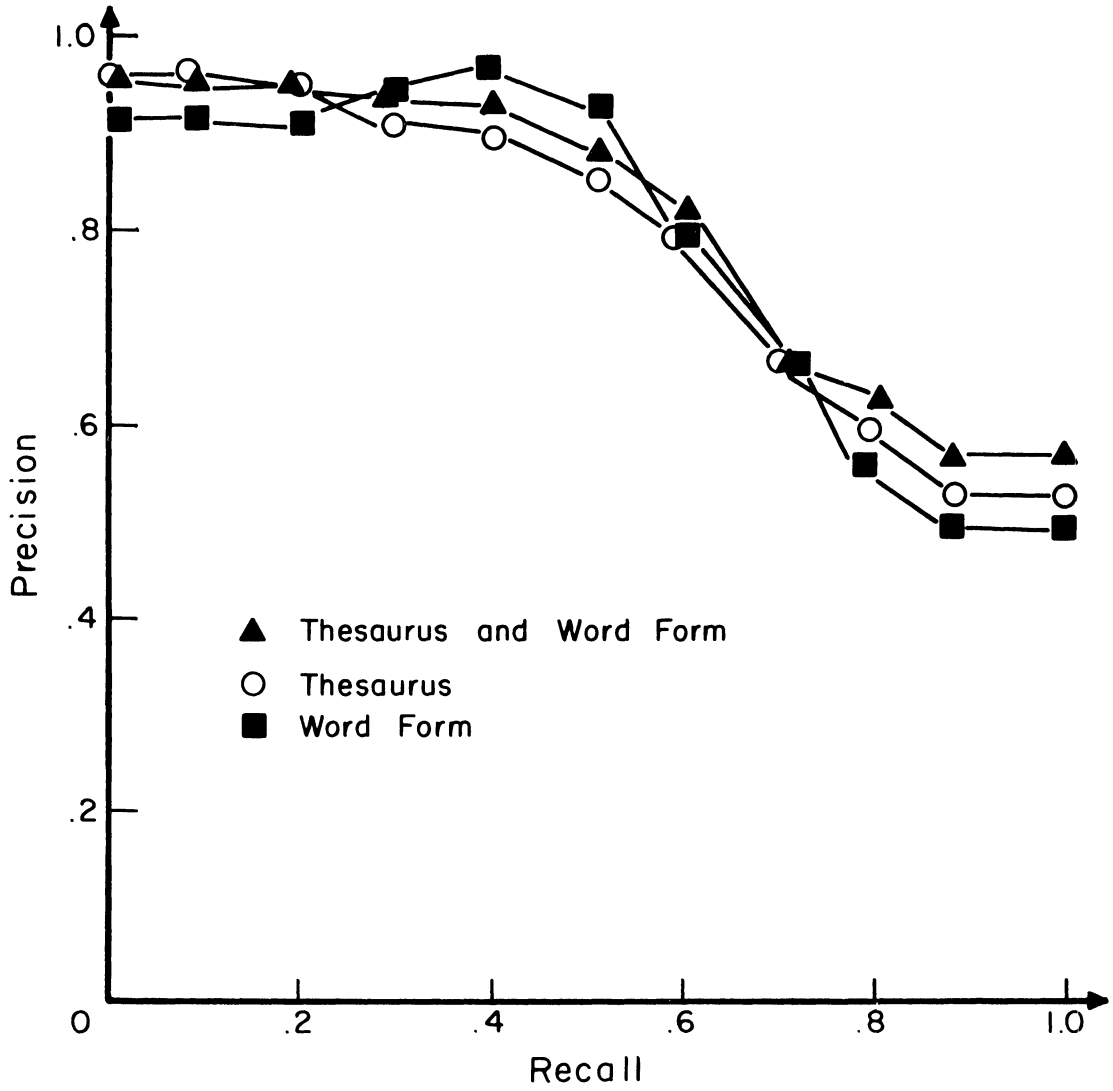
Precision vs. Recall Graphs using
Document Level Averages For Third Iteration

Graph 1



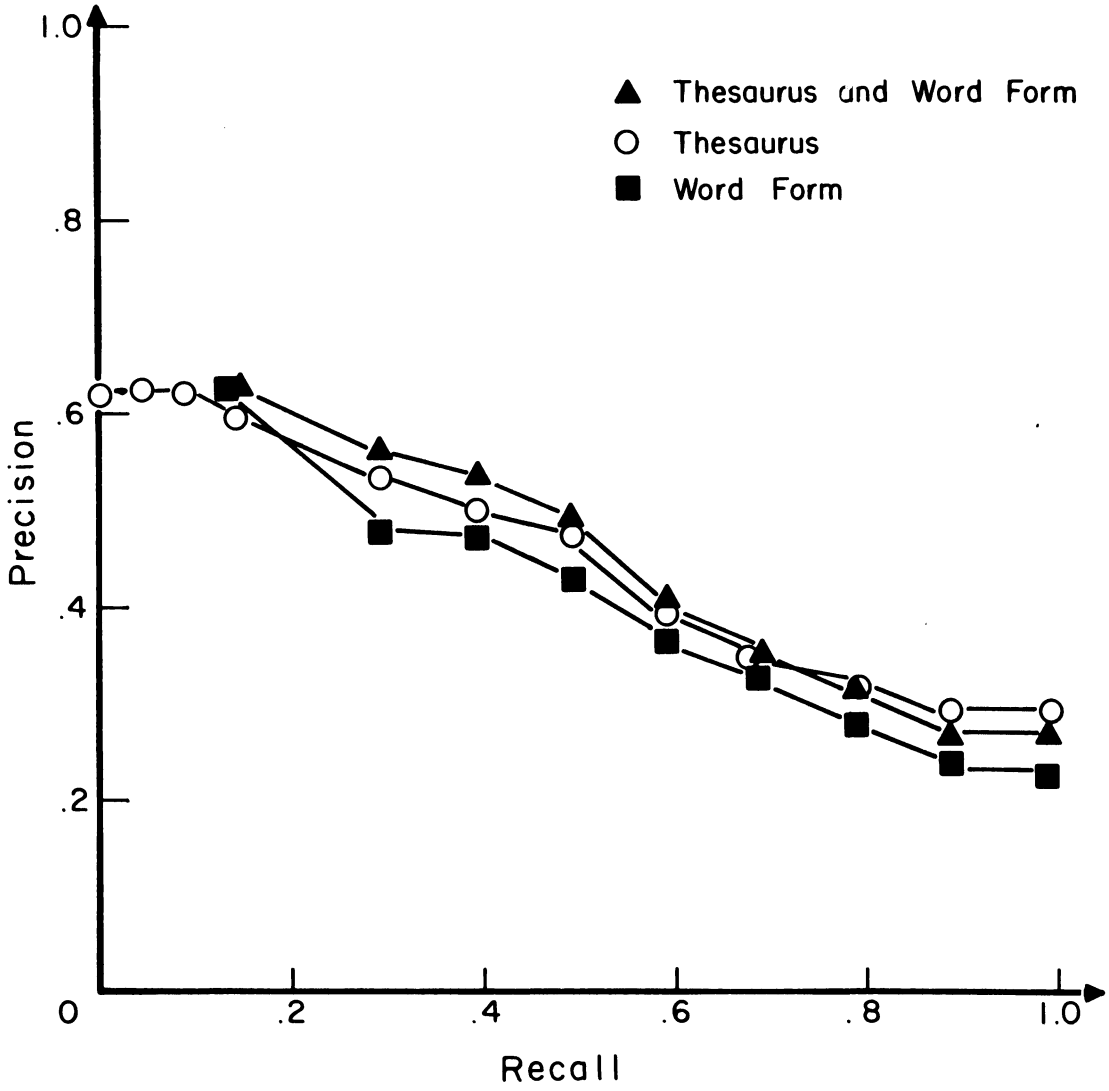
Precision vs. Recall Graphs using
Document Level Averages for "Zero" Iteration
(Full Search)

Graph 2



Precision vs. Recall Graphs using
Recall-Level Averages For Third Iteration

Graph 3



Precision vs. Recall Graphs using
Recall-Level Averages for "Zero" Iteration
(Full Search)

Graph 4

of document 8 is lower than for the other two methods; for document 13, the combined rank is significantly higher than that for the other two dictionaries. The ranks for query 11 are somewhat different:

Relevant Document Number	Combined Rank	Thesaurus Rank	Word Form Rank
16	1	1	63
45	21	23	2
92	41	39	1
44	49	46	50

For this query, the combined rank ignores the word form rank and follows the thesaurus ranking very closely. This seems to lead to problems, since the ranking produced by the combined dictionary does not follow any particular pattern if the thesaurus and word form dictionaries yield radically different results. However, since overall the differences between the rankings by the CRN2S and CRN2TH vectors are very small, these problems do not often occur. A more typical query is number 36:

Relevant Document Number	Combined Rank	Thesaurus Rank	Word Form Rank
34	1	1	2
35	2	5	1
36	3	2	3
37	5	7	5

In this instance, the combined ranking produces a tighter and more accurate ranking than either of the other two.

It is not clear whether the slightly better results of the combined dictionary are worth the extra computer time required. The thesaurus seems to produce almost as good a result, and if cost is important, the thesaurus vector alone might be used.

4. Further Studies

a) The merged dictionary might be improved by the concatenation of all the thesaurus and only low frequency terms of the word form dictionary. The use of low frequency terms would tend to boost the rank of documents which contain relevant, specific terms, thus retrieving those documents relevant to the more important concepts of a query. With the reduction in the amount of word form used, the resulting dictionary would be smaller, and the execution time and cost would be less. Very likely, the results would not be significantly less than that of the full combined dictionary.

b) One might also experiment with weighting either the thesaurus vector or the word form vector. The results, while perhaps not as good as the previous combined dictionary, might at least be more predictable and consistent.

c) Hopefully, if other collections can be located that have both a thesaurus and a word form dictionary, merges can be performed on them to see if different results appear due to a different size or type of collection.

d) Significance tests should be performed on all of these studies.

References

[1] E. M. Keen, Suffix Dictionaries, and Thesaurus, Phrase and Hierarchy Dictionaries, Report ISR-13 to the National Science Foundation, Sections VI and VII, Cornell University, Department of Computer Science, January 1968.