The JHU/APL HAIRCUT System at TREC-8

James Mayfield, Paul McNamee and Christine Piatko
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099 USA
James.Mayfield@jhuapl.edu
Paul.McNamee@jhuapl.edu
Christine.Piatko@jhuapl.edu

Goals

The Johns Hopkins University Applied Physics Laboratory (JHU/APL) is a second-time entrant in the TREC Category A evaluation. The focus of our information retrieval research this year has been on the relative value of and interaction among multiple term types and multiple similarity metrics. In particular, we are interested in examining words and n-grams as indexing terms, and vector models and hidden Markov models as similarity metrics.

Approach

The Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) system was built to explore the use of multiple term types, including words, stemmed words, multi-word phrases and character n-grams of various lengths. The system is implemented in Java for ease of development and portability. It supports both a vector model and a hidden Markov model (HMM) for comparing queries against documents.

Under the vector model, terms are usually weighted by TF/IDF. Okapi BM 25 [Walker *et al.,* 1998] and plain TF weightings are also supported. Cosines can be computed either relative to the origin, or relative to the corpus centroid. Terms that appear in a high percentage of documents are stop-listed. The HAIRCUT HMM is a simple two-state model capturing both document and collection statistics [Miller *et al.,* 1999]. HAIRCUT also supports breaking documents into passages, although to date we have received no significant boost in average precision from doing so.

HAIRCUT performs rudimentary preprocessing on queries to remove stop structure [Allan *et al.,* 1998], *e.g.,* affixes such as "… would be relevant" or "relevant documents should…". Other than this preprocessing, queries are parsed in the same fashion as are documents in the collection.

After the query is parsed and appropriately weighted (TF/IDF for the vector model, TF for the HMM; query section term-weighting was not used), an initial retrieval is performed with a single round of blind relevance feedback. We found a 27.6% relative increase in average precision on the TREC-7 ad hoc task when using relevance feedback. To perform relevance feedback, HAIRCUT first retrieves the top 1000 documents. The top 20 documents are then used for positive feedback, and the bottom 75 documents for negative feedback. Query terms are reweighted using the Rocchio score (α=3, β=2, γ=2) and an affinity score, which is a function of the term's frequency in the retrieved documents and its frequency in the collection as a whole. The top-scoring terms, ignoring very high and very low frequency terms, are then used as the revised query.

After retrieval using this expanded and reweighted query, we have found a slight improvement by penalizing document scores for documents missing many query terms. We multiply document scores by a penalty factor:

$$PF = 1.0 - \left(\frac{\#\ \text{of missing terms}}{\text{number of terms in query}} \right)^{1.25}$$

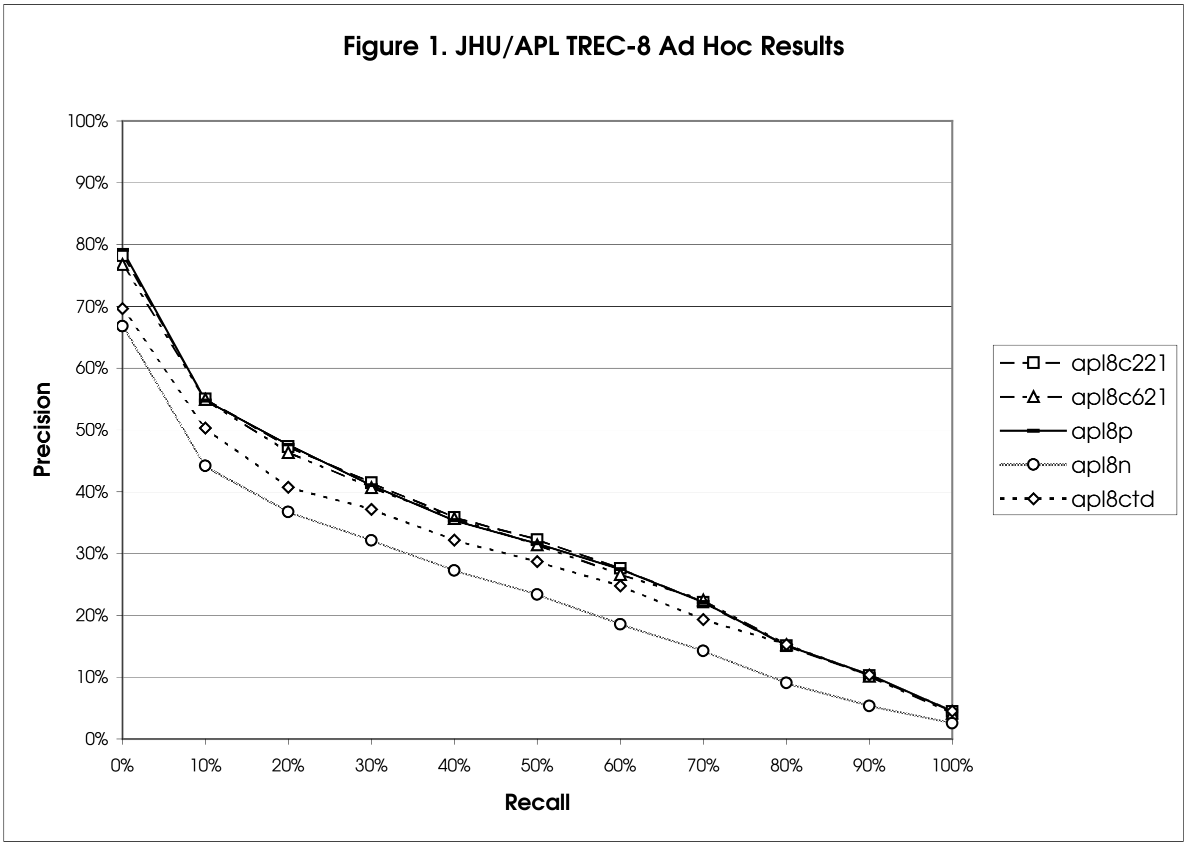
We use only about one-fifth of the terms of the expanded query for this penalty function:

	# Top Terms	*# Penalty Terms*
words	60	12
6-grams	400	75

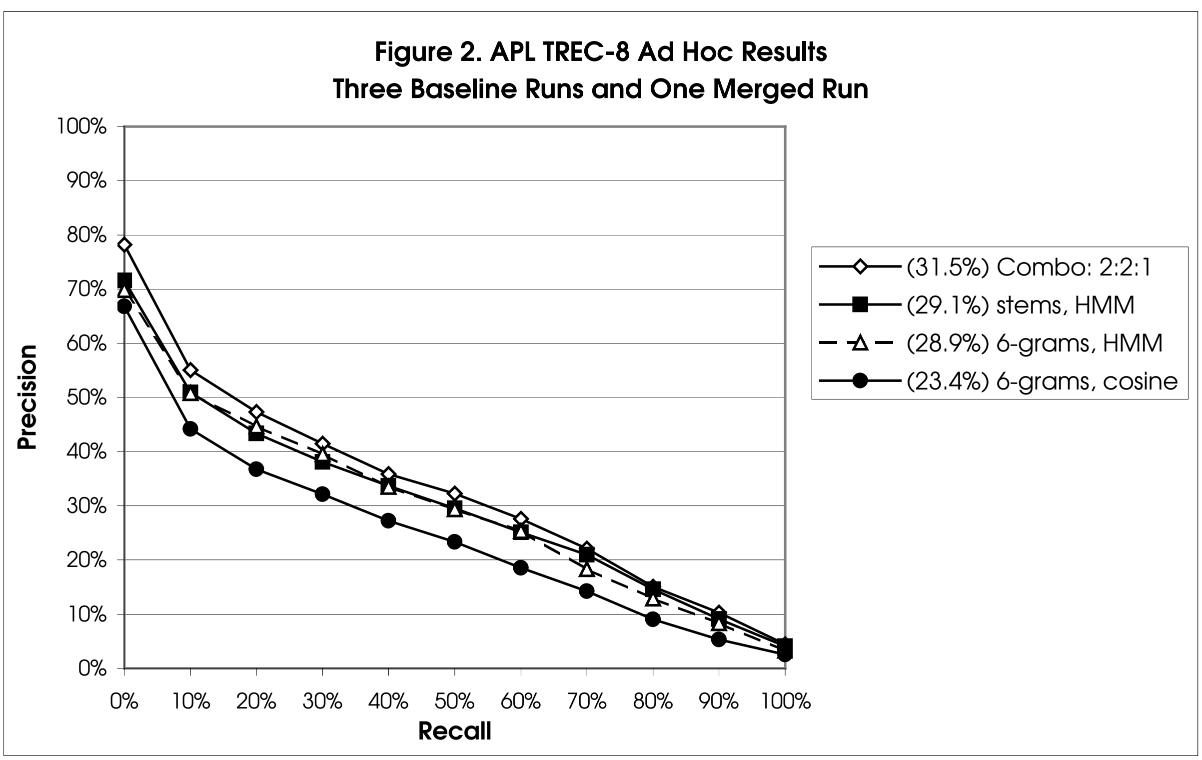
We conducted our work on a shared 4-node Sun Microsystems Ultra Enterprise 450 server. The workstation had 2 GB of physical memory and access to 50 GB of hard disk space.

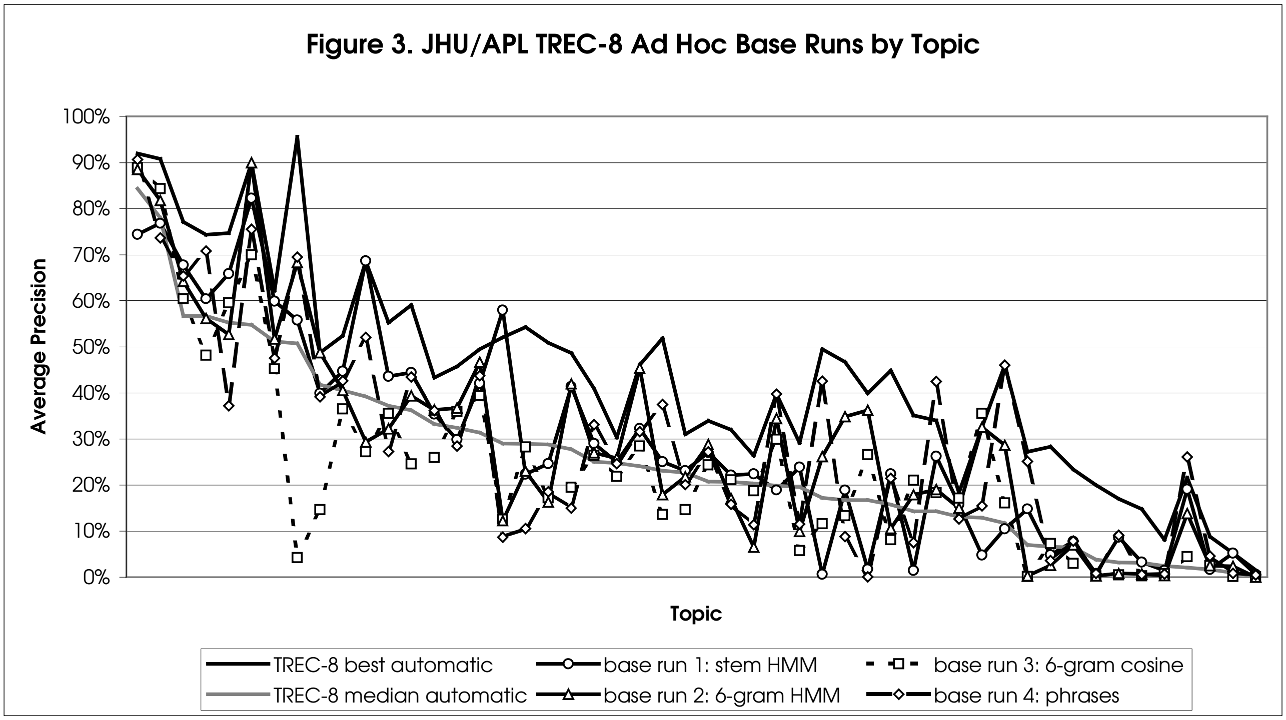
The HAIRCUT system comprises approximately 25,000 lines of Java code.

For TREC-8 we tested three types of terms: stemmed words, 6-grams and phrases. After eliminating



punctuation, downcasing letters, and mapping numbers to a single digit, a word was any remaining blank-delimited sequence of characters. For n-grams we used 6-grams formed from the same character stream used for selecting words. In TREC-7 we used 5-grams [Mayfield & McNamee 1999], but we have found 6-grams to be preferable for English. Our use of 6-grams for other languages is based on convenience, not proven superiority.





Our focus this year was on combining evidence from multiple runs. We varied three system features to obtain runs that were then merged:

- use of a vector model versus use of a hidden Markov model;
- use of n-grams as terms versus use of words or stemmed words as terms; and
- use of phrases.

We used two different methods for merging these runs. In most cases we used linear combination of normalized scores. Scores were normalized simply by scaling the range of scores exhibited by the top 1000 documents to the range 0..1. In one case, we selected the retrieval technique on a per-query basis.

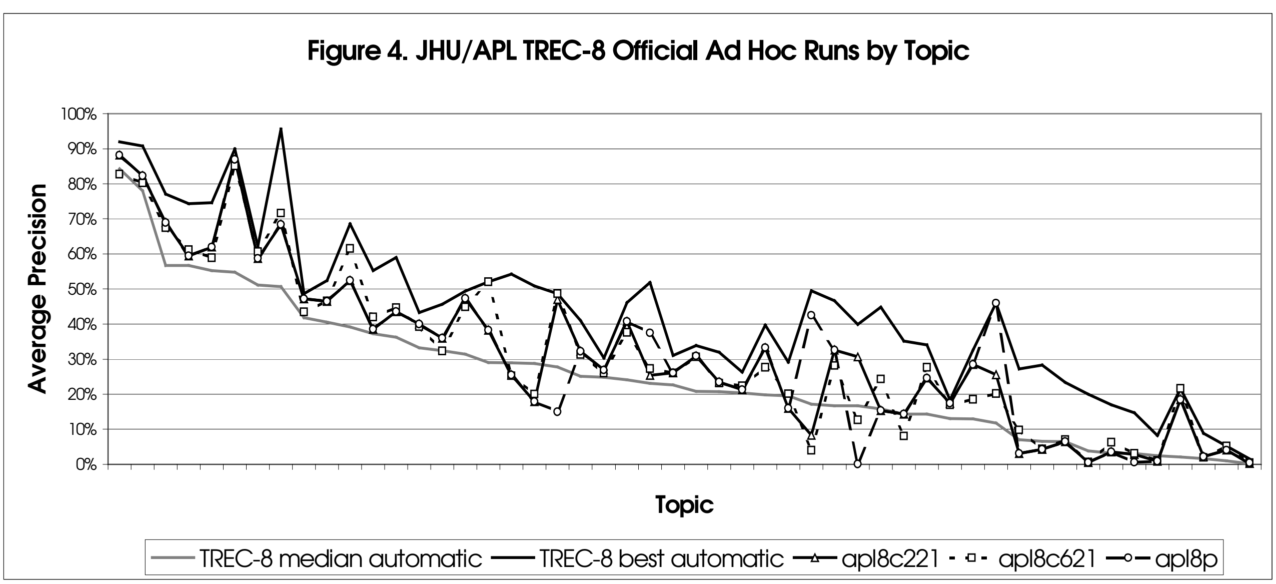
Ad hoc Results

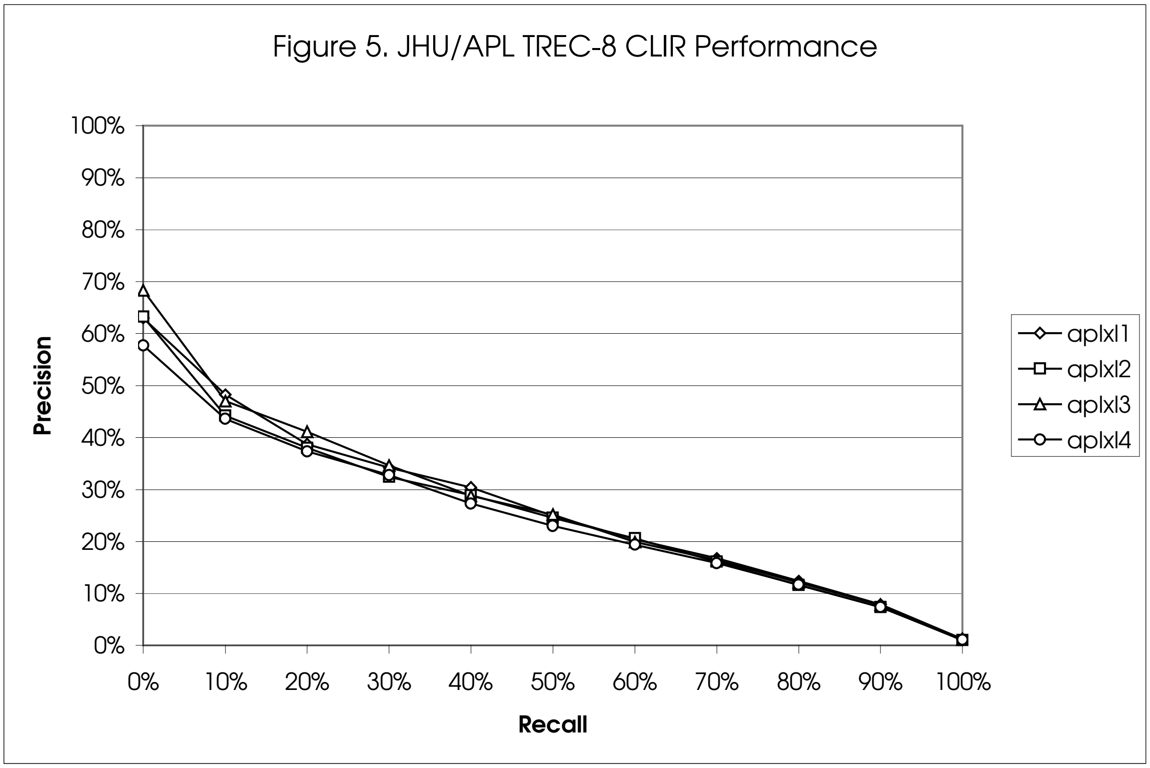
The JHU/APL submissions were based on four underlying retrieval runs:

- **stemhmm:** an HMM run using stemmed words, with α=0.3.
- **sixhmm:** an HMM run using 6-grams, using α=0.15.
- **sixcos:** a vector-based run using 6-grams as terms and TF/IDF term weighting.
- **phrase:** an HMM run using common stem bi-grams, in addition to individual stems, as terms. Our phrase list comprised one million stem pairs.

JHU/APL submitted five ad hoc runs. All underlying runs were first normalized as described above. Four runs were based on topic/description/narrative (TDN); one run, as required by the TREC-8 rules, was based on topic and description (TD) only:

- **apl8c221**: A linear combination of stemhmm, sixhmm, and sixcos, with scores from both stemhmm and sixhmm receiving twice the weight of sixcos.
- **apl8c621**: A linear combination comprising six parts stemhmm, two parts sixhmm, and one part sixcos.





- **apl8p**: A query-by-query combination of apl8c221 and phrase, in which queries that the system judged likely to benefit from phrases were handled exclusively by the phrase run, while the remaining queries were handled exactly as in apl8c221.
- **apl8n**: sixhmm unmodified, submitted as a test of the efficacy of raw n-grams.
- **apl8ctd**: A TD run combining stemhmm, sixhmm, and sixcos in a 2:2:1 ratio.

Our official results for these five runs are shown in Figure 1. Aggregate numbers for the base and submitted runs are as follows:

	Average precision	*Relevant retrieved*	*Precision @100*
stemhmm	0.2914	3156	0.2378
sixhmm	0.2885	3061	0.2436
sixcos	0.2340	2919	0.2106
phrase	0.2850	3153	0.2410
apl8c221	0.3150	3332	0.2558
apl8c621	0.3126	3335	0.2558
apl8p	0.3154	3295	0.2568
apl8n	0.2885	3061	0.2436
apl8ctd	0.2860	3117	0.2324

Our submitted TDN runs exhibited eight of the top scores over the fifty queries; our single TD run exhibited four bests:

	Task pool	*Best*	*At or above median*
apl8c221	automatic TDN	0	40
apl8c621	automatic TDN	4	39
apl8p	automatic TDN	1	39
apl8n	automatic TDN	3	30
apl8ctd	automatic TD	4	37

Figure 2 shows the aggregate improvement obtained by combining three base runs to form apl8c221. Figure 3 shows the average precision for each of the fifty TREC-8 ad hoc topics produced by our four base runs. The chart shows wide variability in the responsiveness of queries to the four techniques. As shown in Figure 4, our linear combination method manages to outperform each of the base methods in aggregate by performing in the upper half of the range of base scores for most queries. Our phrase combination method, also shown in Figure 4, is a first attempt to select a retrieval method on a query-by-query basis.

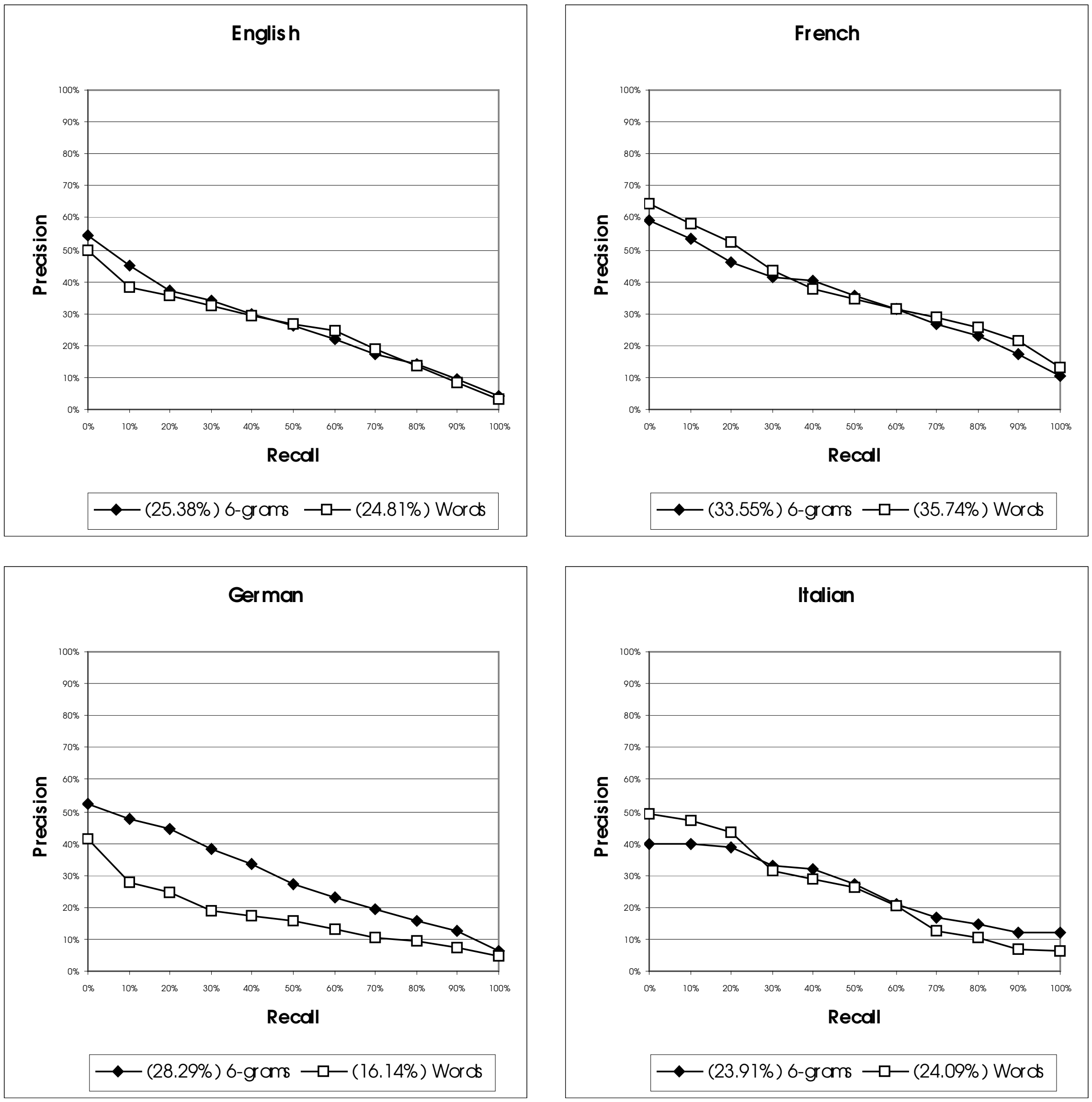
Cross Language Track

JHU/APL was a first-time entrant in the CLIR track at TREC-8. We used SYSTRAN® to translate English queries into German, French and Italian, HAIRCUT to perform both word-based and n-gram-based retrieval on the four collections, and linear combination of normalized scores to combine the eight runs.

We found that different languages responded differently to words and n-grams; thus (except for run aplxl2) *within each language* we used the following weights to combine the two types of runs (which weights were derived from training on the TREC-7 CLIR task):

	English	*French*	*German*	*Italian*
6-grams	1.0	3.3	3.5	1.0
Words	3.0	1.0	1.0	1.2

Figure 6. Words and n-gram base runs for the four CLIR languages.



We submitted four CLIR runs:

1. **aplxl1**: A combination that weighted the languages as follows:

English	*French*	*German*	*Italian*
2.3	2.0	2.0	1.8

These weights were also derived from training on the TREC-7 CLIR task. It is unclear whether these weights optimize average precision on the TREC-7 task because of differing HAIRCUT performance on the various languages, because of the number of relevant documents in the various languages, or because of other factors.

2. **aplxl2**: A two-phase combination, in which the languages were first individually combined using the following weights:

	English	*French*	*German*	*Italian*
6-grams	1.0	2.5	8.5	1.0
Words	6.0	1.0	1.0	1.5

The resulting combined runs were then merged using these weights:

English	*French*	*German*	*Italian*
2.0	1.1	1.1	0.6

As with aplxl1, these weights were trained from the TREC-7 CLIR task.

3. **aplxl3**: A combination that weighted the four languages evenly:

English	*French*	*German*	*Italian*
1.0	1.0	1.0	1.0

4. **aplxl4**: A combination that weighted English as 1.5, and the other languages as 1.0:

English	*French*	*German*	*Italian*
1.5	1.0	1.0	1.0

Results from these four runs are shown in Figure 5. The eight base runs are shown in Figure 6. Aggregate numbers for our submitted CLIR runs are as follows:

	Average precision	*Relevant retrieved*	*Precision @100*
aplxl1	0.2571	1911	0.2714
aplxl2	0.2471	1944	0.2782
aplxl3	0.2542	1890	0.2711
aplxl4	0.2389	1902	0.2575

The similarity of these runs lead us to believe that tuning weights under this approach to merging is not a cost-effective use of one's time.

Our monolingual results for the four languages, using the human-translated queries provided by NIST, were significantly below those seen on the TREC-7 CLIR task:

	English	*French*	*German*	*Italian*
T8 Words	0.2481	0.3780	0.2525	0.3152
T8 6-grams	0.2538	0.3342	0.3933	0.2778
T7 Words	0.4533	0.3715	0.3671	0.3936
T7 6-grams	0.4363	0.3767	0.4143	0.3281

The reasons for this drop in performance are unclear. Given HAIRCUT's average precision of 31.5% on the TREC-8 ad hoc task, it seems unlikely that the drop in English performance is due to some sort of overtraining on the TREC-7 data. We note a significant drop in the number of relevant documents for the TREC-8 CLIR task, which may play a role:

	English	*French*	*German*	*Italian*
TREC-7	1689	991	917	501
TREC-8	956	578	717	170

Query Track

JHU/APL also participated in the query track. We generated four query sets, although space aliens in black helicopters managed to prevent two of them from appearing in the official query track collection. These latter two were generated by hand, by reading the narrative version of each source query and generating a title query and a description query for each. Our first official query set (APL5a) was created using a variant of the mutual information statistic [Church & Hanks 1990] to extract important terms from the top 75 documents retrieved for the source query. Our second set (APL5b) used the same statistic to extract important terms from the query track training set. All terms in these query sets were unstemmed words; we did not anticipate that other systems could make use of n-grams.

We used a single system configuration to process each of the 23 query track query sets. This configuration used unstemmed words as terms, and a hidden Markov model to gauge document similarity.

Our results showed tremendous variability in result quality across the 23 query sets. The following table shows HAIRCUT's performance on each query set, the average performance over all eight runs submitted by five different groups to the query track, and HAIRCUT's percentage above or below the all systems average. Our best results were obtained from APL5b, which was developed using training data. For further details on the query track, see the Query Track Overview paper in these proceedings.

	HAIRCUT Average precision	*Rel. ret.*	*Prec. @100*	*All systs. average prec.*	*Diff. in avg.*
acs1a	0.2521	5553	0.3316	0.2449	2.9%
APL5a	0.2471	5648	0.3300	N/A	N/A
APL5b	0.2912	6242	0.4192	N/A	N/A
INQ1a	0.1697	4261	0.2242	0.1771	-4.2%
INQ1b	0.2089	4902	0.2642	0.2124	-1.6%
INQ1c	0.2261	5195	0.2934	0.2259	0.1%
INQ1d	0.2061	4901	0.2788	0.1919	7.4%
INQ1e	0.2300	4480	0.2796	0.2269	1.4%
INQ2a	0.1622	4201	0.2180	0.1612	0.6%
INQ2b	0.1990	4673	0.2586	0.1869	6.5%
INQ2c	0.2658	5427	0.3386	0.2407	10.4%
INQ2d	0.2208	5367	0.2862	0.2022	9.2%
INQ2e	0.2533	5440	0.3222	0.2388	6.1%
INQ3a	0.1486	3725	0.2050	0.1320	12.6%
INQ3b	0.1292	3766	0.1980	0.1182	9.3%
INQ3c	0.1293	3430	0.1716	0.1299	-0.5%
INQ3d	0.1601	4220	0.2078	0.1578	1.5%
INQ3e	0.1758	4470	0.2382	0.1754	0.2%
pir1a	0.2603	5999	0.3590	0.2464	5.6%
Sab1a	0.2550	5610	0.3312	0.2384	7.0%
Sab1b	0.2791	5946	0.3636	0.2504	11.5%
Sab1c	0.2405	5555	0.3040	0.2533	-5.1%
Sab3a	0.2627	5816	0.3558	0.2364	11.1%

Conclusions

We had good results using a simple linear combination of scores across several HAIRCUT runs. In general, HMMs outperformed the vector model, while n-grams and words were roughly comparable on average. Using run combination, HAIRCUT sees an 8% relative gain on the ad hoc task over the best base run. Such combination is low risk; we have never found a drop in average precision over fifty or more queries. Furthermore, effective combination does not require disparate systems. A single system can produce the required variability simply by using different term types or similarity metrics.

Availability of both words and n-grams also helped us significantly in the cross-language task, for which

HAIRCUT was a first-time participant. HAIRCUT exhibited 79% recall at 1000 on the CLIR task, and a high average precision relative to retrieval using human-translated queries.

References

[Allan *et al.*, 1998] James Allan, Jamie Callan, W. Bruce Croft, Lisa Ballesteros, Don Byrd, Russell Swan, and Jinxi Xu, 'INQUERY does battle with TREC-6.' In E. M. Voorhees and D. K. Harman, eds., *The Sixth Text REtrieval Conference (TREC-6).* NIST Special Publication 500-240, pp. 169-206, 1998.

[Church & Hanks 1990] Kenneth Ward Church and Patrick Hanks, 'Word association norms, mutual information, and lexicography.' *Computational Linguistics* 16(1):22-29, 1990.

[Mayfield & McNamee 1999] James Mayfield and Paul McNamee, 'Indexing using both n-grams and words.' In E. M. Voorhees and D. K. Harman, eds., *The Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-242, pp. 419-423, 1999.

[Miller *et al.*, 1999] David R. H. Miller, Tim Leek and Richard M. Schwartz, 'BBN at TREC7: Using Hidden Markov Models for Information Retrieval.' In E. M. Voorhees and D. K. Harman, eds., *The Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication 500-242, pp. 133-142, 1999.

[Walker *et al.*, 1998] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones and Karen Sparck Jones, 'Okapi at TREC-6, Automatic ad hoc, VLC, routing, filtering and QSDR.' In E. M. Voorhees and D. K. Harman, eds., *The Sixth Text REtrieval Conference (TREC-6).* NIST Special Publication 500-240, pp. 125-136, 1998.