Novel Results and Some Answers

The University of Iowa TREC-11 Results

David Eichmann[1,2] and Padmini Srinivasan[1,3]

[1]School of Library and Information Science
[2]Computer Science Department
[3]Department of Management Sciences
The University of Iowa
{david-eichmann, padmini-srinivasan}@uiowa.edu

The University of Iowa participated in the novelty, adaptive filtering and question answering tracks of TREC-11. The filtering system used was an extension of the one used in TREC-7 [1] and TREC-8 [2]. Question answering was derived from the TREC-10 system. The novelty system was new…

1 – Novelty

With novelty being a new track, we had little prior experience to build upon. Hence we decided to begin our development experiments with a simple similarity match between the topic definition and the candidate sentence. For the available training topics, this proved to be remarkably responsive to tuning between precision-focused runs and recall-focused runs for novelty as well as the more predictable relevance decision. Figure 1 shows relevance tuning runs and Figure 2 new tuning runs. We used these runs as a baseline for our subsequent experiments in both threshold tuning

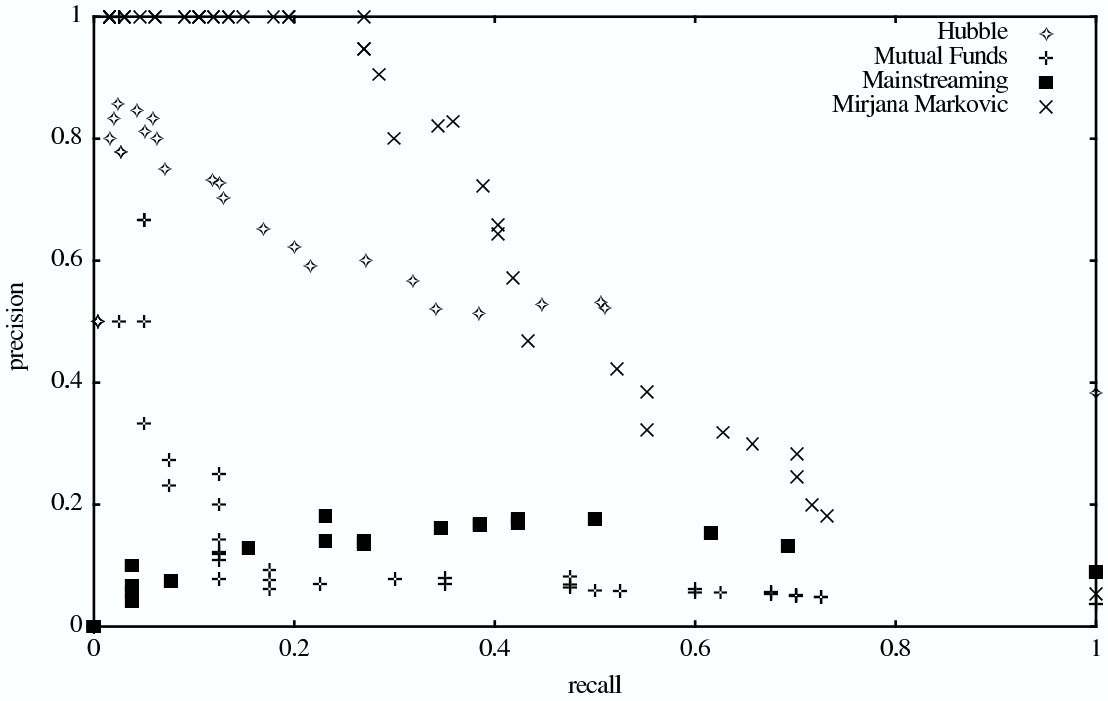

Figure 1: Training on Relevance, Simple Similarity

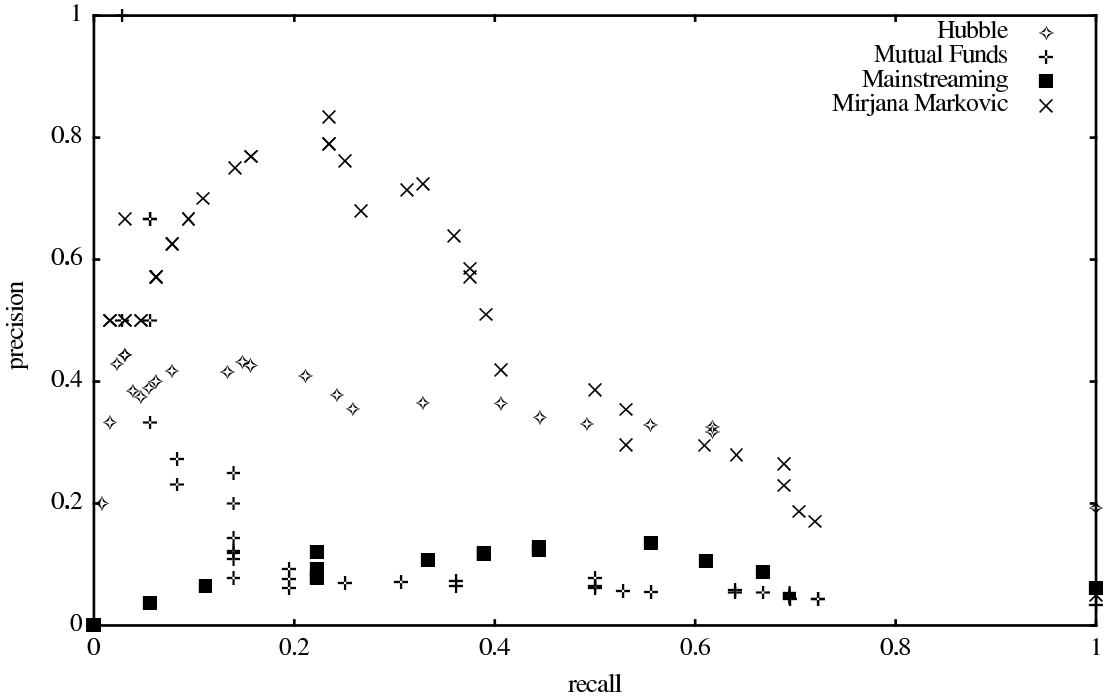

Figure 2: Training on New, Simple Similarity

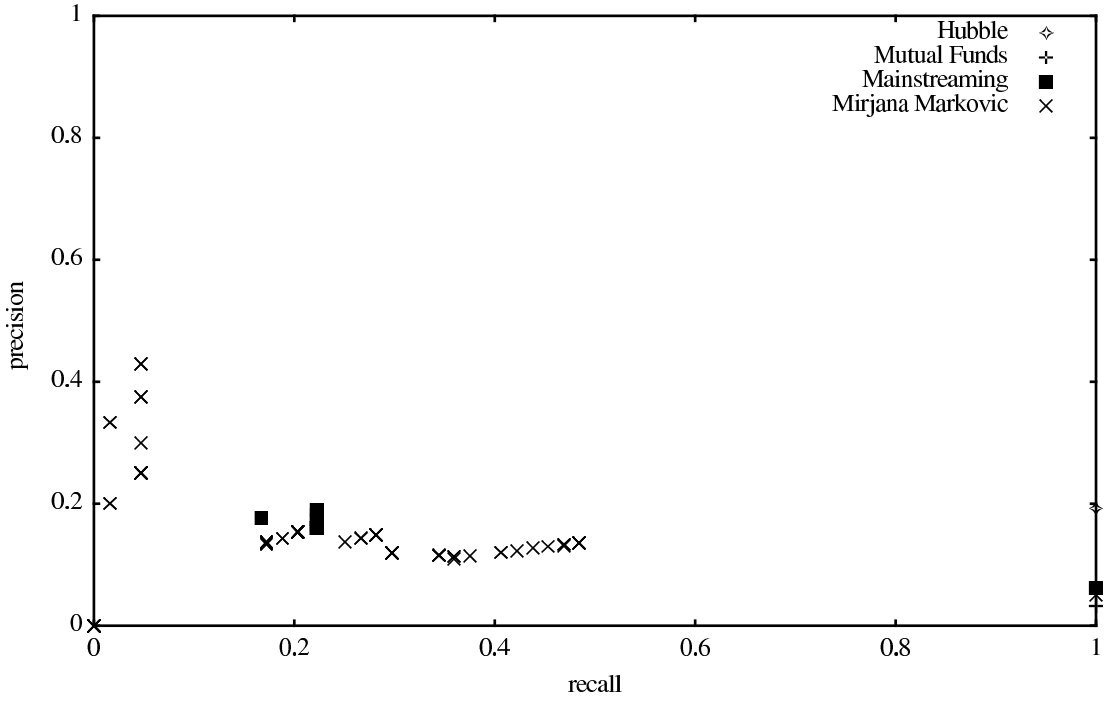

Figure 3: Training on New, New Entity Occurrence

and use of newly detected noun phrases and named entities. We first tried configurations that used single sources of similarity – e.g., just entities or just noun phrases. As shown in Figure 3 and Figure 4 for entities and noun phrases, respectively, this fared poorly

These preliminary experiments resulted in three different configurations:

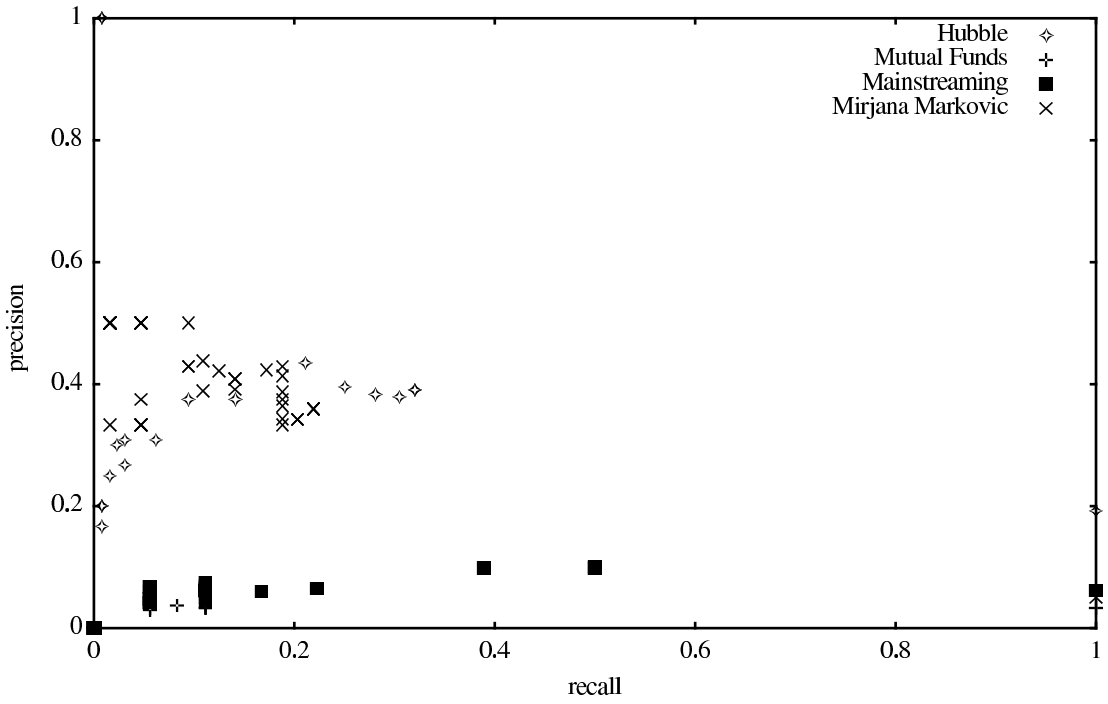

Figure 4: Training on New, New Noun Phrase Occurrence

1. simple stemmed term similarity, threshold-tunable for a range of precision/recall performance
2. new noun phrase triggering, guarded by a dual threshold of sentence similarity and full-document similarity. If the full document is sufficiently similar and the current sentence is sufficiently similar, the number of newly-detected noun phrases is compared against a minimum threshold and if the minimum is met, the current sentence is declared to be novel. Relevance in this configuration is driven by simple term similarity.
3. new named entity and noun phrase triggering, guarded by a dual threshold of sentence similarity and full-document similarity. If the full document is sufficiently similar and the current sentence is sufficiently similar, the number of newly-detected named entities and noun phrases is compared against a minimum threshold and if the minimum is met, the current sentence is declared to be novel. The named entities used in this configuration include persons, organizations and place names. Relevance in this configuration is driven by simple term similarity.

Configurations 2 and 3 hence are tunable both on the guard similarity thresholds and minimum threshold. Experiments with the training data indicated that a single new entity or noun phrase was sufficient for plausible performance against a range of guard thresholds for precision/recall tradeoff. Tuning runs on the configuration 3 yielded Figure 5.

Our official submissions included a run targeting a balance of precision and recall and a run targeting precision only. As shown in Tables 1 and 2, the runs ended up not that different in overall distribution of topic performance, but substantially better than random, as reported at the conference.*

* The track results reported elsewhere in this proceedings uses F = 2RP / (R+P), rather than F = R*P as used at the conference and in this paper.

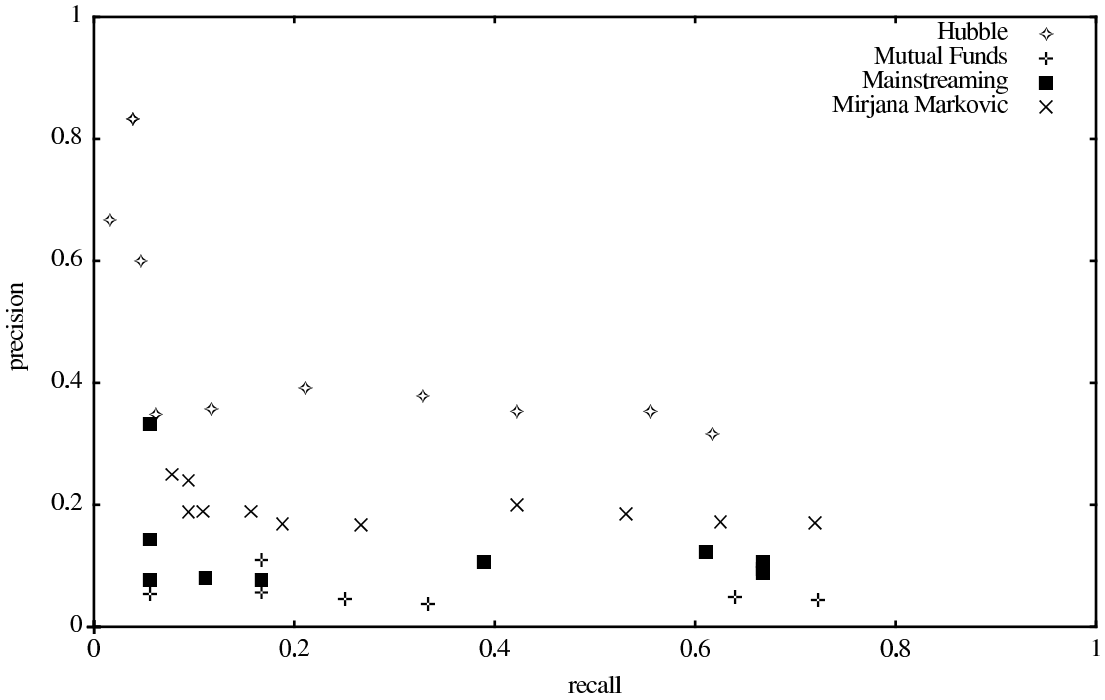

Figure 5: Training on New, New Entities & NPs with Guard

Table 1: Official Novelty Results, Relevant

	Precision & Recall Run	Precision Run	Human	Random
Ave. Prec.	0.12	0.14		
Ave. Recall	0.58	0.36		
Average P*R	0.073	0.059	0.191	0.006

Table 2: Official Novelty Results, New

	Precision & Recall Run	Precision Run	Human	Random
Ave. Prec.	0.12	0.14		
Ave. Recall	0.37	0.12		
Average P*R	0.048	0.020	0.170	0.004

Plotting the official runs' precision and recall, as shown for relevance in Figure 6 and novelty in Figure 7 shows that our tuning for 'precision' on the training data failed to improve precision on the test data to any noticeable degree, while substantially lowering recall. We plan to run further experiments to see if this is just a matter of 'failing to turn the knob sufficiently.'

Our overall impressions regarding novelty as an information retrieval task involve interesting inversions of expectations compared to relevance. For instance, in relevance, polysemy typically leads to false alarms and synonymy leads to misses. For novelty, our conjecture for further research

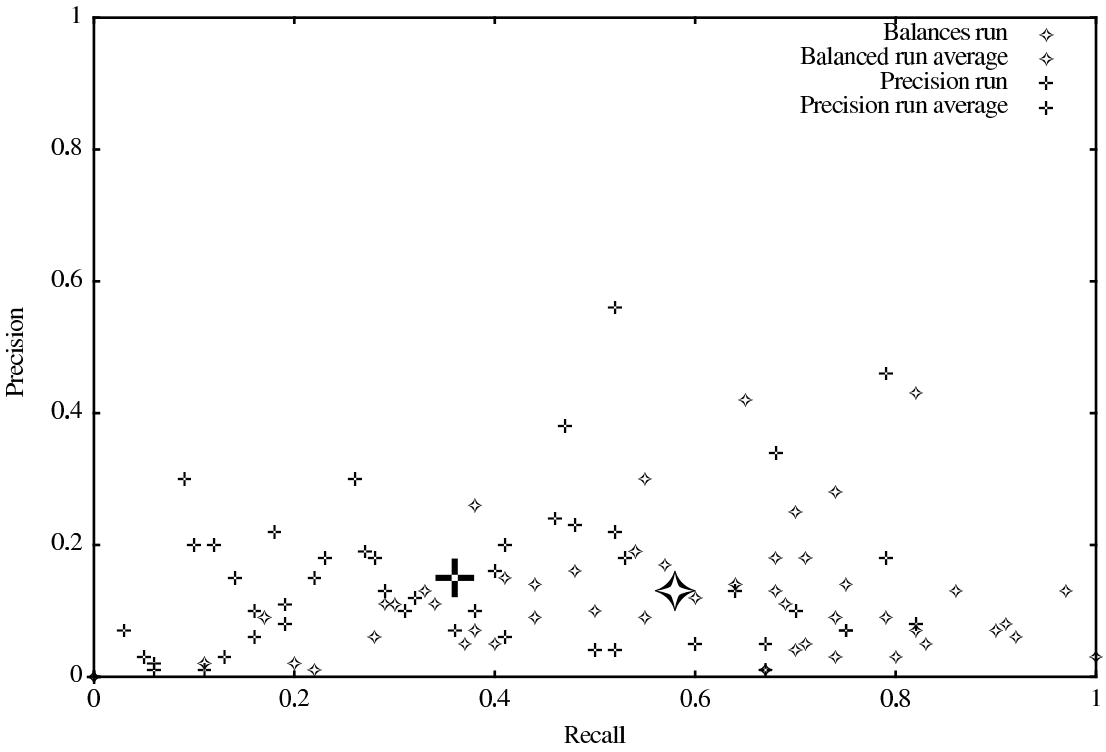

Figure 6: Official Relevance Runs, Precision vs. Recall, by Topic

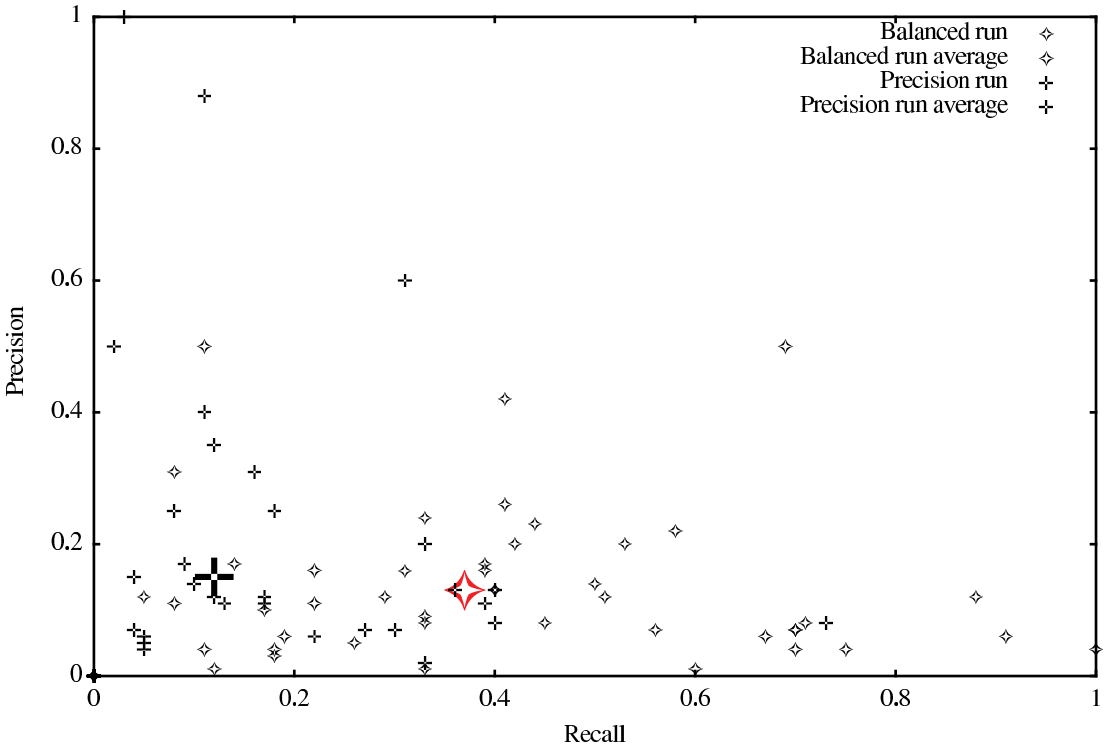

Figure 7: Official Novelty Runs, Precision vs. Recall, by Topic

involves polysemy leading to misses and synonymy leading to false alarms. Tuning our approach also clearly requires additional work, given the substantial differences in behavior between the training and test runs.

2 – Question Answering

Our system for QA is a complete overhaul of one of the two systems used for our previous TREC runs. We first used the previous year's questions to improve our classification of a question into one of a relatively small number of categories (quantity, person/organization, location, etc.) with testing achieving ~92% accuracy. Once the questions are classified, they are parsed using the CMU link grammar parser to extract subject/verb/object structure.

Each document in the corpus is then decomposed into doc-id / sentence pairs, with the sentence being the unit of analysis from that point. Each sentence is then POS-tagged and fed to the grammar parser. The parse tree for the sentence is then attributed with the POS tags for each word. Processing both queries and documents using this scheme allows us to establish both the nature of the query (using a fairly typical taxonomy) and the nature of the needed answer. This is particularly useful with respect to identification of candidate phrases in sentences and scoring of these phrases against the goal of the query.

The availability of the parse tree for the phrase allows for elision of subordinate clauses that can cause answers to span too long a string and for extraction of likely answers through heuristic matching of, for example, a subordinate clause immediately trailing a mention of a candidate named entity.

Table 3: Question Answering Results

	Best Single Score	Summed Scores	Response Frequency
Unsupported	0	2	4
Inexact	6	3	5
Right	18	21	15
Score	0.055	0.042	0.023

3 – Adaptive Filtering:

Our approach to filtering involves a two-level dynamic clustering technique. Each filtering topic is used to create a primary cluster that forms a general profile for the topic. Documents that are attracted into a primary cluster participate in a topic-specific second level clustering process yielding what we refer to as secondary clusters. These secondary clusters, depending upon their status, are responsible for declaring, i.e., retrieving, documents for the topic.

As documents are temporally processed they are attracted to a primary cluster if their similarity with the cluster vector is above a primary threshold. These documents enter the secondary clustering stage where again, based on similarity to cluster vectors and a secondary threshold, they either join an existing secondary cluster or start a new one. If at some point the similarity between a sec-

ondary cluster and the primary cluster exceeds a third declaration threshold then the document most recently added to the secondary cluster is retrieved for the user.

When deriving representations we use TF*IDF weights after stemming the terms using Porter's stemmer. We also limit document vectors and cluster vectors to the best 100 and 200 stems respectively.

References

[1] Eichmann, D., M. E. Ruiz and P. Srinivasan, "Cluster-Based Filtering for Adaptive and Batch Tasks," *Seventh Conference on Text Retrieval,* NIST, Washington, D.C., November 11 - 13, 1998.

[2] Eichmann, D. and P. Srinivasan, "Filters, Webs and Answers: The University of Iowa TREC-8 Results," *Eighth Conference on Text Retrieval,* NIST, Washington, D.C., November, 1999.