System Summary and Timing
Organization Name: Xerox PARC
List of Run ID's: Xerox1, Xerox2, Xerox3, Xerox4

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 951
    - Stemming Algorithm: Two-level Finite-state lexicon
      - Morphological Analysis: yes
    - Term Weighting: applied at run-time, typically tf.idf
    - Phrase Discovery? : yes
      - Kind of Phrase: word pairs
      - Method Used (statistical, syntactic, other): statistical
    - Tokenizer? : Rule-based Finite-state
      - Patterns which are tokenized: dates, numbers, abbreviations, words, model numbers
    - Other Techniques for building Data Structures: automatically constructed word thesaurus from analysis of lexical cooccurrence events; uses SVD matrix decomposition to reduce dimensionality
  - Statistics on Data Structures built from TREC Text
    - Inverted index
      - Run ID : Xerox1, Xerox2, Xerox3, Xerox4
      - Total Storage (in MB): 1200
      - Total Computer Time to Build (in hours): 120
      - Automatic Process? (If not, number of manual hours): yes
      - Use of Term Positions? : yes
      - Only Single Terms Used? : no
    - Special Routing Structures
      - Run ID : Xerox1, Xerox2
      - Type of Structure: Neural net classifier
      - Total Storage (in MB): less than 1
      - Total Computer Time to Build (in hours): 100
      - Automatic Process? (If not, number of manual hours): yes
      - Brief Description of Method: Used standard back-propagation neural net training method
    - Other Data Structures built from TREC text
      - Run ID : xerox1, xerox2, xerox3, xerox4
      - Type of Structure: Document profiles
      - Total Storage (in MB): 700
      - Total Computer Time to Build (in hours): 120
      - Automatic Process? (If not, number of manual hours): yes
      - Brief Description of Method: These are parsed, tokenized, but uninverted document representation that speed subsequent computations
    - Other Data Structures built from TREC text

            - Run ID : xerox1, xerox2, xerox3, xerox4
            - Type of Structure: Word thesaurus vectors
            - Total Storage (in MB): 80
            - Total Computer Time to Build (in hours): 24
            - Automatic Process? (If not, number of manual hours): yes
            - Brief Description of Method: One vector per unique word that captures the cooccurrence patterns for that word. Used to define similarity between words
        * Other Data Structures built from TREC text
            - Run ID : xerox1, xerox2, xerox3, xerox4
            - Type of Structure: Document segments: TextTiles
            - Total Storage (in MB): 20
            - Total Computer Time to Build (in hours): 24
            - Automatic Process? (If not, number of manual hours): yes
            - Brief Description of Method: Segment long texts into topic coherent multi-paragraph sections

- Query construction
    - Automatically Built Queries (Ad-Hoc)
        * Topic Fields Used: yes
        * Average Computer Time to Build Query (in cpu seconds): less than 1
        * Method used in Query Construction
            - Term Weighting (weights based on terms in topics)? : yes
            - Phrase Extraction from Topics? : yes
            - Tokenizer? : rule-based finite-state
                I. Patterns which are Tokenized: dates, numbers, abbreviations, words, model numbers
            - Expansion of Queries using Previously-Constructed Data Structure? :
                I. Structure Used: word thesaurus vectors
    - Automatically Built Queries (Routing)
        * Topic Fields Used: yes
        * Average Computer Time to Build Query (in cpu seconds): 120
        * Method used in Query Construction
            - Terms Selected From
                I. Topics: yes
                II. All Training Documents: yes
            - Term Weighting with Weights Based on terms in
                I. Topics: yes
                II. All Training Documents: yes
            - Phrase Extraction from
                I. Topics: yes
                II. All Training Documents: yes
            - Tokenizer
                I. Patterns which are tokenized (dates, phone numbers, common patterns, etc): dates, numbers, abbreviations, words, model number

II. from Topics: dates, numbers, abbreviations, words, model number

III. from All Training Documents: dates, numbers, abbreviations, words, model number

IV. from Documents with Relevance Judgments: dates, numbers, abbreviations, words, model number

· Expansion of Queries using Previously-Constructed Data Structure:
I. Structure Used: word thesaurus vectors

· Other: query terms were partitioned into topic-cherent groups called query factors. These factors were scored separately over documents and document segments, or TextTiles

- Searching
  - Search Times
    - Run ID : xerox1, xerox2
    - Computer Time to Search (Average per Query, in CPU seconds): 150
    - Component Times : build neural net input representation, score against trained neural net

    - Run ID : xerox3, xerox4
    - Computer Time to Search (Average per Query, in CPU seconds): 60
    - Component Times : score and filter document used standard tf-idf method, apply probabilistic scoring to reorder documents
  - Machine Searching Methods
    - Vector Space Model? : yes, for initial filtering
    - Probabilistic Model? : yes, to reorder results
    - Fuzzy Logic? : yes
    - Neural Networks? : yes, for routing only
  - Factors in Ranking
    - Term Frequency? : yes
    - Inverse Document Frequency? : yes
    - Other Term Weights? : TextTile scores, from document segments and Factor scores from query factors
    - Semantic Closeness? : yes, based on word thesaurus vectors
    - Proximity of Terms? : yes, though phrases
    - Document Length? : yes
    - Percentage of Query Terms which match? : yes
    - Other: We tested various combinations of whole query vs query factors vs whole documents vs document TextTiles to generate scores. For example, one predictor scored documents according to the min score across factor of their maximum score across a document's TextTiles
  - Machine Information
    - Machine Type for TREC Experiment: Sun Sparc Station 10
    - Was the Machine Dedicated or Shared: Dedicated
    - Amount of Hard Disk Storage (in MB): 7000

    * Amount of RAM (in MB): 100
    * Clock Rate of CPU (in MHz): 50
- System Comparisons
    * Amount of "Software Engineering" which went into the Development of the System: variable. Some parts are the result of considerable engineering, other parts were custom for TREC 3
    * Given appropriate resources
        · Could your system run faster? : yes
        · By how much (estimate)? : 100