System Summary and Timing
Organization Name: ETH Zurich, Switzerland
List of Run ID's: ETH001, ETH002 (adhoc), ETH003, ETH004 (routing)

- Construction of Indices, Knowledge Bases, and other Data Structures 011
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 571
    - Stemming Algorithm: suffix stripping (Porter, 1980) 011
    - Term Weighting: (1+log(ff(phi_i,d_j))) (so-called lnc)
    - Phrase Discovery? : yes 011
      - Kind of Phrase: two words without separator (punctuation etc.) in between
      - Method Used (statistical, syntactic, other): weak syntax rules
    - Tokenizer? : yes
      - Patterns which are tokenized: words and phrases
  - Statistics on Data Structures built from TREC Text
    - Other Data Structures built from TREC text
      - Run ID : link method
      - Type of Structure: links and files for direct access to ff of each document; link: phrase and a document list where it is
      - Total Storage (in MB): 670
      - Total Computer Time to Build (in hours): ca. 48h
      - Automatic Process? (If not, number of manual hours): yes
      - Brief Description of Method: extract phrases from documents; invert them; cut all phrases that are in less than three documents or in more than 7000 (about 1phrases as 'UNITED STATES' are cut this way) of all documents; cut all phrases that have a weight less than a limit c
    - Other Data Structures built from TREC text
      - Run ID : passage retrieval method
      - Type of Structure: token files
      - Total Storage (in MB): 1723
      - Total Computer Time to Build (in hours): ca. 14h
      - Automatic Process? (If not, number of manual hours): yes
      - Brief Description of Method: extract features from TIPSTER disks, delete stopwords
  - Data Built from Sources Other than the Input Text
    - Internally-built Auxiliary File
      - Domain (independent or specific): specific, basic method
      - Type of File (thesaurus, knowledge base, lexicon, etc.): mapping of features to a number
      - Total Storage (in MB): 10
      - Number of Concepts Represented: 500'000
      - Type of Representation: hash table
      - Total Computer Time to Build (in hours): ca. 10

- Total Manual Time to Build (in hours): no manual time
  - Internally-built Auxiliary File
    - Domain (independent or specific): specific, passage retrieval method
    - Type of File (thesaurus, knowledge base, lexicon, etc.): HMM
    - Total Storage (in MB): 6KB
    - Number of Concepts Represented: 2
    - Type of Representation: HMM
    - Total Computer Time to Build (in hours): 120
    - Total Manual Time to Build (in hours): no manual time

- Query construction
  - Automatically Built Queries (Ad-Hoc)
    - Topic Fields Used: all
    - Average Computer Time to Build Query (in cpu seconds): basic method: msec link method: 60 passage retrieval method: msec (same query as basic method)
    - Method used in Query Construction
      - Term Weighting (weights based on terms in topics)? : (1+log(ff(phi_i,q)))*nidf(phi_i) (so-called ltn)
      - Tokenizer? : yes
        I. Patterns which are Tokenized: words and phrases
      - Expansion of Queries using Previously-Constructed Data Structure? : yes
        I. Structure Used: links
  - Automatically Built Queries (Routing)
    - Topic Fields Used: all
    - Average Computer Time to Build Query (in cpu seconds): basic method: msec link method: 60 passage retrieval method: msec
    - Method used in Query Construction
      - Terms Selected From
        I. Topics: 101-150
        II. All Training Documents: yes
      - Term Weighting with Weights Based on terms in
        I. Topics: 101-150
        II. All Training Documents: yes
      - Tokenizer
        I. Patterns which are tokenized (dates, phone numbers, common patterns, etc): words and phrases
        II. from Topics: 101-150
        III. from All Training Documents: yes
      - Expansion of Queries using Previously-Constructed Data Structure:
        I. Structure Used: links

- Searching

- Search Times
  * Run ID : ETH001 (adhoc), ETH003 (routing)
  * Computer Time to Search (Average per Query, in CPU seconds): msec
  * Component Times :
    · basic method: msec
    · link method: msec
  * Run ID : ETH002, ETH004
  * Computer Time to Search (Average per Query, in CPU seconds): 600 seconds
  * Component Times :
    · basic method: msec
    · link method: msec
    · passage retrieval: 600 seconds
  * Machine Searching Methods
    · Vector Space Model? : yes
    · Other: yes - HMM based passage extraction
- Factors in Ranking
  * Factors in Ranking
    · Term Frequency? : yes
    · Inverse Document Frequency? : yes
    · Position in Document? : yes
    · Proximity of Terms? : yes
    · Document Length? : yes
- Machine Information
  * Machine Type for TREC Experiment: SPARC MP690
  * Was the Machine Dedicated or Shared: shared
  * Amount of Hard Disk Storage (in MB): 8000
  * Amount of RAM (in MB): 128
  * Clock Rate of CPU (in MHz): 40
- System Comparisons
  * Amount of "Software Engineering" which went into the Development of the System: basic method 200h; link method 380h; passage retrieval 300h
  * Given appropriate resources
    · Could your system run faster? : yes
    · By how much (estimate)? : more than 10 times
  * Features the System is Missing that would be beneficial: effectivity: Rel. feedback, query dependent training (routing) efficiency: optimized data structures (they leave less room for experiments)
- Significant Areas of System
  * Brief Description of features in your system which you feel impact the system and are not answered by above questions: none