Integrating Clustering and Multi-Document Summarization to Improve Document Understanding

Dingding Wang\textsuperscript{1} Shenghuo Zhu\textsuperscript{2} Tao Li\textsuperscript{1} Yun Chi\textsuperscript{2} Yihong Gong\textsuperscript{2}

1. School of Computer Science, Florida International University, Miami, FL
2. NEC Laboratories America, Inc. Cupertino, CA

Motivation

- Lack of interpretation for each document cluster
- Ignore the context dependency of the sentences

Solution

Simultaneously cluster and summarize documents

Document Clustering

Document Summarization

Document Understanding

Motivation

Solution

Simultaneously cluster and summarize documents

Experiments

Data set: TDT_10 (Number of Documents: 7879; Number of Clusters: 10; Number of Frequent Words: 1000)

Illustrative Interpretation:

One-sentence summaries for the top 4 largest topics in TDT2 Corpus:

- **Topic 1**: The Security Council has refused to lift the sanctions until Iraq complies with council resolutions demanding it destroy its weapons of mass destruction.
- **Topic 2**: Clinton says he had a very clear memory of the incident and he stands by the sworn court statement he has made that he did nothing wrong.
- **Topic 3**: The IOC had been expected to approve a new rule that all challenges to Olympic results must be made within three years after the games and settled by the time the next games begins.
- **Topic 4**: HONG KONG (AP): southeast Asian currencies hit new lows Tuesday for a second straight day, unnerving investors and sending regional stock markets tumbling.


Computational Algorithm

Algorithm 1 Model factorization given base language models

Input: \( A \): term-document matrix; \( B \): term-sentence matrix;

Output: \( U \): sentence-topic matrix; \( V \): document-topic matrix.

begin

1. Initialization:
   - Initialize \( U \) and \( V \) follow Dirichlet distribution, with hyper-parameter \( \alpha_U \) and \( \alpha_V \) respectively.

2. Iteration:
   - repeat
     1. Compute \( C_U = A_U \) \(-\) \((B U V)^T\).
     2. Assign \( U = U \) \(-\) \((B^T C V)^T\) \(+\) \( \alpha_U \),
        and normalize each column to 1.
     3. Compute \( C_V = A_V \) \(-\) \((B V U)^T\).
     4. Assign \( V = V \) \(-\) \((C^T B U)^T\) \(+\) \( \alpha_V \),
        and normalize each row to 1.
     - until convergence
   - Return \( U, V \)

end