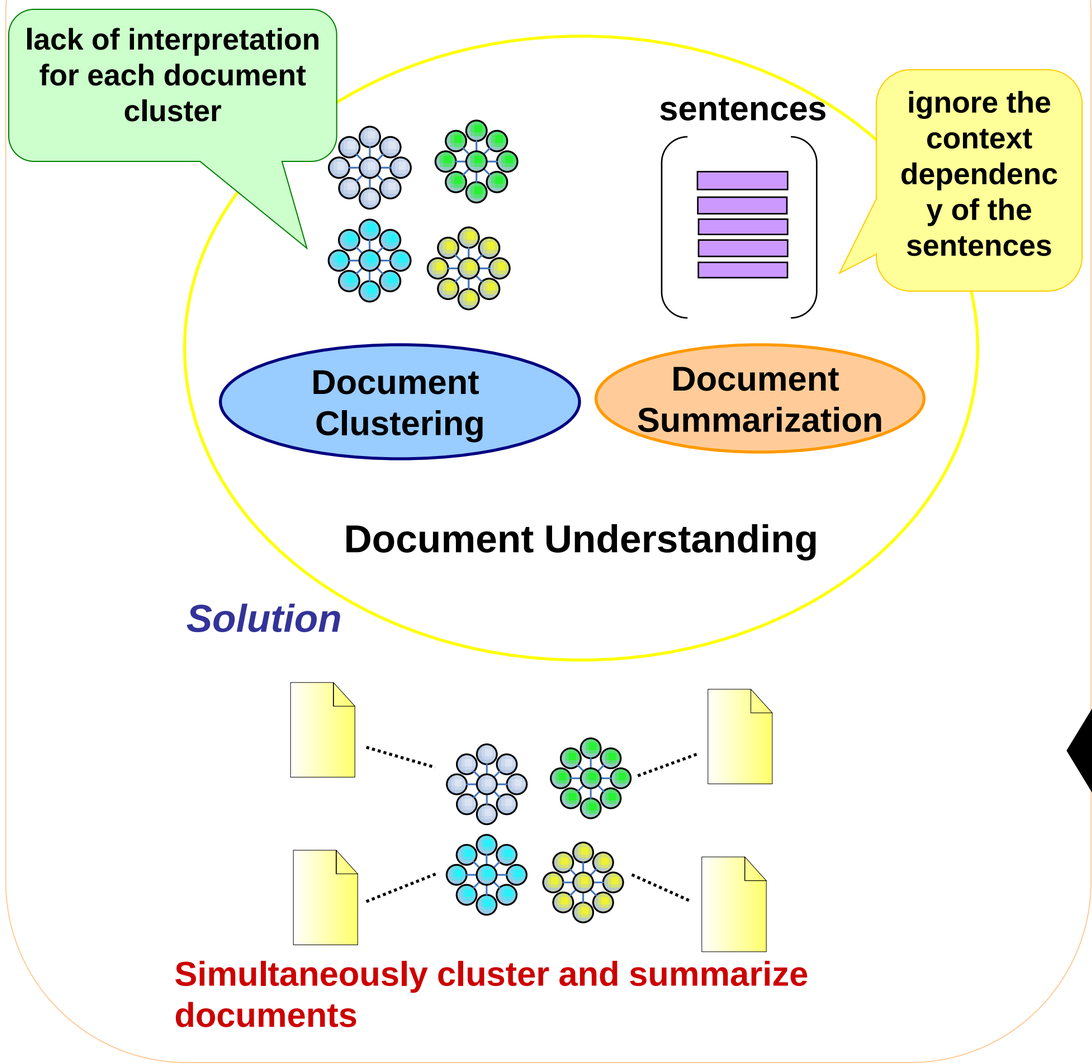
Integrating Clustering and Multi-Document Summarization to Improve Document Understanding

Dingding Wang[1] Shenghuo Zhu[2] Tao Li[1] Yun Chi[2] Yihong Gong[2]

1. School of Computer Science, Florida International University, Miami, FL

2. NEC Laboratories America, Inc. Cupertino, CA

Motivation



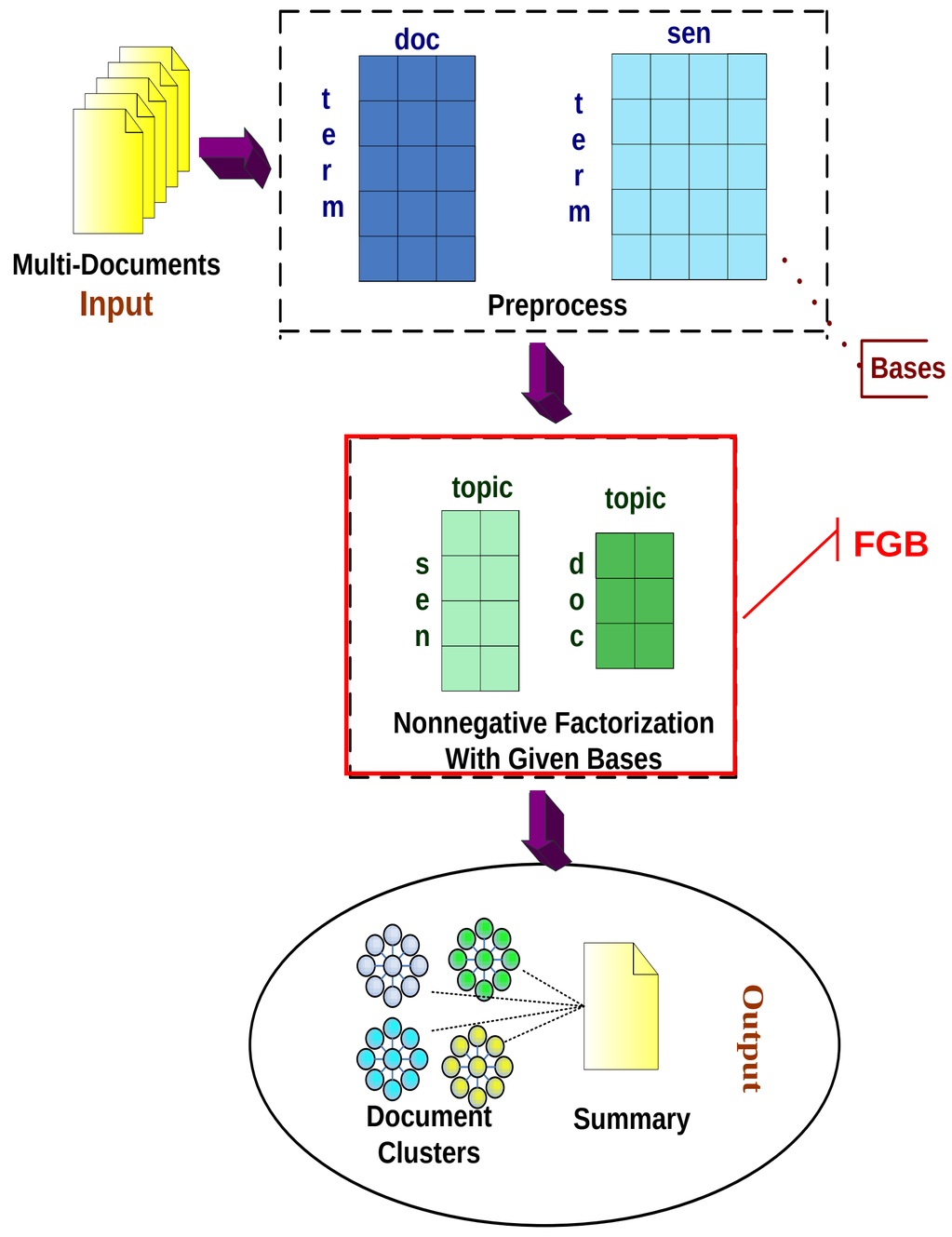
Solution

Simultaneously cluster and summarize documents

YES

HOW

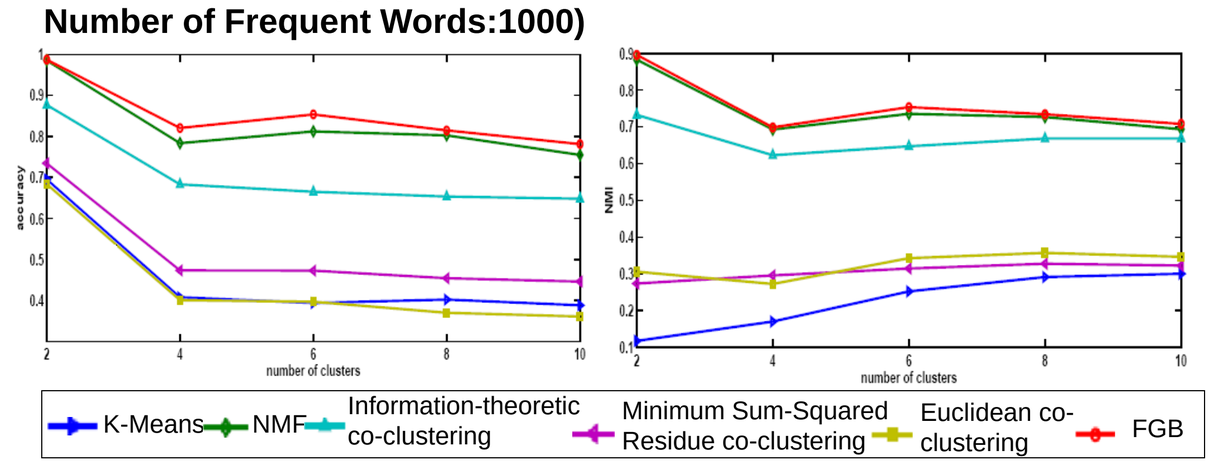
Model



THEN

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 make 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.**

1. Current Conflict with Iraq; 2. Monica Lewinsky Case; 3. 1998 Winter Olympics; 4. Asian Economic Crisis.

WOW

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 α_U and α_V respectively.
2. **Iteration:**
repeat
2.1 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.2 Assign $\mathbf{U}_{st} \leftarrow \mathbf{U}_{st}\left[\mathbf{B}^\top\mathbf{CV}\right]_{st} + \alpha_U$, and normalize each column to 1;
2.3 Compute $\mathbf{C}_{ij} = \mathbf{A}_{ij}/[\mathbf{BUV}^\top]_{ij}$;
2.4 Assign $\mathbf{V}_{dt} \leftarrow \mathbf{V}_{dt}\left[\mathbf{C}^\top\mathbf{BU}\right]_{dt} + \alpha_V$, and normalize each row to 1;
until convergence
3. **Return U, V**

end