Information Distance Based and Graph Based Update Summarization

Dr. Minlie Huang

Joint work with C Long, C Shou, Y Yu, F Jin, Q Li

aihuang@tsinghua.edu.cn

Tsinghua University, Beijing, China

Outline

- Background
- Summarization based on Information Distance
- Summarization based on Graph Centrality
- Conclusion

Background

- Too many information
  - Concise & coherent summary
- Generative summarization
  - Language generation (re-phras
  - Very difficult to express seman
- Extractive summarization
  - Extract key sentences

Previous Studies

- Statistical approaches (Nomoto,2001; …)
- Linguistic techniques (Nakao,2000; …)
- Graph-based methods
  - LexRank (Erkan&Radev, 2004;)
  - TextRank (Mihalcea&Tarau, 2004;)
  - Query specific document summarizer (Varadarajan&Hristidis, 2006)
  - Many more …

System I: Information Distance Based

- Kolmogorov complexity
  - $K(x)$: length of the shortest program that outputs x
  - $K(x|y)$=length of shortest program for x given y.
- Examples:
  - 1^n has a very short program: for i=1 to n, print "1".
    - $K(1^n)$ is very small
  - A competely "random" x has a very long program: print "x"
    - $K(x_n)$ is very large

Information Distance

- Information distance: a universal distance metric, defined as a conversion energy between two objects X and Y (Zhang, KDD'07; Long, CIKM'08);
  - $D_{max}(x,y)= \mathit{max}\{K(x|y),K(y|x)\}$
  - $D_{min}(x,y)= \mathit{min}\{K(x|y),K(y|x)\}$

Problem Reformulation

- Given cluster A with m documents $A_1, A_2, \ldots, A_m$, the update sum. task for cluster B=$\{B_1, B_2, \ldots, B_n\}$ should:

$$\mathrm{Min}\{D_{max}(\mathrm{S}, \mathrm{B}_1\mathrm{B}_2\ldots\mathrm{B}_n \mid \mathrm{A}_1\mathrm{A}_2\ldots\mathrm{A}_m)\},\ |\mathrm{S}| \leq \Theta$$

S=$\{s_1, s_2, \ldots, s_k\}$, each s_i is a sentence selected for the summary

Problem Reformulation

- $K(AB)=K(A \cup B)$ $\quad$ $K(A|B)=K(A\backslash B)$
- $D_{max}(S, B_1B_2 \ldots B_n | A_1A_2 \ldots A_m)$
  $=K((B_1B_2 \ldots B_n \backslash A_1A_2 \ldots A_m)\backslash S_1S_2 \ldots S_k)$
- $Min\{D_{max}(S, B_1B_2 \ldots B_n | A_1A_2 \ldots A_m)\}$
  $=Max\{K(S_1S_2 \ldots S_k)\}$

Approximation

- How to compute $K(S_1S_2...S_k)$?
- Assumption: each important word carries one unit of information, then

$$K(S)=|S| \quad \text{(the cardinality)}$$

- Important words
  - Non stop-words
  - Named entities (person, org., loc., date, …)
  - With high document frequency

Approximation

- Select one representative sentence *s* for each document *D* by:

$$\text{argmin}_s\{D_{max}(s,T),\ s \in D\}$$

T: the union of topic title and narrative

- Remove redundant representative sentences
  - —8 continuous common words
  - —60% common words

Generate Summary

- With the representative sentences, select a subset that could max{K(…)}
- Compute all combinations of sentences with the length limit;

Evaluation Results (RUN id 23)

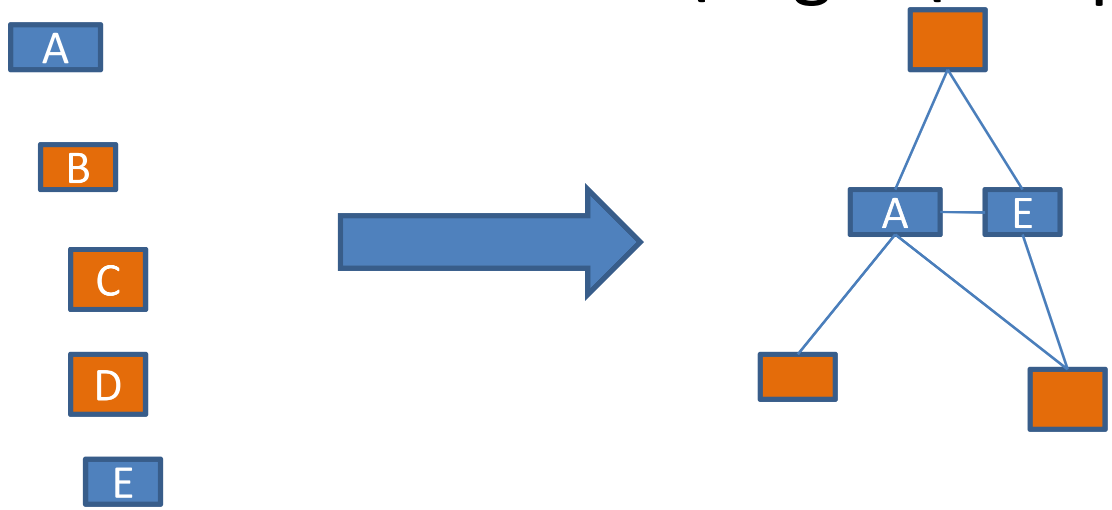
Evaluation Method	Best Result	Our Result	Rank
Average Modified Score	0.336	0.309	5/58
Macroaverage Modified Score with 3 models	0.331	0.304	5/58
Average Linguistic Quality	3.333	2.958	3/58
Average Overall Responsiveness	2.667	2.667	1/58

Outline

- Background
- Summarization based on Information Distance
- Summarization based on Graph Centrality
- Conclusion

Why graph model

- It may be a new solution for text presentation
  - Bag-of-Words
- Iterating on the graph can propagate very distant dependence
- Key points: define nodes\edges\computation



Graph-based Update Summarization

- 1 Select most salient terms
- 2 Build the term-sentence matrix *W*
- 3 Use the LSI sentence-sentence similarity matrix *SIM*
- 4 Construct a graph based on *SIM*
- 5 Compute the graph centrality (power iteration algorithm)
- 6 Select the top 15 sentences with high centrality

Graph-based Update Summarization

- 7 Compute all combinations with the length limit
- 8 Score a summary as a whole, and keep the best
- 9 Re-order the sentences within the summary

Graph centrality

- Centrality measure: which is the most important node in a graph?
  - Degree centrality
  - Eigenvector centrality
- Suppose the centrality of sentence *s* is C_s, then

$$\lambda C_s = \sum_{r \in D, r \neq s} Sim_{r,s} C_r$$

  - Important connections make the node itself more important

Tailored to Update Summarization

- Problem: given cluster A, summarize cluster B

$$SIM = \begin{pmatrix} SIM_{AA} & SIM_{AB} \\ SIM_{BA} & SIM_{BB} \end{pmatrix}$$

- Sentence in cluster B should be penalized

$$\lambda C_s = \left[\left(\sum_{r \in D_B, r \neq s} Sim_{r,s} C_r \right) - \beta \left(\sum_{r \in D_A, r \neq s} Sim_{r,s} C_r \right) \right]$$

- Matrix form:

$$SIM' = \begin{pmatrix} SIM_{AA} & -\beta SIM_{AB} \\ -\beta SIM_{BA} & SIM_{BB} \end{pmatrix}$$

How to score term, summary?

- Score a term
  - The position of a word (headline, first sentence)
  - With manual tuning parameters:
    - $Score(w)=tf(w)^{0.4} * F_d(df(w))$
- Score a summary
  - The term frequency of each word
  - The centrality of each sentence

$$Power(w)=\left(1-0.45\left(\frac{Score(w)}{maxscore}\right)^{0.15}\right)/2 \quad SumScore(S)=\sum_{w\in S}(count_w)^{Power(w)}$$

Evaluation results (RUN id 49)

Evaluation Method	Best Result	Our Result	Rank
Average Modified Score	0.336	0.304	7 /58
Macroaverage Modified Score with 3 models	0.331	0.299	7 /58
Average Linguistic Quality	3.333	3.073	2 /58
Average Overall Responsiveness	2.667	2.667	1 /58

Conclusion

- Information distance based summarization
- Graph centrality based summarization
- Theoretically sound but
  - Too many parameters in the second one;
- We debate the position assumption
  - The first sentence for newswire articles, but others?
- Sophisticated NLP techniques contribute to better results
  - Named entity recognition
  - Topic and Narrative analysis
  - And …

Acknowledgement

- Phd Students Chong Long, Feng Jin, Linjing Qin
- Undergraduate Students: Shouyuan Chen, Yuanming Yu
- Prof. Xiaoyan Zhu

Q&A

- Thanks !