Cover Coefficient based Multi-document Summarization

Gonenc Ercan and Fazli Can
Computer Engineering Department,
Bilkent University Ankara, Turkey

speaker: Ning Li

Outline

- Summarization
- History and Related Work
- Multidocument Summarization (MS)
- Our Approach: MS via C³M
- Datasets
- Evaluation
- Conclusion and Future Work

Summarization

- Information overload problem
- Increasing need for IR and automated text summarization systems
- Summarization: Process of extracting the most salient information from a source/sources for a particular user and task

Summarization Techniques

- Surface level: Shallow features
  - Term frequency statistics, position in text, presence of text from the title, cue words/phrases: e.g. “in summary”, “important”
- Entity level: Model text entities and their relationship
  - Vocabulary overlap, distance between text units, co-occurence, syntactic structure, coreference
- Discourse level: Model global structure of text
  - Document outlines, narrative stucture
- Hybrid

History and Related Work

- in 1950's: First systems surface level approaches
  - Term frequency (Luhn, Rath)
- in 1960's: First entity level approaches
  - Syntactic analysis
  - Surface Level: Location features (Edmundson 1969)
- in 1970's:
  - Surface Level: Cue phrases (Pollock and Zamora)
  - Entity Level
  - First Discourse Level: Stroy grammars
- in 1980's:
  - Entity Level (AI): Use of scripts, logic and production rules, semantic networks (Dejong 1982, Fum et al.1985)
  - Hybrid (Aretoulaki 1994)
- from 1990's-:explosuion of all

Multidocument Summarization (MS)

- Multiple source documents about a single topic or an event.
- Application oriented task, such as;
  - News portal, presenting articles from different sources
  - Corporate emails organized by subjects.
  - Medical reports about a patient.
- Some real-life systems
  - Newsblaster, NewsInEssence, NewsFeed Researcher

Term Frequency and Summarization

- Salient; Obvious, noticeable.
- Salient sentences should have more common terms with other sentences
- Two sentences are talking about the same fact if they share too much common terms. (Repetition)
- Select salient sentences, but inter-sentence-similarity should be low.

CC-Based Multi-document Summarizer

- S matrix 0-1 matrix
  - row : per sentence
  - Column : per term
  - s_{ij} : whether sentence i contains term j (0 , 1)

CC-Based Multi-document Summarizer

- C matrix

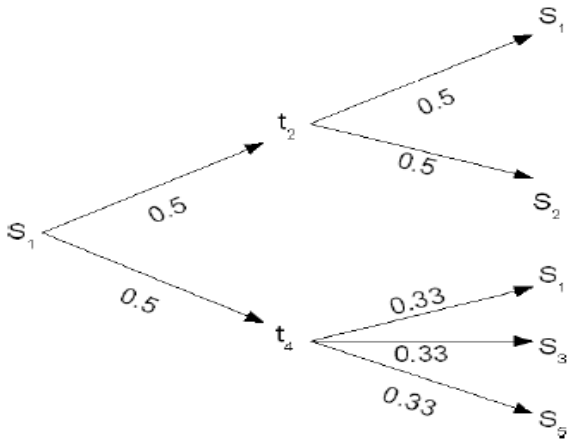
$$c_{ij} = \sum_{k}^{n} \alpha_{ik} * \beta_{kj} \qquad 1 \leq i, j \leq m$$

- c_{ij} : probability as the joint probabilities of α and β probabilities.
- n : the number of terms
- m : the number of sentences
- α_{ik} : probability is the probability of selecting term k from sentence i.
- β_{kj} : probability of term k occurring in sentence j.
- s_i : how much sentence i is covered by other sentences $= 1\text{-}c_{ii}$

An Example

$$S = \begin{vmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{vmatrix}$$

→



(a) Example S Matrix

Fig. 2. Probability graph of s_1

$$\alpha_{12} = S_{12} \div (S_{12} + S_{14}) = 0.5$$

$$\beta_{21} = S_{21} \div (S_{21} + S_{22}) = 0.5$$

An Example

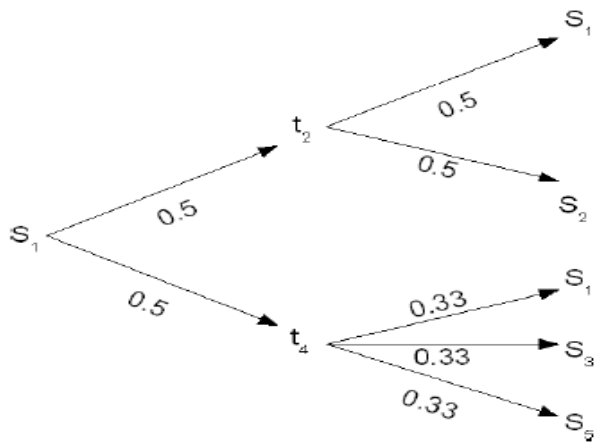


Fig. 2. Probability graph of s_1

→

$$C = \begin{vmatrix} \mathbf{0.42} & 0.25 & 0.17 & 0.00 & 0.17 \\ 0.17 & \mathbf{0.44} & 0.00 & 0.28 & 0.11 \\ 0.17 & 0.00 & \mathbf{0.42} & 0.00 & 0.42 \\ 0.00 & 0.42 & 0.00 & \mathbf{0.42} & 0.17 \\ 0.11 & 0.11 & 0.28 & 0.11 & \mathbf{0.39} \end{vmatrix}$$

(b) Cover Coeficient Matrix

$$c_{12} = \alpha_{12} \times \beta_{21} = 0.25$$

CC-Based Multi-document Summarizer

- selecting candidate sentences that are represented most by other candidate sentences.
  - select max (s_i) or ($1\text{-}c_{ii}$)

- selecting only candidate sentences that are not covered by an already selected sentence.
  - If $c_{ij} > \frac{c_{ii}}{\mu}$ or $c_{ji} > \frac{c_{jj}}{\mu}$,

    meaning repetition (μ is a constant value)

An Example

$$C = \begin{vmatrix} \mathbf{0.42} & 0.25 & 0.17 & 0.00 & 0.17 \\ 0.17 & \mathbf{0.44} & 0.00 & 0.28 & 0.11 \\ 0.17 & 0.00 & \mathbf{0.42} & 0.00 & 0.42 \\ 0.00 & 0.42 & 0.00 & \mathbf{0.42} & 0.17 \\ 0.11 & 0.11 & 0.28 & 0.11 & \mathbf{0.39} \end{vmatrix}$$

(b) Cover Coeficient Matrix

Sorted si values;

s5 ==> 0.61
s1 ==> 0.58
s3 ==> 0.58
s4 ==> 0.58
s2 ==> 0.56

Lets Form a Summary of 3 Sentences!!!

An Example (Step 1)

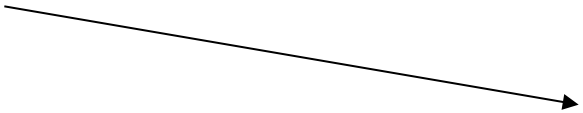
$$C = \begin{vmatrix} \mathbf{0.42} & 0.25 & 0.17 & 0.00 & 0.17 \\ 0.17 & \mathbf{0.44} & 0.00 & 0.28 & 0.11 \\ 0.17 & 0.00 & \mathbf{0.42} & 0.00 & 0.42 \\ 0.00 & 0.42 & 0.00 & \mathbf{0.42} & 0.17 \\ 0.11 & 0.11 & 0.28 & 0.11 & \mathbf{0.39} \end{vmatrix}$$

(b) Cover Coeficient Matrix

Sorted si values;

s5 ==> 0.61
s1 ==> 0.58
s3 ==> 0.58
s4 ==> 0.58
s2 ==> 0.56

Summary Sentences;
s5



Choose the sentence which is most similar to others.

An Example (Step 2)

$$C = \begin{vmatrix} \mathbf{0.42} & 0.25 & 0.17 & 0.00 & 0.17 \\ 0.17 & \mathbf{0.44} & 0.00 & 0.28 & 0.11 \\ 0.17 & 0.00 & \mathbf{0.42} & 0.00 & 0.42 \\ 0.00 & 0.42 & 0.00 & \mathbf{0.42} & 0.17 \\ 0.11 & 0.11 & 0.28 & 0.11 & \mathbf{0.39} \end{vmatrix}$$

(b) Cover Coeficient Matrix

Sorted si values;

s5 ==> 0.61
s1 ==> 0.58
s3 ==> 0.58
s4 ==> 0.58
s2 ==> 0.56

s1 is next according to si values. Check if s1 is too much similar to s3, which is in summary. Include it to summary if s5 does not cover s1.

Summary Sentences;
s5
s1 →

$AC5 = (c_{55}) / 2 = 0.20$
$AC1 = (c_{11}) / 2 = 0.21$
$(c_{51} = 0.17) < (AC5 = 0.20)$
$(c_{15} = 0.11) < (AC3 = 0.21)$

An Example (Step 3)

$$C = \begin{vmatrix} \mathbf{0.42} & 0.25 & 0.17 & 0.00 & 0.17 \\ 0.17 & \mathbf{0.44} & 0.00 & 0.28 & 0.11 \\ 0.17 & 0.00 & \mathbf{0.42} & 0.00 & 0.42 \\ 0.00 & 0.42 & 0.00 & \mathbf{0.42} & 0.17 \\ 0.11 & 0.11 & 0.28 & 0.11 & \mathbf{0.39} \end{vmatrix}$$

(b) Cover Coeficient Matrix

Sorted si values;

s5 ==> 0.61
s1 ==> 0.58
s3 ==> 0.58
s4 ==> 0.58
s2 ==> 0.56

Summary Sentences;
s5
s1
s3

s3 is next.

check with s5.

$AC5 = (c_{55}) / 2 = 0.20$
$AC3 = (c_{33}) / 2 = 0.21$
$(c_{53} = 0.28) > (AC5 = 0.20)$
$(c_{35} = 0.42) > (AC3 = 0.21)$ s3 not ok

Clustering algorithm: C^3M

- D matrix (m * n)
  - row : per document
  - column : per term
  - d_{ij} : indicate the number of occurrences of term j in document i
- D matrix (m * n) → C matrix (m * m)
  - row , column: per document
  - c_{ij} : indicate relation between document i and document j
- some of the documents are selected as cluster seeds and non- seed documents are assigned to one of the clusters initiated by the seed documents

C^3M vs. CC Summarization

Clustering based on C^3M	**Summarization based on Cover Coefficient**
Aim: to cluster	Aim : to select the most representative sentences avoiding redundancy in the summary
Uses document by term matrix	Uses sentence by term matrix
Create document by document C matrix	Create sentence by sentence C matrix
Calculate number of clusters	Calculate the number of summary sentences using compression percentage(i.e, %10)
Seed Power Function: $p_i = \delta_i \times \psi_i \times x_{di}$	Summary Power Function: $s_i = 1 - c_{ii}$
Select seed documents with the highest p_i	Select sentences with the highest si values, that are dissimilar to already selected sentences.

Datasets

- We will use two datasets.
  - DUC (Document Understanding Conferences) dataset for **English** Multidocument Summarization.
  - Turkish New Event Detection and Tracking dataset for **Turkish** Multidocument Summarization.

Evaluation

Two methods for evaluation:

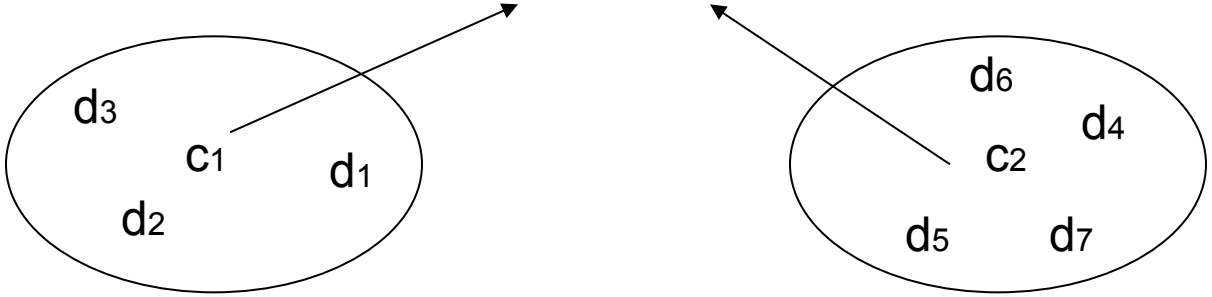
1. We will use this method for English Multidocument Summarization. Overlap between the **model summaries** which are prepared by human judges and the **system generated summary** gives the accuracy of the summary.
    - ROUGE (Recall Oriented Understudy for Gist Evaluation) is the official scoring technique for Document Understanding Conference (DUC) 2004.
    - ROUGE uses different measures. ROUGE-N uses N-Grams to measure the overlap. ROUGE-L uses Longest Common Subsequence. ROUGE-W uses Weighted Longest Common Subsequence.

Evaluation

2. We wil use this method for Turkish Multidocument Summarization.
   - We will add the **extracted summaries** as new documents.
   - Then, we will select these summary documents as the centroids of clusters.
   - Then, a centroid based clustering algorithm is used for clustering.
   - If the documents are attracted by their centroids which is the summary of these documents then we can say our summarization approach is good.

Evaluation

Summary documents are selected as the centroids



c_1 is the summary of d_1, d_2 and d_3.

c_2 is the summary of d_4, d_5 , d_6 and d_7.

Conclusion and Future Work

- Multidocument Summarization using Cover Coefficents of sentences is an intuitive and to our knowledge a new approach.
- This situation has its own advantages and disadvantages. We have fun because it is new. We are anxious about it because we have not seen any result summary yet.

Conclusion and Future Work

- After implementing the CC based summarization, we can try different methods on the same multidocuments set.

- First method:
  - A sentence-by-term matrix from **all** sentences of **all** documents can be formed.
  - Then, CC based Summarization can be applied.

Conclusion and Future Work

- Second method:
  - Cluster the documents using C3M.
  - Then, apply the first method to each cluster.
  - Combine the extracted summaries of each cluster to form one summary.

- Third method:
  - Summarize each document applying the first method. The only difference is that sentence-by-term matrices are constructed for sentences of each document.
  - Then, take the summaries of documents as documents and apply the first approach.

Questions

Thank you.