OPINIONS on OPINOSIS

OPINOSIS - A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions

by Kavita Ganesan, ChengXiang Zhai and Jiawei Han (← GZH)

Cosmina CROITORU
Presentation’s structure

- **Background**
  - The GZH’s paper in a nutshell.
- **OPINOSIS**
- **My Opinions**
Text summarization

Automatic creation of a shortened version of a text maintaining its most important points

Coherent and correctly-developed summaries

A possible approach to Information overload

Useful summary: concise, readable, and fairly well-formed
Text summarization

Types of summaries:

- **Purpose** (Indicative, informative, and critical)
- **Form**
  - Extractive (salient paragraphs, sentences, phrases)
  - Abstractive (concise summary of the central subject)
- **Dimensions** (Single-document vs. multi-document)
- **Context** (Query-specific vs. query-independent)
Information organized around the key aspects to represent a wider diversity of views on the topic.

- Centroid-based, sentence utility
  **MEAD** *(Radev et al. 00)*
- Reformulation
  *(McKeown et al. 99, McKeown et al. 02)*
- Generation by Selection and Repair
  *(DiMarco et al. 97)*
MEAD 1

- Implements **extractive summarization**: selects a subset of highly relevant sentences from the cluster’s overall set of sentences.

- Deep NLP and machine learning techniques.
  - **Decision-tree**, trained on a manually annotated corpus for CST relationships
  - **CST (Cross-document Structure Theory) relationships**: subsumption, identity, paraphrase, elaboration/refinement, etc.
For each sentence computes:

- **centroid score** (a measure of the centrality of a sentence to the overall topic of a cluster)
- **position score** (decreases linearly as the sentence gets farther from the beginning of a document)
- **overlap-with-first score** (the inner product of the weighted vector representation of the sentence and the first sentence (or title, if there is one))

Produces a **cluster centroid**, consisting of words which are central to all of the documents in the cluster.

**Ranks sentences** on their distance to the centroid.
Shallow NLP: mixing simple syntactic feature (word order or location and similarity) with domain-specific interpretation.

Deep NLP: sophisticated syntactic, semantic and contextual processing: named-entity recognition, relation detection, coreference resolution, syntactic alternations, word sense disambiguation, logic form transformation, logical inferences (abduction) and commonsense reasoning, temporal or spatial reasoning, etc.
<table>
<thead>
<tr>
<th>POS Tag</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>cc</td>
<td>coordinating conjunction</td>
<td>and</td>
</tr>
<tr>
<td>dt</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>nn</td>
<td>noun, singular</td>
<td>table</td>
</tr>
<tr>
<td>vb</td>
<td>verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>jj</td>
<td>adjective</td>
<td>red</td>
</tr>
<tr>
<td>rb</td>
<td>adverb</td>
<td>however, here, good</td>
</tr>
<tr>
<td>in</td>
<td>preposition</td>
<td>in, of, like</td>
</tr>
<tr>
<td>to</td>
<td>TO</td>
<td>to go, to him</td>
</tr>
</tbody>
</table>
Presentation’s structure

- Background
- The GZH’s paper in a nutshell.
- OPINOSIS
- My Opinions
A framework for summarizing **highly redundant opinions**

Use **graph representation** to generate concise abstractive summaries

Any corpus with high redundancies (Twitter comments, Blog comments, etc)
GZH’s Approach Schema

Input

SET OF SENTENCES
POS annotated
Topic related

OPINOSIS Graph

The paper is nice and useful for summarization.

Top scoring paths:

Output

Promising paths

OPINOSIS on OPINOSIS
Presentation’s structure

- The GZH’s paper in a nutshell.
- Background
- OPINOSIS
- My Opinions
Each sentence $\rightarrow$ sequence of word units

Word unit: pair $(word, POS annotation)$

Word units $\rightarrow$ nodes of OPINOSIS-Graph

Directed edges: $(v, w)$ iff $v$ and $w$ are successive word units in the same sentence

$v$-node $\leftarrow PRI_v$ Positional Reference Information

$PRI_v$ list of pairs $(SID_v, PID_v)$
OPINOSIS-Graph

- \( SID_v = i \) iff \( v \) is a unit word in sentence \( i \)
- \( PID_v = j \) iff position of \( v \) in sentence \( SID_v \) is \( j \)

Algorithm 1: \( \text{OpinosisGraph}(Z) \)

Input: Topic related sentences to be summarized:
\[ Z = \{ z_i \}_{i=1}^{n} \]

Output: \( G = (V, E) \)
for \( i = 1 \) to \( n \) do {
    \( w \leftarrow \text{Tokenize}(z_i); \) \( \text{sent\_size} \leftarrow \text{SizeOf}(w) \)
    for \( j = 1 \) to \( \text{sent\_size} \) do {
        \( \text{LABEL} \leftarrow w_j; \) \( \text{SID} \leftarrow i; \) \( \text{PID} \leftarrow j \)
        if ExistsNode(\( G, \text{LABEL} \)) then {
            \( v_j \leftarrow \text{GetExistingNode}(G, \text{LABEL}) \)
            \( \text{PRI}_{v_j} \leftarrow \text{PRI}_{v_j} \cup (\text{SID}; \text{PID}) \)
        } else {
            \( v_j \leftarrow \text{CreateNewNode}(G, \text{LABEL}) \)
            \( \text{PRI}_{v_j} \leftarrow (\text{SID}; \text{PID}) \)
        }
        if \( j > 1 \) and not ExistsEdge(\( v_{j-1} \rightarrow v_j, G \)) then
            AddEdge(\( v_{j-1} \rightarrow v_j, G \))
    }
}
Remarks on OPINOSIS-Graph

Property 1. (Redundancy Capture) Highly redundant discussions are naturally captured by subgraphs.

Property 2. (Gapped Subsequence Capture) Existing sentence structures introduce lexical links that facilitate the discovery of new sentences or reinforce existing ones.

Property 3. (Collapsible Structures) Nodes that resemble hubs are possibly collapsible.
Valid Paths

Valid Start Node - VSN: \( v \) s.t. \( \text{Average}(PID_v) \leq \sigma_{vsn} \)

Valid End Node - VEN: \( v \) is a punctuation (period, comma), or any coordinating conjunction (but, yet).

Valid Path: A path connecting a VSN to a VEN satisfying a set of well-formedness POS constraints.

Regular-expression POS constraints:
1. \( .*(/nn) + .*(/vb) + .*(/jj) + .* \)
2. \( .*(/jj) + .*(/to) + .*(/vb).* \)
3. \( .*(/rb)*.*(/jj) + .*(/nn) + .* \)
4. \( .*(/rb) + .*(/in) + .*(/nn) + .* \)
Redundancy and Gap

Path Scoring
Favor a valid path with a high redundancy score, to represent well most of the redundant opinions.

- $W = \{v_1, v_2 \ldots, v_s\}$ a "path" in the Opinosis-Graph.

- Sentence $z_i$ cover $W$ if $\forall j \in \{1, \ldots, s - 1\}$ $\exists (i, p) \in PRI_{v_j}$ and $\exists (i, p') \in PRI_{v_{j+1}}$ such that $p' - p \leq \sigma_{gap}$.

- Path redundancy of the path $W$:
  $$r(W) = |\{z_i | z_i \text{ cover } W\}|.$$
Redundancy and Gap

\[ \sigma_{gap} = 1 \Rightarrow r(W) = 0 \]
\[ \sigma_{gap} = 2 \Rightarrow r(W) = 2 \]
\[ \sigma_{gap} = 4 \Rightarrow r(W) = 3 \]
Scores
\[ W = \{ v_1, v_2 \ldots, v_s \}; \quad |W| = \text{length of } W; \]
\[ W_{i,j} = \text{the subpath } \{ v_i, \ldots, v_j \} \quad (1 \leq i < j \leq s) \]

- \[ S_{basic}(W) = \frac{1}{|W|} \sum_{k=1,s} r(W_{1,k}) \]
- \[ S_{wt\_len}(W) = \frac{1}{|W|} \sum_{k=1,s} r(W_{1,k}) \cdot |W_{1,k}| \]
- \[ S_{wt\_loglen}(W) = \frac{1}{|W|} \left[ r(W_{1,2}) + \sum_{k=2,s} r(W_{1,k}) \cdot \log |W_{1,k}| \right] \]
Path Composition

- Collapsible Node: node $v_c$ with $POS = vb$
- Collapsed candidates: Remaining path after a candidate node $v_c$:

\[
\text{anchor} \left\{ v_0, \ldots, v_c, v_{\text{first}}, \ldots, v_{\text{last}} \right\}
\]

\[
CC(v_c) = \bigcup_P \text{anchor} \left\{ P' | P' \text{ collapsed candidate for } P \right\}.
\]

- Stitched sentence: Logical sentence obtained from an anchor and its collapsed candidates.
Collapsible node: $v_c$ anchor of $v_c$: $P$

Collapsed candidates for $P$: $\{P_1, \ldots, P_{k-1}, P_k\}$

Stitched sentence ($k \geq 2$):

$$PP_1, comma, P_2, comma, \ldots, comma, P_{k-1}, cc, P_k.$$  

Examples:

The paper is nice, deep and useful for summarization.
The paper is nice, interesting but not useful.

coordinating conjunction $cc$: from all parents $u$ in $G$ of the first node $v$ of $P_k$, having POS=$cc$, select $\arg\max_{u:POS(u)=cc} r(\{u, v\})$.  


System’s parameters

- $\sigma_{\text{gap}}$ - controls the maximum allowed gaps in discovering redundancies.
- $\sigma_{\text{vsn}}$ - qualify nodes that tend to occur early on in a sentence.
- $\sigma_{\text{ss}}$ - controls the maximum number of paths to be chosen (summary size).
- $\sigma_{r}$ - a redundancy score threshold, to prune non-promising paths.

Empirically set.
Algorithm 2: $OpinosisSummarization(Z)$

- **Input**: Topic related sentences to be summarized:

  \[ Z = \{z_i\}_{i=1}^n \]

- **Output**: \( O = \{Opinosis Summaries\} \)
OpinosisSummarization(Z)

\[ g \leftarrow \text{OpinosisGraph}(Z) \]
\[ \text{node\_size} \leftarrow \text{SizeOf}(g) \]
\[
\text{for } j = 1 \text{ to } \text{node\_size} \text{ do } \{
    \text{if } \text{VSN}(v_j) \text{ then } \{
        \text{pathLen} \leftarrow 1; \text{score} \leftarrow 0
        \text{cList} \leftarrow \text{CreateNewList}()
        \text{Traverse}(\text{cList}, v_j, \text{score}, \text{PRI}_{v_j}, \text{label}_{v_j}, \text{pathLen})
        \text{candidates} \leftarrow \{\text{candidates} \cup \text{cList}\}
    \}
\}
\]
\[ C \leftarrow \text{EliminateDuplicates}(\text{candidates}) \]
\[ C \leftarrow \text{SortByPathScore}(C) \]
\[
\text{for } i = 1 \text{ to } \sigma_{s,s} \text{ do }
    \mathcal{O} = \mathcal{O} \cup \text{PickNextBestCandidate}(C)
\]
Comments on Traverse

- Traverse is a recursive Depth First Search to find valid paths ($\sigma_r$ used to avoid unuseful paths)

- The PRI overlap information, path length, summary sentence and path score are maintained during recursion

  - when $v$, the node visited, is collapsible, corresponding collapsed candidates are composed with the current path to $v$

  - the stitched sentence and its final score are added to the list of candidate summaries
Evaluation techniques

- Strong focus on evaluation: BLEU, ROUGE...
- ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation)
  1. Calculates n-gram overlaps between system generated summaries and model summaries (human made)
  2. A high level of overlap should indicate a high level of shared concepts between the two summaries.
  3. Unable to provide any feedback on a summary’s coherence
### Evaluation Techniques

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>System relevant</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>System non-relevant</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

- **Precision** \[ P = \frac{A}{A+B} \]
- **Recall** \[ R = \frac{A}{A+C} \]
- **F-score** \[ F = \frac{2PR}{P+R} \]
Experimental Setup

- Reviews from Tripadvisor.com, Amazon.com and Edmunds.com
- 51 review documents each about an entity $E$ and a topic $X$
- 100 sentences per review document
- Best 4 reference (human) summaries for each review document

Performance comparison between humans, Opinosis and the baseline method Mead
Experimental Setup

- For each reference summary compute ROUGE scores over the remaining $4-1=3$ reference summaries.
- Method MEAD selects $2$ most representative sentences as summaries.
- **Readability test:** are Opinosis summaries are readable?
  - Mix $N$ sentences from system summary and $M$ sentences from human summary.
  - Ask a human assessor to pick at most $N$ sentences that are least readable.
Experimental Setup

- **Readability test:**

\[
\text{readability}(O) = 1 - \frac{\#\text{CorrectPick}}{N}
\]

- **Opinosis parameters:** \(\sigma_{ss} = 2; \sigma_{vsn} = 15\)

- **Opinosis_{best}:** \(\sigma_{gap} = 4; \sigma_r = 2; S_{wt\_loglen}\)

- **ROUGE scores reported with the use of stemming and stopword removal**
Performance comparison

Human vs. Opinosis vs. MEAD

**Highest recall**
- HUMAN (17 words): ROUGE-1 = 0.3184, ROUGE-SU4 = 0.1293
- OPINOSIS best (15 words): ROUGE-1 = 0.2831, ROUGE-SU4 = 0.0851
- MEAD (75 words): ROUGE-1 = 0.4932, ROUGE-SU4 = 0.2316

**Lowest precision**
- HUMAN (17 words): ROUGE-1 = 0.3434, ROUGE-SU4 = 0.3088
- OPINOSIS best (15 words): ROUGE-1 = 0.4482, ROUGE-SU4 = 0.3271
- MEAD (75 words): ROUGE-1 = 0.0916, ROUGE-SU4 = 0.1515

**ROUGE Recall**

- More words in MEAD summary

**ROUGE Precision**
Performance comparison

- **Mead**:
  - very high recall scores (explanation: extractive)
  - extremely low precision scores (explanation: sentence extraction)

- **humans**
  - reasonable agreement amongst themselves
  - better than Opinosis
  - comparable to Mead
Performance comparison

Opinosis:
- closer in performance to humans than to Mead
- recall scores slightly lower than that achieved by humans
- improvement of precision by Opinosis over that of humans is more significant than the decrease of recall (Wilcoxon test)
Gap setting

\[ \sigma_{\text{gap}} = 1 : \text{low performance; redundancies veiled.} \]

\[ \sigma_{\text{gap}} = 2 : \text{big jump in performance} \]

\[ \sigma_{\text{gap}} > 2 : \text{slow improving performance} \]
Comparison of scoring functions

$S_{wt\_loglen}$ is the winner!

The effect of heavily favoring redundant subgraphs ($S_{basic}$) over longer but reasonably redundant ones ($S_{wt\_loglen}, S_{wt\_len}$) is not sound.
Readability test

- Human assessor picked the least 2 readable from 565 sentences (102 were Opinosis generated).

- Out of these 102, the human assessor picked only 34, resulting in an average readability score of 0.67.

- 34 sentences with problems:
  - 11 contained no information, incomprehensible
  - 12 were incomplete (false positives of validity check)
  - 8 had conflicting information (e.g. "the hotel room is clean and dirty").
  - 3 considered "poor grammar"
Presentation’s structure

- Background
- The GZH’s paper in a nutshell.
- OPINOSIS
- My Opinions
My Opinions

Positive Opinions

- Nice, interesting and simple approach.
- Carrying ideas from sequential data mining.
- Professional evaluation
- Path-aggregation building of a network!
- Adding OPINOSIS to conference management software systems (e.g. EasyChair)?
My Opinions

Not quite positive Opinions:

- Gap could change an opinion *(Not not).*
- Emphasizes too much on the surface order of words *(not quite abstractive).*
- Shallow NLP must be reward by learning.
- Mathematical inaccuracies in the text.
- Algorithms not optimized for huge corpus.
Opinions?

Thanks!