Experiments in Cross Language Query Focused Multi-Document Summarization

Prasad Pingali, Jagadeesh J, Vasudeva Varma

LANGUAGE TECHNOLOGIES RESEARCH CENTRE,
International Institute of Information Technology,
Hyderabad, India
Agenda

- Introduction
- Problem Statement
- Our System
- Experiments
- Conclusion
What is a Summary?

- *Summary* is a **condensed version of a source information content** having a recognizable genre and a very specific **purpose**: to give the reader an exact and concise idea of the contents of the source.
Various Genre

1. Extract vs. Abstract
2. Indicative vs. Informative vs. Evaluative
3. Generic vs. Query Based
5. Mono lingual vs. Cross lingual
6. Text vs. Audio vs. Video
This Talk

1. Extract vs. Abstract
2. Indicative vs. Informative vs. Evaluative
3. Generic vs. Query Based
5. Mono lingual vs. Cross lingual
6. Text vs. Audio vs. Video
Motivation for CLQ-Summarization

- MT for a CL Information Access is inevitable
- When should we do MT and on what?
- Translating top 10 CLIR results may require MT on at least 200 sentences (in runtime)
- Translating query focused snippets is an option.
- Snippets are indicative summaries (most of the times demands for document access)
- Informative summaries on the other hand may contain the answer to the information need.
- The problem is to build a system which avoids complete document translation.
Other Related Experiments

- In DUC and ACL multilingual summarization task,
- Summaries are query-independent
- All Arabic documents are machine translated apriori.
  - Unrealistic assumption for dynamic document sets.
Agenda

- Introduction
- Problem Statement
- Our System
- Experimentse
- Conclusion
Problem Statement

- **Given Inputs**
  - A topic statement (or query) in source language $L_1$
  - A set of 25-50 documents relevant to the topic in language $L_2$

- **Required Output**
  - Well-organized, fluent answer in $L_2$
  - Of length 250 words
An Example Source Language Topic

<title>అధికారం నటునీ విషయం</title>

<narr>వాగుండి విషయం, వారి ప్రత్యేకిత సందర్భం, లేదా దీని సాధారణ రంగాన సంబంధాన్ని సాధారణంగా గురించాం. మే విషయం మాత్రమే? మే విషయం ప్రపంచంలో విస్తరించాం? మా సంపాదకుల నిపుణత చెంది మన విషయం విస్తరించాం?
</narr> ⇒ Information need, expressed in multiple questions

</topic>
<topic>
<num> d301i </num>

<title> International Organized Crime </title> ⇒ Title of the topic

<narr> Identify and describe types of organized crime that crosses borders or involves more than one country. Name the countries involved. Also identify the perpetrators involved with each type of crime, including both individuals and organizations if possible. </narr> ⇒ Information need, expressed in multiple questions

</topic>
Comparison with TREC and CLEF

- Questions are not of factoid or definitional type.
- In TREC, CLEF answer to be extracted from large corpus.
- Corpora in TREC, CLEF are multi-topic
- Retrieval of documents is not a part of the problem definition. Relevant documents are given.
Agenda

- Introduction
- Problem Statement
- Our System
- Experiments
- Conclusion
Our System

- Our function to score sentences has,
  - Query Translation component
  - Query Expansion component

- Top ranking sentences form summary after eliminating redundancy
  - Redundancy identification using term overlap
The relevance based language model [Lavrenko and Croft, 2001] can be defined as,

\[ P(w|R) \approx P(w|Q) = P(w|q_1, q_2...q_k) = \frac{P(w, q_1 \ldots q_k)}{P(q_1 \ldots q_k)} \]  

(1)

Joint probability can be estimated using conditional sampling [Bruza and Song, 2003]

\[ P(w, q_1, \ldots q_k) = P(w) \prod_{i=1}^{k} P(q_i|w) \]  

(2)
• Since the query \((q_i)\) language and sentence term’s \((w)\) language are different, we factor in the query translation probability as,

\[
P(q_i | w) = \sum_{j=1}^{n} P(q_i | e_j) . P(e_j | w)
\] (3)
Calculation of $P(q_i|e_j)$ and $P(e_j|w)$

- Translation Probability: We assume uniform distributions for all possible translations found in the dictionary.

- Query Expansion using co-occurring terms:
  \[
  \text{HAL}(w'|w) = \sum_{k=0}^{K} W(k) n(w, k, w')
  \]
  \[
  \text{pHAL}(w'|w) = \sum_{k=0}^{K} P(k) P(w'|w, k)
  \] (4)

where $P(w'|w, k) = \frac{n(w, k, w')}{\sum_{w''} n(w, k, w'')}$
Sentence Score

- Each term’s relevance score towards the query

\[ P(w|R) \approx \frac{P(w)}{P(Q)} \prod_{q_j} P(q_j|w) \]

- Thus, score of a sentence can be calculated as,

\[ P(S|R) \approx \prod_{w_i \in S} P(w_i) \prod_{q_j} P(q_j|w_i) \]
Agenda

- Introduction
- Problem Statement
- Our System
- Experiments
- Conclusion
Evaluation Setup

• Used DUC 2005 dataset
• Topics manually translated into Telugu
• Output summary truncated to 250 words (white-space delimited tokens)
• No bonus for a shorter summary
• Document set consists of news articles and reports
• Documents sourced from Financial Times of London and LA Times
• Human written summaries for each topic are treated as model summaries
• System generated summaries are evaluated against model summaries
• ROUGE package [Lin and Hovy, 2003] was used for evaluation

• ROUGE settings were retained to be the same as used in DUC 2005

• ROUGE package generates various string comparison measures between model summaries and machine generated summaries.

• ROUGE-2 and ROUGE-SU4 are shown to correlate most with human evaluation (0.95 and 0.94 Spearman’s rank correlation coefficient respectively) [Dang, 2005]

• Monolingual baseline is a special case of this system, uses only co-occurrence statistics
### ROUGE Scores

<table>
<thead>
<tr>
<th>Sys ID</th>
<th>ROUGE-SU4</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.13307</td>
<td>0.07006</td>
</tr>
<tr>
<td>17</td>
<td>0.1297</td>
<td>0.0717</td>
</tr>
<tr>
<td>8</td>
<td>0.1279</td>
<td>0.0696</td>
</tr>
<tr>
<td>4</td>
<td>0.1277</td>
<td>0.0685</td>
</tr>
<tr>
<td>10</td>
<td>0.1252</td>
<td>0.0698</td>
</tr>
<tr>
<td>5</td>
<td>0.1232</td>
<td>0.0674</td>
</tr>
<tr>
<td>11</td>
<td>0.1225</td>
<td>0.0642</td>
</tr>
<tr>
<td>19</td>
<td>0.1218</td>
<td>0.0632</td>
</tr>
<tr>
<td>Tel-Eng-Sum</td>
<td><strong>0.12058</strong></td>
<td><strong>0.06048</strong></td>
</tr>
<tr>
<td>16</td>
<td>0.1189</td>
<td>0.0632</td>
</tr>
<tr>
<td>7</td>
<td>0.1189</td>
<td>0.0627</td>
</tr>
</tbody>
</table>

Table 1: Comparison of cross-lingual summarization system Tel-Eng-Sum with official scores (F-Measure) of DUC ’05 participants and mono-lingual HAL feature baseline
Agenda

- Introduction
- Problem Statement
- Our System
- Experiments
- Conclusion
Conclusion

- Crosslingual query focused summarization systems can be evaluated using existing monolingual query focused summarization framework.

- Cross-language query focused summarization leads to very little performance drop from monolingual baseline when compared to CLIR.

- This deterioration may be reduced even more if more sophisticated query translation techniques are used.
Thank You!