Automated Summarization System

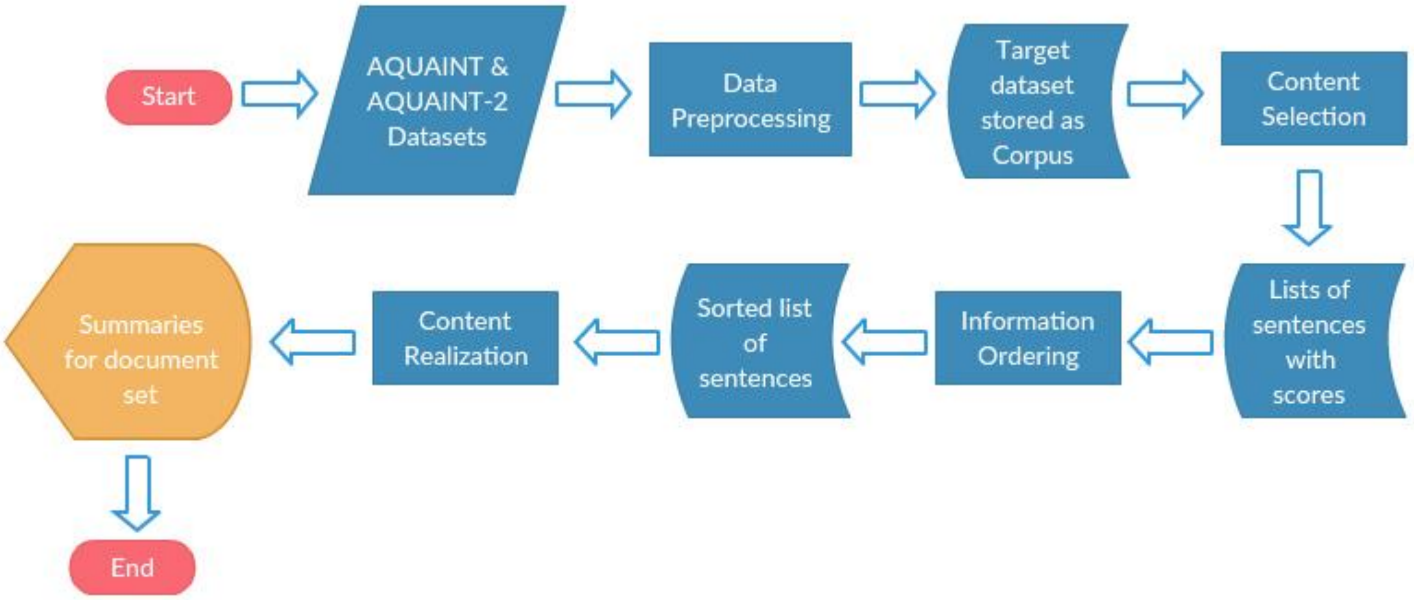
Haobo Gu, Yuanhe Tian, Weifeng Jin, Haotian Zhu

May 17th, 2018

Overview: Improved System

- Combination of both unsupervised and supervised methods
- Extraction-based approach without modifying sentences
- Three parts:
  - Improvement of Content Selection
  - Information Ordering
  - Content Realization

System Flowchart



Improvement of Content Selection

Two Salient Score:

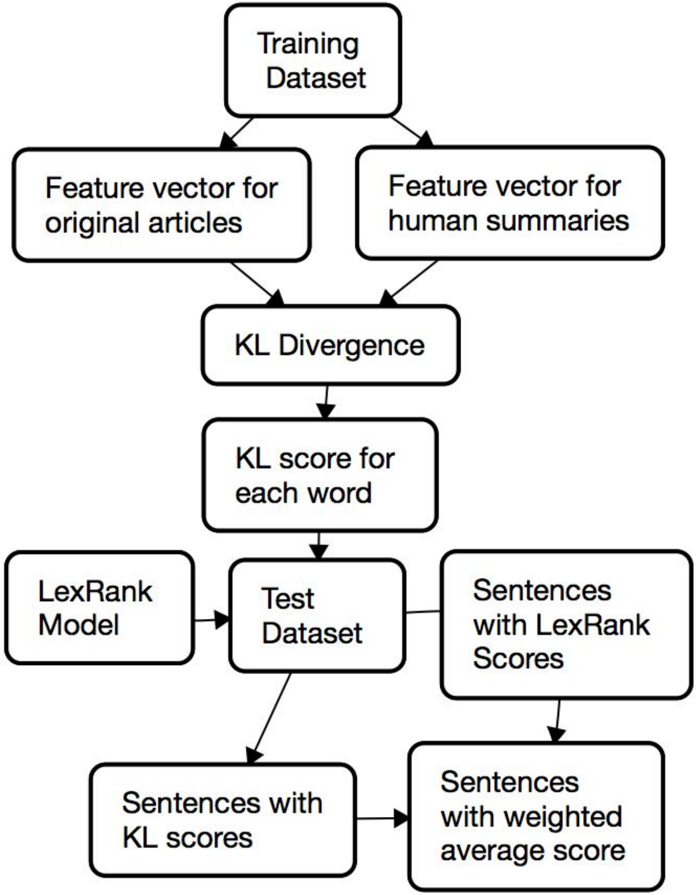
1. Salient score based on LexRank algorithm
2. Salient score based on Kullback–Leibler (KL) divergence

Final Salient Score:

Weighted average of the two salient score

Other Constraints:

Sentence length must be longer than 8



Content Selection Flow chart

Calculate Word KL-Score

- Build model on articles (*A*) and human summaries (*S*)
- Two KL-scores for a word (*w*)

$$\mathrm{KL}(S\,||A)(w) = Pr_S(w) \times ln\left(\frac{Pr_S(w)}{Pr_A(w)}\right) \qquad \mathrm{KL}(A\,||S)(w) = Pr_A(w) \times ln\left(\frac{Pr_A(w)}{Pr_S(w)}\right)$$

- Final KL-Score of a word (*w*)

$$Score_{KL}(w) = \frac{1}{2}\left(\mathrm{KL}(S\,||A)(w) - \mathrm{KL}(A\,||S)(w)\right)$$

Calculate Sentence Salient Score

- KL salient Score of a sentence

$$Score_{KL}(sent) = \frac{\sum_{w \in sent} Score_{KL}(w)}{length(sent)}$$

- LexRank salient Score of a sentence

$$Score_{LR}(sent) = LexRank(sent, docset)$$

- Normalize the two score
- Weighted average salient score

$$Score(sent) = 0.25 \times Score_{LR}(sent) + 0.75 \times Score_{KL}(sent)$$

Information Ordering

$$\mathrm{PREF}_{chro}(u, v, Q) = \begin{cases} 1 & T(u) < T(v) \\ 1 & [D(u) = D(v)] \wedge [N(u) < N(v)] \\ 0.5 & [T(u) = T(v)] \wedge [D(u) \neq D(v)] \\ 0 & otherwise \end{cases}.$$

- Chronological Expert (Bollegala et al.)
  - Define a preference function as:
  - T(u) is the publication time of sentence u.
  - D(u) is the unique id code of sentence u's document.
  - N(u) is the in-text index of sentence u in it document.

Information Ordering

- Chronological Expert (Bollegala et al.)
  - Sentence Ordering Algorithm

Algorithm 1 Sentence Ordering Algorithm.

Input: A set $\mathcal{X}$ of the extracted (unordered) sentences and a total preference function $\mathrm{PREF}_{total}(u, v, Q)$.

Output: Ranking score $\hat{\rho}(t)$ of each sentence $t \in \mathcal{X}$.

```
V = X
Q = ∅
for each v ∈ V do
    π(v) = Σ_{u∈V} PREF_total(v, u, Q) − Σ_{u∈V} PREF_total(u, v, Q)
end for
while V ≠ ∅ do
    t = arg max_{u∈V} π(u)
    ρ̂(t) = |V|
    V = V − {t}
    Q = Q + {t}
    for each v ∈ V do
        π(v) = π(v) + PREF_total(t, v, Q) − PREF_total(v, t, Q)
    end for
end while
return ρ̂
```

Information Ordering

Input: a list of selected sentences.

Do:

- Compute chronological preference of each sentence pair.
- Order sentences using Algorithm 1

Output: a list of sentences with ranking scores.

Content Realization

We improved our **Integer Linear Programming** content realization approach:

- Instead of considering salience score of the sentence, we use the weight sum of bigrams as the new target function
    - By maximizing target function, the redundancy is minimized implicitly
    - For a bigram , we count the number of sentences which contain this bigram as its weight
- Three constraints: one for length limitation and two for establishing the relation between bigram variables and sentence variables
- Rewrite the whole module -- migrate from scipy’s linprog to PuLP

Prune sentences

We tried 3 approaches to compress the sentence:

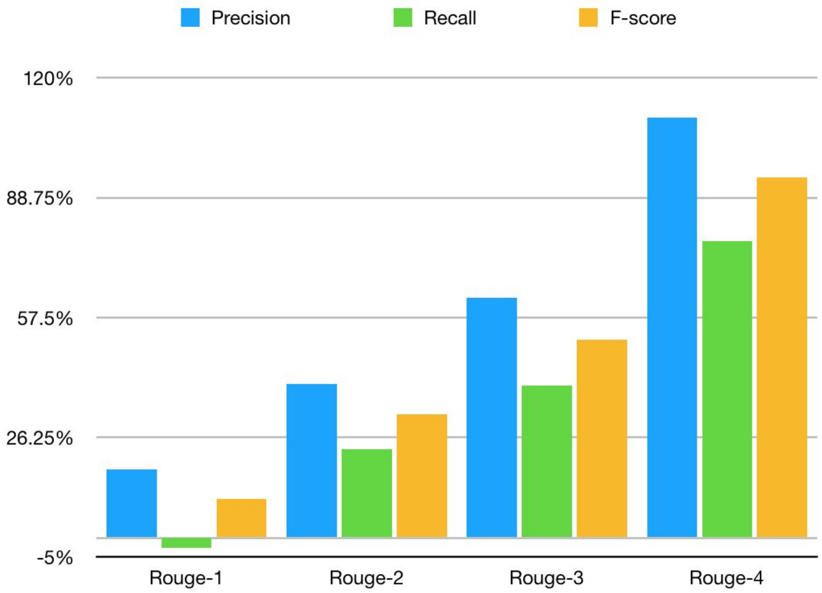
- Remove clauses starting with an adverb
- Remove appositions
- Remove contents in a pair of parenthesis

Unfortunately, all three approaches have negative effect, based on ROUGE score. So we didn’t activate any sentence compression methods in our submission

Results and Analysis

Table 1: ROUGE Results for devtest

ROUGE	Recall	Precision	F-score
ROUGE-1	0.23619	0.24349	0.24384
ROUGE-2	0.06092	0.06541	0.06292
ROUGE-3	0.01981	0.02088	0.02028
ROUGE-4	0.00746	0.00768	0.00755



Issues and Successes

Issues:

- Randomness in small subset of document sets
- Minor fluctuation in Rouge scores
- Extraction-based method, does not modify sentence content or combine several sentences into one.
- The balance between supervised and unsupervised methods

Successes:

- Big improvement over the baseline system, some aspects achieve a 100%+ improvement.
- Combination of both unsupervised and supervised learning methods

Strategies for Group Work

- Weekly meeting on Sunday night, discuss this week's work and coordinate next week's tasks.
- Two members worked on Content Selection, two worked on Information Ordering.
- At the same, each member make some minor improvements for its own part during the baseline system development.
- Mainly used github for code sharing and coordination.

References

- Danushka Bollegala, Naoaki Okazaki and Mitsuru Ishizuka 2012. A Preference Learning Approach to Sentence Ordering for Multi document Summarization. Information Sciences 2012, 217:78-95.

- Dimitrios Galanis, Gerasimos Lampouras and Ion Androutsopoulos.2012. Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression. Proceedings of COLING 2012, 2012: 911-926

- Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing, pages 1018, Boulder, Colorado, June 2009.

References

- John M.Conroy, Judith D.Schlesinger, Jade Goldstein and Dianne P.O'Leary 2004. Left-Brain/Right-Brain Multi-Document Summarization. Proceedings of DUC, 2004

- Kai Hong and Ani Nenkova 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2004, 2004:712-721

- Gunes Erkan and Dragomir R.Radev 2004. LexRank:Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 2004, 2004:457-479

-Ryan A.Georgi 2018. Lecture of Content Selection and Information Ordering. in the course LING 573 Spring 2018 in the University of Washington.