Concept level ILP.

X text unit

X sentence level

✓ concept level.

The important concepts that a ~~summary~~ summary should cover.

(1) Summary benefits by including a particular concept only once!

(2) That means redundancy is implicitly captured.

$c_i \leftarrow$ ~~idica~~ indicator variable for the presence of concept i in the target summary:

$w_i \leftarrow$ weight of the concept i

ConcepNet. { Concept → "word bigrams"

Weight → no. of occurrences of the word bigrams in the input documents.

SIM ←

$t_i \xleftarrow{\text{vectorization of}} \vec{t_i}$ — tf-idf.

$\mathcal{D} \xleftarrow{\text{vectorization of}} \vec{\mathcal{D}}$ — sum up position-wise.

— take position wise avg.

~~$cosine(\vec{t_i}, \vec{\mathcal{D}}$~~

$SIM = cosine(\vec{t_i}, \vec{\mathcal{D}})$

$\boxed{Red(i,j) = SIM(t_i, t_j) \rightarrow cosine(\vec{t_i}, \vec{t_j})}$.

Query focused summarization:

↳ given a query Q → relevant summary.

$Rel(i) = SIM(t_i, Q) + SIM(t_i, \mathcal{D})$.

Sentence level ILP formulation:

Optimation function:

$$\boxed{\text{maximize} \sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j).}$$

Constraints. $(\forall i, \forall j)$

$\alpha_i, \alpha_{ij} \in \{0, 1\}$ (1) ← indicator variables. whether or not a textual unit or a pair are included in the target summary or not.

~~$\sum_i \alpha_i l($~~

$\sum_i \alpha_i l(i) \leq K$ (2) ↖ Budget constraint.

$$\boxed{\begin{array}{l} \alpha_{ij} - \alpha_i \leq 0 \\ \alpha_{ij} - \alpha_j \leq 0 \end{array}}$$

(3),(4) ↖ if t_i & t_j are both included then they must be individually included.

$\alpha_i + \alpha_j - \alpha_{ij} \leq 1$ -(5) ← inverse of (3) & (4).

How to operationalize $Rel(\cdot)$ & $Red(\cdot, \cdot)$.

Domain for summarization.

News summarization.

$\mathcal{D}$ ← Document collection.

D is a single document.

$$Rel(i) = \underline{\overset{\checkmark}{POS}(t_i, D)}^{-1} + \frac{SIM(t_i, \mathcal{D})}{(t_i \in D, D \in \mathcal{D})}$$

$POS(i, D)$ ← position of the text unit t_i in the Document D.

$SIM(t_i, \mathcal{D})$ ← similarity of text unit t_i with the overall collection of documents

→ Empirical obs: Initial sentences in a news doc are usually very important (Headlines)

Optimization based summarization *

Global inferencing method.

D is a document

t_n ~~number~~ textual units

D = $t_1, t_2, \ldots, t_n$. [textual units ↓ individual sentences].

→ Rel(i) : the relevance of t_i to the target summary.

→ Red(i,j) : Redundancy between t_i & t_j

→ $l(i)$ is the length of t_i (in terms of the no. of words).

Inferencing:

Problem is to select a subset S of textual units from D such that the summary score s, i.e., $s(S)$ is maximized.

$$s(S) = \overset{\max}{\cancel{\text{Argmax}}}_{S \subseteq D} \left[\sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i<j} Red(i,j) \right]$$

increase the relevance of t_i to the target summary S

reduce the redundancy ~~betn~~ among the choice of text units.

→ Greedy approaches give us approximate results.

→ Better solutions.

↑

Integer Linear Programming (ILP).

→ GNU ← ILP solver.

Recast the problem into a constraint optimization formulation.

$\underline{Rel(i)}$, $Red(i,j)$ ←

presence of some special entities etc.

enforces diverse information.

Greedy Solution.

```
1. Sort D so that $Rel(i) > Rel(i+1) \ \forall i$
2. $S = \{t_1\}$ ← most relevant text unit
3. while $\sum_{t_i \in S} \underline{l(i)} < K$ . ← budget
4.     $\underline{t_j} = \arg\max_{\boxed{t_j \in D-S}} f(S \cup \{t_j\})$ .
5.     $S = S \cup \{t_j\}$
6. return S.
```

Evaluation of summaries

ROUGE score

(Recall oriented Understudy for Gisting Evaluation)

ROUGE → RUJ

How much ← "coverage"

Precision ← safety-critical system

Medical diagnostics
(Decision support systems).

Revised ILP.

maximize $\sum_i w_i c_i$

subject to:

$\sum_j l_j s_j \le K$. (1)

← length of the target summary (budget)

c_i ← indicator of concept i in the target summary

s_j ← indicator of the sentence i in the target summary.

Occ_{ij} ← occurrence of concept i in sentence j

$s_j Occ_{ij} \le c_i \quad \forall i,j$. (2)

↳ If you select a sentence then it mandate that all concepts in theat sentence should be selected.

$c_i \in \{0,1\} \quad \forall i$ (5)

$s_j \in \{0,1\} \quad \forall j$ (6)

$\sum_j s_j Occ_{ij} \ge c_i$ (3)

If a concept is ever selected then it mandates that the sentence containing it must be selected.

ROUGE-1.

Example:

R: The cat is on the mat.

C: The cat and the dog.

(the, cat, the)

ROUGE-1 precision:

$$\frac{3}{5} = 0.6.$$

ROUGE-1 recall:

$$\frac{3}{6} = 0.5$$

R: {the cat, cat is, is on, on the, the mat}.

C: {the cat, cat and, and the, the dog}

ROUGE-2 precision.

$$\frac{1}{4} = 0.25$$

ROUGE-2 recall.

$$\frac{1}{5} = 0.2.$$

ROUGE - N.

N = 1, 2, ...

N → N-gram.

M/C generated summary (Candidate summary) : C

Reference summary → Human written summary → gold standard summary — R.

R: The cat is on the mat

C: The cat and the dog.

ROUGE-N precision

Ratio of the number of N-grams in C that also appears in R over the number of N-grams in C.

ROUGE-N recall (More important)

Ratio of the number of N-grams in C that also appears in R over the no. of N-grams in R.

$$\text{ROUGE F1.} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}$$

$$= 0.22.$$

ROUGE-L.

L: Longest common subsequence.

longest sequence — not necessarily consecutive.

R: The cat is on the mat } ← nuggets

C: the cat and the dog } ← nuggets.

LCS = the cat the.

|LCS| = 3.

ROUGE-L precision = 3/5 = 0.6.

ROUGE-L recall = 3/6 = 0.5.

Numerator is the LCS of R, C.
Denominator is the unigram count.

ROUGE-S. (Skip-Grams).

R: The cat is on the mat

C: The gray cat and the dog.

- ROUGE-2. (Skip = 1).
  ↳ unigram skipping).

The cat
The gray cat } match.

ILP summary.

Selected some sentences / text units.

Ordering.

News summarization:

Chronological ordering.

Coherence.

→ Choose orderings that make neighboring sentences / text units similar (cosine).

→ Multi entity summary:
Choose orderings that bring similar/same entities closer in the summary).

Topicality.

↓ finding the topics.

Make the summary topically coherent

ILP summary.

Selected some sentences / text units.

Ordering.

News summarization:

Chronological ordering.

Coherence.

→ Choose orderings that make neighboring sentences / text units similar (cosine).

→ Multi entity summary:
Choose orderings that bring similar/same entities closer in the summary).

Topicality.

↓ finding the topics.

Make the ~~[illegible]~~ summary topically coherent

Simplifying sentences.

Parse the output summary & decide based on some rules which parts to clip off.

Initial adverbials: "On the other hand", "as a matter fact".

PPs **without** named entities:

↳ prepositional phrases.

E.g. The commercial fishing restrictions in Washington will not be lifted unless the salmon population increases ~~to a substantial number~~. (PP removed).

Attribution clauses:

E.g. Rebels agreed to talks with govt officials, ~~international observers said Tuesday~~. attribution

Appositives:

Rajan, 28, ~~an artist who was living at the time in Philadelphia~~, found the inspiration in the back of city magazines.