Evaluation Algorithms for Extractive Summaries

Paul Tarau

Department of Computer Science and Engineering
University of North Texas

April 22, 2016

joint work with Fahmida Hamid and David Haraburda
comparing two summaries is sensitive to their lengths and the length of the document they are extracted from

⇒ the overlap between two summaries should be compared against the average intersection size of random sets

a summary for the same document can be quite different when written by different humans

⇒ weighted relatedness to reference summaries

comparing human written abstractive summaries to machine generated extractive ones

⇒ we need an evaluation mechanism using semantic equivalence relations

⇒ a “diamond standard”: scientific documents where author-written summaries provide a baseline for the evaluation of computer generated ones
Evaluating System Generated Summaries: State-of-the-Art

- **ROUGE-N**: n-gram recall between a candidate summary and a set of reference summaries

\[
\frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{count}(\text{gram}_n)}
\]

- Variants of ROUGE: ROUGE-L, ROUGE-W, ROUGE-S
Evaluating System Generated Summaries: State-of-the-Art

- Pyramid: *Summarization Content Unit (SCU)*
  - weighted overlapping instead of simple averaging technique
  - manual vs. automatic detection of SCU
  - no known means to handle the length variation
  - credibility of the human annotator: amazon mechanical turk

- Evaluation based on the Jensen-Shannon Divergence of Distributions
Evaluating computer-generated summaries vs. human-made summaries

- to summarize: we are using computer-based evaluation of computer-generated summaries to compare them to human-made ones
- can one summarize without “understanding”? most likely yes, humans do it all the time : – )
- How different are computer generated summaries from the human ones knowing that the human ones are quite different from each other?
- to devise a scale for evaluation normalized with respect to differences occurring between human-made summaries we need to:
  - make summaries of different sizes comparable
  - propose a ranking approach for machine generated summaries based on the concept of closeness with respect to reference summaries
  - ⇒ human-made reference summaries are compared against each other and also the baseline
Given a set $N$ of size $n$, and two randomly selected subsets with $l$ and $k$ elements, the average size of the intersection is:

$$\text{avg}(n, k, l)_{random} = \frac{\sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{l-i}}{\sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{l-i}}$$  \hspace{1cm} (1)
Simplifying the Baseline

- $|N| = n$
- $|K| = k$, $|L| = l$
- $|I| = i$

- $P(x \in K) = k/n$
- $P(x \in L) = l/n$
- $P(x \in I) = i/n$

$$Pr(x \in I) = Pr(x \in K) \cdot Pr(x \in L)$$

$$i/n = (k/n) \cdot (l/n)$$

$$i = \frac{kl}{n}$$
the i-measure: observed vs. random intersection

\[
i\text{-measure}(N, K, L) = \frac{\text{observed size of intersection}}{\text{random size of intersection}}
\]

\[
= \frac{\omega}{i}
\]

\[
= \frac{\omega}{kl/n}
\]

- less sensitivity towards length
$i$-measure vs. $f$-measure

Two random sets $K_r$ and $L_r$:

$$
r = \frac{|K_r \cap L_r|}{|K_r|} = \frac{i}{k} = \frac{l}{n} \tag{3}
$$

$$
p = \frac{|K_r \cap L_r|}{|L_r|} = \frac{i}{l} = \frac{k}{n} \tag{4}
$$

$$
i = kl/n \tag{5}
$$

$$
f\text{-measure}_{\text{random}} = \frac{2pr}{p + r} = \frac{2(l/n)(k/n)}{(l/n + k/n)}
= \frac{2(lk)}{(n^2) / ((k + l)/n)}
= \frac{2lk}{n(k + l)}
= \frac{2i}{k + l}
= i/((k + l)/2) \tag{6}
$$
i-measure as relativized f-measure

Same computation for observed intersection size \( \omega \)

from i-measure to f-measure

\[
i\text{-measure}(N, K, L) = \frac{\omega}{i}
\]

we get

\[
i\text{-measure}(N, K, L) = \frac{\omega/((k+l)/2)}{i/((k+l)/2)}
\]

\[\text{(7)}\]

\[= \frac{f\text{-measure}_{\text{observed}}}{f\text{-measure}_{\text{random}}}\]

\(\Rightarrow\) the i-measure is just the f-measure normalized with respect to the f-measure computed for random sets
Improving the “gold standard”

- *i*-measure helps with flexibility on length
- ⇒ no need to trim summaries on byte or word length
- Evaluating the Evaluators
  - compare overlaps between each pair with *i*-measure
  - ⇒ devise an algorithm that associates a degree of confidence to each evaluator
- towards a “diamond standard”: set up a repository of trusted summaries - the author-written ones
A data set with multiple human-made summaries

\[ \mathcal{D} = \{ d_1, d_2, \ldots, d_t \} \]
\[ \mathcal{H} = \{ h_1, h_2, \ldots, h_z \} \]
\[ \mathcal{S} = \{ s_1, s_2, \ldots, s_\lambda \} \]

for each document \( d \), a subset of annotators (say, \( \mathcal{H}_d = \{ h_1, h_2, \ldots, h_m \} \)) write summaries independently.

---

**DUC 2004**

\[ \mathcal{D} = \{ d_1, d_2, \ldots, d_t \} \]
\[ \mathcal{H} = \{ h_1, h_2, \ldots, h_z \} \]
\[ \mathcal{S} = \{ s_1, s_2, \ldots, s_\lambda \} \]

for each document \( d \), a subset of annotators (say, \( \mathcal{H}_d = \{ h_1, h_2, \ldots, h_m \} \)) write summaries independently.
Confidence-based scoring

**Step 01**

normalize i-measure (based on best pair)

\[
\begin{align*}
    w_d(h_p, h_q) &= \frac{i\text{-}measure(d, h_p, h_q)}{\mu_d} \\
    w_d(s_j, h_p) &= \frac{i\text{-}measure(d, s_j, h_p)}{\mu(d, h_p)}
\end{align*}
\]

\[
\begin{align*}
    \mu_d &= \max \{ i\text{-}measure(d, h_p, h_q) \}, \forall (h_p, h_q) \in H_d \times H_d, h_p \neq h_q \\
    \mu(d, h_p) &= \max \{ i\text{-}measure(d, s, h_p) \}, \forall s \in S
\end{align*}
\]

(8)
Confidence-based scoring - continued

**Step 02**
define a degree of confidence to each reference

\[
c_d(h_p) = \frac{\sum_{q=1, p\neq q}^m w_d(h_p, h_q)}{m - 1}.
\]  

(9)

**Step 03**
assign a weighted score for each system-generated summary

\[
\text{score}(s_j, d) = \sum_{p=1}^m c_d(h_p) \times w_d(s_j, h_p)
\]

(10)

**Step 04**
average the score

\[
i\text{-score}(s_j) = \frac{\sum_{i=1}^t \text{score}(s_j, d_i)}{t}
\]

(11)
Analysis through an example

Summary of Reference B and G

B: Clinton arrives in Israel, to go to Gaza, attempts to salvage Wye accord.
G: Mid-east Wye Accord off-track as Clintons visit; actions stalled, violence

\[i\text{-measure}(d, B, G) = \frac{3}{10\times9/282}\] which is 9.4

Summary of Reference G and F

G: Mid-east Wye Accord off-track as Clintons visit; actions stalled, violence
F: Clinton meets Netanyahu, says peace only choice. Office of both shaky

\[i\text{-measure}(d, G, F) = \frac{1}{10\times8/282}\] which is 3.525
The 4 human-made summaries

normalize \textit{i-measure}

<table>
<thead>
<tr>
<th>Reference</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Clinton arrives in Israel, to go to Gaza, attempts to salvage Wye accord.</td>
</tr>
<tr>
<td>G</td>
<td>Mid-east Wye Accord off-track as Clintons visit; actions stalled, violence</td>
</tr>
<tr>
<td>E</td>
<td>President Clinton met Sunday with Prime Minister Netanyahu in Israel</td>
</tr>
<tr>
<td>F</td>
<td>Clinton meets Netanyahu, says peace only choice. Office of both shaky</td>
</tr>
</tbody>
</table>

Table: reference summaries (B,G,E,F) on document \textit{D30053.APW19981213.0224}
### Normalized i-measures

<table>
<thead>
<tr>
<th>Pair $(p, q)$</th>
<th>$n$</th>
<th>$k$</th>
<th>$l$</th>
<th>$\omega$</th>
<th>$i$</th>
<th>i-measure</th>
<th>$w_d(h_p, h_q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(G, F)</td>
<td>282</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>0.28</td>
<td>3.52</td>
<td>0.375</td>
</tr>
<tr>
<td>(G, B)</td>
<td>282</td>
<td>10</td>
<td>9</td>
<td>3</td>
<td>0.32</td>
<td>9.40</td>
<td>1.0</td>
</tr>
<tr>
<td>(G, E)</td>
<td>282</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>0.28</td>
<td>3.52</td>
<td>0.375</td>
</tr>
<tr>
<td>(F, B)</td>
<td>282</td>
<td>8</td>
<td>9</td>
<td>1</td>
<td>0.25</td>
<td>3.91</td>
<td>0.4166</td>
</tr>
<tr>
<td>(F, E)</td>
<td>282</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>0.22</td>
<td>8.81</td>
<td>0.9375</td>
</tr>
<tr>
<td>(E, B)</td>
<td>282</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>0.25</td>
<td>7.83</td>
<td>0.8333</td>
</tr>
</tbody>
</table>

Table: normalized i-measure of all possible reference pairs
Confidence associated to a reference human made summary

Confidence associated to a reference for a specific document \( d \) is the average of its normalized i-measure

\[
\begin{align*}
c_d(G) &= \frac{0.375 + 0.375 + 0.375}{3} = 0.583 \\
c_d(B) &= \frac{0.375 + 0.4166 + 0.833}{3} = 0.75
\end{align*}
\]

<table>
<thead>
<tr>
<th>reference: ( h_p )</th>
<th>confidence: ( c_d(h_p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.583</td>
</tr>
<tr>
<td>F</td>
<td>0.576</td>
</tr>
<tr>
<td>B</td>
<td>0.75</td>
</tr>
<tr>
<td>E</td>
<td>0.715</td>
</tr>
</tbody>
</table>

Table: Confidence Score
Calculate scores for a computer-made summary: a good one

31: Clinton met Israeli Netanyahu put Wye accord

- **B**: Clinton arrives in Israel, to go to Gaza, attempts to salvage Wye accord.
- **G**: Mid-east Wye Accord off-track as Clintons visit; actions stalled, violence
- **E**: President Clinton met Sunday with Prime Minister Netanyahu in Israel
- **F**: Clinton meets Netanyahu, says peace only choice. Office of both shaky

<table>
<thead>
<tr>
<th>pair(s_j, h_p)</th>
<th>n</th>
<th>l</th>
<th>k</th>
<th>ω</th>
<th>i</th>
<th>i-measure</th>
<th>w_d(s_j, h_p)</th>
<th>h_p</th>
<th>c_d(h_p)</th>
<th>μ(d, h_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(31, F)</td>
<td>282</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>0.198</td>
<td>10.07</td>
<td>0.285</td>
<td>F</td>
<td>0.576</td>
<td>35.25</td>
</tr>
<tr>
<td>(31, B)</td>
<td>282</td>
<td>7</td>
<td>9</td>
<td>3</td>
<td>0.223</td>
<td>13.42</td>
<td>0.428</td>
<td>B</td>
<td>0.75</td>
<td>31.33</td>
</tr>
<tr>
<td>(31, E)</td>
<td>282</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>0.198</td>
<td>15.1</td>
<td>0.428</td>
<td>E</td>
<td>0.715</td>
<td>35.25</td>
</tr>
<tr>
<td>(31, G)</td>
<td>282</td>
<td>7</td>
<td>10</td>
<td>3</td>
<td>0.248</td>
<td>12.08</td>
<td>0.476</td>
<td>G</td>
<td>0.583</td>
<td>25.38</td>
</tr>
</tbody>
</table>

\[\text{score}(31) = 0.285 \times 0.576 + 0.428 \times 0.75 + 0.428 \times 0.715 + 0.476 \times 0.583 = 1.608\]
ISRAELI FOREIGN MINISTER ARIEL SHARON TOLD REPORTERS DURING PICTURE-TAKING

B :: Clinton arrives in Israel, to go to Gaza, attempts to salvage Wye accord.

G :: Mid-east Wye Accord off-track as Clintons visit; actions stalled, violence

E :: President Clinton met Sunday with Prime Minister Netanyahu in Israel

F :: Clinton meets Netanyahu, says peace only choice. Office of both shaky

<table>
<thead>
<tr>
<th>pair(s_j,h_p)</th>
<th>n</th>
<th>l</th>
<th>k</th>
<th>ω</th>
<th>i</th>
<th>i-measure</th>
<th>w_d(s_j,h_p)</th>
<th>h_p</th>
<th>c_d(h_p)</th>
<th>μ(d,h_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(90, F)</td>
<td>282</td>
<td>9</td>
<td>8</td>
<td>0</td>
<td>0.255</td>
<td>0.00</td>
<td>0.00</td>
<td>F</td>
<td>0.576</td>
<td>35.25</td>
</tr>
<tr>
<td>(90, B)</td>
<td>282</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0.287</td>
<td>0.00</td>
<td>0.00</td>
<td>B</td>
<td>0.75</td>
<td>31.33</td>
</tr>
<tr>
<td>(90, E)</td>
<td>282</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0.255</td>
<td>3.91</td>
<td>0.11</td>
<td>E</td>
<td>0.715</td>
<td>35.25</td>
</tr>
<tr>
<td>(90, G)</td>
<td>282</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>0.319</td>
<td>0.00</td>
<td>0.00</td>
<td>G</td>
<td>0.583</td>
<td>25.38</td>
</tr>
</tbody>
</table>

score for 90 = .11 * .715 = 0.0786
<table>
<thead>
<tr>
<th>Score</th>
<th>Summary</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>ISRAELI FOREIGN MINISTER ARIEL SHARON TOLD REPORTERS DURING PICTURE-TAKING</td>
<td>0.0786</td>
</tr>
<tr>
<td>31</td>
<td>Clinton met Israeli Netanyahu put Wye accord</td>
<td>1.608</td>
</tr>
</tbody>
</table>
Correlation with ROUGE-1

Evaluation Tasks:

- Task 01: single doc. summarization
- Task 02: multi doc. summarization
- Task 05: question specific multi doc. summarization

<table>
<thead>
<tr>
<th>i-score vs. ROUGE-1</th>
<th>Spearman’s $\rho$</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>0.786</td>
<td>0.638</td>
</tr>
<tr>
<td>Task 2</td>
<td>0.713</td>
<td>0.601</td>
</tr>
<tr>
<td>Task 5</td>
<td>0.720</td>
<td>0.579</td>
</tr>
</tbody>
</table>

Table: Rank Correlations

- **Spearman’s Rank Correlation Coefficient**
  
  assesses how well the relationship between two variables (X and Y) can be described using a monotonic function. A positive (negative) Spearman correlation coefficient corresponds to an increasing (decreasing) monotonic trend between X and Y.

- **Kendall’s Rank Correlation Coefficient**
  
  measures the association between two measured quantities. A $\tau$-test is a non-parametric hypothesis test for statistical dependence.
Correlation with Human Judgement

Responsiveness score (DUC 2004, Task 5)

- For each doc. cluster, a single human was assigned to score each participants on the scale of 0 to 4.

A histogram divides the \( i \)-score based space into categories

<table>
<thead>
<tr>
<th>sys. id</th>
<th>given_score</th>
<th>guess_score</th>
</tr>
</thead>
<tbody>
<tr>
<td>147</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>122</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>86</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>109</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

normalized root mean square error (RMSE) = 0.303

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y - \hat{y}_i)^2}
\]
A “Heisenberg effect”: summaries are distorted by the way we evaluate them

- syntactic well-formedness is not part of evaluator algorithms
- the “bag of words” view (or n-grams, to a lesser extent) misses relevant information hidden in word ordering (subject versus complement position)
- site-words including negation are removed to make room to nouns and verbs
- rhetorical structures implying negative sentiment are not detected
- \(\Rightarrow\) negation and modality information tends to be missed
- more generally, sentiment analytics are ignored (and they are critical for things like a product or movie review)
Some remedies

- use i-measure to allow for flexibility for both human and computer-made summaries
- weight positively syntactic well-formedness
- interpret some logical elements like modality, negation, quantifiers
- use a more abstract representation for words (e.g. word2vec vectors) that encapsulates context information
- add sentiment analysis: the summary should reflect key sentiment elements, especially if product descriptions, media reviews, political believes are involved
Extractive vs. abstractive summaries

- human-made summaries are abstractive
- computer-made summaries (for now) are mostly extractive
- ⇒ semantic equivalences are needed to compare them fairly
  - replace words with Wordnet synsets
  - define equivalence relations using common Wordnet hypernyms
  - replace words with word2vec vectors, encapsulating context information learned from a large corpus like Wikipedia
  - a “distributed representation” for words as vectors obtained from the hidden layer of a shallow neural network trained with
    - the “continuous bag of words” architecture predicts the current word based on the context
    - the “skip-gram” architecture predicts surrounding words given the current word
- ⇒ graph-based methods could be used to test overall semantic connectivity between summaries in the context of the document they are extracted or abstracted from
- relativize summaries to natural context (ontology, domain) of a given document set
The case of scientific papers

- not a good idea to have your favorite category-theory, genomics or string-theory paper summarized by the Mechanical Turk
- fortunately, scientific papers come with an author-written abstract
- ⇒ building a “diamond standard” from (PDF-extracted) author-written abstracts and unicode approximations of the documents
- adding to it an implementation of a fair and flexible evaluation algorithm
- adding reference implementations of “classic algorithms” (e.g. TextRank)
- should we use some graph-based techniques not only to generate but also to evaluate computer generated summaries
Revisiting TextRank

what can we use as nodes?
- words, synsets
- word2vec vectors
- sentences
- semantic frames, conceptual graphs

what can we use as edges?
- equality
- equivalences
- distances
  - wordnet tree-walk steps
  - word2vec vectors: cosine similarity provides weights
How can we improve existing computer generated summaries?

- **ontology driven summarizers**
  - detecting the overall context the document is about - placing it on a concept map
  - prioritizing sentences that match key elements of the concept map (via semantic distances and via graph ranking)
  - abstractive aspects: text simplification, using dominant words of the ontology

- **identify “natural sources” for training machine learning algorithms** (possibly ontology dependent)
  - 1 star 5 stars product or media reviews
  - number of followers on social media
  - up-down votes for forums like stack exchange
  - impact factors for scientific papers (hIndex, number of downloads etc.)
  - causal explanations in online media for stock marked fluctuations
  - factual information accuracy: e.g. the Onion vs. Google News
Conclusions

- accurate computer-based evaluation of computer-generated summaries is far from being obvious or easy
- most of the shortcomings might come from the (unavoidable) simplifications that statistical measures need to assume
- accurate evaluation is useful - including for their use in machine learning
- tools like the i-measure introduce some flexibility
- evaluation of summaries needs to be relativized w.r.t. human-to-human variations
- trusting human-made summaries is ontology-dependent: questionable for scientific documents or even for fact checking or media reviews
- small steps of progress are happening: from natural language “processing” to natural language understanding