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

Outline

- comparing two summaries is sensitive to their lengths and the length of the document they are extracted from
- $\Rightarrow$ 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
- $\Rightarrow$ weighted relatedness to reference summaries
- comparing human written abstractive summaries to machine generated extractive ones
- $\Rightarrow$ we need an evaluation mechanism using semantic equivalence relations
- $\Rightarrow$ 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

$$\frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} count(gram_n)}$$

- 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

Average size of the intersection of two subsets

Given a set N of size n, and two randomly selected subsets with l and k elements, the average size of the intersection is:

$$avg(n,k,l)_{random} = \frac{\sum_{i=0}^{k} i \binom{k}{i} \binom{n-k}{l-i}}{\sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{l-i}} \qquad (1)$$

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

$$i\text{-}measure(N,K,L) = \frac{observed\_size\_of\_intersection}{random\_size\_of\_intersection} = \frac{\omega}{i} = \frac{\omega}{kl/n} \qquad (2)$$

- less sensitivity towards length

i-measure vs. *f-measure*

Two random sets K_r and L_r:

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

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

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

$$\begin{aligned}
f\text{-}measure_{random} &= 2pr/(p+r) \\
&= 2(l/n)(k/n)/(l/n+k/n) \\
&= 2(lk)/(n^2)/((k+l)/n) \\
&= 2(lk)/(n(k+l)) \\
&= 2i/(k+l) \\
&= i/((k+l)/2)
\end{aligned} \tag{6}$$

i-measure as relativized f-measure

Same computation for observed intersection size ω

from i-measure to f-measure

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

we get

$$\begin{aligned} i\text{-}measure(N,K,L) &= \frac{\omega/((k+l)/2)}{i/((k+l)/2)} \\ &= \frac{f\text{-}measure_{observed}}{f\text{-}measure_{random}} \end{aligned} \tag{7}$$

$\Rightarrow$ 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
- $\Rightarrow$ no need to trim summaries on byte or word length
- Evaluating the Evaluators
  - compare overlaps between each pair with *i-measure*
  - $\Rightarrow$ 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

DUC 2004

$D = \{d_1, d_2, \ldots, d_t\}$
$H = \{h_1, h_2, \ldots, h_z\}$
$S = \{s_1, s_2, \ldots, s_\lambda\}$
for each document d, a subset of annotators (say, $H_d = \{h_1, h_2, \ldots, h_m\}$) write summaries independently

Confidence-based scoring

Step 01

normalize i-measure (based on best pair)

$$
\begin{aligned}
w_d(h_p, h_q) &= \frac{i\text{-}measure(d, h_p, h_q)}{\mu_d} \\
w_d(s_j, h_p) &= \frac{i\text{-}measure(d, s_j, h_p)}{\mu_{(d,h_p)}}
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
\mu_d &= max(i\text{-}measure(d, h_p, h_q)), \forall (h_p, h_q) \in H_d \times H_d, h_p \neq h_q \\
\mu_{(d,h_p)} &= max(i\text{-}measure(d, s, h_p)), \forall s \in S
\end{aligned}
$$

Confidence-based scoring - continued

Step 02

define a degree of confidence to each reference

$$c_d(h_p) = \frac{\sum_{q=1,p\neq q}^{m}(w_d(h_p,h_q))}{m-1}. \tag{9}$$

Step 03

assign a weighted score for each system-generated summary

$$score(s_j,d) = \sum_{p=1}^{m} c_d(h_p) \times w_d(s_j,h_p) \tag{10}$$

Step 04

average the score

$$i\text{-}score(s_j) = \frac{\sum_{i=1}^{t} score(s_j,d_i)}{t}. \tag{11}$$

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

$i\text{-}measure(d, B, G)$ is $\frac{3}{10*9/282}$ 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

$i\text{-}measure(d, G, F)$ is $\frac{1}{10*8/282}$ which is 3.525

The 4 human-made summaries

normalize *i-measure*

Reference	Summary
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: reference summaries (B,G,E,F) on document $D30053.APW19981213.0224$

Normalized i-measures

$Pair(p, q)$	n	k	l	ω	i	i-measure	$w_d(h_p, h_q)$
(G , F)	282	10	8	1	0.28	3.52	0.375
(G, B)	282	10	9	3	0.32	9.40	1.0
(G, E)	282	10	8	1	0.28	3.52	0.375
(F, B)	282	8	9	1	0.25	3.91	0.4166
(F, E)	282	8	8	2	0.22	8.81	0.9375
(E, B)	282	8	9	2	0.25	7.83	0.8333

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

$c_d(G) = \frac{0.375+1+0.375}{3}$ which is 0.583

$c_d(B) = \frac{0.375+0.4166+0.833}{3}$ which is 0.75

reference: h_p	confidence: $c_d(h_p)$
G	0.583
F	0.576
B	0.75
E	0.715

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

$pair(s_j, h_p)$	n	l	k	ω	i	i-measure	$w_d(s_j, h_p)$
(31 , F)	282	7	8	2	0.198	10.07	0.285
(31, B)	282	7	9	3	0.223	13.42	0.428
(31, E)	282	7	8	3	0.198	15.1	0.428
(31, G)	282	7	10	3	0.248	12.08	0.476

h_p	$c_d(h_p)$	$\mu(d, h_p)$
F	0.576	35.25
B	0.75	31.33
E	0.715	35.25
G	0.583	25.38

$$score(31) = .285 * .576 + .428 * .75 + .428 * .715 + .476 * .583 = 1.608$$

Calculate scores for a computer-made summary: a bad one

90: ISRAELI FOREIGN MINISTER ARIEL SHARON TOLD REPORTERS DURING PICTURE-TAKIN=

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

$pair(s_j, h_p)$	n	l	k	ω	i	i-measure	$w_d(s_j, h_p)$
(90 , F)	282	9	8	0	0.255	0.00	0.00
(90, B)	282	9	9	0	0.287	0.00	0.00
(90, E)	282	9	8	1	0.255	3.91	0.11
(90, G)	282	9	10	0	0.319	0.00	0.00

h_p	$c_d(h_p)$	$\mu(d, h_p)$
F	0.576	35.25
B	0.75	31.33
E	0.715	35.25
G	0.583	25.38

$score$ for 90 = $.11 * .715$ = 0.0786

Compare scores

90

summary: *ISRAELI FOREIGN MINISTER ARIEL SHARON TOLD REPORTERS DURING PICTURE-TAKIN=*

score := 0.0786

31

summary: *Clinton met Israeli Netanyahu put Wye accord*

score := 1.608

Correlation with ROUGE-1

Evaluation Tasks:

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

i-score vs. ROUGE-1	Spearman's ρ	Kendall's τ
Task 1	0.786	0.638
Task 2	0.713	0.601
Task 5	0.720	0.579

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 τ-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

sys. id	*given_score*	*guess_score*
147	3	2
122	2	2
B	4	4
86	2	0
109	3	3
H	3	4
F	4	4

normalized root mean square error (RMSE) = 0.303

$$RMSE = \sqrt{1/n\sum_{i=1}^{n}(y - \hat{y}_i)^2}$$

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
- $\Rightarrow$ 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 a being obvious or easy
- most of the shortcoming 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