Evaluating Text Summarization Systems with a Fair Baseline from Multiple Reference Summaries

Fahmida Hamid(✉), David Haraburda, and Paul Tarau

Department of Computer Science and Engineering,
University of North Texas, Denton, TX, USA
{fahmida.hamid,dharaburda,ptarau}@gmail.com

Abstract. Text summarization is a challenging task. Maintaining linguistic quality, optimizing both compression and retention, all while avoiding redundancy and preserving the substance of a text is a difficult process. Equally difficult is the task of evaluating such summaries. Interestingly, a summary generated from the same document can be different when written by different humans (or by the same human at different times). Hence, there is no convenient, complete set of rules to test a machine generated summary. In this paper, we propose a methodology for evaluating extractive summaries. We argue that the overlap between two summaries should be compared against the *average intersection size* of two random generated *baselines* and propose ranking machine generated summaries based on the concept of *closeness* with respect to reference summaries. The key idea of our methodology is the use of *weighted relatedness* towards the reference summaries, normalized by the relatedness of reference summaries among themselves. Our approach suggests a relative scale, and is tolerant towards the length of the summary.

Keywords: Evaluation technique · Baseline · Summarization · Random average · Reference summary · Machine-generated summary

1 Introduction

Human quality text summarization systems are difficult to design and even more difficult to evaluate [1]. The extractive summarization task has been most recently portrayed as *ranking sentences* based on their likelihood of being part of the summary and their salience. However different approaches are also being tried with the goal of making the ranking process more semantically meaningful, for example: using synonym-antonym relations between words, utilizing a semantic parser, relating words not only by their co-occurrence, but also by their semantic relatedness. Work is also on going to improve anaphora resolution, defining dependency relations, etc. with a goal of improving the *language understanding* of a system.

A series of workshops on text summarization (WAS 2000-2002), special sessions in ACL, CoLING, SIGIR, and government sponsored evaluation efforts in

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 351–365, 2016.
DOI: 10.1007/978-3-319-30671-1_26

United States (DUC 2001-DUC2007) have advanced the technology and produced a couple of experimental online systems [15]. However there are no common, convenient, and repeatable evaluation methods that can be easily applied to support system development and comparison among different summarization techniques [8].

Several studies ([9,10,16,17]) suggest that *multiple human gold-standard summaries would provide a better ground for comparison*. Lin [5] states that multiple references tend to increase evaluation stability although human judgements only refer to single reference summary.

After considering the evaluation procedures of ROUGE [6], Pyramid [12], and their variants e.g., ParaEval [19], we present another approach to evaluating the performance of a summarization system which works with one or many reference summaries.

Our major contributions are:

- We propose *the average or expected size of the intersection of two random generated summaries* as a generic *baseline* (Sects. 3 and 4). Such a strategy was discussed briefly by Goldstein et al. [1]. However, to the best of our knowledge, we have found no direct use of the idea while scoring a summarization system. We use the baseline to find a *related (normalized)* score for each reference and machine-generated summaries.
- Using this baseline, we outline an approach (Sect. 5) to evaluating a summary. Additionally, we outline the rationale for a new measure of summary quality, detail some experimental results and also give an alternate derivation of the average intersection calculation.

2 Related Work

Most of the existing evaluation approaches use absolute scales (e.g., precision, recall, f-measure) to evaluate the performance of the participating systems. *Such measures can be used to compare summarization algorithms, but they do not indicate how significant the improvement of one summarizer over another is* [1].

ROUGE (Recall Oriented Understudy for Gisting Evaluation) [6] is one of the well known techniques to evaluate single/multi-document summaries. ROUGE is closely modelled after BLEU [14], a package for machine translation evaluation. ROUGE includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the machine-generated summary and the reference summaries.

Among the major variants of ROUGE measures, e.g., ROUGE-N, ROUGE-L, ROUGE-W, and, ROUGE-S, three have been used in the Document Understanding Conference (DUC) 2004, a large-scale summarization evaluation sponsored by NIST. Though ROUGE shown to correlate well with human judgements, it considers fragments, of various lengths, to be equally important, a factor that rewards low informativeness fragments unfairly to relative high informativeness ones [3].

Nenkova [12] made two conclusions based on their observations:

- DUC scores cannot be used to distinguish a good human summarizer from a bad one
- The DUC method is not powerful enough to distinguish between systems

Another piece of work that we would like to mention is the Pyramid method [11]. A key assumption of the method is the need for multiple models, which taken together, yield a gold standard for system output. A pyramid represents the opinions of multiple human summary writers each of whom has written a model summary for the multiple set of documents. Each tier of the pyramid *quantitively* represents the agreements among human summaries based on *Summary Content Units (SCU)* which are content units, not bigger than a clause. SCUs that appear in more of the human summaries are weighted more highly, allowing differentiation between important content from less important one.

The original pyramid score is similar to a *precision metric*. It reflects *the number of content units that were included in a summary* under evaluation as highly weighted as possible and it penalizes the content unit when a more highly weighted one is available but not used. We would like to address following important aspects here -

- Pyramid method does not define a *baseline* to compare the degree of (dis)agreement between human summaries.
- High frequency units receive higher weights in the Pyramid method. Nenkova [13], in another work, stated that the frequency feature is not adequate for capturing all the contents. To include less frequent (but more informative) content into machine summaries is still an open problem.
- There is no clear direction about the summary length (or compression ratio).

Our method uses a unique baseline for all (system, and reference summaries) and it does not need the absolute scale (like f, p, r) to score the summaries.

3 A *Generic Baseline* for All

We need to ensure a single rating for each system unit [7]. Besides, we need a common ground for comparing available multiple references to reach a unique standard. Precision, Recall, and F-measure are not exactly good fit in such case.

Another important task for an evaluation technique is defining a *fair baseline*. Various ways (first sentence, last sentence, sentences overlapped mostly with the title, etc.) are being tried to generate the baseline. Nenkova [13] designed a baseline generation approach: *SumBasic*. It was applied over DUC 2004 dataset. But we need a generic way to produce the baseline for all types of documents. *The main task of a baseline is to define (quantify) the least possible result that can be compared with the competing systems to get a comparative scenario.*

Compression ratio plays an important role in summarization process. If the compression ratio is too high, it is difficult to cover the stated topic(s) in the summary. Usually the compression ratio is set to a fixed value (100 words, 75 bytes, etc.) So, the participants are not free to generate the summary as they might want. We believe the participants should be treated more leniently on selecting the size of summary. When it is allowed, we need to make sure *the evaluation is not affected due to the length.*

The following two sections discuss about our proposed *baseline*, its relationship with precision, recall, f-measure and how to use it for computing the integrity of a (both reference and system generated) summary.

3.1 Average (Expected) Size of Intersection of Two Sets

Given a set N of size n, and two randomly selected subsets $K_1 \subseteq N$ and $K_2 \subseteq N$ with k elements each, the average or expected size of the intersection ($|K_1 \cap K_2|$) is

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

For *two* randomly selected subsets $K \subseteq N$ and $L \subseteq N$ of sizes k and l (say, $k \leq l$) respectively this formula generalises to

$$avg(n,k,l)_{random} = \frac{\sum_{i=0}^{k} i \binom{k}{i} \binom{n-k}{l-i}}{\sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{l-i}}. \tag{2}$$

For each possible size $i = \{0..k\}$ of an intersecting subset, the numerator sums the product of i and the number of different possible subsets of size i, giving the total number of elements in all possible intersecting subsets. For a particular size i there are $\binom{k}{i}$ ways to select the i intersecting elements from K, which leaves $n-k$ elements from which to choose the $k-i$ non-intersecting elements (or $l-i$ in the case of two randomly selected subsets). The denominator simply counts the number of possible subsets, so that the fraction itself gives the expected number of elements in a randomly selected subset.

Simplifying Equation 2: Equation 2 is expressed as a combinatorial construction, but the probabilistic one is perhaps simpler: the probability of any element x being present in both subset K and subset L is the probability that x is contained in the intersection of those two sets $I = L \cap K$.

$$\begin{aligned} &\Pr(x \in K) \cdot \Pr(x \in L) \\ &\qquad = \Pr(x \in (L \cap K)) \\ &\qquad = \Pr(x \in I) \end{aligned} \tag{3}$$

Putting another way, the probability that an element x is in K, L, or I is k/n, l/n and i/n respectively (where i is the number of elements in I). Then from Eq. 3 accordingly,

$$(k/n)(l/n) = i/n \tag{4}$$

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

A combinatorial proof, relying on identities involving binomial coefficients shows that Eqs. 2 and 5 are equivalent and is contained in Appendix A.

3.2 Defining f-$measure_{expected}$

Recall and *Precision* are two re-known metrics to define the performance of a system. Recall (r) is the ratio of *number of relevant information received to the total number of relevant information in the system.* Precision (p), on the other hand, is the ratio of *number of relevant records retrieved to the total number (relevant and irrelevant) of records retrieved.* Assuming the subset with size k as the gold standard, we define recall, and precision for the randomly generated sets as:

$$r = \frac{i}{k} \qquad p = \frac{i}{l} \qquad f\text{-}measure = \frac{2pr}{p+r}$$

Therefore, f-measure (the balanced harmonic mean of p and r) for these two random sets is:

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

3.3 Defining f-$measure_{observed}$, with Observed Size of Intersection 'ω'

Let, for a machine generated summary L and a reference summary K, the observed size of intersection, $|K \cap L|$ is ω.

$$r = \frac{|K \cap L|}{|K|} = \frac{\omega}{k} \qquad p = \frac{|K \cap L|}{|L|} = \frac{\omega}{l}$$

f-$measure$, in this case, can be defined as,

$$\begin{aligned} f\text{-}measure_{observed} &= 2pr/(p+r) \\ &= \tfrac{2\cdot\omega^2}{k\cdot l} / \tfrac{(k+l)\cdot\omega}{k\cdot l} \\ &= 2\omega/(k+l) \\ &= \omega/((k+l)/2) \end{aligned} \tag{7}$$

4 The i-measure: A Relative Scale

A more direct comparison of an observed overlap, seen as the intersection size of two sets K and L, consisting of lexical units like unigrams or n-grams drawn from a single set N is provided by the *i-measure*:

$$\begin{aligned} i\text{-}measure(N,K,L) &= \tfrac{observed\_size\_of\_intersection}{expected\_size\_of\_intersection} \\ &= \tfrac{|K \cap L|}{\frac{|K|\cdot|L|}{|N|}} = \tfrac{\omega}{\left(\frac{kl}{n}\right)} = \tfrac{\omega}{i} \end{aligned} \tag{8}$$

By substituting ω and i using Eqs. 7 and 6, we get,

$$i\text{-}measure(N,K,L) = \frac{f\text{-}measure_{observed}}{f\text{-}measure_{expected}}$$

Interestingly, *i-measure* turned out as a ratio between the observed *f-measure* and the expected/ average *f-measure*. The *i-measure* is a form of *f-measure* with some tolerance towards the length of the summaries.

In the next section, we prepare an example to explain how *i-measure* adjusts the variation on lengths, yet produces comparable score.

4.1 Sample Scenario: *i-measure* Balances the Variation in Length

Suppose we have a document with $n = 200$ unique words, a reference summary composed of $k = 100$ unique words, and a set of machines $\{a, b, \ldots, h, i\}$. Each machine generates a summary with l unique words. Table 1 outlines some sample scenarios of *i-measure* scores that would allow one to determine a comparative performance of each of the systems.

For system b, e, and h, ω is the same, but the *i-measure* is highest for h as its summary length is smaller than the other two. On the other hand, systems e

Table 1. Sample cases: *i-measure*

case	n	k	l	i	ω	*i-measure*	sys. id
$k = l$	200	100	100	50	30	0.6	a
	200	100	100	50	45	0.9	b
	200	100	100	50	14	0.28	c
$k < l$	200	100	150	75	30	0.4	d
	200	100	150	75	45	0.6	e
	200	100	150	75	14	0.186	f
$k > l$	200	100	80	40	30	0.75	g
	200	100	80	40	45	1.125	h
	200	100	80	40	14	0.35	i

and a receive the same *i-measure*. Although ω is larger for e, it is penalized as its summary length is larger than a. We can observe the following properties of the *i-measure*:

- The system's summary size (l) does not have to be exactly same as the reference' summary size size (k); which is a unique feature. Giving this flexibility encourages systems to produce more informative summaries.
- If k and l are equal, *i-measure* follows the observed intersection, for example b wins over a and c. In this case i-measure shows a compatible behavior with *recall* based approaches.
- For two systems with different l values, but same intersection size, the one with smaller l wins (e.g., a,d, and g). It indicates that system g (in this case) was able to extract important information with greater compression ratio; this is compatible with the *precision* based approaches.

5 Evaluating a System's Performance with Multiple References

When multiple reference summaries are available, a fair approach is to compare the machine summary with each of them. *If there is a significant amount of disagreement among the reference (human) summaries, this should be reflected in the score of a machine generated summary. Averaging* the overlaps of machine summaries with human written ones does not *weigh* less informative summaries differently than more informative ones. Instead, the evaluation procedure should be modified so that it first compares the reference summaries among themselves in order to produce some weighting mechanism that provides a fair way to judge all the summaries and gives a unique measure to quantify the machine generated ones. In the following subsections we introduce the dataset, weighting mechanism for references, and finally, outline the scoring process.

5.1 Introduction to the Dataset and System

Our approach is generic and can be used for any summarization model that uses multiple reference summaries. We have used *DUC*-2004 structure as a model. We use *i-measure*(d, x_j, x_k) to denote the i-measure calculated for a particular document d using the given summaries x_j and x_k.

Let λ machines ($S = \{s_1, s_2, \ldots, s_\lambda\}$) participate in a *single document summarization task*. For each document, m reference summaries ($H = \{h_1, h_2, \ldots, h_m\}$) are provided. We compute the *i-measure* between $\binom{m}{2}$ pairs of reference summaries and normalize with respect to the best pair. We also compute the *i-measure* for each machine generated summary with respect to each reference summary and then normalize it. We call these *normalized i-measures* and denote them as

$$\begin{aligned} w_d(h_p, h_q) &= \frac{i\text{-}measure(d,h_p,h_q)}{\mu_d} \\ w_d(s_j, h_p) &= \frac{i\text{-}measure(d,s_j,h_p)}{\mu_{(d,h_p)}} \end{aligned} \tag{9}$$

where,

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

The next phase is to build a heterogeneous network of systems and references to represent the relationship.

5.2 Confidence Based Score

We assign each reference summary h_p a "confidence" $c_d(h_p)$ for document d by taking the average of its *normalized i-measure* with respect to every other reference summary:

$$c_d(h_p) = \frac{\sum_{q=1,p\neq q}^{m}(w_d(h_p, h_q))}{m-1}. \tag{10}$$

Taking the confidence factor associated with each reference summary allows us to generate a score for s_j:

$$score(s_j, d) = \sum_{p=1}^{m} c_d(h_p) \times w_d(s_j, h_p) \tag{11}$$

Given t different tasks (single documents) for which there are reference and machine generated summaries from the same sources, we can define the total performance of system s_j as

$$i\text{-}score(s_j) = \frac{\sum_{i=1}^{t} score(s_j, d_i)}{t}. \tag{12}$$

Table 2. Reference summaries (B,G,E,F) and three machine summaries on document $D30053.APW19981213.0224$

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
90	ISRAELI FOREIGN MINISTER ARIEL SHARON TOLD REPORTERS DURING PICTURE-TAKIN=
6	VISIT PALESTINIAN U.S. President Clinton met to put Wye River peace accord
31	Clinton met Israeli Netanyahu put Wye accord

Table 3. Normalized i-measure of all possible reference pairs for document: $D30053.APW19981213.0224$

$Pair(p, q)$	n	k	l	ω	i	i-$measure$	$w_d(h_p, h_q)$
(G , F)	282	10	8	1	0.283687	3.525	0.375
(G, B)	282	10	9	3	0.319148	9.4	1.0
(G, E)	282	10	8	1	0.283687	3.525	0.375
(F, B)	282	8	9	1	0.255319	3.916	0.4166
(F, E)	282	8	8	2	0.226950	8.8125	0.9375
(E, B)	282	8	9	2	0.255319	7.8333	0.8333

Table 4. Confidence score

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

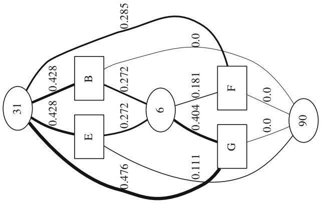

Fig. 1. System-reference graph: edge-weights represent the normalized *i-measure*

Table 5. Confidence based system score

System Id(s_j)	$score(s_j, d_i)$
31	0.2676
6	0.1850
90	0.0198

Table 2 shows four reference summaries (B, G, E, F) and three machine summaries $(31, 90, 6)$ for document *D30053.APW19981213.0224*. Table 3 shows the normalized *i-measure* for each reference pair. While comparing the summaries, we ignored the *stop-words* and *punctuations*. Tables 4 and Table 5, and Fig. 1 represents some intermediate calculation using Eqs. 10 and 11 for document D30053.APW19981213.0224.

6 Evaluating Multi-document Summary

Methodology defined in Sect. 5.2 can be adapted for evaluating *multi-document summaries* with minor modifications. Let, there are q clusters of documents, i.e. $D = \{D_1, D_2, \ldots, D_q\}$. Each cluster D_i contains t number of documents, $D_i = \{d_1, \ldots, d_t\}$. The system has a set of humans $(H = \{h_1, h_2, \ldots, h_z\})$ to generate gold summaries. For each D_i, a subset of humans $(H_{D_i} = \{h_1, h_2, \ldots, h_m\}, m \leq z)$ write m different *multi-document summaries*.

We need to compute a score for system s_j among λ participating systems $(S = \{s_1, s_2, \ldots, s_\lambda\})$. We, first, compute $score(s_j, D_i)$ for each D_i using formula 11. Then we use formula 12 to find the rank of s_j among all participants.

The only difference is at defining the *i-measure*. The value of n (total number of units like unigram, bi-gram etc.) comes from all the participating documents in D_i, other than a single document.

7 Experimental Results

We perform different experiments over the dataset. Section 7.1 describes how *i-measure* among the reference summaries can be used to find the confidence/

nearness/ similarity of judgements. In Sect. 7.2, we examine two types of rank-correlations (pair-based, distance based) generated by *i-score* and *ROUGE*-1. Section 7.3 states the correlation of *i-measure* based ranks with human assessors.

7.1 Correlation Between Reference Summaries

The *i-measure* works as a preliminary way to address some intuitive decisions. We discuss them in this section with two extreme cases.

- If the *i-measure* is too low (Table 6) for most of the pairs, some of the following issues might be true:-
  - The document discusses about diverse topics.
  - The compression ratio of the summary is too high even for a human to cover all the relevant topics discussed in the document.
  - The probability of showing high performance by a system is fairly low in this case.
- If the *i-measure* is fairly close among most of the human pairs (Table 3), it indicates:-
  - The compression ratio is adequate
  - The document is focused into some specific topic.
  - If a system shows good performance for this document, it is highly probable that the system is built on good techniques.

Therefore, the *i-measure* could be an easy technique to select ideal documents that are good candidates for summarization task. For example, Table 3 shows that all of the reference pairs have some words in common, hence their confidence score (Table 4) is fairly high. But Table 7 shows that most of the references do not share common words, hence *confidence* values of the references for document $D30015.APW19981005.1082$ is quite different from each other.

Table 6. Normalized i-measure of all possible reference pairs for $D30015.APW19981005.1082$

$Pair(p,q)$	n	k	l	ω	i	*i-measure*	$w_d(h_p, h_q)$
(A, H)	357	9	10	0	0.25210	0.0	0.0
(A, B)	357	9	10	3	0.25210	11.9	1.0
(A, E)	357	9	7	1	0.17647	5.66	0.4761
(H, B)	357	10	10	1	0.2801	3.57	0.3
(H, E)	357	10	7	0	0.19607	0.0	0.0
(B, E)	357	10	7	0	0.19607	0.0	0.0

Table 7. Confidence score

reference: h_p	confidence: $c_d(h_p)$
A	0.492
B	0.433
H	0.099
E	0.158

7.2 Correlation of Ranks: ROUGE-1 Vs. I-Score

To understand how the confidence based i-measures compare to the ROUGE-1 metric, we calculated *Spearman's* ρ [18] and *Kendall's* τ [4], (both of which are rank correlation coefficients) by ranking the machine and reference summary scores. Spearman's ρ considers the squared difference between two rankings while

Table 8. Rank correlations

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
i-score vs. f-measure		
Task 1	0.896	0.758
Task 2	0.955	0.838
Task 5	0.907	0.772

Table 9. Guess score for $D188$, assessor F

sys. id	*given_score*	*guess_score*
147	3	2
43	2	3
122	2	2
B	4	4
86	2	0
24	1	1
109	3	3
H	3	4

Kendall's τ is based on the number of concordant/discordant pairs (Table 8). Since the list of stopwords used by us can be different from the one used by ROUGE system, we also calculate pure *f-measure* based rank and report the correlation of with *i-score*. The results show, for both cases, i-measure is positively correlated, but not completely.

7.3 Correlation with Human Judgement: Guess the ***RESPONSIVENESS*** score

For multi-document summarization (DUC2004, task5), the special task (RESPONSIVENESS) was to assess the machine summaries per cluster (say, D_i) by a single human-assessor (h_a) and score between 0 to 4, to reflect the *responsiveness* on a given topic (question). We have used a *histogram* to divide the *i-score* based space into 5 categories ($\{0, 1, 2, 3, 4\}$). We found 341 decisions out of 947 responsiveness scores as an exact match (36.008 % accuracy) to the human assessor. Table 9 is a snapshot of the scenario.

The *Root Mean Square Error (RMSE)* based on *i-score* is 1.212 at the scale of 0 to 4. Once normalized over the scale, the error is 0.303

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

7.4 Critical Discussion

After carefully analyzing the system generated summaries, rouge based scores, and i-score, we noticed that most of the systems are not producing well-formed sentences. Scoring based on weighted/un-weighted overlapping of *bag-of-important-phrases* is not the best way to evaluate a summarizer. Constraint on the length of the summary (byte/word) might be a trigger. As *i-measure* is lenient on lengths, we can modify Eq. 11 with the following to apply *extraction/generation of proper sentences* within a maximum word/sentence window as an impact factor.

$$score(s_j, d) = \left(\sum_{p=1}^{m} c_d(h_p) \times w_d(s_j, h_p) \right) \times \frac{c\_sen}{t\_sen} \tag{13}$$

where, t_sen is the total number of sentences produced/ extracted by s_j and c_sen is the number of grammatically *well-formed* sentences. For example, *"This is a meaningful sentence. It can be defined using english grammar."* is a delivered summary. Suppose, the allowed word-window-size is 8. So the output is chopped as *"This is a meaningful sentence. It can be"*. Now it contains 1 well-formed sentence out of 2. Then the *bag of words/phrases* model (e.g., *i-measure*) can be applied over it and a score can be produced using Eq. 13.

Standard sentence tokenizers, POS taggers, etc. can be used to analyze sentences. The word/ sentence window-size can be determined by some ratio of sentences (words) present in the original document. As we could not find any summary-evaluation conferences who follow similar rules (TREC, DUC, etc.), we were unable to generate results based on this hypothesis.

8 Conclusion

We present a mathematical model for defining a generic *baseline*. We also propose a new approach to evaluate machine-generated summaries with respect to multiple reference summaries, all normalized with the baseline. The experiments show comparable results with existing evaluation techniques (e.g., ROUGE). Our model correlates well with human decision as well.

The *i-measure* based approach shows some flexibility with summary length. Instead of using average overlapping of words/phrases, we define pair based $confidence$ calculation between each reference. Finally, we propose an extension of the model to evaluate the quality of a summary by combining the bag-of-words like model to accredit *sentence structure* while scoring.

We will be extending the model, in future, so it works with semantic relations (e.g. synonym, hypernym etc.) We also need to investigate some more on the confidence defining approach for question-based/ topic-specific summary evaluation task.

A Appendix

The equivalence of Eqs. 2 and 5 can be shown using the following elementary identities on binomial coefficients: the *symmetry rule*, the *absorption rule* and *Vandermonde's convolution* [2].

Proof. Consider first the denominator of Eq. 2. The introduction of new variables makes it easier to see that identities are appropriately applied, and we do so here by letting $s = n - k$ and then swapping each binomial coefficient for its symmetrical equivalent (*symmetry rule*).

$$\sum_{i=0}^{k} \binom{k}{i} \binom{s}{l-i} = \sum_{i=0}^{k} \binom{k}{k-i} \binom{s}{s-l+i}$$

Substituting $j = s - l$ for clarity shows that *Vandermonde's convolution* can be applied to convert the sum of products to a single binomial coefficient, after which we back substitute the original variables, and finally apply the *symmetry rule*.

$$\begin{aligned}\sum_{i=0}^{k} \binom{s}{j+i}\binom{k}{k-i} &= \binom{k+s}{j+k} \\ &= \binom{k+n-k}{n-k-l+k} \\ &= \binom{n}{n-l} \\ &= \binom{n}{l}\end{aligned}$$

The numerator of Eq. 2 can be handled in a similar fashion, after the i factor is removed using the *absorption rule*.

$$\sum_{i=0}^{k} i\binom{k}{i}\binom{n-k}{l-i} = k\sum_{i=0}^{k} \binom{k-1}{i-1}\binom{n-k}{l-i}$$

Applying *Vandermonde's convolution* yields:

$$k\sum_{i=0}^{k} \binom{k-1}{-1+i}\binom{n-k}{l-i} = k\binom{n-1}{l-1}$$

Eq. 2 has now been reduced to

$$k\binom{n-1}{l-1}\Big/\binom{n}{l}$$

A variation of the *absorption rule* allows the following transformation

$$k\binom{n-1}{l-1}\Big/\binom{n}{l} = k\binom{l}{1}\Big/\binom{n}{1}$$

which reduces to kl/n.

References

1. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.:Summarizing text documents: sentence selection and evaluation metrics. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 121–128. ACM, New York (1999). http://doi.acm.org/10.1145/312624.312665
2. Graham, R., Knuth, D., Patashnik, O.: Concrete Mathematics: A Foundation for Computer Science. Addison-Wesley, Boston (1994)
3. Hovy, E., Lin, C.-Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006) (2006)
4. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938). http://www.jstor.org/stable/2332226

5. Lin, C.Y.: Looking for a few good metrics: automatic summarization evaluation - how many samples are enough? In: Proceedings of the NTCIR Workshop 4 (2004)
6. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries, pp. 25–26 (2004)
7. Lin, C.Y., Hovy, E.: Manual and automatic evaluation of summaries. In: Proceedings of the ACL-2002 Workshop on Automatic Summarization, AS 2002, vol. 4, pp. 45–51. Association for Computational Linguistics, Stroudsburg, PA, USA (2002). http://dx.doi.org/10.3115/1118162.1118168
8. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, vol. 1, pp. 71–78. Association for Computational Linguistics, Stroudsburg, PA, USA (2003). http://dx.doi.org/10.3115/1073445.1073465
9. Mani, I., Maybury, M.T.: Automatic summarization. In: Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Companion Volume to the Proceedings of the Conference: Proceedings of the Student Research Workshop and Tutorial Abstracts, p. 5, Toulouse, France, 9-11 July 2001
10. Marcu, D.: From discourse structures to text summaries. In: Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 82–88 (1997)
11. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process. 4(2) (2007). http://doi.acm.org/10.1145/1233912.1233913
12. Nenkova, A., Passonneau, R.J.: Evaluating content selection in summarization: the pyramid method. In: HLT-NAACL, pp. 145–152 (2004). http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/91_Paper.pdf
13. Nenkova, A., Vanderwende, L.: The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Technical report MSR-TR-2005-101 (2005)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 311–318. Association for Computational Linguistics, Stroudsburg, PA, USA (2002). http://dx.doi.org/10.3115/1073083.1073135
15. Radev, D., Blair-Goldensohn, S., Zhang, Z., Raghavan, R.: Newsinessence: a system for domain-independent, real-time news clustering and multi-document summarization. In: Proceedings of the First International Conference on Human Language Technology Research (2001). http://www.aclweb.org/anthology/H01-1056
16. Rath, G.J., Resnick, A., Savage, T.R.: The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. Am. Documentation **12**, 139–141 (1961). http://dx.doi.org/10.1002/asi.5090120210
17. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Inf. Process. Manage. **33**(2), 193–207 (1997). http://dx.doi.org/10.1016/S0306-4573(96)00062-3

18. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **15**(1), 72–101 (1904). http://www.jstor.org/stable/1412159
19. Zhou, L., Lin, C.Y., Munteanu, D.S., Hovy, E.: Paraeval: using paraphrases to evaluate summaries automatically. Association for Computational Linguistics, April 2006. http://research.microsoft.com/apps/pubs/default.aspx?id=69253