Summarization Evaluation for Text and Speech: Issues and Approaches

Ani Nenkova

Stanford University
anenkova@stanford.edu

Abstract

This paper surveys current text and speech summarization evaluation approaches. It discusses advantages and disadvantages of these, with the goal of identifying summarization techniques most suitable to speech summarization. Precision/recall schemes, as well as summary accuracy measures which incorporate weightings based on multiple human decisions, are suggested as particularly suitable in evaluating speech summaries.
Index Terms: evaluation, text and speech summarization

1. Introduction

Much has been said about the information overload chacterizing modern-day life, the constant need for timely access and digest of news, email, scientific publications and other information sources. Such concerns have sparked interest in automatic summarization as early as the late 1950s [1] and have lead to the development of numerous summarization applications for news, email threads, discussion lists and chat-rooms, lay and professional medical information, scientific articles, spontaneous dialogues, voicemail, broadcast news and video, and meeting recordings.

In addition to the research challenges in developing these summarization systems, the question of how best to evaluate their results has emerged as a research issue in itself. Ideally, summarization results need to be assessed in a task-based setting, determining their usefulness as part of an information browsing and access interface (extrinsic evaluation) [2, 3, 4]. But such extrinsic evaluations are time-consuming, expensive and require a considerable amount of careful planning. They are thus not very suitable for system comparisons and evaluation during development. Intrinsic evaluations are normally employed in such cases, either by soliciting human judgments on the goodness and utility of a given summary, or by a comparision of the summary with a human-authored gold-standard. When comparisions with a gold-standard are involved, it is desirable that these be done automatically to further reduce the need for human involvement.

In this paper, we provide an overview of the main approaches used in summarization evaluation and the motivation for developing certain evaluation methods and for abandoning others. We also discuss some of the lessons learned in past evaluations.

2. A brief history of news summarization evaluation

2.1. Precision and recall

Most summarization systems select the most representative sentences in the input to form a *extractive* summary; the selected sentences are strung together to form a summary without any modification of their original wording. In such settings, the commonly used information retrieval metrics of precision and recall can be used: a person is asked to select sentences that seem to best convey the meaning of the text to be summarized and then the sentences selected automatically by a system are evaluated against the human selections. Recall is the fraction of sentences chosen by the person that were also correctly identified by the system

$$Recall = \frac{|\text{system-human choice overlap}|}{|\text{sentences chosen by human}|} \quad (1)$$

and precision is the fraction of system sentences that were correct

$$Precision = \frac{|\text{system-human choice overlap}|}{|\text{sentences chosen by system}|} \quad (2)$$

The appeal of precision and recall as evaluation measure is obvious: after a human defines the gold-standard sentence selection, it can be repeatedly used to evaluate automatically produced summaries by a simple comparison of sentence ids. Unfortunately, there are also several problems.

Human variation Different people tend to choose different sentences. Research as early as [5] reported that extracts selected by six different human judges for 10 articles from Scientific American had only 8% overlap on average. It is thus unclear how to define a gold-standard. It has been shown [6] that the same summary can obtain a recall score that is between 25% and 50% different depending on which of two available human extracts are used for evaluation. Thus, a system can choose a good sentence, but sill be penalized in P/R evaluation. In light of this observation, it also seems that in summarization evaluation it might be more beneficial to concentrate on recall rather then precision. Precision might be overly strict—some of the sentences chosen by the system might be good, even if they have not been chosen by the gold-standard creator. Recall, on the other hand, measures the overlap with already observed sentence choices.

Granularity Another problem with the P/R measures is the fact that sentences are not the best granularity for measuring content. Different sentences might differ in word length and convey different amounts of information. Selecting a longer and more informative sentence can be more desirable than selecting a short sentence. Imagine, for example, a human extract consisting of the sentences "(1) We need urgent help. (2) Fires have spread in the nearby forest, and threaten several villages in this remote area." Now imagine two systems, each choosing only one sentence appearing in the human extract, one choosing sentence (1) and the other choosing sentence (2). Both summaries will have the same P/R score, but can hardly be perceived as equally informative.

Semantic equivalence Yet another problem with using sentences as the selection unit is that two distinct sentences can express the same meaning. This situation is very common in news, and is particularly pertinent in multi-document summarization of

news, in which the input to the system consists of many articles on the same topic. Again, a human would select only one of the equivalent sentences but a system will be penalized for choosing an alternate sentence that expresses the same meaning.

Many of the subsequently developed evaluation measures were designed to address the issues that were raised regarding P/R. For example, it has been suggested to use multiple human models rather than a single person's judgment [7], smaller, more-semantically oriented units of analysis have been proposed, and more emphasis has been given on recall.

2.2. Relative utility

Relative utility [8] has been proposed as a way to address the human variation and semantic equivalence problems in P/R evaluation. In this method, multiple judges score *each sentence in the input* on a scale from 0 to 10 as to its suitability for inclusion in a summary; highly ranked sentences are very suitable for a summary, and low ranked sentences should not be included in a summary. The judges also explicitly mark which sentences are mutually substitutable because of semantic equivalence. Thus, each possible selection of sentences by a system can be assigned a score showing how good a choice of sentences it represents.

The approach seems intuitive and quite appealing, but requires a good deal of manual effort in sentence tagging. Moreover, it does not seem to be very good at discriminating between human and automatic summaries, a distinction which a good evaluation measure should be able to do. Particularly when applied to the evaluation of SWITCHBOARD summaries [9], automatic summarizers achieved a score higher than that of the humans, indicating that this approach for evaluation is not a good choice for evaluation of summarization of conversational speech.

2.3. DUC manual evaluation

The Document Understanding Conference (DUC)[1] has been carrying out large-scale evaluations of summarization systems on a common dataset since 2001. On average, over 20 different sites participate in this NIST-run evaluation each year and a lot of effort has been invested by the conference organizers to improve evaluation methods. DUC content evaluations are still based on a single human model. However, in order to mitigate the bias coming from using gold-standards from only one person, different annotators create the models for different subset of the test data [10].

In order to address the need for better analysis granularity than the sentence level, DUC used elementary discourse units (EDUs) as the basis for evaluation. These EDUs roughly correspond to clauses. Each human model was automatically split into EDUs, and machine summaries were evaluated by the degree to which they cover each EDU in the model. The average score, called *coverage* was the average EDU score for the summary under evaluation. The measure was recall-oriented, in essence measuring what fraction of the model EDUs were covered by a summary.

In an attempt to encourage research in abstractive summarization[2], where the system alters the original wording of sentences by merging information from different sentences, or removing parts of the sentences, DUC also started using human abstracts as model, rather than human selection of sentences. The above-described evaluation method supported this transition, at the expense of requiring more human involvement.

[1] `http://duc.nist.gov`

[2] Which is the typical summarization approach for people.

The availability of the output of many systems over many test inputs (varying between 20 and 50 in different years) has allowed researchers to study the factors that influence summarization performance. It has been reported that in ANOVA analysis of coverage scores with system, input and model creator as factors, the model creator turned out to be the most significant factor [11]. This once again raised concerns about the advisability of using a single human model for evaluation. The input document to be summarized was also a significant factor [12], suggesting that some inputs are easier to summarize than others.[3] Summary length was also a significant factor, with summary coverage tending to increase as the length of the summary increases.

Two lines of research on evaluation emerged in an effort to address some of the issues raised by the DUC evaluation protocol: developing cheap automatic methods for comparing human gold-standards with automatic summaries, better analysis of human variation of content selection variation, and using multiple models to avoid result dependence on the gold-standard.

2.4. Automatic evaluation and ROUGE

Automatic evaluation measures have been known even before the widely used BLEU technique for machine translation evaluation [13] and the ROUGE technique derived from it [14] (see for example [15]). The problem has been that different automatic evaluation approaches give different results, so it was not clear what the scores mean and which automatic measure is to be preferred. In using BLEU for machine translation evaluation, however, researchers developed methods to validate automatic approaches. They took manual evaluations generally accepted in the research community, and looked for automatic measures which correlated well with the human scores *over a large set of test points*, especially when multiple human models were used. Inspired by the success of the BLEU n-gram overlap based measure, similar n-gram matching was tried for summarization. Using DUC coverage scores to validate the method, the ROUGE[4] system for automatic evaluation of summarization was developed. ROUGE is also based on the computation of n-gram overlap between a summary and a set of models. ROUGE is recall-oriented, unlike BLEU, which emphasizes precision. The new recall-oriented n-gram counting was shown to correlate better than BLEU with DUC coverage scores. ROUGE has numerous parameters, including words stemming, stopword removal and n-gram size. Different settings work best for different summarization tasks as can be seen from the detailed tables is [14]. This means that different parameters need to be tested for new tasks, such as speech summarization of spontaneous conversations or recordings of meetings. Certain ROUGE configurations has been shown to correlate well with DUC coverage, although this does not necessarily mean that it will correlate well with other human evaluation methods.

2.5. Pyramid Method

The Pyramid Method [16] was concerned with analysis of the variation in human summaries, as well as how evaluation results can be made less dependent on the model used for evaluation. Multiple

[3] This finding shows that paired test such as paired t-test, Wilcoxon sign rank test, or paired permutation tests should be used when comparing the performance of two systems on the same test set. These tests eliminate the variation that is due to the input difficulty and lead to better assessment of the significance of difference between the systems.

[4] Recall-Oriented Understudy for Gisting Evaluation.

human abstracts are analyzed manually to derive a gold-standard for evaluation. The analysis is semantically driven: information with the same meaning, even when expressed using different wording in different summaries, is marked as expressing the same summary content unit (SCU). Each SCU is assigned a weight equal to the number of human summarizers who expressed the SCU in their summaries. The distribution of SCU weights is Zipffian, with few SCUs being included by many summarizers and a heavy tail of low-weight SCUs[5]. SCU analysis shows that summaries that differ in content can be equally good and assign a score that is stable with respect of the models when 4 or 5 human summaries are used. The actual pyramid score is equal to the ratio between the weight of content expressed in a summary and the weight of an ideally informative summary with the same number of SCUs.

A drawback of this approach is that it is very labor intensive, despite the fact that a special annotation tool (DUCView[6]) has been developed to facilitate the process. Also, the method was developed for evaluation of abstractive summaries, and requires analysis that is unnecessary for extractive summaries, as we will see in later sections.

2.6. Readability evaluation

All the evaluation methods discussed so far have been focused on evaluating the information content of a summary, its overall informativeness. But summary readability is also an important factor in summary evaluation, albeit often neglected by summarization researchers. In DUC, a separate set of questions were developed to evaluate readability aspects of summaries. Are they ungrammatical? Do they contain redundant information? Are the references to different entities clear? Does the summary build up sentence by sentence? While much progress has been seen in improving system content selection, most automatic summaries score rather poorly on readability aspects such as coherence and referential clarity [17]. Improving automatic summary readability is an open problem in news summarization, and undoubtedly will be relevant for speech summarization applications as well.

Recent interest in the topic of sentence ordering and referential cohesion have lead to a proposal for automatic evaluation of cohesion [18]. Hopefully, more effort will be focused on readability issues and evaluation in the near future.

3. Intrinsic evaluation for speech summarization

While many speech summarization researchers have used precision/recall of utterances [19, 20] or automatic measures such as ROUGE to evaluate their results, there have been two proposals for evaluation methods specifically designed for the new genre.

Summary accuracy Summary accuracy was defined by Zechner and Waibel [21]: for each word in an utterance they define a weight, which they call a relevance score, equal to the average number of times the word occurred in a phrase selected for inclusion in the summary by a human annotator.[7] So, if five annotators are asked to construct a summary, and exactly three of them select the same span of text, all the words in this span will be assigned a relevance score equal to $3/5$. Summary accuracy is then defined as equal to the sum of relevance scores of the (correctly recognized by the ASR system) words in a system-selected utterance, divided by the *maximum achievable relevance score* with the same number of words somewhere in the text. This definition of word relevance (weight) and overall summary score is very similar to the idea on which the pyramid evaluation method for news is based. In fact, while attempting to apply the pyramid method for evaluation of meeting transcripts, Galley [22] observed that in this domain human summaries formed by sentence extraction convey the same information only when the two annotators extracted exactly the same sentence.[8] He then computed pyramid scores based on words rather than content units, with the restriction that a given word is assigned a non-zero (relevance) score only when it is part of *the same* utterance that the humans selected. This scoring worked out quite well for the meeting domain, and is almost equivalent to summary accuracy. Such reinvention of scoring metrics is very indicative of the need for closer interaction between researchers tackling different types of summarization genres.

Summarization accuracy Summarization accuracy has been defined in the context of the evaluation of a summarization system performing sentence compaction [23]. The sentence compaction task is to eliminate 'unnecessary' words from a sentence. Again, multiple annotators are asked to produce possible compactions. Even in this case, many possible compactions for a given sentence can be produced. In order to extrapolate more possible compactions from those produced by the annotators, all human productions for a sentence are merged into a single network and different traversals of the network can produce new compaction variants that were not produced by any of the humans, but that are considered possible. The thus enriched network is then used to evaluate the summarization accuracy of the automatic compaction. This evaluation procedure also allows for weighting of words that are included in the summary by many humans.

The summarization accuracy measure has been found to work well for high compression ratios, but results in problems for summaries at small ratios such as 10% [24]. In such cases, the authors propose the use of a score based on individual comparisons between the automatic summary and all the manual summaries, choosing the best score among all the individual comparisons. This idea is very interesting and has not been explored in news summarization: it suggest that rather than using the multiple human summaries for weighting, one can find the human summary that is most similar to the produced machine summary.

What about using ROUGE? The use of a generally agreed on and automatic metric such as ROUGE is hugely appealing. It allows for cheap evaluation and ease in comparing results from different research efforts. For these reasons, researchers have investigated the degree to which ROUGE scores correlate with human judgments of informativeness of such summaries. In [25], subjective human judgments were collected for summaries of meetings (6 test meetings), and compared with several of the popular ROUGE variants. ROUGE scores were not found to correlate with the human judgments on this data. More disturbingly, when [22] compared automatic and human summaries for the same test meetings, ROUGE scores were not able to distinguish between the two types. Both results suggest that the use of ROUGE is not advisable for

[5] Hence the name of the method. If SCUs are ordered in tiers from low to high weight, we get a pyramid

[6] `http://www1.cs.columbia.edu/~ani/DUCView.html`

[7] Such a definition addresses the *granularity* problem discussed in Section 2.1 for precisions/recall, because using word-by-word comparison accounts for the possibly different informativeness of utterances.

[8] This fact suggests that the *semantic equivalence problem* of precision/recall might not be an issue for meeting summarization evaluation.

this type of data and with so few test points.[9] In a separate, much larger study on a different type of data [24], ROUGE-2 (as well as summarization accuracy and F-score) measures were found to highly correlate with human judgments on a five point scale. Such findings suggest that ROUGE should be used only when a large number of test points is available.

4. Discussion

The evaluation methods surveyed in this paper suggest a strong tendency in the summarization community, especially in text summarization, to favor the use of multiple human models for intrinsic evaluation, which allow for an importance weighting of information. Weighted precision and recall as used in [26], or summary accuracy measures such as those in [21, 22] seem particularly suitable for speech summarization. The use of widely available automatic metrics such as ROUGE could also be possibly used, but only given a large number of test points. Task-based evaluation and the integration of speech summarization in information browsing and access interfaces [27, 28, 29] also present interesting opportunities for assessing the usefulness of automatic speech summaries.

5. References

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[2] I. Mani, G. Klein, D. House, L. Hirschman, and T. Firmin abd B. Sundheim, "Summac: a text summarization evaluation," *Natural Language Engineering*, vol. 8, no. 1, pp. 43–68, 2002.

[3] K. McKeown, R. Passonneau, D. Elson, A. Nenkova, and J. Hirschberg, "Do summaries help? a task-based evaluation of multi-document summarization," in *SIGIR*, 2005.

[4] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, 2005, In press.

[5] G. J. Rath, A. Resnick, and R. Savage, "The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines," *American Documentation*, vol. 2, no. 12, pp. 139–208, 1961.

[6] R. Donaway, K. Drummey, and L. Mather, "A comparison of rankings produced by summarization evaluation measures," in *NAACL-ANLP Workshop on Automatic Summarization*, 2000.

[7] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad, "Summarization evaluation methods: Experiments and analysis," in *AAAI Symposium on Intelligent Summarization*, 1998.

[8] D. Radev and D. Tam, "Single-document and multi-document summary evaluation via relative utility," in *Poster session, CIKM'03*, 2003.

[9] X. Zhu and G. Penn, "Evaluation of sentence selection for speech summarization," in *Proceedings of RANLP workshop on Crossing Barriers in Text Summarization Research*, 2005.

[10] D. Harman and P. Over, "The effects of human variation in duc summarization evaluation," in *ACL Text summarization branches out workshop*, 2004.

[11] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, B. Schiffman, and S. Teufel, "Columbia multi-document summarization: Approach and evaluation," in *DUC*, 2001.

[12] Ani Nenkova, "Automatic text summarization of newswire: lessons learned from the document understanding conference," in *Proceedings of AAAI'05*, 2005.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002.

[14] C. Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *ACL Text Summarization Workshop*, 2004.

[15] D. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Çelebi, D. Liu, and E. Drabek, "Evaluation challenges in large-scale multi-document summarization: the mead project," in *Proceedings of ACL 2003*, Sapporo, Japan, 2003.

[16] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *HLT/NAACL*, 2004.

[17] Ani Nenkova, *Understanding the process of multi-document summarization: content selection, rewrite and evaluation*, Ph.D. thesis, Columbia University, January 2006.

[18] M. Lapata and R. Barzilay, "Automatic evaluation of text coherence: Models and representations," in *IJCAI'05*, 2005.

[19] I. Gurevych and M. Strube, "Semantic similarity applied to spoken dialogue summarization," in *COLING*, 2004.

[20] S. Maskey and J. Hirschberg, "Summarizing speech without text using hidden markov models," in *Short paper, HLT-NAACL*, 2006.

[21] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *NAACL*, 2000.

[22] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *EMNLP*, 2006.

[23] C. Hori, T. Hori, and S. Furui, "Evaluation methods for automatic speech summarization," in *EUROSPEECH*, 2003.

[24] S. Furui, M. Hirohata, Y. Shinnaka, and K. Iwano, "Sentence extraction-based automatic speech summarization and evaluation techniques," in *Proceedings of Symposium on Large-Scale Knowledge Resources (LKR'05)*, 2005.

[25] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *ACL Workshop on Evaluation Measures for MT/Summarization*, 2005.

[26] G. Murray, S. Renals, J. Carletta, and J. Moore, "Incorporating speaker and discourse features into speech summarization," in *Proceedings of HLT/NAACL'06*, 2006.

[27] J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, G. Zamchick, and S. Whittaker, "SCANMail: Browsing and searching speech data by content," in *EUROSPEECH*, 2001.

[28] N. Papernick and A. Hauptmann, "Summarization of broadcast news video through link analysis of named entities," in *In AAAI-05 Workshop on link analysis*, 2005.

[29] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker, "A meeting browser evaluation test," in *CHI extended abstracts*, 2005.

[9] The results are also consistent with results reported in [9], which report that ROUGE metrics were not correlated with summary accuracy for evaluation of SWITCHBOARD conversations.