Summarization and Evaluation; Where are we today?![*]

Mehrnoush Shamsfard[a], Amir Saffarian[a], Samaneh Ghodratnama[b]

[a] NLP Laboratory, Electrical and Computer Engineering Department,
Shahid Beheshti University, Velenjak, Tehran, Iran
m-shams@sbu.ac.ir
a_saffarian@std.sbu.ac.ir

[b] Department of Computer Science and Engineering,
Shiraz University, Molasadra, Shiraz, Iran
ghodratnama@cse.shirazu.ac.ir

Abstract. The rapid growth of the online information services causes the problem of information explosion. Automatic text summarization techniques are essential for dealing with this problem. There are different approaches to text summarization and different systems have used one or a combination of them. Considering the wide variety of summarization techniques there should be an evaluation mechanism to assess the process of summarization. The evaluation of automatic summarization is important and challenging, since in general it is difficult to agree on an ideal summary of a text. Currently evaluating summaries is a laborious task that could not be done simply by human so automatic evaluation techniques are appearing to help this matter. In this paper, we will take a look at summarization approaches and examine summarizers' general architecture. The importance of evaluation methods is discussed and the need to find better automatic systems to evaluate summaries is studied.

Keywords: Summarization, Evaluation, Recall, Precision, Pyramid, ROUGE

1. Introduction

Today, having access to information summaries is one of the important needs for humans which may affect their lives. These summaries generally are produced with the help of other humans. The question arises here is that in this era where we are encountering a huge volume of increasing data and information, with limited resources to decide on, may it be really possible to use humans as information summarizers? We believe not.

So, researchers have created methods to automatically do summarization task and extract the most important concepts of the information without any need to involve humans. Now a second question emerges. How could we know about the correctness and the completeness of the results generated in an automatic manner? The answer is hidden in evaluation techniques on which we are going to focus in this paper. Therefore besides the growth of summarization methods, evaluation techniques must improve and mature so as to provide acceptable scores for summaries according to some sort of evaluation metrics.

Evaluation approaches mostly evaluate created summaries based on (1) ideal (gold) summary, (2) use in an application and (3) original document. Techniques from the first category compare system generated summaries with a summary known as the best possible summary! This ideal summary currently is created by humans and it could be influenced by subjective effects from judges. Application driven techniques evaluate content of summary by analyzing the level of information that could be obtained from it for a specific task. In this approach the performance of the application is analyzed when using the original document and its summary separately. On the other hand evaluation by original document is really ambiguous because specifying evaluation parameters and metrics is hard but it is more natural.

In this paper we focus on the first category of evaluation approaches; comparing with an ideal (gold) summary. Section 2 of this paper provides a brief explanation about the summarization methods and their high level architecture. Then in the following sections some recent methods for summary evaluation are introduced.

2. Document Summarization

Text summarization is the process of extracting the most important parts of information from source document(s) to produce a reduced version for a particular user or a particular task. According to Mani (2001) automatic summarization is an automated process in which a computer takes a piece of information, also called the source (i.e. an electronic document), selects the most important content in it, and presents that content to the user in a condensed form.

Automatic text summarization can be used in various areas of applications such as telecommunications industry, intelligent tutoring systems, text mining, and filters for web-based information retrieval and word processing tools. Researchers are also investigating the application of this technology to a variety of new and challenging problems, including Multilingual Summarization, Multimedia News Broadcast, Summarization of Online Medical Literature of a Patient, Audio Scanning Services for the Blind and Providing Captions for TV Programs.

2.1. High Level Architecture

As a general view for summarizer architecture, the input to the system could be one or more documents in different forms of text or multimedia. The process of summarization has three main phases:

1. Analyzing the input text or *Topic Identification*
2. Transforming it into a summary representation or *Interpretation*
3. Synthesizing an appropriate output form or *Generation*

Any summary can be characterized by (at least) three major classes of characteristics: (1) Input: characteristics of the source text(s), (2) Output: characteristics of the summary as a text (3) Purpose: characteristics of the summary usage which is discussed in more details in Hovy (2000).

The output may be an extract of the source, or an abstract. Extracts consist of portions of text extracted verbatim, but abstracts consist of novel phrasings which describe the content of the original document(s). In general, producing abstracts requires stages of topic fusion and text generation, but producing extracts requires only the stage of topic identification. Moreover, summarizers can usually produce either generic or user–focused summaries. A user-focused summary is the one in which specific information is selected in order to satisfy the requirements of a particular user group. As opposed to that, a generic summary is more suitable for the average reader. There are also two types of summaries depending on their function. There can be informative and indicative summaries. The purpose of the first type is to deliver as much information as possible to the user and can also serve as a substitute for the source. Indicative summaries on the other hand are only meant to help the user decide whether or not to read the source document.

2.2. Approaches

As defined by Mani (1999) there can be surface-level, corpus based, and discourse structure based and knowledge based approaches for single document summarization. All these approaches could successfully be applied to multi-document summarization.

The surface-level approach requires shallow understanding of the text and this usually involves analysis of the syntactic structure of sentences. It is used to extract salient information by taking into account some key features of a sentence. Corpus based approaches involve statistical analysis of large bodies of text (corpora) to find specific features about the documents in them.

Human abstractors create a mental discourse model of a document while reading it. Discourse-level approaches try to create similar model of the discourse structure of a document which can later be used for the generation of a summary. Knowledge based approaches are used for the creation of summarization systems which act in a specific domain (i.e. domain dependent approaches). Such systems usually produce high quality summaries but do not have the ability to adapt to different types of documents (domains).

Almost all of the summarizer systems in the world use a combination of approaches mentioned above. The final report of SUMMAC project by Mani (1998) listed some the systems in its time. For example, BT’s ProSum used statistical techniques based on the co-occurrences of word stems, the length of sentences and their position in the original text to calculate the importance of a sentence in the context of the overall text in which it occurs. The most important sentences are then used to construct the summary. CIR created a thematic representation of a text that included nodes of thematically related terms simulating topics of the text. Related terms were identified using a thesaurus specially constructed for this task. CGI_CMU used a technique called “Maximal Marginal Relevance” (MMR) which produces summaries of very long documents by identifying key relevant, non-redundant information found within the document. This technique is also used for eliminating or clustering redundant information in multi-document summarization applications. Cornell SabIR used the document ranking and passage retrieval capabilities of the SMART IR engine to effectively identify relevant related passages in a document. GE identified the discourse macro structure for each document and selected the passages from each component that scored well using both content and contextual clues. Currently, we are working to create a more updated list of single- and multi-document summarizer systems and classifying them by the methods they use to create summaries.

3. Summary Evaluation

Text summarization is still an emerging field and serious questions remain concerning the appropriate methods and types of evaluation. There are a variety of possible bases for comparison of summarization system performance e.g., summary to source, machine to human generated, system to system.

The problem with matching a system summary against a best of breed summary is that it is really hard and maybe impossible to find an ideal summary. Indeed, the human summary may be supplied by the author of the article, by a judge asked to construct an abstract, or by a judge asked to extract sentences. There can be a large number of generic and user-focused abstracts that could summarize a given document, just as there can be many ways of describing something.

3.1. Categories of Methods

Methods for evaluating text summarization approaches can be broadly classified into two categories.

1. Intrinsic methods which are based on the comparison of the automatically produced summary with an ideal summary usually produced by human abstractors.
2. Extrinsic methods which are mostly based on tasks that use the output results of summarization systems.

From another point of view we can divide evaluation methods into manual and automatic ones. In manual techniques an individual or a group of individuals compare machine or human generated summaries (peers) to an ideal summary (model) or source text to find out about the extent of coverage of the main concepts between peer and model summaries. At the end an average of the scores given by group of judgments are used for the diversity of ideal summaries mentioned in the previous paragraph.

In the other hand, automatic methods, without any human interference, try to reduce subjective effects that may influence evaluation scores. Members of this family may have different levels of complexity according to the level of details they use for comparison. Those that use lexical similarities as their main approach are the simplest and those which use semantics to find the coverage of concepts are the most complex techniques here.

3.2. Evaluation Metrics

The comparison between summaries is best carried out by humans, but it can also be computed automatically. A variety of different measures can be used. Evaluation metrics can be grouped in at least three categories: (1) Sentence Recall measures, (2) Utility-based measures and (3) Content-Based measures (Mani, 2001-2).

There are many metrics to evaluate summaries but precision and recall measures are used extensively in ideal summary based evaluation of summarization systems. However, they are not appropriate for the summarization task due to their binary nature and the fact that there is no single correct summary. An example of how the binary precision and recall measures fail the task is as follows: suppose there are two sentences which are interchangeable in terms of producing a summary.

If five human subjects extract sentences to build a summary, two of subjects chose sentence 1, two of them chose sentence 2 and the 5th chose sentence 1 by chance. By majority method, sentence 1 will be an ideal sentence. In evaluation, if a summarization system chooses sentence 1, it wins, if it chooses sentence 2, it loses, although sentence 1 and 2 are interchangeable. This is obviously not a good measure (Jing, 1998).

A tested summary should agree with model not only in content but also in length, an aspect which is measured by Precision. Combining both informativeness and brevity is possible through F-Score, a measure often used in Information Retrieval:

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (1)$$

There could be situations that a combinational metrics is used based on the several parameters. For example in Rigouste (2003) a recall metric is introduced which is using the unigram and bigram matching techniques mentioned by Lin and Hovy:

$$F - Score = \frac{1}{3}\frac{Overlap_{Unigrams}}{N} + \frac{21}{3}\frac{Overlap_{Bigrams}}{N} \qquad (2)$$

4. Pyramid, a Manual Evaluation Technique

The motivation behind pyramid method (Passonneau, 2003) is that most of the contents which appear in human summaries are conceptions from the source text, expressed by different sentences or words. The fact behind generating subjective summaries that avoids any two summaries to be unique is that information units in them could be prioritized based on their importance in people views. Pyramid method is introduced to abstract and prioritize content units in source text considering this human behavior.

Actually, a pyramid is made up of summary content units (SCU). There is no accurate definition of SCU because the granularity of information comprising it, is not clear enough. In this technique functional or semantic specifications of SCUs are not very important; however focus is on the way in which summaries are compared to find similar and non-similar SCUs. All

SCUs are weighted according to the frequency they appear in different summaries and after that each summary is scored based on the SCUs it is comprised of. More information about SCU and how to identify them could be found in DUC2005 SCU Annotation Guide[1].

Suppose the pyramid has n tiers, with tier T_n on top and T_1 on the bottom. The weight of SCUs in tier T_i will be i . Let | T_i | denote the number of SCUs in tier T_i. Let D_i be the number of SCUs in the summary that appear in T_i . Other SCUs in a summary that do not appear in the pyramid are assigned weight zero. The total SCU weight D is computed as in equation 3 and the optimal content score for a summary with X SCUs is shown in equation 4.

$$D = \sum_{i=1}^{n} D_i \qquad (3)$$

$$Max = \left(\sum_{i=j+1}^{n} w_{T_i} + |T_i|\right) + w_{T_i} + \left(X - \sum_{i=j+1}^{n} |T_i|\right) \qquad (4)$$

$$where\ j = \max_i\left(\sum_{i=j+1}^{n} |T_i| \geq X\right)$$

With the help of Pyramid, score of a candidate summary (*D*) is computed with dividing the frequency of SCUs appeared in D by the value acquired by the best distribution of SCUs in an ideal summary (*Max*).

The strengths of pyramid scores are that they are reliable, predictive, and diagnostic. There are also two problems with Pyramid method. First, pyramid scores ignore interdependencies among content units, including ordering. Second, creating an initial pyramid is laborious so large-scale application of the method would require an automated or semi-automated approach. DUC conference data of year 2005 and 2006 are analyzed using Pyramid method and the results are available in Nenkova (2005) and Passonneau (2006).

There are two tasks involved in Pyramid evaluation: creating a pyramid by annotating model summaries, and evaluating a new summary (peer) against a pyramid. Ideally, an automated evaluation component would address both tasks. However, the task of creating a pyramid is far more complex than the task of scoring a new summary against existing (hand created) pyramid, and the automated scoring component is useful when doing a large amount of evaluation (of multiple summarizers, or different versions of the same summarizer). Pyramid creators have proposed a four step algorithm to score summaries according to the manually created pyramid. Steps are as follows (Harnly, 2005):

- **Enumerate** Enumerates all candidate contributors (contiguous phrases) in each sentence of the peer summary.
- **Match** For each candidate contributor, find the most similar SCU in the pyramid. In the process, the similarity between the candidate contributor and all pyramid SCUs is computed.
- **Select** From the set of candidate contributors, find a covering, and disjoint set of contributors that have maximum overall similarity with the pyramid.
- **Score** Calculate the pyramid score for the summary, using the chosen contributors and their SCU weights.

More details including the reliability and robustness of the Pyramid method could be found in Nenkova (2007).

5. ROUGE, an Automatic Summary Evaluation System

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans (Lin ,2004).

[1] http://www1.cs.columbia.edu/~ani/DUC2005/AnnotationGuide.htm

There are five metrics introduced in ROUGE that we will briefly explain. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-L, computes the ratio between the length of the two summaries' longest common sub-sequence (LCS) and the length of the reference summary. One advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order as n-grams. The basic LCS also has a problem that it does not differentiate LCSs of different spatial relations within their embedding sequences. To improve the basic LCS method, another metric called ROUGE-W or weighted longest common sub-sequence that favors LCS with consecutive matches is introduced.

Skip-bigram co-occurrence statistics, ROUGE-S, measure the overlap ratio of skip-bigrams between a candidate summary and a set of reference summaries. Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. One potential problem for ROUGE-S is that it does not give any credit to a candidate sentence if the sentence does not have any word pair co-occurring with its references. To accommodate this, ROUGE-S is extended with the addition of unigram as counting unit. The extended version is called ROUGE-SU.

To assess the effectiveness of ROUGE measures, the correlation between ROUGE assigned summary scores and human assigned mean coverage scores is computed. The intuition is that a good evaluation measure should assign a good score to a good summary and a bad score to a bad summary. The ground truth is based on human assigned scores. The effectiveness of ROUGE measures are examined on 2001-2003 three years of DUC data and the result could be found in Lin (2004).

6. A Framework for Summary Evaluation Systems

Researchers have always been looking for a summary evaluation system that provides stable and reliable scores. Almost all current systems do their job by comparing peer summaries with some reference (human generated) ones but even without considering different styles of evaluation, comparison based on summary content does not have enough precision (subjective effects).

Experiences have shown that evaluation by sentence units is not good enough because sentences may have subparts with different importance. Even though comparison at the word level does not consider the effects of the context in which the words are used.

Basic Elements (BEs), as a new concept, was introduced by Hovy (2005) toward establishing a framework for automatic summary evaluation. They addressed the problem of unit size by automatically producing a series of increasingly larger units, starting at the single word level. Experimentally Basic Elements could be defined as:

1. The head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item, or
2. A relation between a head-BE and a single dependent, expressed as a triple (head | modifier | relation).

In order to implement Basic Elements as a method of evaluating summary content, four core questions must be addressed (Hovy, 2005):

- What or how large is a Basic Element? The answer to this is strongly conditioned by: How can BEs be created automatically?
- How important is each BE? What basic score should each BE have?
- When do two BEs match? What kinds of matches should be implemented, and how?
- How should an overall summary score be derived from the individual matched BEs' scores?

Different answers to each of these questions provide a different summary evaluation method. The Pyramid Method, takes approximately clause-length semantic units shared by the reference summaries as BEs; gives each unit a score equal to the number of reference summaries containing it; allows two units to match when they express all or most of the same semantic content, as judged by the assessors; and derives the overall score by summing the scores of each unit of the candidate summary and normalized by the overall score of an ideal summary of equal size.

In contrast, ROUGE uses various n-grams (for example, unigrams) as BEs; scores each unigram by a function that depends on the number of reference summaries containing that unigram; allows unigrams to match under various conditions (for example, exact match only, or root form match); and derives the overall summary score by some weighted combination function of unigram matches.

The BE Package is an overall framework in which various solutions to the four core questions are provided, and therefore serves as a generalization over the particular methods and as an environment to compare them. BE package contains four subsystems:

- **Breaker Units** which accepts a sentence as input and produces a list of BEs as output. Different BE Breakers produce different BEs.
- **Scoring Units** which could assign each BE some points for each reference summary it participates in.
- **Comparing and Matching Units** that use a range of increasingly sophisticated matching strategies like lexical identity, lemma identity, synonym identity to match phrases.
- **Combining Scores and Ranking Units** that add the point values of each BE in the summary to be evaluated and use some optimizations in the score integration task.

The major problem is developing powerful BE matching routines; if one can match minimal BEs (and paraphrases) accurately then building matchers for compound BEs should be an interesting but not an impossible difficult exercise. Similarly, determining optimal weighting functions for individual BEs and for their combination to maximize correlations with human judgments requires careful but not an impossible hard work, and resembles the work done by Lin (2004). The results for DUC[2] 2005 evaluation using basic elements could be found in Hovy (2005).

7. Conclusion

In this paper we have talked about document summarization and the new techniques recently used for their evaluation. It is clear that the current trend of summary evaluation approaches is toward automatic methods and this field of research is still immature mainly because:

1. Lack of accurate and clear definition for a summary which should be independent of the influences which come from summarizer (mainly humans).
2. The way humans try to evaluate summaries is not completely understood and clearly formulated.

Today, researchers are moving forward to find more important metrics in order to formulate the complete human judgment. They are accepting the subjective influences that affects summarizations as a fact and try to provide a basic framework based on these remarks that matches real world conditions. Pyramid as a tool could help us in evaluating summaries but the main problem was the effort needed to construct it. We reviewed some ideas about making this construction task automatic. Others suggested using machine translation techniques in the evaluation. Using paraphrases concerning subjective effects that appear in summaries is discussed in Zhou (2006).

[2] Document Understanding Conferences, http://duc.nist.gov/

References

Lin, C.-Y. 2004. Looking for a Few Good Metrics: *Automatic Summarization Evaluation - How Many Samples Are Enough*?, In Proceedings of NTCIR Workshop 4, Tokyo, Japan.

Harnly, A., A. Nenkova, R. Passonneau and O. Rambow. 2005. *Automation of Summary Evaluation by the Pyramid Method. Recent Advances in Natural Language Processing (RANLP-2005)*, Borovets, Bulgaria.

Hovy, E. and C.Y. Lin. 2000. *Automated Text Summarization and the SUMMARIST System ", Information Sciences Institute of the University of Southern California,* 197-214.

Hovy, E., C.-Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 Using Basic Elements. Document Understanding Workshop, Vancouver, B.C., Canada.

Jing, H., R. Barzilay, K. McKeown and M. Elhadad. 1998. *Summarization evaluation methods experiments and analysis. In AAAI Intelligent Text Summarization Workshop (Stanford, CA, Mar. 1998*), 60--68.

Mani, I., D. House and G. Klein. 1998. *The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report MTR 98W0000138,* MITRE Corporation, Virgina.

Mani, I. and Mark T. Maybury eds., 1999. Advances in Automatic Text Summarization. The MIT Press.

Mani, I. 2001. Automatic Summarization. John Benjamins Publishing Co.

Mani, I. 2001. Summarization Evaluation: An Overview. *Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. National Institute of Informatics*, Tokyo, Japan

Nenkova, A., R. Passonneau, K. McKeown and S. Sigelma. 2005. Applying the Pyramid Method in DUC 2005. Document Understanding Workshop, Vancouver, B.C., Canada.

Nenkova, A., R. Passonneau and K. McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. ACM Transactions on Speech and Language Processing, 4, 2 (May. 2007), 4. DOI= http://doi.acm.org/10.1145/1233912.1233913

Passonneau, Rebecca and Ani Nenkova. 2003. Evaluating content selection in human- or machine-generated summaries: The pyramid method. Technical Report CUCS025 -03, Columbia University.

Passonneau, R., K. McKeown, S. Sigelma and A. Goodkind. 2006, Applying the Pyramid Method in the 2006 Document Understanding Conference. Document Understanding Workshop, Brooklyn, New York USA.

Zhou, L., C.-Y . Lin, D. S. Munteanu and E. Hovy. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (New York, New York, June 04 - 09, 2006). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 447-454. DOI= http://dx.doi.org/10.3115/1220835.1220892

Rigouste L. 2003. *Evolution of a Text Summarization System in an Automatic Evaluation Framework. Master`s Thesis, Ottawa-Carleton Institute for Computer Science, School of Information Technology and Engineering,* University of Ottawa.