Summarization of Technical Articles: Modeling User's Expectation from a Summary Using Specificity Score

Shailesh Deshpande and Athiappan G.

Tata Research Development and Design Centre
{shailesh.deshpande, athiappan.g}@tcs.com

Abstract. Conventional importance based extractive summarization methods face many difficulties as notion of importance is malleable. Instead, users expectation from a summary could possibly be defined in more precise manner – say by some of the discourse properties. In this paper we use specificity score – a measure of how specific or generic a particular text is – to characterize the types of documents, and further encode the expectation from a summary. We further demonstrate use of specificity score to summarize technical articles. Our hypothesis is: users expect summary sentences to convey more specific information from a technical article.

1 Introduction

Extractive summarization techniques create summaries by selecting sentences that are important - in some sense - to the document. In abstractive summarization also, important sentences are extracted and then subsequently paraphrased to the required length of summary. Many techniques assign the importance to the sentences and order them accordingly to select top scoring *k* sentences as a summary. There are many methods for deciding the importance of the sentences: simple word frequency based, key words based, centrality based and so on.

Conventional - importance based summarization - has two difficulties: First, deciding important information for summary is a nontrivial task. Notion of importance is malleable and subjected to change based on point of view. Disagreement between summaries by experts is a well-studied behavior [**1**], [**2**], [**3**], [**4**]. Disagreement between experts on potential best summary affects the summary evaluation task as well. So instead of taking one model summary for evaluation, summaries from two three experts (may be more) are taken for comparison with peer summary [**5**], [**6**], [**7**]. Second, the sentences extracted might be important but not the expected one. The expectation from the summary sentences can be expressed in the form of discourse relations (or any other suitable property). For example; contradictory sentences should not be extracted, sentences should provide specification and so on. Traditional summarization techniques do not incorporate the mentioned expectation measure in the process. The mismatch between summaries by experts and summaries by importance based algorithms [**8**] is because of expectation mismatch.

A case in this point: Technical articles. Generally speaking, we expect more specific sentences from a summary of technical articles such as scientific publications, research report *etc*. Table 1 shows summary of Relative Utility (RU) paper by Radev *et. al.* [**2**] generated by one of the standard algorithms. These selected sentences are from the list of 15 (~15 %) sentences extracted as a summary by Lexrank algorithm [**9**]. They are ordered according to the sequence number of the sentence in the original document.

Table 1. Lexrank summary of RU paper

The main problem with traditional co-selection metrics (thus named because they measure the degree of overlap between the list of sentences selected by a judge and an automatically produced extract) such as Precision, Recall, and Percent Agreement for evaluating extractive summarizers is that human judges often disagree about which the top n% most important sentences in a document or cluster are and yet, there appears to be an implicit importance value for all sentences which is judge-independent.
We have measured the utility correlation for three judges on 3,932 sentences from 200 documents from the HK News corpus.
We will call this observation the principle of Summary Sentence Substitutability (SSS).

From the extracted sentences we can see that the sentences are important and introductory (introducing topic to the reader) or generic. If reader is interested in knowing most prevailing topic, or generic discussion in the paper, then this summary would perform reasonably well. But on the other hand, if reader expects more detailed information such as research findings, comparative assessment of method and so on then the sentences do not convey required information. Reader would expect sentences such as these (Table 2):

Table 2. Alternative expected summary sentences for RU paper

The average value of R across all documents at the 5% target length is 0.598 while the average value of J is 0. 799. The corresponding values for the 20% target length are R = 0 635 and J = 0.835.
Second, MEAD and WEBSUM score approximately the same on all metrics with MEAD doing slightly better on the Word overlap, Bigram overlap, and longest common subsequence measures and WEBSUM on the cosine metric.

In this paper we propose to use specificity of the sentences as a criterion to select the sentences for summary. We take two important types of documents; news articles and technical articles (papers) and show how specificity score can be used for summarization and for modeling expectation from a summary by the reader. The features for specificity (of a sentence) used in this paper are very similar to work by Deshpande *et. al.* [**10**], [**11**] and the work by Louis *et. al.* [**12**].

What is generic (information) and what is specific is less subjective than importance. The disagreement between experts on classifying sentences into specific and generic class is quite low [**11**] compared to the disagreement on important sentence (selected for summary). Moreover, summary can be determined by nature of infor-

mation imparted by the sentences that is if generic, or more specific, or a healthy mix of generic and specific sentences is required.

Predominantly, research in summarization is driven by intrinsic properties (for example, importance of a sentence) of the document. Whereas, using extrinsic criteria for driving summary process completely might be still impractical, deciding how specific or how generic information the summary should have - for performing a particular task - is a viable option. Only specificity score might not be sufficient to convey all the aspects of extrinsic measure, but it certainly helps in expressing the expectation from summary in more objective way – which in turn can be used to drive the summarization process.

2 Related work

The feature model for specificity for document understanding tasks first appear in [**10**], [**11**], and then in [**8**], [**12**]. Both of these models use very similar features such as length, semantic depth, named entities (NE) to characterize the specificity of a sentence. Louis and Nenkova [**12**] use supervised learning approach to classify sentences from model and peer summaries into specific and generic. Whereas, Deshpande *et. al.* [**10**], [**11**] use unsupervised approach to rank the sentences according to their specificity and then select top k sentences as very specific feedback from customer comments. Though these studies indicate complete model for specificity of the sentences and direct usage of specificity score for some of the document understanding tasks, earlier researchers indicate – explicitly or otherwise - need for studying nature of information (in sentences) for its generic and specific tendencies.

Jones [**13**] in her important work on the term indexing for information retrieval task, argues that indexing term specificity should not be decided semantically but rather should be defined statistically. Thus, highlights that the words appearing less frequently tend to be specific.

Hassel [**14**] tests his hypothesis, that sentences containing NE would be more important for summary without much success. The summaries created with NE feature do not show improvement in recall. The lack of improvement in recall rather decrease in recall can be explained using specificity scores: NE indicates more specific sentences (in general) than generic sentences. Naturally, evaluation of such summaries with model summaries with more generic sentences is bound to perform poorly - as in this case.

Halteran *et al.* [**1**] propose use of factoids – self-contained information units – for summary evaluation and study extensively how factoids from different reference summaries can be used to create consensus summary. In the study, they found more general factoids in reference summaries than more specific factoids (similar observation is reported by Louis-Nenkova [**8**]). The study reveals two observations that are important in present context: First, human tendency towards expressing facts in the documents (news articles in this case) in more general way than specific and second, intuitive hypothesis that importance and generalization are inversely proportional to each other (for some type of documents).

Jing et. al. [**15**] use three step process in producing summary – sentence extraction, sentence reduction, and sentences recombination. During the sentence recombination step, one of the substitution operations they suggest is to replace sentence (or its part) with more general or more specific information. They identify rules for these substitutions by manually analyzing human summaries.

Further discussion is organized as follows: Section 3 provide higher level approach for summarization studies using specificity score. Section 4 discusses summary generated by algorithm and its analysis. Section 5 concludes the work.

3 Methodology

3.1 Objectives

Broad level goal for the specificity experiments for summarization is to find the mapping between summarization factors and specificity – that is given a summarization factor (say PURPOSE – audience and use) [**16**], can we define specification of summarization system in terms of specificity? Specific goal for present work is to study a) how specificity score vary for summaries of different types of documents, b) how specificity vary for news articles with single lead (single story and its details) and multiple leads (single story with multiple sub-stories with its own details)? c) What are the characteristics of lead sentences (which are good summary sentences) in terms of specificity? We want test our hypothesis: a) more specific sentences are expected as a summary for some types of news reports (such as finance, interesting court or electoral cases, natural calamities), b) generic sentences are expected for research paper summary which is very close to the abstract of a paper, c) specific sentences form good summary for information such as research findings, scientific claims and so on.

3.2 Algorithm

In this section we provide glimpses of the specificity features and algorithm (Table 3) and describe how the specificity score is used for document understanding and summarization tasks. We show the results on news articles and technical papers. First, we calculate specificity scores for sentences of article of our interests. The plot (Fig. 1) of specificity score and sentences number reveals structure of the document in terms of specificity of information provided by each section. These plots are used for analyzing intentional structure of the document.

We begin with extracting various semantic and statistical features of a word and then of a sentence. Semantic Depth (SD) measures number of edges, in the hypernym tree from the WordNet [**17**], between the root word and a given word. For example, apple is more specific word than the word “fruit”. Average Semantic Depth (ASD) is a sentence level metric that measures average semantic depth of all the words in that sentence. Semantic Height (SH) is reciprocal of Semantic Depth and it measures number of edges from the leaf node in the hypernym tree from WordNet [**17**]. SH is

averaged over all the words in a given sentences to give Average Semantic Height (ASH). Total Occurrence Count (TOC) measures how many times a word occurs in the ontology such as WordNet [**17**]. More specific words tend to occur less frequently (generally speaking with a few exceptions) in the WordNet like ontology. We take three lowest count words and sum TOC of them to indicate TOC for a given sentence. Named Entity (NE) count and length are simple measures indicating number of NEs, and length (number of words) for a given sentence respectively (Please see [**11**] for detailed implementation). After calculating specificity scores for each sentence, summarization can be approached as follows:

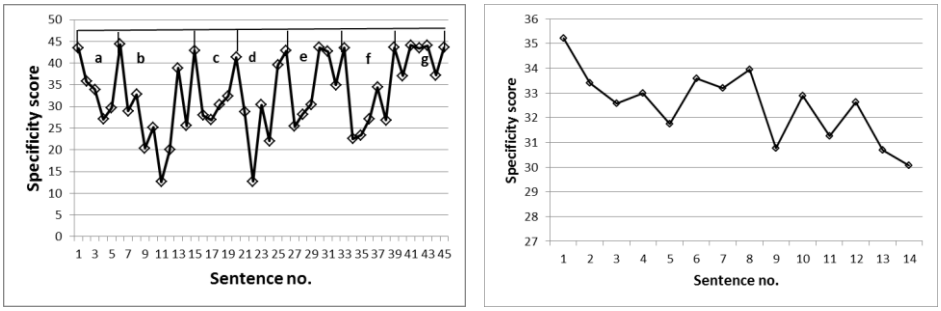
1. Specify the expectation from a summary using specificity score, that is, if user expects more generic information or specific information. For technical articles if abstract like summary is required extract generic sentences. If the technical summary is required, extract more specific sentences.
2. Sort the sentences according to the expectation set by user: For technical summary sort sentences by descending order of the specificity score (larger values first).
3. Choose top *k* sentences as per the length or set the threshold for specificity as per the requirement. Absolute maximum score is 10 times number of features used in calculating specificity score. For 5 features, maximum score is 50. Hence threshold can be set a front. Our experience shows that threshold of 40 performs well for creating shorter summaries.
4. Further, the sentences can be reordered to keep the original sequence in the document. Reordering would improve readability of the summary if desired.

Table 3. Specificity score calculation

```
for each record ri=1…n(r) do
       form sentences
        for each sentence si=1…n(s) do
       POS tagg the sentence si
       Tokenize the sentence si
              for each token tk=1…n(t) do
                     If (tk is not a stopword)
                            then added as a valid token
                            Identify semantic depth of tk
                            Identify semantic height of tk
                            Identify whether it is a NE tk
                            Identify total occurrence count of tk
                            Identify whether tk is a Proper noun
                      end if
              end for //for tokens
              aggregate the average semantic depth, average se-
mantic height, NE count, average total occurrence count, Sen-
tence length, Number of proper nouns of the sentence si
              Based upon the above aggregated values identified,
calculate the specific score of the sentence
       end for //for sentences
end for//for records
Store sentences according to the specificity in descending or-
der.
Extract the top % of sentences to represent as summary
```

4 Experiment Results and Discussion

4.1 News Articles



a b

Fig. 1. Sentence number vs specificity score a) for news article d01a/SJMN91-06290185, b) average specificity score from 50 news articles (DUC 2001)

The Fig. 1 shows how specificity score vary with sentence order (number). The chart appropriately models intentional structure of news article: The example news article can be divided into ~8 blocks (Fig. 1a, a, b, c, d *etc.*) – each block covering sentences from high specificity score to the next sentence with high specificity score (excluding the next peak). The lead sentences are marked by the sentences with high specificity score and subsequent sentences providing further background in the context of leads are indicated by gradually decreasing specificity score. Interestingly, all the lead sentences are having very high specificity score. Close inspection of these sentences (Table 4) reveals that these sentences introduce sub stories around the main story which is introduced by first sentence.

Table 4. Lead sentences and their specificity score (sco.) for SJMN91-06290185

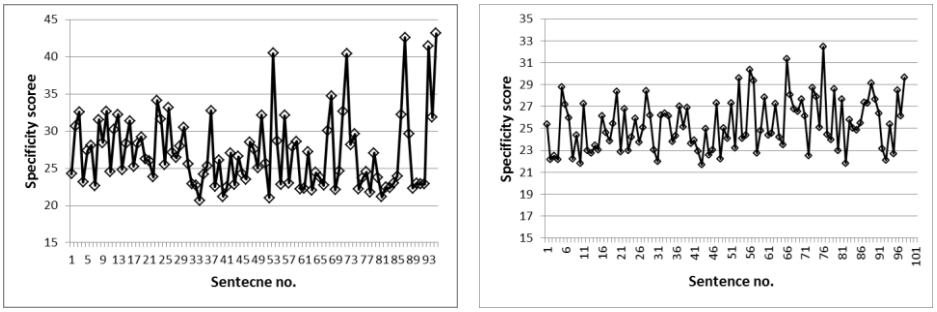
Sentence	sco.
Clarence Thomas, triumphing over eleventh-hour charges of sexual harassment, won Senate confirmation by only four votes Tuesday night to become the youngest member of the Supreme Court and its first black conservative.	43.46
It was the closest Senate confirmation of a Supreme Court nominee since Lucius Q. C. Lamar, an appointee of President Grover Cleveland, also squeezed through by four votes in 1888.	44.36
“Today the Senate sacrificed the integrity of the Supreme Court, its own reputation and the rights of American women to the Bush-Reagan agenda," the Women's Legal Defense Fund said in a statement released after the vote.	42.95
Law Professor Anita Hill, once an aide to Thomas, declined to comment specifically about the Senate vote	41.32

Model summary of this article contains only one sentence from the above list that is 1st sentence and other supporting sentences for this lead sentence. The specificity score of this sentence is second highest - second to the sentence number 6 ("*It was the closest Senate confirmation of a Supreme Court nominee since Lucius Q.C. Lamar, an appointee of President Grover Cleveland, also squeezed through by four votes in 1888*"). One can argue in this case – the very reason the Clarence Thomas wining senate confirmation is a news (apart from its own merit) because it was similar to earlier event (someone winning by 4 votes).

In one embodiment summary can have all the lead sentences without any supporting sentences (as reflected in news story structure). The example (Table 4) can be one of such summary that picks up first *k=4* sentences above some threshold specificity score *S=40*. Variety of such algorithms can be devised very easily with specificity score as the parameter. Current analysis of specificity score for summarization is performed using 5 features (excludes length) and the absolute maximum score a sentence can have is 50. Hence setting up threshold a front or fine turning it for the given set of news stories won't be very difficult. News articles (DUC 2001) about other types of events show similar structure (Fig. 1b). Further comprehensive exploration is required to cover all other types of news.

4.2 Technical Articles

Summary of technical articles also can be seen from specificity perspective. Whereas news article exhibit multiple alternatives of choosing sentences using specificity score (4.1), technical article summary might have limited options. Most of the time summary of technical article is expected to have only specific or only generic sentences. Generally, reader of the technical articles is interested in details such as results, conclusion of the work *etc.* For example, we want to create a bulleted list of research findings. Such expectation from the sentences of a summary can be expressed using high specificity scores directly and hence can be used for generating summary of a technical document. Same is true for technical reports. We begin our investigation with structure of the document as revealed by specificity score.



a b

Fig. 2. Specificity score vs sentence number for a) [2] b) average from [**2**], [**18**], [**19**], [**20**], [**21**]

Structure of the document – in terms of information imparted to the reader – is nicely revealed by the specificity scores. One can easily identify smaller blocks of sentences beginning with low specificity scores and then ending with high specificity score sentences. This is natural given that the technical article would begin with introducing topics to the reader with increasing details as we go on. In this case structure indicated by specificity score overlaps with paragraph structure and can be expected to be a trend for the other technical articles too (Fig. 2b). The difference in technical article and news article is: each block in technical article begin with more generic information and ends with sentences with more and more specific information (further examination with larger dataset is required for higher confidence on this observation, especially for news articles if they begin with more specific sentences). With increase in sentence number specificity score also increases – that is subsequent paragraphs are providing more specific information than the beginning paragraphs. Considering the difference between studies of two types of documents (that is news, and scientific publications), we are tempted to suggest that specificity score can be effectively utilized for identifying genre (technical/news *etc.*) of the document too.

Table 5 shows summary of Relative Utility paper by Radev et. al. [**2**] generated by selecting 10% of sentences. Table 6 shows first 5 sentences of summary of same article generated by lexical chaining algorithm [**20**]. Note that sentences are not reordered according to the original sequence.

Table 5. Summary of paper by Radev *et. al.* [**2**] using specificity score (sco.)

Sentences	Sco.
Second, MEAD and WEBSUM score approximately the same on all metrics with MEAD doing slightly better on the Word overlap , Bigram overlap , and Longest common subsequence measures and WEBSUM on the cosine metric.	42.60
We used the Hong Kong News summary corpus created at Johns Hopkins University in 2001.	40.45
Third , even though the performances of MEAD and WEBSUM S also increase with summary length , MEAD normalized version D decreases slowly with summary length until the two summarizers score about the same on both S and D for longer summaries.	40.36
The single document results tables compare MEAD with WEBSUM and the two baselines RAND and LEAD.	34.72
In fact , the interjudge agreement as measured by RU for this example is0. 76. RU agreement see next section is defined as the relative score that one judge would get given his own extract and the other judge sentence judgements.	34.12
A summarizer could have an RU agreement with judge J1 as low as 0.14 and an agreement with judge J2 as low as0. 38. In other words , given that interjudge agreement is significantly less than 1.0 but significantly more than the worst score possible , an automatic summarizer might score as low as .70 and still be almost as good as the judges themselves.	33.13
Using metrics such as P&R or PA [1 , 2] to evaluate summaries creates the possibility that two equally good extracts are judged very differently.	32.67
The average value of R across all documents at the 5% target length is 0.598 while the average value of J is0. 799. The corresponding values for the 20% target length are R = 0 635 and J = 0 835.	32.66

We will address some advantages of RU over existing co selection metrics such as Precision , Recall , percent agreement , and Kappa.	32.57
Using P&R or PA , system A will be ranked much higher than systemB. It is quite possible however , that for the purpose of summarization , sentences 2 and 3 are equally important , in which case the two systems should get the same score.	32.24

Table 6. First 5 sentences of lexical chaining summary of the RU paper

The main problem with traditional co-selection metrics (thus named because they measure the degree of overlap between the list of sentences selected by a judge and an automatically produced extract) such as Precision, Recall, and Percent Agreement for evaluating extractive summarizers is that human judges often disagree about which the top n% most important sentences in a document or cluster are and yet, there appears to be an implicit importance value for all sentences which is judge-independent.

These include word based cosine between two summaries, word overlap, bigram overlap, and LCS (longest common subsequence). These metrics are all based on the actual text of the extracts (unlike P/R/Kappa/RU, which are all computed on the sentence co-selection vectors).

In the formula for U0, "j (multi-judge summary characteristic function) is 1 for the top e sentences according to the sum of utility scores from all judges.

Relative Utility provides an intuitive mechanism which takes into account the fact that even though human judges may disagree on exactly which sentences belong in a summary, they tend to agree on the overall salience of each sentence.

The Relative Utility (RU) method [3] allows ideal summaries to consist of sentence sets with variable membership.

Table 7. Summary of paper by *Helteran et. al.* [**1**] and specificty score (sco.)

Sentences	Sco.
Some of the generalisation links are part of 3- or 4-link hierarchies, e.g. "FV40 Victim outspoken about/campaigning on immigration issues" (26 mentions) to "FV41 Victim was anti immigration" (23) to "FV42 Victim wanted to close borders to immigration" (9), or "FV50 Victim outspoken about race/religion issues" (17 mentions) to "FV51 Victim outspoken about Islam/Muslims" (16) to "FV52 Victim made negative remarks about Islam" (14) to "FV53 Victim called Islam a backward religion" (9).	43.18
In principle , the comparison can be done via coselection of extracted sentences Rath et al. , 1961; Jing et al. , 1998; Zechner , 1996 , by string based surface measures Lin and Hovy , 2002; Saggion et al. , 2002 , or by subjective judgements of the amount of information overlap DUC , 2002 .	40.00
In the past years, there has been quite a lot of summarisation work that has effectively aimed at finding viable evaluation strategies Sparck Jones , 1999; Jing et al. , 1998; Donaway et al. , 2000 .	36.48
The factoid approach can capture much finer shades of meaning differentiations than DUC style information overlap does - in an example from Lin and Hovy (2002), an assessor judged some content overlap between "Thousands of people are feared dead and "3, 000 and perhaps . 5, 000 people have been killed."	36.21
Pim Fortuyn , a Dutch right wing politician , was shot dead at a radio station in Hilversum.	35.67

The text used for the experiment is a BBC report on the killing of the Dutch politician Pim Fortuyn.	35.10
However, Lin and Hovy 2002 report low agreement for two tasks producing the human summaries around 40% , and assigning information overlap between them.	35.02
Largescale conferences like SUMMAC Mani et al., 1999 and DUC 2002 have unfortunately shown weak results in that current evaluation measures could not distinguish between automatic summaries - though they are effective enough to distinguish them from human written summaries.	35.01
In summarisation there appears to be no “one truth”, as is evidenced by a low agreement between humans in producing gold standard summaries by sentence selection Rath et al. , 1961; Jing et al. , 1998; Zechner , 1996 , and low overlap measures between humans when gold standards summaries are created by reformulation in the summarisers' own words e.g. the average overlap for the 542 single document summary pairs in DUC-02 was only about 47% .	34.41
Lin and Hovy 2002 examine the use of a multiple gold standard for summarisation evaluation, and conclude \we need more than one model summary although we cannot estimate how many model summaries are required to achieve reliable automated summary evaluation .	33.50

Table 8. First 5 sentences of lexical chaining summary of *Helteran et. al.* [**1**]

We present a new approach to summary evaluation which combines two novel aspects, namely (a) content comparison between gold standard summary and system summary via factoids, a pseudo-semantic representation based on atomic information units which can be robustly marked in text, and (b) use of a gold standard consensus summary, in our case based on 50 individual summaries of one text.

If we decide to use a single human summary as a gold standard, we in fact assume that this human's choice of important material is acceptable for all other summary users, which it the wrong assumption, as the lack of consensus between the various human summaries shows.

All in all, the use of consensus summaries and factoid analysis, even though expensive to set up for the moment, provides a promising alternative which could well bring us closer to a solution to several problems in summarisation evaluation.

In summarisation there appears to be no \one truth", as is evidenced by a low agreement between humans in producing gold standard summaries by sentence selection (Rath et al, 1961; Jing et al, 1998; Zechner, 1996), and low overlap measures between humans when gold standards summaries are created by reformulation in the summarisers' own words (eg the average overlap for the 542 single document summary pairs in DUC-02 was only about 47%).

3, There is no such thing as overall consensus, but there is a difference in perceived importance between the various factoids, We can determine whether this is the case by examining how often each factoid is used in the summaries, Factoids that are more important ought to be included more often, In that case, it is still possible to create a consensus-like reference summary for any desired summary size.

5 Conclusion

We demonstrated how summarization can be driven by a parameter other than importance. Summary produced by such a method provide mechanism for choosing

right sentences as per the users expectation from the summary. Summary using specificity score outperforms (assessed using sample cases) summaries by some popular summarization techniques in case detailed information from summary is expected by more informed reader. Evaluation of such a summary is not possible by existing summary evaluation methods that use model summaries by experts as such summaries tend to provide introductory information. Some of the specific observations are:

- Specificity score per sentence provides easy way to assess the structure of the document from information perspective and could be used further for identifying type of the document.
- Specificity score based approach can create a summary to have detailed or introductory information in the given document by setting a threshold for the score, or using sorted list and then by selecting top k sentences as required. Further, complex strategies for choosing right mix of specific and generic sentences can be devised for appropriate summary: Let's say we have a budget of score S (say 150) and task is to choose right sentences within the limits – one extreme case would be choosing many low scoring sentences and another a few high scoring ones.

Design of summarization system need to consider three context factors namely INPUT, PURPOSE, and OUTPUT[1]. INPUT factors and OUTPUT factors (Material, style, expression *etc.*) characterize input material and output material respectively, and PURPOSE factors (audience, use *etc.*) are related to the usage of summary [**16**], [**22**]. It is natural to think that robust summarization system then needs some parameters characterizing each of the above mentioned factors - mainly purpose and output. Many of these factors influence each other in complex way with varying degree (for example, style and expression, brevity and use). The influence on each other can be leveraged to create system specifications (for summary) with only limited number of factors. Especially in some cases PURPOSE fully determines the OUTPUT [**22**], for example, if the reader is reviewing papers for literature survey then abstract like summary might be fine but if he is more informed reader then he would be more interested in scientific claims, results and so on. Such couplings between context factors are usually reflected in more observable parameters such as discourse relations. For the technical paper example in this paper, specificity controlled the summarization process as one would be looking for more specific information. Thus, specificity shows potential to characterize some of the PURPOSE and/or OUTPUT factors of summary. Consensus on summary factors (and their definitions) is required for further elaborate investigations on how summarization can be driven by factors like specificity.

References

1. Halteren, V., Teufel, S.: Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis. HLT-NAACL-DUC '03 Proceedings of the HLT-NAACL 03 on Text summarization workshop (2003)

[1] See DUC roadmap 2005-2007 http://duc.nist.gov/RM0507/rm.html for details

2. Radev, D., Tam, D.: Summarization Evaluation Using Relative Utility. In: Proceeding CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management, p.508/511 (Nov 2003)
3. Lin, C.-Y.: Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples are Enough? In: NTCIR Workshop 4, Tokyo, Japan (June 2-4, 2004)
4. Nenkova, A.: Summarization Evaluation for Text and Speech: Issues and Approaches. (Accessed on Jan 2015) Available at: http://www.cis.upenn.edu/~nenkova/papers/sumEval.pdf.
5. Nenkova, A., Passonneau, R.: Evaluating Content Selection in Summarization: The Pyramid Method. North American Chapter of the Association for Computational Linguistics - NAACL, 145/152 (2004)
6. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain (2004)
7. Hovy, E., Lin, C.-Y., Zhou, L., Fukumoto, J.: Automated Summarization Evaluation with Basic Elements. In: The Fifth Conference on Language Resources and Evaluation (LREC) (2006)
8. Louise, A., Nenkova, A.: Text Specificity and Impact on Quality of News Summaries. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp.34-42 (2011)
9. Erkan, G., Radev, D.: Lexrank: Graphbased Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479 (2004)
10. Palshikar, G., Deshpande, S., Bhat, S.: Quest: Discovering Insights from Survey Responses. In: Proceedings of 8th Australasian Data Mining Conf. (AusDM09), pp.83-92 (2009)
11. Deshpande, S., Palshikar, G., G, Athiappan.: An Unsupervised Approach to Sentence Classification. In: Proceedings of International Conference on Management of Data", COMAD 2010, Nagpur, India (2010)
12. Louis, A., Nenkova, A.: General versus Specific Sentences: Automatic Identification and Application to Analysis of News Summaries.
13. Jones, K.: A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation 28(1), 11-21 (1972)
14. Hassel, M.: Evaluation of Automatic Text Summarization, Licentiate Thesis., Stockholm, Sweden (2004)
15. Jing, H., McKeown, K.: Cut and Paste Based Text Summarization. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, Seattle, Washington (2000)
16. Sparck Jones, K.: Automatic Summarizing: Factors and Directions., Cambridge, MA: MIT Press (1999)
17. Fellbaum, C.: WordNet: an On-line Lexical Database and its Applications. MIT Press (1998)
18. Doran, W., Stokes, N., Carthy, J., Dunnion, J.: Comparing Lexical Chain-based Summarisation Approaches Using an Extrinsic Evaluation. (2004)
19. Apte, H.: Building a Trainable Multi-document Summarizer. (Accessed Jan 2015) Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.4063&rep=rep1&type=pdf
20. Palshikar, G., Deshpande, S., Athiappan, G.: Combining Summaries Using Unsupervised Rank Aggregation. Computational Linguistics and Intelligent Text Processing 7182, 378-389 (2012)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp.311-318 (2002)
22. Sparck Jones, K.: Automatic Summarizing: The State of the Art. (2007)