Text specificity and impact on quality of news summaries

Annie Louis
University of Pennsylvania
Philadelphia, PA 19104
lannie@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104
nenkova@seas.upenn.edu

Abstract

In our work we use an existing classifier to quantify and analyze the level of specific and general content in news documents and their human and automatic summaries. We discover that while human abstracts contain a more balanced mix of general and specific content, automatic summaries are overwhelmingly specific. We also provide an analysis of summary specificity and the summary quality scores assigned by people. We find that too much specificity could adversely affect the quality of content in the summary. Our findings give strong evidence for the need for a new task in abstractive summarization: identification and generation of general sentences.

1 Introduction

Traditional summarization systems are primarily concerned with the identification of important and unimportant content in the text to be summarized. Placing the focus on this distinction naturally leads the summarizers to completely avoid the task of text-to-text generation and instead just select sentences for inclusion in the summary. In this work, we argue that the general and specific nature of the content is also taken into account by human summarizers; we show that this distinction is directly related to the quality of the summary and it also calls for the use and refinement of text-to-text generation techniques.

General sentences are overview statements. Specific sentences supply details. An example general and specific sentence from different parts of a news article are shown in Table 1.

[1] *The first shock let up as the eye of the storm moved across the city.*
[2] The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

Table 1: General (in italics) and specific sentences

Prior studies have advocated that the distinction between general and specific content is relevant for text summarization. Jing and McKeown (2000) studied what edits people use to create summaries from sentences in the source text. Two of the operations they identify are *generalization* and *specification* where the source content gets changed in the summary with respect to specificity. In more recent work, Haghighi and Vanderwende (2009) built a summarization system based on topic models, where both topics at general document level as well as those at specific subtopic levels were learnt. The underlying idea here is that summaries are generated by a combination of content from both these levels. But since the preference for these two types of content is not known, Haghighi and Vanderwende (2009) use some heuristic proportions.

Many systems that deal with sentence compression (Knight and Marcu, 2002; McDonald, 2006; Galley and McKeown, 2007; Clarke and Lapata, 2008) and fusion (Barzilay and McKeown, 2005; Filippova and Strube, 2008), do not take into account the specificity of the original or desired sentence. However, Wan et al. (2008) introduce a generation task where a summary sentence is created by combining content from a key (general) sentence and its supporting sentences in the source. More

Workshop on Monolingual Text-To-Text Generation, pages 34–42,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 34–42,
Portland, Oregon, 24 June 2011.

recently, Marsi et al. (2010) manually annotated the transformations between source and compressed phrases and observe that *generalization* is a frequent transformation.

But it is not known what distribution of general and specific content is natural for summaries. In addition, an analysis of whether this aspect is related to quality of the summary has also not been done so far. We address this issue in our work, making use of an accurate classifier to identify general and specific sentences that we have developed (Louis and Nenkova, 2011).

We present the first quantitative analysis of general and specific content in a large corpus of news documents and human and automatic summaries produced for them. Our findings reveal that human-written abstracts have much more general content compared to human and system produced extractive summaries. We also provide an analysis of how this difference in specificity is related to aspects of summary quality. We show that too much specificity could adversely affect the quality of summary content. So we propose the task of creating general sentences for use in summaries. As a starting point in this direction, we discuss some insights into the identification and generation of general sentences.

2 Data

We obtained news documents and their summaries from the Document Understanding Conference (DUC) evaluations. We use the data from 2002 because they contain the three different types of summaries we wish to analyze—abstracts and extracts produced by people, and automatic summaries. For extracts, the person could only select complete sentences, without any modification, from the input articles. When writing abstracts people were free to write the summary in their own words.

We use data from the generic multi-document summarization task. There were 59 input sets, each containing 5 to 15 news documents on a topic. The task is to provide a 200 word summary. Two human-written abstracts and two extracts were produced for each input by trained assessors at NIST. Nine automatic systems participated in the conference that year and we have 524 automatic summaries overall.

3 General and specific sentences in news

Before we present our analysis of general and specific content in news summaries, we provide a brief description of our classifier and some example predictions. Our classifier is designed to predict for a given *sentence*, its class as general or specific.

As in our example in Table 1, a general sentence hints at a topic the writer wishes to convey but does not provide details. So a reader expects to see more explanation and specific sentences satisfy this role. We observed that certain properties are prominent in general sentences. They either express a strong sentiment, are vague or contain surprising content. Accordingly our features were based on word specificity, language models, length of syntactic phrases and the presence of polarity words. Just the words in the sentences were also a strong indicator of general or specific nature. But we found the combination of all non-lexical features to provide the best accuracy and is the setup we use in this work.

We trained our classifier on general and specific sentences from news texts. Initially, we utilized existing annotations of discourse relations as training data. This choice was based on our hypotheses that discourse relations such as exemplification relate a general with a specific sentence. Later, we verified the performance of the classifier on human annotated general and specific sentences, also from two genre of news articles, and obtained similar and accurate predictions. Detailed description of the features and training data can be found in Louis and Nenkova (2011).

Our classifier uses logistic regression and so apart from hard prediction into general/specific classes, we can also obtain a confidence (probability) measure for membership in a particular class. In our tests, we found that for sentences where there is high annotator agreement for placing in a particular class, the classifier also produces a high confidence prediction on the correct class. When the agreement was not high, the classifier confidence was lower. In this way, the confidence score indicates the level of general or specific content. So for our experiments in this paper, we choose to use the confidence score for a sentence belonging to a class rather than the classification decision.

The overall accuracy of the classifier in binary

[G1] "The crisis is not over".
[G2] No casualties have been reported, but experts are concerned that a major eruption could occur soon.
[G3] Seismologists said the volcano had plenty of built-up magma and even more severe eruptions could come later.
[G4] Their predictions might be a false alarm – the volcano may have done its worst already.

[S1] (These volcanoes – including Mount Lassen in Shasta County, and Mount Rainier and Mount St. Helens in Washington, all in the Cascade Range – arise where one of the earth's immense crust plates is slowly diving beneath another.); Pinatubo's last eruption, 600 years ago, is thought to have yielded at least as much molten rock – half a cubic kilometer – as Mount St. Helens did when it erupted in 1980.
[S2] The initial explosions on Mount Pinatubo at 8:51 a.m. Wednesday sent a 10-mile-high mushroom cloud of swirling ash and rock fragments into the skies over Clark Air Base, forcing the Air Force to evacuate hundreds of American volunteers who had stayed behind to guard it and to tend sensitive communications equipment.
[S3] Raymundo Punongbayan, director of the Philippine Institute of Vulcanology and Seismology, said Friday's blasts were part of a single eruption, the largest since Mount Pinatubo awoke Sunday from its 600-year slumber.

Table 2: General (G) and specific (S) sentences from input d073b

classification is 75%. More accurate predictions are made on the examples with high annotator agreement reaching over 90% accuracy on sentences where there was complete agreement between five annotators. So we expect the predictions from the classifier to be reliable for analysis in a task setting.

In Table 2, we show the top general and specific sentences (ranked by the classifier confidence) for one of the inputs, d073b, from DUC 2002. This input contains articles about the volcanic eruption at Mount Pinatubo. Here, the specific sentences provide a lot of details such as the time and impact of the eruption, information about previous volcanoes and about the people and organizations involved.

In the next section, we analyze the actual distribution of specific and general content in articles and their summaries for the entire DUC 2002 dataset.

4 Specificity analysis

For each text—input, human abstract, human extract and automatic summary—we compute a measure of specificity as follows. We use the classifier to mark for each sentence the confidence for belonging to the *specific* class. Each token in the text is assigned the confidence level of the sentence it belongs to. The *average specificity of words* is computed as the mean value of the confidence score over all the tokens.

The histogram of this measure for each type of text is shown in Figure 1.

For inputs, the average specificity of words ranges between 50 to 80% with a mean value of 65%. So, news articles tend to have more specific content than generic but the distribution is not highly skewed towards either of the extreme ends.

The remaining three graphs in Figure 1 represent the amount of specific content in summaries for the same inputs. Human abstracts, in contrast to the inputs, are spread over a wider range of specificity levels. Some abstracts have as low as 40% specificity and a few actually score over 80%. However, the sharper contrast with inputs comes from the large number of abstracts that have 40 to 60% specificity. This trend indicates that abstracts contain more general content compared to inputs. An unpaired two-sided t-test between the specificity values of inputs and abstracts confirmed that abstracts have significantly lower specificity. The mean value for abstracts is 62% while for inputs it is 65%.

The results of the analysis are opposite for human extracts and system summaries. The mean specificity value for human extracts is 72%, 10% higher compared to abstractive summaries for the same inputs. This difference is also statistically significant. System-produced summaries also show a similar trend as extracts but are even more heavily biased towards specific content. There are even examples of automatic summaries where the average specificity level reaches 100%. The mean specificity value is 74% which turned out significantly higher than all other types of texts, inputs and both types of human summaries. So system summaries appear to be overwhelmingly specific.

The first surprising result is the opposite characteristics of human abstracts and extracts. While abstracts tend to be more general compared to the input texts, extracts are more specific. Even though

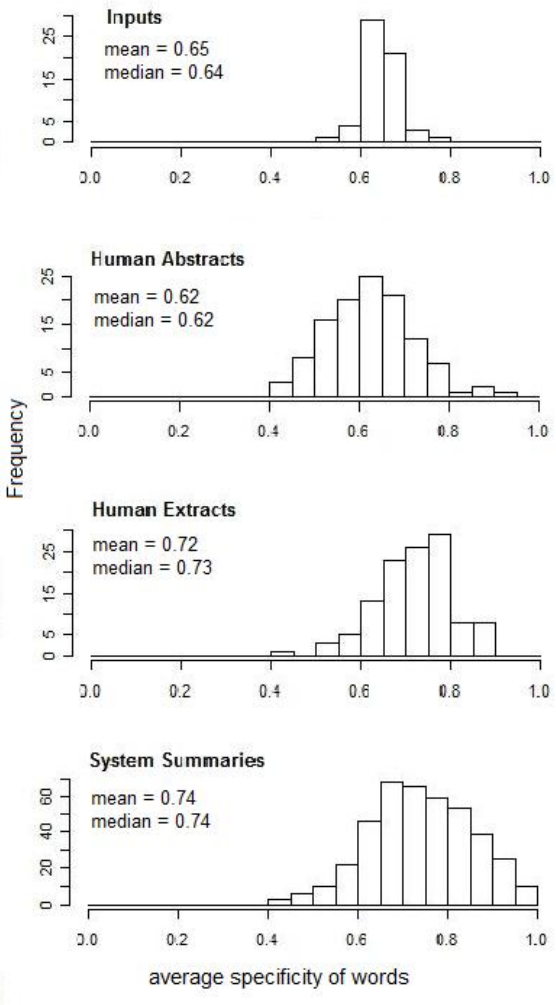

Figure 1: Specific content in inputs and summaries

both types of summaries were produced by people, we see that the summarization method deeply influences the nature of the summary content. The task of creating extractive summaries biases towards more specific content. So it is obvious that systems which mainly use extractive techniques would also create very specific summaries. Further, since high specificity arises as a result of the limitations associated with extractive techniques, perhaps, overly specific content would be detrimental to summary quality. We investigate this aspect in the next section.

5 Specificity and summary quality

In this section, we examine if the difference in specificity that we have observed is related to the perceived quality of the summary. Haghighi and Vanderwende (2009) report that their topic model based system was designed to use both a general content distribution and distributions of content for specific subtopics. However, using the general distribution yielded summaries with better content than using the specific topics. Here we directly study the relationship between specificity of system summaries and their content and linguistic quality scores. We also examine how the specificity measure is related to the quality of specialized summaries where people were explicitly told to include only general content or only specific details in their summaries. For this analysis, we focus on *system produced summaries*.

5.1 Content quality

At DUC, each summary is evaluated by human judges for content and linguistic quality. The quality of content was assessed in 2002 by means of a coverage score. The coverage score reflects the similarity between content chosen in a system summary and that which is present in a human-written summary for the same input. A human abstract is chosen as the reference. It is divided into clauses and for each of these clauses, judges decide how well it is expressed by the system produced summary (as a percentage value). The average extent to which the system summary expresses the clauses of the human summary is considered as the coverage score. So these scores range between 0 and 1.

We computed the Pearson correlation between the specificity of a summary and its coverage score, and obtained a value of -0.16. The correlation is not very high but it is significant (pvalue 0.0006). So specificity does impact content quality and more specific content indicates decreased quality.

We have seen from our analysis in the previous section that when people produce abstracts, they keep a mix of general and specific content but the abstracts are neither too general nor too specific. So it is not surprising that the correlation value is not very high. Further, it should be remembered that the notion of general and specific is more or less independent of the importance of the content itself. Two summaries can have the same level of generality but vary greatly in terms of the importance of the content present. So we performed an analysis to check the contribution of generality to the content scores in addition to the importance factor.

We combine a measure of content importance

Predictor	Mean β	Stdev. β	t value	p-value
(Intercept)	0.212	0.03	6.87	2.3e-11 *
rouge2	1.299	0.11	11.74	$<$ 2e-16 *
avgspec	-0.166	0.04	-4.21	3.1e-05 *

Table 3: Results from regression test

ling score	sums.	avg specificity
1, 2	202	0.71
5	400	0.72
9, 10	79	0.77

Table 4: Number of summaries at extreme levels of linguistic quality scores and their average specificity values

from the ROUGE automatic evaluation (Lin and Hovy, 2003; Lin, 2004) with generality to predict the coverage scores. We use the same reference as used for the official coverage score evaluation and compute ROUGE-2 which is the recall of bigrams of the human summary by the system summary. Next we train a regression model on our data using the ROUGE-2 score and specificity as predictors of the coverage score. We then inspected the weights learnt in the regression model to identify the influence of the predictors. Table 3 shows the mean values and standard deviation of the beta coefficients. We also report the results from a test to determine if the beta coefficient for a particular predictor could be set to zero. The p-value for rejection of this hypothesis is shown in the last column and the test statistic is shown as the 't value'. We used the *lm* function in the R toolkit[1] to perform the regression.

From the table, we see that both ROUGE-2 and average specificity of words (avgspec) turn out as significant predictors of summary quality. Relevant content is highly important as shown by the positive beta coefficient for ROUGE-2. At the same time, it is preferable to maintain low specificity, a negative value is assigned to the coefficient for this predictor.

So too much specificity should be avoided by systems and we must find ways to increase the generality of summaries. We discuss this aspect in Sections 6 and 7.

5.2 Linguistic quality

We have seen from the above results that maintaining a good level of generality improves content quality. A related question is the influence of specificity on the linguistic quality of a summary. Does the amount of general and specific content have any relationship with how clear a summary is to read? We briefly examine this aspect here.

In DUC 2002 linguistic quality scores were only mentioned as the number of errors in a summary, not a holistic score. Moreover, it was specified as a range–errors between 1 and 5 receive the same score. So we use another dataset for this analysis only. We use the system summaries and their linguistic quality scores from the TAC '09 query focused summarization task[2]. Each summary was manually judged by NIST assessors and assigned a score between 1 to 10 to reflect how clear it is to read. The score combines multiple aspects of linguistic quality such as clarity of references, amount of redundancy, grammaticality and coherence.

Since these scores are on an integer scale, we do not compute correlations. Rather we study the specificity, computed in the same manner as described previously, of summaries at different score levels. Here there were 44 inputs and 55 systems. In Table 4, we show the number of summaries and their average specificity for 3 representative score levels—best quality (9 or 10), worst (1 or 2) and mediocre (5). We only used summaries with more than 2 sentences as it may not be reasonable to compare the linguistic quality of summaries of very short lengths.

From this table, we see that the summaries with greater score have a higher level of specificity. The specificity of the best summaries (9, 10) are significantly higher than that with medium and low scores (two-sided t-test). This result is opposite to our finding with content quality and calls attention to an important point. General sentences cannot stand alone and need adequate support and details. But currently, very few systems even make an attempt to organize their summaries. So overly general content and general content without proper context can be detrimental to the linguistic quality. Such summaries can appear uncontentful and difficult to read as the example in Table 5 demonstrates. This summary has an average specificity of 0.45 and its linguistic quality score is 1.

So we see an effect of specificity on both content

[1] http://www.r-project.org/

[2] http://www.nist.gov/tac/2009/Summarization/update.summ.09.guidelines.html

"We are quite a ways from that, actually."
As ice and snow at the poles melt, the loss of their reflective surfaces leads to exposed land and water absorbing more heat.
It is in the middle of an area whose population–and electricity demands–are growing.
It was from that municipal utility framework, city and school officials say, that the dormitory project took root.
"We could offer such a plan in Houston next year if we find customer demand, but we have n't gone to the expense of marketing the plan."
"We get no answers."

Table 5: Example general summary with poor linguistic quality

and linguistic quality though in opposite directions.

5.3 Quality of general and specific summaries

So far, we examined the effect of specificity on the quality of generic summaries. Now, we examine whether this aspect is related to the quality of summaries when they are optimized to be either general or specific content. We perform this analysis on DUC 2005[3] data where the task was to create a general summary for certain inputs. For others, a specific summary giving details should be produced. The definitions of a general and specific summary are given in the task guidelines.[4]

We tested whether the degree of specificity is related to the content scores[5] of system summaries of these two types—general and specific. The Pearson correlation values are shown in Table 6. Here we find that for specific summaries, the level of specificity is significantly positively correlated with content scores. For the general summaries there is no relationship between specificity and content quality.

These results show that specificity scores are not consistently predictive of distinctions within the *same* class of summaries. Within general summaries, the level of generality does not influence the scores obtained by them. This finding again highlights the disparity between content relevance and specific nature. When all summaries are specific or general, their levels of specificity are no longer indicative of quality. We also computed the regression models for these two sets of summaries with ROUGE scores and specificity, and specificity level was not a significant predictor of content scores.

Our findings in this section confirm that general sentences are useful content for summaries. So we face the issue of creating general sentences which are summary-worthy. We concentrate on this aspect for the rest of this paper. In Section 6, we provide an analysis of the types of general sentences extracted from the source text and used in human extracts. We move from this limited view and examine in Section 7, the possibility of generating general sentences from specific sentences in the source text. Our analysis is preliminary but we hope that it will initiate this new task of using general sentences for summary creation.

Summaries	**correlation**	**p-value**
DUC 2005 general	-0.03	0.53
DUC 2005 specific	0.18*	0.004

Table 6: Correlations between content scores and specificity for general and specific summaries in DUC 2005

6 Extraction of general sentences

We examine general sentences that were chosen in human extracts to understand what properties systems could use to identify such sentences from the source text. We show in Table 7, the ten extract sentences that were predicted to be general with highest confidence. The first sentence has a 0.96 confidence level, the last sentence has 0.81.

These statements definitely create expectation and need further details to be included. Taken out of context, these sentences do not appear very contentful. However despite the length restriction while creating summaries, humans tend to include these general sentences. Table 8 shows the full extract which contains one of the general sentences ([9] "Instead it sank like the Bismarck.").

When considered in the context of the extract, we see clearly the role of this general sentence. It introduces the topic of opposition to Bush's nomination for a defense secretary. Moreover, it provides a comparison between the ease with which such a proposition could have been accepted and the strikingly

[3] http://duc.nist.gov/duc2005/

[4] http://duc.nist.gov/duc2005/assessor.summarization.instructions.pdf

[5] We use the official scores computed using the Pyramid evaluation method (Nenkova et al., 2007)

opposite situation that arose—the overwhelming rejection of the candidate by the senate. So sentence [9] plays the role of a topic sentence. It conveys the main point the author wishes to make in the summary and further details follow this sentence.

But given current content selection methods, such sentences would rank very low for inclusion into summaries. So the prediction of general sentences could prove a valuable task enabling systems to select good topic sentences for their summaries. However, proper ordering of sentences will be necessary to convey the right impact but this approach could be a first step towards creating summaries that have an overall theme rather than just the selection of sentences with important content.

We also noticed some other patterns in the general sentences chosen for extracts. A crude categorization was performed on the 75 sentences predicted with confidence above 0.65 and are shown below:

first sentence : 6 (0.08)
last sentence : 13 (0.17)
comparisons : 4 (0.05)
attributions : 14 (0.18)

A significant fraction of these general sentences (25%) were used in the extracts to start and end the summary, likely positions for topic sentences. Some of these (5%) involve comparisons. We detected these sentences by looking for the presence of connectives such as "but", "however" and "although". The most overwhelming pattern is presence of quotations, covering 18% of the sentences we examined. These quotations were identified using the words "say", "says", "said" and the presence of quotes. We can also see that three of the top 10 general sentences in Table 7 are quotes.

So far we have analyzed sentences chosen by summary authors directly from the input articles. In the next section, we analyze the edit operations made by people while creating abstractive summaries. Our focus is on the generalization operation where specific sentences are made general. Such a transformation would be the generation-based approach to obtain general sentences.

7 Generation of general sentences

We perform our analysis on data created for sentence compression. In this line of work (Knight and Marcu, 2002; McDonald, 2006; Galley and McKeown, 2007), compressions are learnt by analyzing pairs of sentences, one from the source text, the other from human-written abstracts such that they both have the same content. We use the sentence pairs available in the Ziff-Davis Tree Alignment corpus (Galley and McKeown, 2007). These sentences come from the Ziff-Davis Corpus (Harman and Liberman, 1993) which contains articles about technology products. Each article is also associated with an abstract. The alignment pairs are produced by allowing a limited number of edit operations to match a source sentence to one in the abstract. In this corpus, alignments are kept between pairs that have *any* number of deletions and upto 7 substitutions. There are 15964 such pairs in this data. It is worth noting that these limited alignments only map 25% of the abstract sentences, so they do not cover all the cases. Still, an analysis on this data could be beneficial to observe the trends.

[1] Folksy was an understatement.
[2] "Long live democracy"!
[3] The dogs are frequent winners in best of breed and best of show categories.
[4] Go to court.
[5] Tajikistan was hit most hard.
[6] Some critics have said the 16-inch guns are outmoded and dangerous.
[7] Details of Maxwell's death are sketchy.
[8] "Several thousands of people who were in the shelters and the tens of thousands of people who evacuated inland were potential victims of injury and death".
[9] Instead it sank like the Bismarck.
[10] "The buildings that collapsed did so because of a combination of two things: very poor soil and very poor structural design," said Peter I. Yanev, chairman of EQE Inc., a structural engineering firm in San Francisco.

Table 7: Example general sentences in humans extracts

We ran the classifier individually on each source sentence and abstract sentence in this corpus. Then we counted the number of pairs which undergo each transformation such as general-general, general-specific from the source to an abstract sentence. These results are reported in Table 9. The table also provides the average number of deletion and substitution operations associated with sentence pairs in that category as well as the length of the uncompressed sentence and the compression rate. Compression rate is defined as the ratio between the

Summary d118i-f:

- President-elect Bush designated Tower as his defense secretary on Dec. 16. [Specific]
- Tower's qualifications for the job –intelligence, patriotism and past chairmanship of the Armed Services Committee –the nomination should have sailed through with flying colors. [Specific]
- *Instead it sank like the Bismarck. [General]*
- In written testimony to the Senate panel on Jan. 26, Tower said he could "recall no actions in connection with any defense activities" in connection with his work for the U.S. subsidiary. [Specific]
- Tower has acknowledged that he drank excessively in the 1970s, but says he has reduced his intake to wine with dinner. [General]
- The Democratic-controlled Senate today rejected the nomination of former Texas Sen. John Tower as defense secretary, delivering a major rebuke to President Bush just 49 days into his term.[Specific]
- The Senate's 53-47 vote came after a bitter and divisive debate focused on Tower's drinking habits, behavior toward women and his business dealings with defense contractors. [General]

Table 8: Example extract with classifier predictions and a general sentence from Table 7

Type	Total	% total	Avg deletions	Avg subs.	Orig length	Compr. rate
SS	6371	39.9	16.3	3.9	33.4	56.6
SG	5679	35.6	21.4	3.7	33.5	40.8
GG	3562	22.3	9.3	3.3	21.5	60.8
GS	352	2.2	8.4	4.0	22.7	66.0

Table 9: Types of transformation of source into abstract sentences

length in words of the compressed sentence and the length of the uncompressed sentence. So lower compression rates indicate greater compression.

We find that the most frequent transformations are specific-specific (SS) and specific-general (SG). Together they constitute 75% of all transformations. But for our analysis, the SG transformation is most interesting. One third of the sentences in this data are converted from originally specific content to being general in the abstracts. So abstracts do tend to involve a lot of generalization.

Studying the SG transition in more detail, we can see that the original sentences are much longer compared to other transitions. This situation arises from the fact that specific sentences in this corpus are longer. In terms of the number of deletions, we see that both SS and SG involve more than 15 deletions, much higher than that performed on the general sentences. However, we do not know if these operations are proportional to the original length of the sentences. But looking at the compression rates, we get a clearer picture, the SG sentences after compression are only 40% their original length, the maximum compression seen for the transformation types. For GG and GS, about 60% of the original sentence words are kept. For the SG transition, long sentences are chosen and are compressed aggressively. In Table 10, we show some example sentence pairs undergoing the SG transition.

Currently, compression systems do not achieve the level of compression in human abstracts. Sentences that humans create are shorter than what systems produce. Our results predict that these could be the cases where specific sentences get converted into general. One reason why systems do not attain this compression level could be because they only consider a limited set of factors while compressing, such as importance and grammaticality. We believe that generality can be an additional objective which can be used to produce even shorter sentences which we have seen in our work, will also lead to summaries with better content.

8 Conclusion

In this work, we have provided the first quantitative analysis of general and specific content as relevant to the task of automatic summarization. We find that general content is useful for summaries however, current content selection methods appear to not include much general content. So we have proposed the task of identifying general content which could be used in summaries. There are two ways of achieving this—by identifying relevant general sentences from the input and by conversion from specific to

[1] American Mitac offers free technical support for one year at a toll-free number from 7:30 to 5:30 P.S.T.
American Mitac offers toll-free technical support for one year.
[2] In addition to Yurman, several other government officials have served on the steering committee that formed the group.
Several government officials also served on the steering committee.
[3] All version of the new tape drives, which, according to Goldbach, offer the lowest cost per megabyte for HSC-based 8mm tape storage, are available within 30 days of order.
The products are available within 30 days of order.
[4] In a different vein is Edward Tufte 's "The Visual Display of Quantitative Information" (Graphics Press, 1983), a book covering the theory and practice of designing statistical charts, maps, tables and graphics.
Tufte 's book covers the theory and practice of designing statistical charts, maps, tables and graphics.
[5] In addition, Anderson said two Ada 9X competitive procurements–a mapping and revision contract and an implementation and demonstration contract–will be awarded in fiscal 1990.
Two competitive procurements will be awarded in fiscal 1989.

Table 10: Example specific to general (in italics) compressions

general content. We have provided a brief overview of these two approaches.

Our work underscores the importance of compression and other post-processing approaches over extractive summaries. Otherwise system content could contain too much extraneous details which take up space where other useful content could have been discussed.

Our study also highlights a semantic view of summary creation. Summaries are not just a bag of important sentences as viewed by most methods today. Rather a text should have a balance between sentences which introduce a topic and those which discuss them in detail. So another approach to content selection could be the joint selection of a general sentence with its substantiation. In future work, it would be interesting to observe if such summaries are judged more responsive and of better linguistic quality than summaries which do not have such a structure.

References

R. Barzilay and K. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3).

J. Clarke and M. Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31(1):399–429.

K. Filippova and M. Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of EMNLP*, pages 177–185.

M. Galley and K. McKeown. 2007. Lexicalized markov grammars for sentence compression. In *Proceedings NAACL-HLT*.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of NAACL-HLT*, pages 362–370.

D. Harman and M. Liberman. 1993. Tipster complete. *Corpus number LDC93T3A, Linguistic Data Consortium, Philadelphia*.

H. Jing and K. McKeown. 2000. Cut and paste based text summarization. In *Proceedings of NAACL*.

K. Knight and D. Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1).

C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*.

C. Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *ACL Text Summarization Workshop*.

A. Louis and A. Nenkova. 2011. General versus specific sentences: automatic identification and application to analysis of news summaries. Technical Report No. MS-CIS-11-07, University of Pennsylvania Department of Computer and Information Science.

E. Marsi, E. Krahmer, I. Hendrickx, and W. Daelemans. 2010. On the limits of sentence compression by deletion. In E. Krahmer and M. Theune, editors, *Empirical methods in natural language generation*, pages 45–66. Springer-Verlag, Berlin, Heidelberg.

R. McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *EACL'06*.

A. Nenkova, R. Passonneau, and K. McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4.

S. Wan, R. Dale, M. Dras, and C. Paris. 2008. Seed and grow: augmenting statistically generated summary sentences using schematic word patterns. In *Proceedings of EMNLP*, pages 543–552.