Verbose, Laconic or Just Right: A Simple Computational Model of Content Appropriateness under Length Constraints

Annie Louis*
School of Infomatics
University of Edinburgh
Edinburgh EH8 9AB
alouis@inf.ed.ac.uk

Ani Nenkova
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19103
nenkova@seas.upenn.edu

Abstract

Length constraints impose implicit requirements on the type of content that can be included in a text. Here we propose the first model to computationally assess if a text deviates from these requirements. Specifically, our model predicts the appropriate length for texts based on content types present in a snippet of constant length. We consider a range of features to approximate content type, including syntactic phrasing, constituent compression probability, presence of named entities, sentence specificity and inter-sentence continuity. Weights for these features are learned using a corpus of summaries written by experts and on high quality journalistic writing. During test time, the difference between actual and predicted length allows us to quantify text verbosity. We use data from manual evaluation of summarization systems to assess the verbosity scores produced by our model. We show that the automatic verbosity scores are significantly negatively correlated with manual content quality scores given to the summaries.

1 Introduction

In dialog, the appropriate length of a speaker turn and the amount of detail in it are hugely influenced by the pragmatic context. For example what constitutes an appropriate answer to the question "How was your vacation?" would be very different when the question is asked as two acquaintances pass each other in the corridor or right after two friends have ordered dinner at a restaurant. Similarly in writing, content is tailored to explicitly defined or implicitly inferred constraints on the appropriate length of text. Many academics have experienced the frustration of needing to adjust their writing when they need to write a short abstract of two hundred words or an answer to reviewer in no more than five hundred words.

50 word summary:
The De Beers cartel has kept the diamond market stable by matching supply to demand. African nations have recently demanded better terms from the cartel. After the Soviet breakup, De Beers contracted for diamonds with the Yukutian Republic. The US remains the largest diamond market, followed by Japan.

100 word summary:
The De Beers cartel, controlled by the Oppenheimer family controls 80% of the uncut diamond market through its Central Selling Organization. The cartel has kept the diamond market stable by maintaining a buffer pool of diamonds for matching supply to demand. De Beers opened a new mine in 1992 and extended the life of two others through underground mining. Innovations have included automated processing and bussing workers in daily from their homes. African nations have recently demanded better terms. After the Soviet breakup, De Beers contracted for diamonds with the Yukutian Republic. The US remains the largest diamond market, followed by Japan.

Table 1: 50 and 100 word summaries written by the same person for the same set of documents

For a specific application-related example consider the texts in Table 1. These are summaries of a set of news articles discussing the De Beers diamond cartel, written by the same person.[1] The first text is written with the instruction to produce a summary of about 50 words while the latter is in response to a request for a 100 word summary. Obviously the longer summary contains more details. It doesn't however simply extend the shorter summary with more sentences; additional details

*Work done while at University of Pennsylvania.

[1]These summaries come from the Document Understanding Conference dataset (year 2001).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 636–644,
Gothenburg, Sweden, April 26-30 2014.

are interspersed with the original shorter summary.

The performance of a range of human-machine applications can be enhanced if they had the ability to predict the appropriate length of a system contribution and the type of content appropriate for that length. Such applications include document generation (O'Donnell, 1997), soccer commentator (Chen and Mooney, 2008) and question answering with different compression rates for different types of questions (Kaisser et al., 2008). Predicting the type of content appropriate for the given length alone would be highly desirable, for example in automatic essay grading, summarization and even in information retrieval, in which verbose writing is particularly undesirable. In this respect, our work supplements recent computational methods to predict varied aspects of writing quality, such as popular writing style and phrasing in novels (Ganjigunte Ashok et al., 2013), science journalism (Louis and Nenkova, 2013), and social media content (Danescu-Niculescu-Mizil et al., 2012; Lakkaraju et al., 2013).

Our work is the first to explore text verbosity. We introduce a simple application-oriented definition of verbosity and a model to automatically predict verbosity scores. We start with a brief overview of our approach in the next section.

2 Text length and content appropriateness

In this first model of verbosity, we do not carry out an elaborate annotation experiment to create labels for verbosity. There are two main reasons for this choice: a) People find it hard to distinguish between individual aspects of quality and often the ratings for different aspects are highly correlated (Conroy and Dang, 2008; Pitler et al., 2010) b) Moreover, for verbosity in particular, the most appropriate data for annotation would be concise and verbose versions of the same text (possibly of similar lengths). It is more likely that people can distinguish between verbosity of these controlled pairs compared to ratings on an individual article. Such writing samples are not easily available. So we have avoided the uncertainties in annotation in this first work by adopting a simpler approach based on three key ideas.

(i) We define a concise article of length l as "an article that has the appropriate types of content expected in an article of length l". Note that length is not equal to verbosity in our model. Our definition allows for articles of different lengths to be considered concise. Verbosity depends on the appropriateness of content for the article length.

(ii) We model this appropriateness of content for the given length restriction via a set of easily computable features that serve as proxies for (a) type of content and level of detail (syntactic features and sentence specificity) (b) sentence complexity (simple readability-related features), (c) secondary details (syntactic structures with high compression probability) and (d) structure (discourse relations and inter-sentence continuity).

(iii) Forgoing any explicit annotation, we simply train the model on professionally written text in which we assume content is appropriately tailored to the length requirements. We train a regression model on the well-written texts to predict the length of an article based on a single snippet of fixed (short) length from the article. For a new test article, we can obtain a predicted length from this model (length supposing the article is written concisely) based on a short snippet. We use the mismatch between the predicted and actual text length of the article to determine if it is verbose.

We believe that this definition of verbosity has natural uses in applications such as summarization. For example, current systems do not distinguish the task of summary creation for different target lengths. They simply try to maximize estimated sentence importance and to minimize repetitive information. They pay no attention to the fact that the same type of sentences are unlikely to be an optimal selection for both a 50 word and a 400 word summary.

We now briefly present the formal definition of the problem of content appropriateness for a specified text length. Let $T = (t_1, t_2, ...t_n)$ be a collection of concisely-written texts and let $l(t_i)$ denote the length of text t_i. The learning task is to obtain a function based on the content type properties of t_i which helps to predict $l(t_i)$. More specifically, we are given a snippet from t_i, called s_{t_i}, of a constant length k where k is a parameter of our model and $k < \min_{t_j} l(t_j)$. The mapping f is learned based on the constant length snippet only and the aim is to predict the original text length.

$$f(s_{t_i}) \rightarrow \hat{l}(t_i)$$

In our work we choose to work with topical segments from documents rather than the complete documents themselves.

Once the model is trained, we identify the verbosity for a test article as follows: Let us consider a new topic segment t_x during test time. Let the length of the segment be l. We obtain a snippet s_{t_x} of size k from t_x. Now assume that our model predicts $f(s_{t_x}) = \hat{l}$.

Case 1: $\hat{l} \simeq l$, the content type in t_x matches the content types generally present in articles of length l. We consider such articles as concise.

Case 2: $\hat{l} \gg l$, the type of content included in t_x is really suitable for longer and detailed topic segments. Thus t_x is likely conveying too much detail given its length i.e. it is verbose.

Case 3: $\hat{l} \ll l$, the content in t_x is of the type that a skillful writer would include in a much shorter and less detail-oriented text. Thus t_x is likely lacking appropriate details (laconic).

We compute the following scores to quantify verbosity:

Predicted length. is the model prediction $\hat{l}$.

Verbosity degree. This score is the difference between the predicted length and the actual length of the text, $\hat{l} - l$. Positive values of the score indicate the degree of verbosity, negative values indicate that the text is laconic.

Deviation score. Since both being verbose and being laconic is potentially problematic for text, we define a score which does not differentiate the type of mismatch. This score is given by the absolute magnitude $|\hat{l} - l|$.

The next section describes the features used for indicating the content type of a snippet. In Section 4, we test the features on a four-way classification task to predict the length of a human-written summary based on a snippet of the summary. In Section 5, we extend our model to a regression setting by learning feature weights on news articles of varied lengths from the New York Times (NYT), which we consider to be a sample in which content is chosen appropriately for each article length. Finally in Section 6 we evaluate the model trained on NYT articles on machine-produced summaries and confirm that summaries scored with higher verbosity by our model also receive poor content quality scores during manual evaluation.

3 Features mapping content type to appropriate length

We propose a diverse set of 87 features for characterizing content type. These features are computed over the constant length snippet sampled from an article. All the syntax based features are computed from the constituency trees produced from the Stanford Parser (Klein and Manning, 2003).

Length of units (10 features).

This set of features captures basic word and sentence length, and redundancy properties of the snippet. It includes number of sentences, average sentence length in words, average word length in characters, and type to token ratio. We also include the counts of noun phrases, verb phrases and prepositional phrases and the average length in words of these three phrase types.

Syntactic realization (30 features).

We compute the grammatical productions in a set of around 47,000 sentences taken from the AQUAINT corpus (Graff, 2002) We select the most frequent 15 productions in this set that involve a description of entities, i.e the LHS (left-hand side) of the production is a noun phrase. The count of each of these productions is added as a feature allowing us to track what type of information about the entities is conveyed in the snippet. We also add features for the most frequent 15 productions whose LHS is not a noun phrase.

Discourse relations (5 features).

These features are based on the hypothesis that different discourse relations would vary in their appropriateness for articles of different lengths. For example causal information may be included only in more detailed texts.

We use a tool (Pitler and Nenkova, 2009) to identify all explicit discourse connectives in our snippets, along with the general semantic class of the connective (temporal, comparison, contingency and expansion). We use the number of discourse connectives of each of the four types as features, as well as the total number of connectives.

Continuity (6 features).

These features capture the degree to which adjacent sentences in the snippet are related and continue the topic. The amount of continuity for subtopics is likely to vary for long and short texts.

We add the number of pronouns and determiners as two features. Another feature is the average word overlap value between adjacent sentences. For computing the overlap measure, we represent every sentence as a vector where each dimension represents a word. The number of times the word appears in the sentence is the value for that dimension. Cosine similarity is computed between

the vectors of adjacent sentences and the average value of the similarity across all pairs of adjacent sentences is the feature value.

We also run the Stanford Coreference tool (Raghunathan et al., 2010) to identity pronoun and entity coreference links within the snippet. The number of total coreference links, and the number of intra- and inter-sentence links are added as three separate features.

Amount of detail (7 features).

To indicate descriptive words, we compute the number of adjectives and adverbs (two features). We also include the total number of named entities (NEs), average length of NEs in words and the number of sentences that do not have any NEs. The named entities were identified using the Stanford NER recognition tool (Finkel et al., 2005).

We also use the predictions of a classifier trained to identify general versus specific sentences. We use a data set of general and specific sentences and features described in Louis and Nenkova (2011) to implement a sentence specificity model. The classifier produces a binary prediction and also a graded score for specificity. We add two features—the percentage of specific sentences and the average specificity score of words.

Compression likelihood (29 features).

These features use an external source of information about content importance. Specifically, we use data commonly employed to develop statistical models for sentence compression (Knight and Marcu, 2002; McDonald, 2006; Galley and McKeown, 2007). It consists of pairs of sentences in an original text and a professional summary of that text. In every pair, one of the sentences (source) appeared in the original text and the other is a shorter version with the superfluous details deleted. Both sentences were produced by people.

We use the dataset created by Galley and McKeown (2007). The sentences are taken from the Ziff Davis Corpus which contains articles about technology products. This data also contains alignment between the constituency parse nodes of the source and summary sentence pair. Through the alignment it is possible to track nodes that where preserved during compression.

On this data, we identify for every production in the source sentence whether it undergoes deletion in the compressed sentence. A production (LHS $\rightarrow$ RHS) is said to undergo deletion when either the LHS node or any of the nodes in the RHS do not appear in the compressed sentence. Only productions which involve non-terminals in the RHS are used for this analysis as lexical items could be rather corpus-specific. The proportion of times a production undergoes deletion is called the *deletion probability*. We also incorporate frequency of the production with the deletion probability to obtain a representative set of 25 productions which are frequently deleted and also occur commonly. This *deletion score* is computed as: *deletion probability* * log(*frequency of production in source sentences*)

Parentheticals appear in the list as would be expected and also productions involving conjunctions, prepositional phrases and subordinate clauses. We expect that such productions will indicate the presence of details that are only appropriate for longer texts.

To compute the compression-related features for a snippet, we first obtain the set of all productions in the sentences from the snippet. We add features that indicate the number of times each of the top 25 'most deleted' productions was used in the snippet. We also use the sum, average and product of deletion probabilities for set of snippet productions as features. The product feature gives the likelihood of the text being deleted. We also add the perplexity value based on this likelihood, $P^{-1/n}$ where P is the likelihood and n is the number of productions from the snippet for which we have deletion information in our data.[2]

For training a model, we need texts which we can assume are written in a concise manner. We use two sources of data—summaries written by people and high quality news articles.

4 A classification model on expert summaries

Here we use a collection of news summaries written by expert analysts for four different lengths and build a classification model to predict given a snippet what is the length of the summary from which the snippet was taken. This task only differentiates four lengths but is a useful first approach for testing our assumptions and features.

4.1 Data

We use human written summaries from the Document Understanding Conference (DUC[3]) evalua-

[2]Some productions may not have appeared in the Ziff Davis Corpus.

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

tion workshops conducted in 2001 and 2002. An input given for summarization contains 10 to 15 documents on a topic. The person had to create 50, 100, 200 and 400 word summaries for each of the inputs. These summary writers are retired information analysts and we can assume that their summaries are of high quality and concise nature. Further, the four different length summaries for an input are produced by the same person.[4] Therefore differences in length are not confounded by differences in writing style of different people.

The 2001 dataset has 90 summaries for each of the four lengths. In 2002, there are 116 summaries for each length. All of the summaries are abstracts, i.e. people wrote the summary in their own words, with the exception of one set. In 2002, abstracts were only created for 50, 100 and 200 lengths. However, extracts created by people are available for 400 words. In extracts, the summary writer is only allowed to choose complete sentences (no edits can be done), however, the sentences can be ordered in the summary and people tend to create coherent extractive summaries as well. Since it is desirable to have data for another length, we also include the 400-word extracts from the 2002 data.

4.2 Snippet selection

We choose 50 words as the snippet length for our experiment since the length of the shortest summaries is 50. We experiment with multiple ways to select a snippet: the first 50 words of the summary (START), the last 50 words (END) and 50 words starting at a randomly chosen sentence (RANDOM). However, we do not truncate any sentence in the middle to meet the constraint for 50 words. We allow a leeway of 20 words so that snippets can range from 30 to 70 words. When a snippet could not be created within this word limit (eg. the summary has one sentence which is longer than 70 words), we ignore the example.

4.3 Classification results

The task is to predict the length of the summary from which the fixed length snippet was taken, i.e. 4-way classification—50, 100, 200 or a 400 word summary. We trained an SVM classifier with a radial basis kernel on the 2001 data. The regularization and kernel parameters were tuned using 10-fold cross validation on the training set. The accuracies of classification on the 2002 data are shown in Table 2. Since there are four equal classes, the random baseline performance is 25%.

snippet position	accuracy
START	38.4
RANDOM	34.4
END	39.3

Table 2: Length prediction results on DUC summaries

The START and END position snippets gave the best accuracies, 38% and 39% which are 13-14% absolute improvement above the baseline. At the same time, there is much scope for improvement. The confusion matrices showed that 50 and 400 word lengths, the extreme ones in this dataset, were the easiest to predict. Most of the confusions occur with the 100 and 200 word summaries.

The overall accuracy is slightly better when snippets from the END of the summary are chosen compared to those from the START. However, with START snippets, better prediction of different length summaries was obtained, whereas the accuracy in the END case comes mainly from correct prediction of 50 and 400 word summaries. So we use the START selection for further experiments.

5 A regression approach based on New York Times editorials

We next build a model where we predict a wider range of lengths compared to just the four classes we had before. Here our training set comprises news articles from the New York Times (NYT) based on the assumption that edited news from a good source would be of high quality overall.

5.1 Data

We obtain the text of the articles from the NYT Annotated Corpus (Sandhaus, 2008). We choose the articles from the opinion section of the newspaper since they are likely to have good topic continuity and related content compared to general news which often contain lists of facts. We further use only the editorial articles to ensure that the articles are of high quality.

We collect 10,724 opinion articles from years 2000 to 2007 of the NYT. We divide each article into topic segments using the unsupervised topic segmentation method developed by Eisenstein and Barzilay (2008). We use the following heuristic to decide on the number of topic segments for each article. If the article has fewer than 50 sentences, we create segments such that the expected length

[4] Different inputs however may be summarized by different assessors.

of a segment is 10 sentences, i.e, we assign the number of segments as number of sentences divided by 10. When the article is longer, we create 5 segments. This step gives us 18,167 topic segments, ranging in length from 14 to 773 words.

We use a stratified sampling method to select training and test examples. Starting from 90 words and upto a maximum length of 500 words, we divide the range into bins in increments of 30 words. From each bin we select 100 texts for training and around 35 for testing. There are 2,100 topic segments in the training set and 681 for testing.

5.2 Training approach

We use 100 word snippets for our experiments. We learn a linear regression model on the training data using *lm* function in R (R Development Core Team, 2011). The features which turned out significant in the model are shown in Table 3. The significance value shown is associated with a t-test to determine if the feature can be ignored from the model. We report the coefficients for the significant features under column 'Beta'. The R-squared value of the model is 0.219.

Many of the most significant features are related to entities. Longer texts are associated with larger number of noun phrases but they tend not to be proper names. Average word and sentence length also increase with article length, at the same time, longer articles have shorter verb phrases. Specific sentences and determiners are also positively related to article length. At the discourse level, comparison relations increase with length.

5.3 Accuracy of predictions

On the test data, the lengths predicted by the model have a Pearson correlation of 0.44 with the true length of the topic segment. The correlation is highly significant (p-value $<$ 2.2e-16). The Spearman correlation value is 0.43 and the Kendall Tau is 0.29, both also highly significant. These results show that our model can distinguish content types for a range of article lengths.

6 Text quality assessment for automatic summaries

In the models above, we learned weights which relate the features to the length of concisely written human summaries and NYT articles. Now we use the model to compute verbosity scores and assess how well they correlate with text quality scores assigned by people.

Feature	Beta	p-value
Positive coefficients		
total noun phrases	6.052e+00	***
avg. word length	3.201e+01	***
avg. sent. length	3.430e+00	**
avg. NP length	6.557e+00	*
no. of adverbs	4.244e+00	**
% specific sentences	4.773e+01	**
comparison relations	9.296e+00	.
determiners	2.955e+00	*
NP → NP PP	4.305e+00	*
NP → NP NP	1.174e+01	*
PP → IN S	7.268e+00	.
WHNP → WDT	1.196e+01	**
Negative coefficients		
NP → NNP	-8.630e+00	***
no. of sentences	-2.498e+01	**
no. of relations	-1.128e+01	**
avg. VP length	-2.982e+00	**
type token ratio	-1.784e+02	*
NP → NP , SBAR	-1.567e+01	*
NP → NP , NP	-9.582e+00	*
NP → DT NN	-3.423e+00	.
VP → VBD	-1.189e+01	.
S → S : S .	-1.951e+01	.
ADVP → RB	-4.198e+00	.

Table 3: Significant regression coefficients in the length prediction model on NYT editorials. '***' indicates p-value $<$ 0.001, '**' is p-value $<$ 0.01, '*' is $<$ 0.05 and '.' is $<$ 0.1

We perform this evaluation for the system summaries produced during the 2006 DUC evaluation workshop. There are 22 automatic systems in that evaluation.[5] Each system produced 250 word summaries for each of 20 multidocument inputs. Each summary was evaluated by DUC assessors for multiple dimensions of quality. We examine how the verbosity predictions from our model are related to these summary scores. In this experiment, we use automatic summaries only.

6.1 Gold-standard summary scores

Two kinds of manual scores—content and linguistic quality—are available for each summary from the DUC dataset. One type of content score, the 'pyramid score' (Nenkova et al., 2007) computes the overlap of semantic units of the system summary with that present in human-written summaries for the same input. For the other content score, called 'content responsiveness', assessors directly provide a rating to summaries on a scale from 1 (very poor) to 5 (very good) without using any reference human summaries.

[5] We use only the set of systems for which pyramid scores are also available.

Verbosity scores	Corr. with actual length
predicted length	-0.01
verbosity degree	-0.29
deviation score	-0.27

Table 4: Relationship between verbosity scores and summary length

Linguistic quality is evaluated separately from content for different aspects. Manually assigned scores are available for *non-redundancy* (absence of repetitive information), *focus* (well-established topic), and *coherence* (good flow from sentence to sentence). For each aspect, the summary is rated on a scale from 1 (very poor) to 5 (very good).

This dataset is less ideal for our task in some ways as system summaries often lack coherent arrangement of sentences. Some of our features which rely on coreference and adjacent sentence overlaps when computed on these summaries could be misleading. However, this data contains large scale quality ratings for different quality aspects which allow us to examine our verbosity predictions across multiple dimensions.

6.2 Verbosity scores and summary quality

We choose the first 100 words of each summary as the snippet. No topic segmentation was done on the summary data. We use the NYT regression model to predict the expected lengths of these summaries and compute its verbosity and deviation scores as defined in Section 2.

We also compute two other measures for comparison.

Actual length. To understand how the verbosity scores are related to the length of the summary, we also keep track of the actual number of words present in the summary.

Redundancy score: We also add a simple score to our analysis to indicate redundancy between adjacent sentences in the summary. It is simple measure of verbosity since repetitive information leads to lower informativeness overall. The score is the cosine similarity based sentence overlap measure described in Section 3.

For each of the 22 automatic systems, the scores of its 20 summaries (one for each input) are averaged. (We ignore empty summaries and those which are much smaller than the 100 word snippet that we require). We find the average values for both our verbosity based scores above and the gold-standard scores (pyramid, content responsiveness, focus, non-redundancy and coherence). We also compute the average value of the summary lengths for each system.

	Content quality	
scores	**Pyramid**	**Resp.**
actual length	0.64*	0.43*
predicted length	-0.29	-0.11
verbosity degree	-0.47*	-0.23
deviation score	-0.44*	-0.29
redundancy score	-0.01	-0.06

	Linguistic quality		
scores	**Non-red**	**Focus**	**Coher.**
actual length	-0.32	-0.25	-0.32
predicted length	0.48*	0.39$^{\cdot}$	0.38$^{\cdot}$
verbosity degree	0.55*	0.44*	0.46*
deviation score	0.53*	0.40$^{\cdot}$	0.42$^{\cdot}$
redundancy score	0.06	0.32	0.23

Table 5: Pearson correlations between verbosity scores and gold standard summary quality scores

First we examine the relationship between verbosity scores and the actual summary lengths. The Pearson correlations between the three verbosity measures and true length of the summaries are reported in Table 4. The verbosity scores are not significantly related to summary length. They seem to have an inverse relationship but the correlations are not significant even at 90% confidence level. This result supports our hypothesis that verbosity scores based on expected length are different from the actual summary length.

Next Table 5 presents the Pearson correlations of the verbosity measures with gold standard summary quality scores. Since the number of points (systems) is only 22, we indicate whether the correlations are significant at two levels, 0.05 (marked by a '*' superscript) and 0.1 (a '.' superscript).

The first line of the table indicates that longer summaries are associated with higher content scores both according to pyramid and content responsiveness evaluations. This result also supports our hypothesis that length alone does not indicate verbosity. Longer summaries on average have better content quality. The length is not significantly related to linguistic quality scores but there is a negative relationship in general.

On the other hand, all the three verbosity scores have a negative correlation with content scores. The verbosity degree score is the strongest indicator of summary quality with -0.47 (significant) correlation with pyramid score. At the same time however, verbosity is preferred for linguistic quality. This effect could arise due to the fact these summaries are bags of unordered sentences. Therefore verbose style could be perceived as hav-

System 23's summary: Actual length = 253 words, Predicted length = 343 words, Verbosity degree = 90 A senior Scotland Yard police officer apologized to the parents of a black teenager slain five years ago in a race killing that has become the focus of debate over relations between police and ethnic minorities. Black teenager Stephen Lawrence was stabbed to death at a bus-stop in Eltham, south London by five white youngsters six years ago. The parents of the murdered black teenager Stephen Lawrence began legal action against the men suspected of his killing. Two suspects in the Stephen Lawrence murder case and one other man were arrested on suspicion of theft by Kent Police. The five men suspected of killing Stephen Lawrence were thumped and pelted with bottles by an enraged crowd Tuesday after a day of evasive and implausible evidence that made a mockery of their appearance before the public inquiry. The dawn raids came as police questioned three men in connection with the country's most notorious racist crime: the unsolved 1993 murder of black teenager Stephen Lawrence. A public inquiry after the Lawrence case found London police institutionally racist, prompting a government pledge to take a more active role in combating racial intolerance. The report, commissioned after police botched the investigation into the 1993 racially motivated murder of a black teenager, Stephen Lawrence has put pressure on Sir Paul Condon, the Metropolitan Police chief, to resign. British authorities and police have learned from the 1993 murder of black teen-ager Stephen Lawrence by a gang of white youths and the failure of the police to
System 18's summary: Actual length = 244 words, Predicted length = 597 words, Verbosity degree = 353 The government, which has received praise from backers of the Lawrence family for its pursuit of the case, came in for criticism on Monday for actions it took this weekend to prevent publication of a leaked version of the report, which is due to be made public on Wednesday. Sir William Macpherson, a retired High Court justice who was the author of the report and chairman of the eight-month government inquiry, defined institutional racism as 'the collective failure of an organization to provide an appropriate professional service to people because of their color, culture or ethnic origin' reflected, he said, in 'processes, attitudes and behavior which amounts to discrimination through unwitting prejudice ignorance, thoughtlessness and racist stereotyping.' Richard Norton-Taylor, whose play about Lawrence's killing, 'The Color of Justice,' has been playing to rave reviews in London, said that the attention paid to the Lawrence case and others was a sign that British attitudes toward the overarching authority of the police and other institutions were finally being called into question. She said British authorities and police have learned from the 1993 murder of black teenager Stephen Lawrence by a gang of white youths and the failure of the police to investigate his death adequately A senior Scotland Yard police officer Wednesday apologized to the parents of a black teenager slain five years ago in a race killing that has become the focus of debate over relations between police and ethnic minorities.

Table 6: Summaries produced by two systems for input D0624 (DUC 2006) shown with the verbosity scores from our model

ing greater coherence compared to short and succinct sentences which are jumbled such that it is hard to decipher the full story.

The simple redundancy score (last row of the table) does not have any significant relationship to quality scores. One reason could be that most summarization systems make an effort to reduce redundant information (Carbonell and Goldstein, 1998) and therefore a simple measure of word overlap is not helpful for distinguishing quality.

As examples of the predictions from our model, Table 6 shows two summaries produced for the same input by two different systems. They both have almost the same actual length but the first received a prediction close to its actual length while the other is predicted with a much higher verbosity degree score. Intuitively, the second example is more verbose compared to the first one. According to the manual evaluations as well, the first summary receives a higher score of 0.4062 (pyramid) compared to 0.2969 for the second summary.

7 Conclusions

There are several ways in which our approach can be improved. In this first work, we have avoided the complexities of manual annotation. In future, we will explore the feasibility of human annotations of verbosity on a suitable corpus, such as news articles on the same topic from different sources. In addition, our current approach only considers a snippet of the text or topic segment during prediction but ignores the writing in the remaining text. In future work, we plan to use a sliding window to obtain and aggregate length predictions while considering the full text.

Acknowledgements

This work was partially supported by a NSF CAREER 0953445 award. We also thank the anonymous reviewers for their comments.

References

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336.

D. L. Chen and R. J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of ICML*, pages 128–135.

J. M. Conroy and H. T. Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of COLING*, pages 145–152.

C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of ACL*, pages 892–901.

J. Eisenstein and R. Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

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

V. Ganjigunte Ashok, S. Feng, and Y. Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP*, pages 1753–1764.

D. Graff. 2002. The AQUAINT Corpus of English News Text. *Corpus number LDC2002T31, Linguistic Data Consortium, Philadelphia*.

M. Kaisser, M. A. Hearst, and J. B. Lowe. 2008. Improving search results quality by customizing summary lengths. In *Proceedings of ACL-HLT*, pages 701–709.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

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

H. Lakkaraju, J. J. McAuley, and J. Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *ICWSM*.

A. Louis and A. Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*, pages 605–613.

A. Louis and A. Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *TACL*, 1:341–352.

R. McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *Proceedings of EACL*.

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.

M. O'Donnell. 1997. Variable-length on-line document generation. In *Proceedings of the 6th European Workshop on Natural Language Generation*.

E. Pitler and A. Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of ACL-IJCNLP*, pages 13–16.

E. Pitler, A. Louis, and A. Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of ACL*.

R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP*, pages 492–501.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*.