What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain

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

Great writing is rare and highly admired. Readers seek out articles that are beautifully written, informative and entertaining. Yet information-access technologies lack capabilities for predicting article quality at this level. In this paper we present first experiments on article quality prediction in the science journalism domain. We introduce a corpus of great pieces of science journalism, along with typical articles from the genre. We implement features to capture aspects of great writing, including surprising, visual and emotional content, as well as general features related to discourse organization and sentence structure. We show that the distinction between great and typical articles can be detected fairly accurately, and that the entire spectrum of our features contribute to the distinction.

1 Introduction

Measures of article quality would be hugely beneficial for information retrieval and recommendation systems. In this paper, we describe a dataset of New York Times science journalism articles which we have categorized for quality differences and present a system that can automatically make the distinction.

Science journalism conveys complex scientific ideas, entertaining and educating at the same time. Consider the following opening of a 2005 article by David Quammen from Harper's magazine:

One morning early last winter a small item appeared in my local newspaper announcing the birth of an extraordinary animal. A team of researchers at Texas A&M University had succeeded in cloning a whitetail deer. Never done before. The fawn, known as Dewey, was developing normally and seemed to be healthy. He had no mother, just a surrogate who had carried his fetus to term. He had no father, just a "donor" of all his chromosomes. He was the genetic duplicate of a certain trophy buck out of south Texas whose skin cells had been cultured in a laboratory. One of those cells furnished a nucleus that, transplanted and rejiggered, became the DNA core of an egg cell, which became an embryo, which in time became Dewey. So he was wildlife, in a sense, and in another sense elaborately synthetic. This is the sort of news, quirky but epochal, that can cause a person with a mouthful of toast to pause and marvel. What a dumb idea, I marveled.

The writing is clear and well-organized but the text also contains creative use of language and a clever story-like explanation of the scientific contribution. Such properties make science journalism an attractive genre for studying writing quality. Science journalism is also a highly relevant domain for information retrieval in the context of educational as well as entertaining applications. Article quality measures can hugely benefit such systems.

Prior work indicates that three aspects of article quality can be successfully predicted: **a)** whether a text meets the acceptable standards for spelling (Brill and Moore, 2000), grammar (Tetreault and Chodorow, 2008; Rozovskaya and Roth, 2010) and discourse organization (Barzilay et al., 2002; Lapata, 2003); **b)** has a topic that is interesting to a particular user. For example, content-based recommendation systems standardly represent user interest using frequent words from articles in a user's history and retrieve other articles on the same topics (Paz-

Transactions of the Association for Computational Linguistics, 1 (2013) 341–352. Action Editor: Mirella Lapata.
Submitted 12/2012; Revised 3/2013, 5/2013; Published 7/2013.

zani et al., 1996; Mooney and Roy, 2000); and **c)** is easy to read for a target readership. Shorter words (Flesch, 1948), less complex syntax (Schwarm and Ostendorf, 2005) and high cohesion between sentences (Graesser et al., 2004) typically indicate easier and more 'readable' articles.

Less understood is the question of what makes an article interesting and beautifully written. An early and influential work on readability (Flesch, 1948) also computed an interest measure with the hypothesis that interesting articles would be easier to read. More recently, McIntyre and Lapata (2009) found that people's ratings of interest for fairy tales can be successfully predicted using token-level scores related to syntactic items and categories from a psycholinguistic database. But large scale studies of interest measures for adult educated readers have not been carried out.

Further, there have been little attempts to measure article quality in a genre-specific setting. But it is reasonable to expect that properties related to the unique aspects of a genre should contribute to the prediction of quality in the same way that domain-specific spelling and grammar correction (Cucerzan and Brill, 2004; Bao et al., 2011; Dale and Kilgarriff, 2010) techniques have been successful.

Here we address the above two issues by developing measures related to interesting and well-written nature specifically for science journalism. Central to our work is a corpus of science news articles with two categories: written by popular journalists and typical articles in science columns (Section 2). We introduce a set of genre-specific features related to beautiful writing, visual nature and affective content (Section 3) and show that they have high predictive accuracies, 20% above the baseline, for distinguishing our quality categories (Section 4). Our final system combines the measures for interest and genre-specific features with those proposed for identifying readable, well-written and topically interesting articles, giving an accuracy of 84% (Section 5).

2 Article quality corpus

Our corpus[1] contains chosen articles from the larger New York Times (NYT) corpus (Sandhaus, 2008), the latter containing a wealth of metadata about each article including author information and manually assigned topic tags.

2.1 General corpus

The articles in the VERY GOOD category include all contributions to the NYT by authors whose writing appeared in "The Best American Science Writing" anthology published annually since 1999. Articles from the science columns of leading newspapers are nominated and expert journalists choose a set they consider exceptional to appear in these anthologies. There are 63 NYT articles in the anthology (between years 1999 and 2007) that are also part of the digital NYT corpus; these articles form the seed set of the VERY GOOD category.

We further include in the VERY GOOD category all other science articles contributed to NYT by the authors of the seed examples. Science articles by other authors not in our seed set form the TYPICAL category. We perform this expansion by first creating a *relevant set* of science articles. There is no single meta-data tag in the NYT which refers to all the science articles. So we use the topic tags from the seed articles as an initial set of *research tags*. We then compute the minimal set of *research tags* that cover all best articles. We greedily add tags into the minimal set, at each iteration choosing the tag that is present in the majority of articles that remain uncovered. This minimal set contains 14 tags such as 'Medicine and Health', 'Space', 'Research', 'Physics' and 'Evolution'.

We collect articles from the NYT which have at least one of the minimal set tags. However, even a cursory mention of a research topic results in a research-related tag being assigned to the article. So we also use a dictionary of research-related terms to determine whether the article passes a minimum threshold for research content. We created this dictionary manually and it contains the following words and their morphological variants (total 63 items). We used our intuition about a few categories of research words to create this list. The category is shown in capital letters below.

PEOPLE: researcher, scientist, physicist, biologist, economist, anthropologist, environmentalist, linguist, professor, dr, student

PROCESS: discover, found, experiment, work, finding, study, question, project, discuss

TOPIC: biology, physics, chemistry, anthropology, primatology

[1] Available from http://www.cis.upenn.edu/~nlp/corpora/scinewscorpus.html

PUBLICATIONS: report, published, journal, paper, author, issue
OTHER: human, science, research, knowledge, university, laboratory, lab
ENDINGS: -ology -gist, -list, -mist, -uist, -phy

The items in the ENDINGS category are used to match word suffixes. An article is considered science-related if at least 10 of its tokens match the dictionary and in addition, at least 5 unique words from the dictionary are matched. Since the time span of the best articles is 1999 to 2007, we limit our collection to this timespan. In addition, we only consider articles that are at least 500 words long. This filtered set of 23,709 articles form the *relevant set* of science journalism.

The 63 seed samples of great writing were contributed by about 40 authors. Some authors have multiple articles selected for the best writing book series, supporting the idea that these authors produce high quality pieces that can be considered distinct from typical articles. Separating the articles from these authors gives us 3,467 extra samples of VERY GOOD writing. In total, the VERY GOOD set has 3,530 articles. The remaining articles from the relevant set, 20,242, written by about 3000 other authors form the TYPICAL category.

2.2 Topic-paired corpus

The general corpus of science writing introduced so far contains articles on diverse topics including biology, astronomy, religion and sports. The VERY GOOD and TYPICAL categories created above allow us to study writing quality without regard to topic. However a typical information retrieval scenario would involve comparison between articles of the same topic, i.e. relevant to the same query. To investigate how quality differentiation can be done within topics, we created another corpus where we paired articles of VERY GOOD and TYPICAL quality.

For each article in the VERY GOOD category, we compute similarity with all articles in the TYPICAL set. This similarity is computed by comparing the topic words (computed using a loglikelihood ratio test (Lin and Hovy, 2000)) of the two articles. We retain the most similar 10 TYPICAL articles for each VERY GOOD article. We enumerate all pairs of VERY GOOD with matched up TYPICAL ARTICLES (10 in number) giving a total of 35,300 pairs.

There are two distinguishing aspects of our corpus. First, the average quality of articles is high. They are unlikely to have spelling, grammar and basic organization problems allowing us to investigate article quality rather than the detection of errors. Second, our corpus contains more realistic samples of quality differences for IR or article recommendation compared to prior work, where system produced texts and permuted versions of an original article were used as proxies for lower quality text.

2.3 Tasks

We perform two types of classification tasks. We divide our corpus into development and test sets for these tasks in the following way.

Any topic: Here the goal is to separate out VERY GOOD versus TYPICAL articles without regard to topic. The setting roughly corresponds to picking out an interesting article from an archive or a day's newspaper. The test set contains 3,430 VERY GOOD articles and we randomly sample 3,430 articles from the TYPICAL category to comprise the negative set.

Same topic: Here we use the topic-paired VERY GOOD and TYPICAL articles. The goal is to predict which article in the pair is the VERY GOOD one. This task is closer to a information retrieval setting, where articles similar in topic (retrieved for a user query) need to be distinguished for quality. For test set, we selected 34,300 pairs.

Development data: We randomly selected 100 VERY GOOD articles and their paired (10 each) TYPICAL articles from the topic-normalized corpus. Overall, these constitute 1,000 pairs which we use for developing the same-topic classifier. From these selected pairs we take the 100 VERY GOOD articles and sample 100 unique articles from the TYPICAL articles making up the pairs. These 200 articles are used to tune the any-topic classifier.

3 Facets of science writing

In this section, we discuss six prominent facets of science writing which we hypothesized will have an impact on text quality. These are the presence of passages of highly visual nature, people-oriented content, use of beautiful language, sub-genres, sentiment or affect, and the depth of research description. Several other properties of science writing could also be relevant to quality such as the use of

humor, metaphor, suspense and clarity of explanations and we plan to explore these in future work.

We describe how we computed features related to each property and tested how these features are distributed in the VERY GOOD and TYPICAL categories. To do this analysis, we randomly sampled 1,000 articles from each of the two categories as representative examples. We compute the value of each feature on these articles and use a two-sided t-test to check if the mean value of the feature is higher in one class of articles. A p-value less than 0.05 is taken to indicate significantly different trend for the feature in the VERY GOOD versus TYPICAL articles.

Note that our feature computation step is not tuned for the quality prediction task in any way. Rather we aim to represent each facet as accurately as possible. Ideally we would require manual annotations for each facet (visual, sentiment nature etc.) to achieve this goal. At this time, we simply check some chosen features' values on a random collection of snippets from our corpus and check if they behave as intended without resorting to these annotations.

3.1 Visual nature of articles

Some texts create an image in the reader's mind. For example, the snippet below has a high visual effect.

> When the sea lions approached close, seemingly as curious about us as we were about them, their big brown eyes were encircled by light fur that looked like makeup. One sea lion played with a conch shell as if it were a ball.

Such vivid descriptions can engage and entertain a reader. Kosslyn (1980) found that people spontaneously form images of concrete words that they hear and use them to answer questions or perform other tasks. Books written for student science journalists (Blum et al., 2006; Stocking, 2010) also emphasize the importance of visual descriptions.

We measure the visual nature of a text by counting the number of visual words. Currently, the only resource of imagery ratings for words is the MRC psycholinguistic database (Wilson, 1988). It contains a list of 3,394 words rated for their ability to invoke an image, so the list contains both words that are highly visual along with words that are not visual at all. With a cutoff value we adopted, of 4.5 for the Gilhooly-Logie and 350 for the Bristol Norms[2] we obtain 1,966 visual words. So the coverage of that lexicon is likely to be low for our corpus.

We collect a larger set of visual words from a corpus of tagged images from the ESP game (von Ahn and Dabbish, 2004). The corpus contains 83,904 total images and 27,466 unique tags. The average number of tags per picture is 14.5. The tags were collected in a game setting where two users individually saw the same image and had to guess words related to it. The players increased their scores when the word guessed by one player matched that of the other. Due to the simple annotation method, there is considerable noise and non-visual words assigned as tags. So we performed filtering to find high precision image words and also group them into topics.

We use Latent Dirichlet Allocation (Blei et al., 2003) to cluster image tags into topics. We treat each picture as a document and the tags assigned to the picture are the document's contents. We use symmetric priors set to 0.01 for both topic mixture and word distribution within each topic. We filter out the 30 most common words in the corpus, words that appear in less than four pictures and images with fewer than five tags. The remaining words are clustered into 100 topics with the Stanford Topic Modeling Toolbox[3] (Ramage et al., 2009). We did not tune the number of topics and choose the value of 100 based on the intuition that the number of visual topics is likely to be small.

To select clean visual clusters, we make the assumption that visual words are likely to be clustered with other visual terms. Topics that are not visual are discarded altogether. We use the manual annotations available with the MRC database to determine which clusters are visual. For each of the 100 topics from the topic model, we examine the top 200 words with highest probability in that topic. We compute the precision of each topic as the proportion of these 200 words that match the MRC list of visual words (1,966 words using the cutoff mentioned above). Only those topics which had a precision of at least 25% were retained, resulting in 68 visual topics. Some example topics, with manually created headings, include:

landscape. grass, mountain, green, hill, blue, field, brown, sand, desert, dirt, landscape, sky

[2]The visual words resource in MRC contains two lists—Gilhooly-Logie and Bristol Norms.

[3]http://nlp.stanford.edu/software/tmt/tmt-0.4/

jewellery. silver, white, diamond, gold, necklace, chain, ring, jewel, wedding, diamonds, jewelry

shapes. round, ball, circles, logo, dots, square, dot, sphere, glass, hole, oval, circle

Combining these 68 topics, there are 5,347 unique visual words because topics can overlap in the list of most probable words. 2,832 words from this set are not present in the MRC database. Some examples of new words in our list are 'daffodil', 'sailor', 'helmet', 'postcard', 'sticker', 'carousel', 'kayak', and 'camouflage'. For later experiments we consider the 5,347 words as the visual word set and also keep the information about the top 200 words in the 68 selected topics. We compute two classes of features one based on all visual words and the other on visual topics. We consider only the adjectives, adverbs, verbs and common nouns in an article as candidate words for computing visual quality.

Overall visual use: We compute the proportion of candidate words that match the visual word list as the TOTAL_VISUAL feature. We also compute the proportions based only on the first 200 words of the article (BEG_VISUAL), the last 200 words (END_VISUAL) and the middle region (MID_VISUAL) as features. We also divide the article into five equally sized bins of words where each bin captures consecutive words in the article. Within each bin we compute the proportion of visual words. We treat these values as a probability distribution and compute its entropy (ENTROPY_VISUAL). We expected these position features to indicate how the placement of visual words is related to quality.

Topic-based features: We also compute what proportion of the words we identify as visual matches the list under each topic. The maximum proportion from a single topic (MAX_TOPIC_VISUAL) is a feature. We also compute a greedy cover set of topics for the visual words in the article. The topic that matches the most visual words is added first, and the next topic is selected based on the remaining unmatched words. The number of topics needed to cover 50% of the article's visual words is the TOPIC_COVER_VISUAL feature. These features capture the mix of visual words from different topics. Disregarding topic information, we also compute a feature NUM_PICTURES which is the number of images in the corpus where 40% of the image's tags are matched in the article.

We found 8 features to vary significantly between the two types of articles. The features with significantly higher values in VERY GOOD articles are: BEG_VISUAL, END_VISUAL, MAX_TOPIC_VISUAL. The features with significantly higher values in the TYPICAL articles are: TOTAL_VISUAL, MID_VISUAL, ENTROPY_VISUAL, TOPIC_COVER_VISUAL, NUM_PICTURES.

It appears that the simple expectation that VERY GOOD articles contain more visual words overall does not hold true here. However the great writing samples have a higher degree of visual content in the beginning and ends of articles. Good articles also have lower entropy for the distribution of visual words indicating that they appear in localized positions in contrast to being distributed throughout. The topic-based features further indicate that for the VERY GOOD articles, the visual words come from only a few topics (compared to TYPICAL articles) and so may evoke a coherent image or scene.

3.2 The use of people in the story

We hypothesized that articles containing research findings that directly affect people in some way, and therefore involve explicit references to people in the story, will make a bigger impact on the reader. For example, the most frequent topic among our VERY GOOD samples is 'medicine and health' where articles are often written from the view of a patient, doctor or scientist. An example is below.

Dr. Remington was born in Reedville, Va., in 1922, to Maud and P. Sheldon Remington, a school headmaster. Charles spent his boyhood chasing butterflies alongside his father, also a collector. During his graduate studies at Harvard, he founded the Lepidopterists' Society with an equally butterfly-smitten undergraduate, Harry Clench.

We approximate this facet by computing the number of explicit references to people, relying on three sources of information about animacy of words. The first is named entity (NE) tags (PERSON, ORGANIZATION and LOCATION) returned by the Stanford NE recognition tool (Finkel et al., 2005). We also created a list of personal pronouns such as 'he', 'myself' etc. which standardly indicate animate entities (*animate_pronouns*).

Our third resource contains the number of times different noun phrases (NP) were followed by each of the relative pronouns 'who', 'where' and 'which'.

These counts for 664,673 noun phrases were collected by Ji and Lin (2009) from the Google Ngram Corpus (Lin et al., 2010). We use a simple heuristic to obtain a list of animate (*google_animate*) and inanimate nouns (*google_inanimate*) from this list. The head of each NP is taken as a candidate noun. If the noun does not occur with 'who' in any of the noun phrases where it is the head, then it is inanimate. In contrast, if it appears only with 'who' in all noun phrases, it is animate. Otherwise, for each NP where the noun is a head, we check whether the count of times the noun phrase appeared with 'who' is greater than each of the occurrences of 'which', 'where' and 'when' (taken individually) with that noun phrase. If the condition holds for at least one noun phrase, the noun is marked as animate.

When computing the features for an article, we consider all nouns and pronouns as candidate words. If the word is a pronoun and appears in our list of *animate_pronouns*, it is assigned an 'animate' label and 'inanimate' otherwise. If the word is a proper noun and tagged with the PERSON NE tag, we mark it as 'animate' and if it is a ORGANIZATION or LOCATION tag, the word is 'inanimate'. For common nouns, we check it if appears in the *google_animate* and *inanimate* lists. Any match is labelled accordingly as 'animate' and 'inanimate'. Note that this procedure may leave some nouns without any labels.

Our features are counts of animate tokens (ANIM), inanimate tokens (INAMIN) and both these counts normalized by total words in the article (ANIM_PROP, INANIM_PROP). Three of these features had significantly higher mean values in the TYPICAL category of articles: ANIM, ANIM_PROP, INANIM_PROP. We found upon observation that several articles that talk about government policies involve a lot of references to people but are often in the TYPICAL category. These findings suggest that the 'human' dimension might need to be computed not only based on simple counts of references to people but also involve finer distinctions between them.

3.3 Beautiful language

Beautiful phrasing and word choice can entertain the reader and leave a positive impression. Multiple studies in the education genre (Diederich, 1974; Spandel, 2004) note that when teachers and expert adult readers graded student writing, word choice and phrasing always turn out as a significant factors influencing the raters' scores.

We implement a method for detecting creative language based on a simple idea that creative words and phrases are sometimes those that are used in unusual contexts and combinations or those that sound unusual. We compute measures of unusual language both at the level of individual words and for the combination of words in a syntactic relation.

Word level measures: Unusual words in an article are likely to be those with low frequencies in a background corpus. We use the full set of articles (not only science) from year 1996 in the NYT corpus as a background (these do not overlap with our corpus for article quality). We also explore patterns of letters and phoneme sequences with the idea that unusual combination of characters and phonemes could create interesting words. We used the CMU pronunciation dictionary (Weide, 1998) to get the phoneme information for words and built a 4-gram model of phonemes on the background corpus. Laplace smoothing is used to compute probabilities from the model. However, the CMU dictionary does not contain phoneme information for several words in our corpus. So we also compute an approximate model using the letters in the words and obtain another 4-gram model.[4] Only words that are longer than 4 characters are used in both models and we filter out proper names, named entities and numbers.

During development, we analyzed the articles from an entire year of NYT, 1997, with the three models to identify unusual words. Below is the list of words with lowest frequency and those with highest perplexity under the phoneme and letter models.

Low frequency. undersheriff, woggle, ahmok, hofman, volga, oceanaut, trachoma, baneful, truffler, acrimal, corvair, entomopter

High perplexity-phoneme model. showroom, yahoo, dossier, powwow, plowshare, oomph, chihuahua, ionosphere, boudoir, superb, zaire, oeuvre

High perplexity-letter model. kudzu, muumuu, qipao, yugoslav, kohlrabi, iraqi, yaqui, yakuza, jujitsu, oeuvre, yaohan, kaffiyeh

For computing the features, we consider only nouns, verbs, adjectives and adverbs. We also require that the words are at least 5 letters long

[4] We found that higher order n-grams provided better predictions of unusual nature during development.

and do not contain a hyphen[5]. Three types of scores are computed. FREQ_NYT is the average of word frequencies computed from the background corpus. The second set of features are based on the phoneme model. We compute the average perplexity of words under the model, AVR_PHONEME_PERP_ALL. In addition, we also order the words in an article based on decreasing perplexity values and the average perplexity of the top 10, 20 and 30 words in this list are added as features (AVR_PHONEME_PERP_10, 20, 30). We obtain similar features from the letter n-gram model (AVR_CHAR_PERP_ALL, AVR_CHAR_PERP_10, 20, 30). In phoneme features, we ignore words that do not have an entry in the CMU dictionary.

Word pair measures: Next we attempt to detect unusual combinations of words. We do this calculation only for certain types of syntactic relations–a) nouns and their adjective modifiers, b) verbs with adverb modifiers, c) adjacent nouns in a noun phrase and d) verb and subject pairs. Counts for co-occurrence again come from NYT 1996 articles. The syntactic relations are obtained using the constituency and dependency parses from the Stanford parser (Klein and Manning, 2003; De Marneffe et al., 2006). To avoid the influence of proper names and named entities, we replace them with tags (NNP for proper names and PERSON, ORG, LOC for named entities).

We treat the words for which the dependency holds as a (auxiliary word, main word) pair. For adjective-noun and adverb-verb pairs, the auxiliary is the adjective or adverb; for noun-noun pairs, it is the first noun; and for verb-subject pairs, the auxiliary is the subject. Our idea is to compute usualness scores based on frequency with which a particular pair of words appears in the background.

Specifically, we compute the conditional probability of the auxiliary word given the main word as the score for likelihood of observing the pair. We consider the main word as related to the article topic, so we use the conditional probability of auxiliary given main word and not the other way around. However, the conditional probability has no information about the frequency of the auxiliary word. So we apply ideas from interpolation smoothing (Chen and Goodman, 1996) and compute the conditional probability as a interpolated quantity together with the unigram probability of the auxiliary word.

$$\hat{p}(aux|main) = \lambda * p(aux|main) + (1-\lambda) * p(aux)$$

The unigram and conditional probabilities are also smoothed using Laplace method. We train the lambda values to optimize data likelihood using the Baum Welch algorithm and use the pairs from NYT 1997 year articles as a development set. The lambda values across all types of pairs tended to be lower than 0.5 giving higher weight to the unigram probability of the auxiliary word.

Based on our observations on the development set, we picked a cutoff of 0.0001 on the probability (0.001 for adverb-verb pairs) and consider phrases with probability below this value as unusual. For each test article, we compute the number of unusual phrases (total for all categories) as a feature (SURP) and also this value normalized by total number of word tokens in the article (SURP_WD) and normalized by number of phrases (SURP_PH). We also compute features for individual pair types and in each case, the number of unusual phrases is normalized by the total words in the article (SURP_ADJ_NOUN, SURP_ADV_VERB, SURP_NOUN_NOUN, SURP_SUBJ_VERB).

A list of the top unusual words under the different pair types are shown in Table 1. These were computed on pairs from a random set of articles from our corpus. Several of the top pairs involve hyphenated words which are unusual by themselves, so we only show in the table the top words without hyphens.

ADJ-NOUN	**ADV-VERB**
hypoactive NNP	suburbs said
plasticky woman	integral was
psychogenic problems	collective do
yoplait television	physiologically do
subminimal level	amuck run
ehatchery investment	illegitimately put

NOUN-NOUN	**SUBJ-VERB**
specification today	blog said
auditory system	briefer said
pal programs	hr said
steganography programs	knucklehead said
wastewater system	lymphedema have
autism conference	permissions have

Table 1: Unusual word-pairs from different categories

[5]We noticed that in this genre several new words are created using hyphen to concatenate common words.

Most of these features are significantly different between the two classes. Those with higher values in the VERY GOOD set include: AVR_PHONEME_PERP_ALL, AVR_CHAR_PERP_(ALL, 10), SURP, SURP_PH, SURP_WD, SURP_ADJ_NOUN, SURP_NOUN_NOUN, SURP_SUBJ_VERB. The FREQ_NYT feature has higher value in the TYPICAL class.

All these trends indicate that unusual phrases are associated with the VERY GOOD category of articles.

3.4 Sub-genres

There are several sub-genres within science writing (Stocking, 2010): short descriptions of discoveries, longer explanatory articles, narratives, stories about scientists, reports on meetings, review articles and blog posts. Naturally, some of these sub-genres will be more appealing to readers. To investigate this aspect, we compute scores for some sub-genres of interest—narrative, attribution and interview.

Narrative texts typically have characters and events (Nakhimovsky, 1988), so we look for entities and past tense in the articles. We count the number of sentences where the first verb in surface order is in the past tense. Then among these sentences, we pick those which have either a personal pronoun or a proper noun before the target verb (again in surface order). The proportion of such sentences in the text is taken as the NARRATIVE score.

We also developed a measure to identify the degree to which the article's content is attributed to external sources as opposed to the author's own statements. Attribution to other sources is frequent in the news domain since many comments and opinions are not the views of the journalist (Semetko and Valkenburg, 2000). For science news, attribution becomes more important since the research findings were obtained by scientists and reported in a second-hand manner by the journalists. The ATTRIB score is the proportion of sentences in the article that have a quote symbol, or the words 'said' and 'says'.

We also compute a score to indicate if the article is the account of an interview. There are easy clues in NYT for this genre with paragraphs in the interview portion of the article beginning with either 'Q.' (question) or 'A.' (answer). We count the total number of 'Q.' and 'A.' prefixes combined and divide the value by the total number of sentences (INTERVIEW). When either the number of 'Q.' tags is zero or 'A.' tags is zero, the score is set to zero.

All three scores are significantly higher for the TYPICAL class.

3.5 Affective content

Some articles, for example those detailing research on health, crime, ethics, can provoke emotional reactions in readers as shown in the snippet below.

> Medicine is a constant trade-off, a struggle to cure the disease without killing the patient first. Chemotherapy, for example, involves purposely poisoning someone – but with the expectation that the short-term injury will be outweighed by the eventual benefits.

We compute affect-related features using three lexicons. The MPQA (Wilson et al., 2005) and General Inquirer (Stone et al., 1966) give lists of positive and negative sentiment words. The third resource is emotion-related words from FrameNet (Baker et al., 1998). The sizes of these lexicon are 8,221, 5,395, and 653 words respectively. We compute the counts of positive, negative, polar, and emotion words, each normalized by the total number of content words in the article (POS_PROP, NEG_PROP, POLAR_PROP, EMOT_PROP). We also include the proportion of emotion and polar words taken together (POLAR_EMOT_PROP) and the ratio between count of positive and negative words (POS_BY_NEG).

The features with higher values in the VERY GOOD class are NEG_PROP, POLAR_PROP, EMOT_POLAR_PROP. In TYPICAL articles, POS_BY_NEG, EMOT_PROP have higher values.

VERY GOOD articles have more sentiment words, mostly skewed towards negative sentiment.

3.6 Amount of research content

For a lay audience, a science writer presents only the most relevant findings and methods from a research study and interleaves research information with details about the relevance of the finding, people involved in the research and general information about the topic. As a result, the degree of explicit research descriptions in the articles varies considerably.

To test how this aspect is associated with quality, we count references to research methods and researchers in the article. We use the research dictionary that we introduced in Section 2 as the source of research-related words. We count the total num-

ber of words in the article that match the dictionary (RES_TOTAL) and also the number of unique matching words (RES_UNIQ). We also normalize these counts by the total words in the article and create features RES_TOTAL_PROP and RES_UNIQ_PROP.

All four features have significantly higher values in the VERY GOOD articles which indicate that great articles are also associated with a great amount of direct research content and explanations.

4 Classification accuracy

We trained classifiers using all the above features for for the two settings–'any-topic' and 'same-topic' introduced in Section 2.3. The baseline random accuracy in both cases is 50%. We use a SVM classifier with a radial basis kernel (R Development Core Team, 2011) and parameters were tuned using cross validation on the development data.

The best parameters were then used to classify the test set in a 10 fold cross-validation setting. We divide the test set into 10 parts, train on 9 parts and test on the held-out data. The average accuracies in the 10 experiments are 75.3% accuracy for the 'any-topic' setup, and 68% accuracy for the topic-paired 'same-topic' setup. These accuracies are considerable improvements over the baseline.

The 'same-topic' data contains article pairs with varying similarity. So we investigate the relationship between topic similarity and accuracy of prediction more closely for this setting. We divide the article pairs into bins based on the similarity value. We compute the 10-fold cross validation predictions using the different feature classes above and collect the predicted values across all the folds. Then we compute accuracy of examples within each bin. These results are plotted in Figure 1. *int-science* refers to the full set of features and the results from the six feature classes are also indicated.

As the similarity increases, the prediction task becomes harder. The combination of all features gives 66% accuracy for pairs above 0.4 similarity and 74% when the similarity is less than 0.15. Most individual feature classes also show a similar trend. This result is understandable because articles on similar topics could exhibit similar properties. For example, two articles about 'controversies surrounding vaccination' are likely to have similar levels of people-oriented nature or written in a narrative style

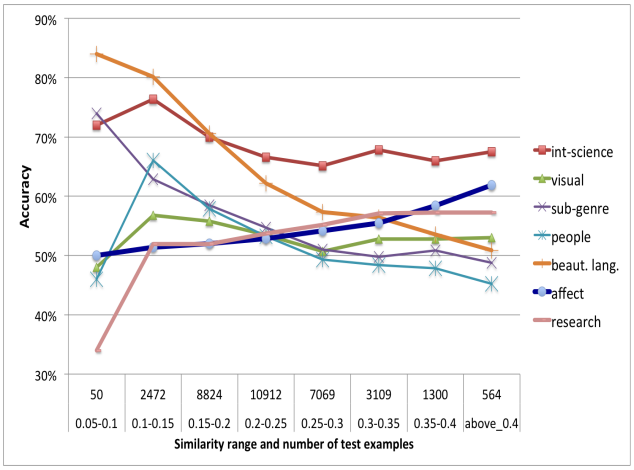

Figure 1: Accuracy on pairs with different similarity

in the same way as two space-related articles are both likely to contain high visual content. There are however two exceptions—affect and research. For these features, the accuracies improve with higher similarity; affect features give 51% for pairs with similarity 0.1 and 62% for pairs above 0.4 similarity, accuracy of research features goes from 52% to 57% for the same similarity values. This finding illustrates that even articles on very similar topics can be written differently, with the articles by the excellent authors associated with greater degree of sentiment, and deeper study of the research problem.

5 Combining aspects of article quality

We now compare and combine the genre-specific *interest-science* features (41 total) with those discussed in work on readability, well-written nature, interest and topic classification.

Readability (16 features): We test the full set of readability features studied in Pitler and Nenkova (2008), involving token-type ratio, word and sentence length, language model features, cohesion scores and syntactic estimates of complexity.

Well-written nature (23 features): For well-written nature, we use two classes of features, both related to discourse. One is the probabilities of different types of entity transitions from the Entity Grid model (Barzilay and Lapata, 2008) which we compute with the Brown Coherence Toolkit (Elsner et al., 2007). The other class of features are those defined in Pitler and Nenkova (2008) for likelihoods and counts of explicit discourse relations. We identified the relations for texts in our corpus using the

AddDiscourse tool (Pitler and Nenkova, 2009).

Interesting fiction (22 features): are those introduced by McIntyre and Lapata (2009) for predicting interest ratings on fairy tales. They include counts of syntactic items and relations, and token categories from the MRC psycholinguistic database. We normalize each feature by the total words in the article.

Content: features are based on the words present in the articles. Word features are standard in content-based recommendation systems (Pazzani et al., 1996; Mooney and Roy, 2000) where they are used to pick out articles similar to those which a user has already read. In our work the features are the most frequent *n* words in our corpus after removing the 50 most frequent ones. The word's count in the article is the feature value. Note that word features indicate topic as well as other content in the article such as sentiment and research. A random sample of the word features for *n = 1000* is shown below and reflects this aspect. "matter, series, wear, nation, account, surgery, high, receive, remember, support, worry, enough, office, prevent, biggest, customer".

Table 2 compares the accuracies of SVM classifiers trained on features from different classes and their combinations.[6] The readability, well-written nature and interesting fiction classes provide good accuracies 60% and above. The genre-specific *interesting-science* features are individually much stronger than these classes. Different *writing aspects* (without content) are clearly complementary and when combined give 76% to 79% accuracy for the 'any-topic' task and 74% for the topic pairs task.

The simple bag of words features work remarkably well giving 80% accuracy in both settings. As mentioned before these word features are a mix of topic indicators as well as other content of the articles, ie., they also implicitly indicate animacy, research or sentiment. But the high accuracy of word features above all the writing categories indicates that topic plays an important role in article quality. However, despite the high accuracy, word features are not easily interpretable in different classes related to writing as we have done with other writing features. Further, the total set of writing features is only 102 in contrast to 1000 word features. In our interest-science feature set, we aimed to highlight how well some of the aspects considered important to good science writing can predict quality ratings.

[6]For classifiers involving content features, we did not tune the SVM parameters because of the small size of development data compared to number of features. Default SVM settings were used instead.

Feature set	**Any Topic**	**Same**
Interesting-science	75.3	68.0
Readable	65.5	63.0
Well-written	59.1	59.9
Interesting-fiction	67.9	62.8
Readable + well-writ	64.7	64.3
Readable + well-writ + Int-fict	71.0	70.3
Readable + well-writ + Int-sci	79.5	73.2
All writing aspects	76.7	74.7
Content (500 words)	81.7	79.4
Content (1000 words)	81.2	82.1
Combination: Writing (all) + Content (1000w)		
In feature vector	82.6*	84.0*
Sum of confidence scores	81.6	84.9
Oracle	87.6	93.8

Table 2: Accuracy of different article quality aspects

We also combined writing and word features to mix topic with writing related predictors. We do the combination in three ways a) word and writing features are included together in the feature vector; b) two separate classifiers are trained (one using word features and the other using writing ones) and the sum of confidence measures is used to decide on the final class; c) an oracle method: two classifiers are trained just as in option (b) but when they disagree on the class, we pick the correct label. The oracle method gives a simple upper bound on the accuracy obtainable by combination. These values are 87% for 'any-topic' and a higher 93.8% for 'same-topic'. The automatic methods, both feature vector combination and classifier combination also give better accuracies than only the word or writing features. The accuracies for the folds from 10 fold cross validation in the feature vector combination method were also found to be significantly higher than those from word features only (using a paired Wilcoxon signed-rank test). Therefore both topic and writing features are clearly useful for identifying great articles.

6 Conclusion

Our work is a step towards measuring overall article quality by showing the complementary benefits of general and domain-specific writing measures as well as indicators of article topic. In future we plan to focus on development of more features as well as better methods for combining different measures.

References

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL*, pages 86–90.

Z. Bao, B. Kimelfeld, and Y. Li. 2011. A graph approach to spelling correction in domain-centric search. In *Proceedings of ACL-HLT*, pages 905–914.

R. Barzilay and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

R. Barzilay, N. Elhadad, and K. McKeown. 2002. Inferring strategies for sentence ordering in multidocument summar ization. *Journal of Artificial Intelligence Research*, 17:35–55.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

D. Blum, M. Knudson, and R. M. Henig, editors. 2006. *A field guide for science writers: the official guide of the national association of science writers*. Oxford University Press, New York.

E. Brill and R.C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of ACL*, pages 286–293.

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, pages 310–318.

S. Cucerzan and E. Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of EMNLP*, pages 293–300.

R. Dale and A. Kilgarriff. 2010. Helping our own: text massaging for computational linguistics as a new shared task. In *Proceedings of INLG*, pages 263–267.

M. C. De Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

P. Diederich. 1974. *Measuring Growth in English*. National Council of Teachers of English.

M. Elsner, J. Austerweil, and E. Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of NAACL-HLT*, pages 436–443.

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.

R. Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221 – 233.

A.C. Graesser, D.S. McNamara, M.M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers*, 36(2):193–202.

H. Ji and D. Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person name detection. In *Proceedings of PACLIC*.

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

S.M. Kosslyn. 1980. *Image and mind*. Harvard University Press.

M. Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*, pages 545–552.

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.

D. Lin, K. W. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, K. Dalwani, and S. Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of LREC*.

N. McIntyre and M. Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*, pages 217–225.

R. J. Mooney and L. Roy. 2000. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204.

A. Nakhimovsky. 1988. Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43, June.

M. Pazzani, J. Muramatsu, and D. Billsus. 1996. Syskill & Webert: Identifying interesting web sites. In *Proceedings of AAAI*, pages 54–61.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, pages 186–195.

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

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

D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*, pages 248–256.

A. Rozovskaya and D. Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of EMNLP*, pages 961–970.

E. Sandhaus. 2008. The new york times annotated corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*.

S. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*, pages 523–530.

H.A. Semetko and P.M. Valkenburg. 2000. Framing european politics: A content analysis of press and television news. *Journal of communication*, 50(2):93–109.

V. Spandel. 2004. *Creating Writers Through 6-Trait Writing: Assessment and Instruction*. Allyn and Bacon, Inc.

S. H. Stocking. 2010. *The New York Times Reader: Science and Technology*. CQ Press, Washington DC.

P. J. Stone, J. Kirsh, and Cambridge Computer Associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

J. R. Tetreault and M. Chodorow. 2008. The ups and downs of preposition error detection in esl writing. In *Proceedings of COLING*, pages 865–872.

L. von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of CHI*, pages 319–326.

R. L. Weide. 1998. The cmu pronunciation dictionary, release 0.6. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354.

M. Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, 20(1):6–10.