What makes Writing Great? First experiments on Article Quality in the Science Journalism Domain

Annie Louis
University of Edinburgh

Ani Nenkova
University of Pennsylvania

A new task

- A lot of work on identifying what is wrong with a text
  - Spelling mistakes, grammar errors, incoherent writing

- It is not known how to characterize writing that is engaging, interesting and nice
  - Some work on predicting articles on a topic of interest for a user
  - Predict interest value of short fairy tales (McIntyre and Lapata, 09)

This work

- A corpus of science journalism for text quality
  - Suitable for identifying beautiful and interesting writing
- Genre-specific measures for identifying great writing
  - Features for visual language, surprisal, animacy, affect and text structure
  - Accuracies much above baseline
- Complementary nature of aspects of writing
  - Indicators for beautiful writing add to the strengths of readability and coherence measures to predict article quality

Science journalism genre

- informative and entertaining at the same time

“Cell phone tracking study shows we are creatures of habit

News flash: We are boring.

New research .. (based on cell phone tracking)... suggests that most people can be found in one of just a few locations at any time. “

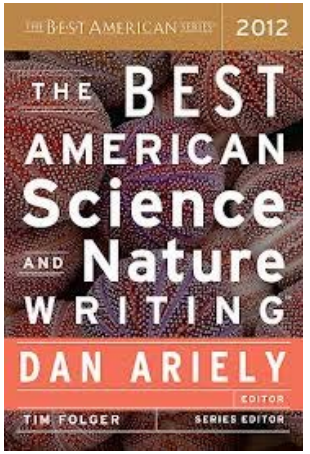
Source: New York Times

>> Corpus of article quality

Category 1 : VERY GOOD articles

- Seed set = 63 New York Times articles that appeared in the Best American Science Writing series

- We choose only the NYT articles
  - We use the NYT Corpus to expand our category
  - Normalize for differences in writing due to source



Expanding the VERY GOOD set

- Assume: ~40 authors of the seed set are excellent writers

- Other articles from the NYT written by the same authors
  - which are research related
  - during the same 10 year period
  - on similar topics
  - similar lengths

Category 2: TYPICAL writing in the NYT

- Other science articles around the same time, but not written by the popular authors

The general corpus:

Category	Total Articles
VERY GOOD	3,530
TYPICAL	20,242

A topic-paired corpus

- The general categories we have mix different topics
  - geography, biology, astronomy, linguistics…
- But an IR system should compare articles on the same topic

- For each VERY GOOD article, get 10 most similar TYPICAL articles (based on the content)
- Enumerate all pairs of (VERY GOOD, TYPICAL)

- 35,300 pairs

Two quality prediction tasks

2 categories
GOOD (~3500)
TYPICAL (~3500)

→ `Any-topic'

– is this article VERY GOOD or TYPICAL?

Topically similar pairs
<VERY GOOD, TYPICAL>
~35,000 pairs

→ `Same-topic'

– which article in the pair is the VERY GOOD one?

Properties of the dataset

- Allow to focus on aspects such as beautiful writing
  - Less likely to have spelling and grammar errors
- Realistic sample of writing differences
  - Previous work often used machine generated text or artificially manipulated text
- Large scale

>> How to predict interesting writing?

Approach to feature development

- Features related to six aspects of writing in science news
  - Surprising words and phrases, visual language, text structure, people-oriented nature, affect and explicit research content
- Focus on interpretable features
  - Each feature is a composite one: indicates an aspect directly
  - As a result, the number of features is rather small -- 41
  - Linguistically interesting
- Confirming that features represent the intended aspect
  - Tune by checking feature values on random snippets of text

1. Unusual words and phrases

- Word-based
  - low frequency
  - high perplexity under a phoneme n-gram model
  - high perplexity under a letter n-gram model
  - Eg: ‘undersheriff’, ‘powwow’, ‘chihuahua’, ‘qipao’

- Word pairs—based
  - adjective-noun, noun-noun, adverb-verb, subject-verb pairs
  - perplexity under a language model
  - language model uses interpolation smoothing

$$p(adjective|noun) = \lambda p(adjective|noun) + (1-\lambda) p(adjective)$$

  - Eg: ‘plasticky woman’, ‘cavernous soundstage’, ‘so-called superkids’

Features = average perplexity of words, count of unusual phrases ...

2. Visual nature

- Existing resources are small
  - MRC database = 3,400 words manually annotated for visual nature
  - Actual number of visual terms even smaller

- Creating a larger lexicon of visual terms
  - Source: an image-tagged corpus
  - Large number of potentially visual words, but quite noisy

Visual nature

- Filtering approach
  - Create LDA-based topics on the tag set
  - Use the manually annotated MRC terms to filter out non-visual topics
  - Resulting in visual terms as well as categories of visual terms

grass, mountain, green, hill, blue, field, sand...
round, ball, circles, logo, dots, square, sphere...
silver, white, diamond, gold, necklace, chain...

- Features for an article
  - total visual terms, position of visual terms (beginning, middle, end), ...
  - number of topics making up the visual terms

Human interest and text structure

3. Use of people in the story
   - whether the story revolves around a person or persons
   - animacy information from NEs, pronouns, ngram patterns
   - Features: no. of animate and inanimate words…

4. Sub-genre
   - whether the article is a narrative, interview or dialog
   - Eg: narrative score ~ past tense verbs, pronouns, proper names
   - Features: one score each for narrative, interview and dialog

Sentiment and Research

5. Affect
   - whether there is an emotional angle to the story
   - using sentiment word dictionaries
   - Features: positive, negative words, positive to negative ratio

6. Research content
   - how much explicit research description is present
   - using a hand-built dictionary of research words
   - Features: number of total research words…

>> Strengths in predicting article quality

How the features vary in a random sample of very good and typical articles (t-test)

Higher values in VERY GOOD set

- ✓ Visual words in beginning and end of articles
- ✓ Unusual words and phrases
- ✓ Sentiment words, negative polarity
- ✓ Research words

- ✗ Total visual words
- ✗ Animacy counts
- ✗ Narrative, interview or dialog format

Accuracies on the two tasks

Any Topic: Given an article, is it "VERY GOOD" or "TYPICAL" ?

System	Accuracy
Baseline (random)	50%
Interesting-science features	75%

- 10 fold cross validation results
- SVM classifier

Same Topic: Given a pair of articles on the same topic, which one is "VERY GOOD"?

System	Accuracy
Baseline (random)	50%
Interesting-science features	68%

Combining reader interest with other aspects

Feature set	any topic	same topic
Interesting science	**75.3**	**68.0**

Genre-specific measures are stronger than generic ones

Different aspects of writing have complementary strengths

>> Comparing with a bag of words system

- Generally work well for text classification
- But not easily interpretable

- Standard in recommendation systems for indicating topic
- How much influence of topic on text quality categories?

A word based classifier

- Features = most frequent 1000 words in the corpus after stop word filtering
  - E.g. “matter”, “series”, “customer”, “worry”, “surgery”

Accuracy of the Bag of Words system

Feature set	any topic	same topic
1000 words	81.2	82.1

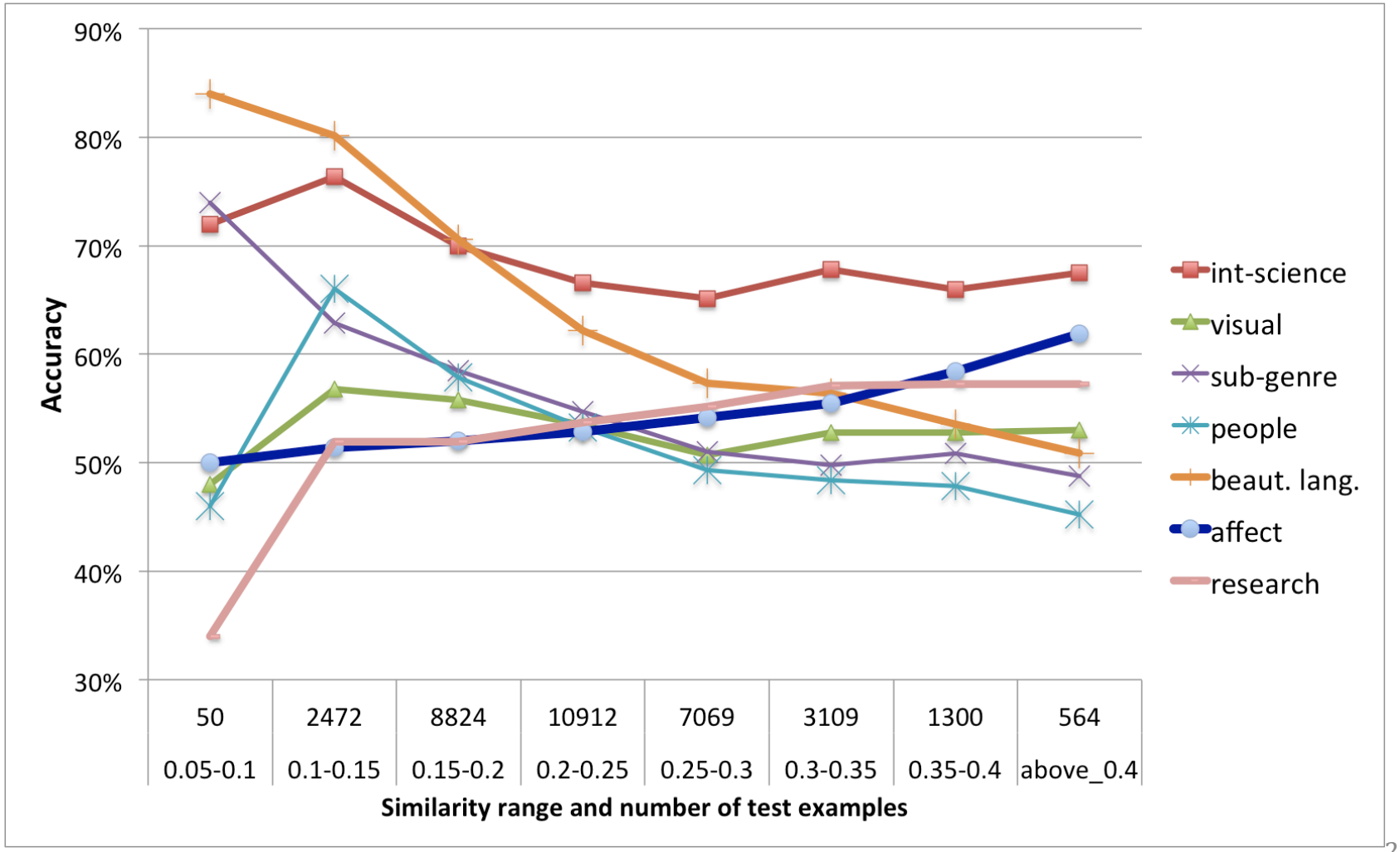
Conclusions

- New task settings: identifying beautifully written text and genre-specific text quality
  - Based on a corpus obtained using simple heuristics
- A first work showing how different aspects of writing are complementary for predicting overall article quality
- Future work
  - Approaches for combining different aspects of quality
  - Measuring and incorporating the role of topic

Thank you !

(Corpus is also available from our website)

Some features more useful for differentiating articles on the same topic: `same-topic' task



Topics in the seed set

Distribution of topics in the seed set	
Medicine and Health	22
Space	14
Physics	10
Biology and Biochemistry	8
Genetics and Heredity	8
Archaeology and Anthropology	7
Reproduction (Biological)	7
Animals	5
Diseases and Conditions	5
Ethics	5
Finances	5
Women	5
Computers and the Internet	4
Doctors	4

Corpus summary

Category	No. articles	No. sentences	No. tokens
GREAT	63	7,212	177,775
VERY GOOD	4,190	232,824	1,453,570
TYPICAL	19,520	1,213,534	36,254,539