Information density, Heaps' Law, and perception of factiness in news (REVISED 06-23-2014)

Miriam Boon
Technology and Social Behavior, Northwestern University
Evanston, IL 60208, USA
MiriamBoon2012@u.northwestern.edu

Abstract

Seeking information online can be an exercise in time wasted wading through repetitive, verbose text with little actual content. Some documents are more densely populated with factoids (fact-like claims) than others. The densest documents are potentially the most efficient use of time, likely to include the most information. Thus some measure of "factiness" might be useful to readers. Based on crowdsourced ratings of the factual content of 806 online news articles, we find that after controlling for widely varying document length using Heaps' Law, a significant positive correlation exists between perceived factual content and relative information entropy.

1 Introduction

In today's information-based society, finding accurate information is of concern to everyone. There are many obstacles to this goal. Not all people are equally skilled at judging the veracity of a factoid (a term used here to indicate something that is stated as a fact, but that may or may not actually be true.). Nor is it always easy to find the single drop of content you need amidst the oceans of the Internet. Even for those equipped with both skill and access, time is always a limiting factor.

It is this last problem with which this paper is concerned. How can we identify content that most efficiently conveys the most information, given that any information seeker's time is limited?

1.1 The difficulty with factoids

Imagine that we must select from a set of documents those that efficiently convey the most information in the fewest words possible; that is, those with the highest factoid rate, $count(factoids)/count(words)$. A human doing this by hand would count the factoids and words in each document. Automating this exact approach would require 'teaching' the computer to identify unique factoids in a document, which requires being able to recognize and discard redundant factoids, which requires at least a rudimentary understanding of each factoid's meaning. These are all difficult tasks for a computer.

Luckily, to achieve our goal, we don't need to know which sentences are factoids. What we need is a good heuristic estimate of information density that computers can easily calculate.

1.2 Linking vocabulary to factoids

To insert new information into a text, an author must add words, making the document longer. While the new information can sometimes be conveyed using the same vocabulary as the rest of the text, if the information is sufficiently different from what is already present, it will also likely introduce new vocabulary words.

The result is that the introduction of a new factoid into a text is likely to also introduce new vocabulary, unless it is redundant. Thus, the more non-redundant factoids a text contains, the more varied the vocabulary of the text is likely to be.

1.3 From vocabulary to relative entropy

Vocabulary is commonly used in connection with Shannon's information entropy to measure such things as surprisal, redundancy, perplexity, and, of course, information density (Shannon, 1949; McFarlane et al., 2009).

Entropy models text as being created via a Markov process. In its most basic form, it can be written as:

$$H = -K \sum_{i=0}^{L} p_i log_2 p_i \qquad (1)$$

where K is a constant chosen based on the units, L is the length of the document, and p_i is the probability of the i^{th} word. This equation works

equally well whether it is used for unigrams, bigrams, or trigrams.

Consider for a moment the relationship between entropy and length, vocabulary, and the probability of each vocabulary word. Entropy increases as both document length and vocabulary increase. Words with lower probability increase entropy less than those with higher probabilities, and entropy for a document of given length is maximized when all words have equal probability summing to one. For this study, probabilities were calculated based on corpus-wide frequencies, and then normalized such that the sum of these probabilities would equal one for each document.

Given two documents of equal length on the same topic, only one of which is rich in information, we might wonder why the information-poor document is, relatively speaking, so long or the information-rich document is so short. This can be explained by noting that: 1. "translation" into simpler versions of a language often leads to a longer text, 2. simple versions of languages generally consist of the most common words in that language, and 3. words that are less common often have more specific, information-dense, complex meanings. Similarly, inefficient word choices typically make excessive use of highly probable function words, which increase the entropy more than less common words. Thus, we can expect the entropy to be lower for the denser document.

1.4 Controlling for document length with Heaps' Law

While entropy may not rise as fast with the repetition or addition of low probability words, every word added does still increase the entropy. We can try to compensate by dividing by document length. But dividing by document length doesn't remove this dependency. I propose that this is because, as Heap's Law tells us, the vocabulary used in a document has a positive relationship with document size (Heaps, 1978). To control for this effect, I fit a curve for unigrams, bigrams, and trigrams to create a model for these relationships; an example can be seen in Figure 1.

I then used that model to calculate the expected document length, expected entropy, and residual entropy, as follows:

$$L_{exp} = 10^{(log_{10}v - b)/m} \quad (2)$$

$$H_{exp} = H\frac{L_{exp}}{L} \quad (3)$$

$$H_{res} = H - H_{exp} \quad (4)$$

Here the subscript 'exp' stands for 'expected' and the subscript 'res' for 'residual.' This calculation eliminates the dependency on document length for bigram and trigram-based entropy, and decreases it from an R^2 of 0.99 to only 0.011 for unigram-based entropy.

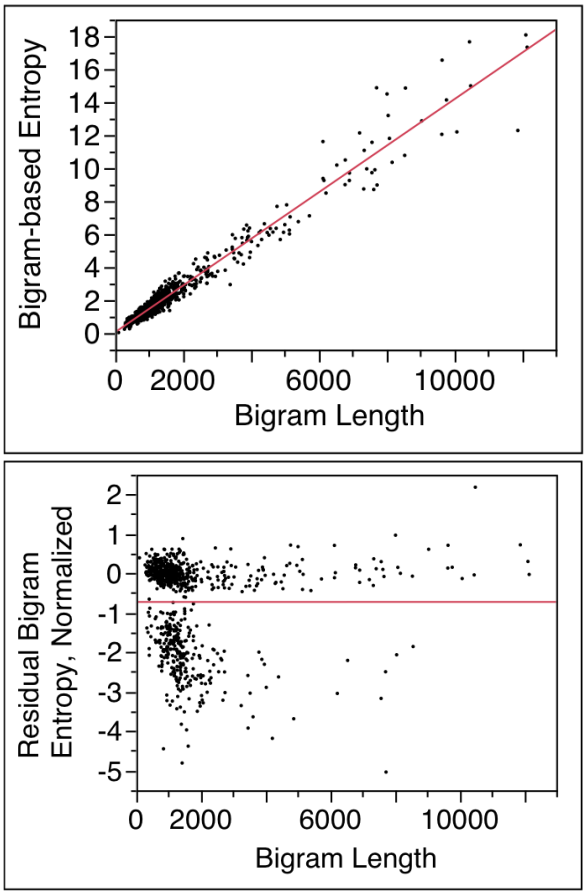

Figure 1: Top: As you can see there is a strong relationship between document length and entropy. R^2=0.957, p > F: < 0.0001. Bottom: Residual entropy, which controls for that relationship, no longer has a significant nor strong relationship with document length. R^2=5.32 * 10^{-9}, p > F: 0.998

2 Data and Analysis

To further pursue the hypothesis that residual entropy could be used to identify news articles with lots of factoids, and thus, a sense of 'factiness,' a labeled data set is necessary. Lots of websites allow users to rate articles, but those ratings don't have anything to do with the presence of factoids. Labeling a data set of adequate size by hand would be tedious, time-consuming, and costly.

Figure 2: Mousing over the question makes the text "Is it based on facts or opinions?" appear in pale grey text. Clicking on the question mark icon next to the question, "Is this story factual?" reveals an explanation of what the user should be rating.

2.1 Crowdsourcing with NewsTrust

Fortunately, a project called NewsTrust provided a feasible alternative. NewsTrust, founded in 2005 by Fabrice Florin, created four sets of specific review questions designed to inspire reviewers to think critically about the quality of articles they review. NewsTrust partnered with independent academic researchers Cliff Lampe and Kelly Garrett to validate the review questions. They jointly administered a survey in which respondents were asked to complete one of the review instruments regarding either the original version of an article or blog post, or a degraded version.

The independent study found that even the less experienced, less knowledgeable readers were able to distinguish between the two versions of the story. The shortest review instrument, with only one question, had the most discriminating power, while the slightly longer normative review instrument (which added five more questions) yielded responses from non-experts that most closely matched those of NewsTrust's expert journalists (Lampe and Garrett, 2007; Florin et al., 2006; Florin, 2009).

Using their validated survey instrument, NewsTrust created a system that allowed users to read articles elsewhere, rate them using one of the four review instruments, and even rate other NewsTrust users' reviews of articles. Each user has a trustworthiness rating (which can be bolstered by becoming validated as a journalist expert), and each article has a composite rating, a certainty level for that rating, reviews, and ratings of reviews.

One of the dimensions of journalistic quality for which NewsTrust users rate articles is called 'facts'. This can be taken as an aspect of "factiness": the extent to which people perceived the article as truthful and factual. It follows that, to the extent that the users are making a good-faith attempt to rate articles based on facts regardless of the soundness of their judgment about what is or is not true, articles with a high rating for 'facts' should have more factoids, and therefore a higher density of information.

2.2 Data acquisition

When this research project was launched, NewsTrust had recently been acquired by the Poynter Institute. Although they were open to making their data available for research purposes, they were not yet able to access the data in order to do so. Instead, the review data for over 11000 stories from NewsTrust's political section were retrieved using Python, Requests, and Beautiful Soup. A combination of Alchemy API, digital library archives, and custom scrapers for 19 different publication websites were used to harvest the corresponding article texts.

It quickly became clear, however, that it would not be possible to completely capture all 11,000 articles. Some of the independent blogs and websites no longer existed. Others had changed their link structure, making it difficult to find the correct article. A great deal of content was behind paywalls, or simply did not have a webpage structure that lent itself to clean extraction. As the text would be used for automated analysis, it was essential that the extracted text be as clean of detritus as possible. As a result, the dataset shrank from a potential 11,000 rated articles to only 3300 for which I could be confident of having clean text. Approximately 2600 of those articles have been rated by at least one NewsTrust user based on factiness, and after removing any with fewer than four facts ratings, the data set shrank further to only 806 articles. Unigrams, bigrams, and trigrams were extracted from these articles using the Natural Language Toolkit, NLTK; all text was lowercased, and only alphanumeric words were included.

2.3 Analysis

The relationship between length and vocabulary was modeled using the optimize toolkit from SciPy, and visualized with MatPlotLib. The resulting relationship was used to calculate the residual, normalized entropy for each document.

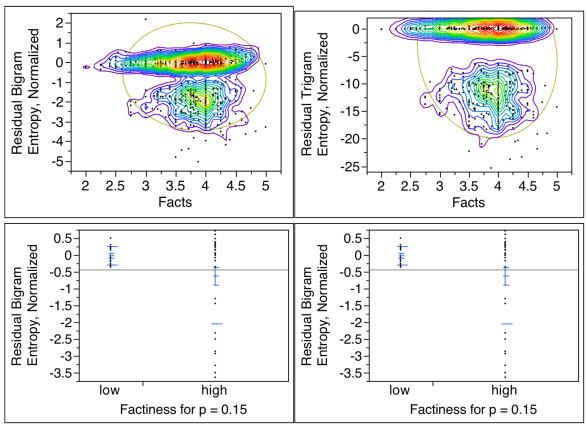

Figure 3: Top: Bivariate fit for bigrams (corr = -0.06, p = 0.096) and trigrams (corr = -0.0975, p = 0.0056). Colors are quantile density contours. Bottom: Oneway ANOVA for bigrams (diff = -0.60, p = 0.033) and trigrams (diff = -4.3, p = 0.0024).

In order to conduct oneway analysis, documents needed to be separated into two distinct clusters. I used Weka's K-Means clustering algorithm to find the mean of three clusters; the lowest and highest means served as the estimates for the means of my "low fact" and "high fact" categories. Membership was determined by using a T-test with an 85% confidence interval; points that could come from both or neither category were discarded (43 remaining cases). This process was repeated for 80 and 70% confidence levels; they yielded more data points (51 and 97), a higher level of significance, and similar effect sizes. A 90% confidence level did not yield enough articles to analyze properly (n = 28).

3 Results and Discussion

As you can see in Figure 3, the ANOVA found that the residual normalized entropy was significantly lower for the high versus low factiness classification for bigrams, and the effect strengthened for trigrams. The bivariate analysis shows something more interesting happening. With bigrams, there is a cluster of documents that are almost separated from the main body of documents, having lower entropy and higher factiness. This transition is completed for the trigrams, resulting in a completely distinct cluster.

This cluster has a higher factiness rating (diff = 0.088, p = 0.0079), and shows no significant relationship with any of the other dimensions measured by NewsTrust review instruments. Furthermore, none of the cluster members are in the low-factiness category (this is true when constructing the categories with p-values ranging from 0.05 to 0.3).

The migrating cluster pictured above is not quite in keeping with expectations given the hypothesis of less probable words being more dense in meaning. If the original hypothesis were correct, this relationship would hold for unigrams as well. Yet it does not. Future work must examine this cluster to better understand the mechanism that is leading reviewers to rate these stories as high in factiness.

If we accept the assumption that the articles rated by NewsTrust users as highly factual will contain a higher density of factoids, then this result supports the hypothesis that residual entropy is positively correlated with that characteristic. Conversely, if we accept the assumption that entropy should be correlated with factoid density, then this result supports the claim that NewsTrust users effectively identify articles that are more information dense.

Other work on the fact-rated sub-corpus has two obvious directions. First, and most closely related to the work described in this paper, is the goal of proving either assumption in a more controlled experiment. If one of these assumptions can be supported, then it strengthens the claim about the other, which will be interesting from both a linguistic perspective, and a human-computer interaction perspective. The other avenue of inquiry that follows naturally from this work is to look for other textual features that might, in combination, enable the automatic prediction of fact ratings based on article text.

Acknowledgments

This work was partly supported by the Technology and Social Behavior program at Northwestern University, the National Science Foundation (grant numbers IIS-0856058 and IIS-0917261/001), the Knight Foundation, and Google. Many thanks to Dr. Darren Gergle for his insight on the larger NewsTrust data set, to Dr. Janet Pierrehumbert for her guidance on entropy and factiness, and to Dr. Larry Birnbaum for his intellectual guidance as well as his assistance on this paper.

References

Fabrice Florin, Cliff Lampe, and Kelly Garrett. 2006. Survey report summary - NewsTrust.

Fabrice Florin. 2009. NewsTrust communications 2009 report. Technical report.

Harold Stanley Heaps. 1978. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc.

Cliff Lampe and R. Kelly Garrett. 2007. It's all news to me: The effect of instruments on ratings provision. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, page 180b180b.

Delano J. McFarlane, Noemie Elhadad, and Rita Kukafka. 2009. Perplexity analysis of obesity news coverage. *AMIA Annual Symposium Proceedings*, 2009:426–430. 00001.

Claude E. Shannon. 1949. *The mathematical theory of communication.* Urbana, University of Illinois Press.