ASSESSING THE QUALITY OF ONLINE NEWS ARTICLES AS REFERENCES FOR AN ENCYCLOPAEDIA ENTRY

Filipo Sharevski[1]

[1]Purdue Polytechnic Institute, Purdue University, West Lafayette, USA

ABSTRACT

The quality of online news articles is decisive both for a reliable perception of their informativeness and for including them as a reference when creating an encyclopaedia entry for a public attention event. Tackling the enormous volume, variety and complexity of different articles disseminated online, several natural language processing techniques have been developed for the purpose of capturing the quality of the web content based on the concepts of objectivity classification and stylometric features, knowledge maturing, factual density, or simple word count. This paper explores utilizes the factual density as a quality measure of the information reported on the missing Malaysia Airliners Flight 370 as a public attention event in two instances, considering the coverage during the initial investigation and the reporting marking the one year anniversary since the plane got missing. The results suggest that the factual density can be utilized in creating a high-quality encyclopaedia entry, however, under strict conditions in terms of increased confidence level of the automated factual extraction.

KEYWORDS

Factual Density, Natural Language Processing, Text Informativeness

1. INTRODUCTION

In the effort to create a factual encyclopaedia entry for an event of public attention, the online news selection must not rely not just on the plausibility of the content itself, but should also take "credibility and reliability of the source and the quality aspects characteristic for this content" should also be taken in consideration[1]. In this context, of practical interest is to assess the quality of online media coverage of such an instance, together with the reliability of the respective information as a reference. However, the quality assessment of the online media coverage in a whole is "too general and practically impossible", given the text, video, audio, picture elements included in the associated articles[2]. Therefore, it is useful to restrict the online news quality assessment only to the textual portion of the news' web feeds, since the text is the skeleton bearing the critical information about given event, supported by the additional set of pictorials, graphics or videos. As such, the focus of the undertaken effort rests on the idea that "the quality of a text document can be related to its informativeness, i.e. the amount of useful information contained in a document"[2]. The "informativeness of a text can be measured through factual density of the source, i.e. the number of facts contained in a given document, normalized by its length". The factual density is useful in that it assess the influential aspects of the textual portion in more general terms compared to the objectivity classification, and that the online news reception precedes the transformation process of "contextualized information artefacts into explicitly linked and formalized learning objects"[3]. Moreover, it outperforms the word count measure as a web content quality measurement[4].

DOI : 10.5121/ijaia.2015.6407

With this in place, the online news' informativeness assessment is directed towards the estimation of their appropriateness in using as a reference when creating an encyclopaedia entry for an public attention event. Given the unpreceded nature of the incident involving the Malaysia Airlines Flight 370, which despite "the largest-ever multinational air-sea search" [5] is still missing after a year, the event and its associated online reporting are taken at the centre of the web news' quality analysis. For this purpose, two sets of online news reports are compiled: One containing the coverages from two different sources – Fox News and Thompson-Reuters – during the initial incident, and another one containing the articles marking the anniversary from the incident after one year. In the first case, each subset includes a daily review article within the period of 33 days after the aircraft disappeared from the regular route, elaborating on the investigation progress by bringing additional facts and potential evidence leads. The factual density of each of the respective articles – extracted using the ReVerb tool [6] – is compared relative to the overall ranking of each of the articles in two instances: one where the threshold for confidence scores of the automatic factual determination is not taken into consideration, and another where the confidence threshold is considered in a context of achieving an optimal balance between the recall and the precision of the extracted factual relations. On the basis of this information, a second subset of ten articles summarizing the incident from a yearlong perspective are taken in bringing the cumulative facts corpora relating to the incident. The outcome of the preliminary comparisons and the final assessment is discussed respective to the reference article selection criteria based on the factual density as a measurement instrument.

2. FACTUAL DENSITY AS AN INFORMATIVENESS MEASURE

The factual density $f_d(t)$ of an arbitrary textual resource t is the ratio between the fact count $f_c(t)$ and the size of the text size(t). The most applicative approach for operationalization of the factual density is the utilization of the "Open IE methods, justified by the fact that the extraction output brings the "relational tuples representing facts from text [subject of the extraction] without requiring a pre-specified vocabulary or manually tagged training corpora"[2], which in turn "makes such systems scalable to a a far larger corpora as the Web"[7]. Assuming a model of textual resource "as a bag-of-relations occurring in all facts extracted from it", the relational tuple as a unit output is structured as a "triplet consisting of two arguments in the form of noun phrases and a relational phrase expressed as a verb phrase"[4], i.e. f = (*10 nations*, *failed to turn up*, *any trace of the Boeing 777 – 200R*). Having the unit dividend in the factual density formula, the actual value of the fact count can be derived by simple summation operation of all relational tuples produced as an outcome of the Open IE system.

Since the basic definition brings no explicit details on the actual granularity of the size variable, there are three possibilities for text quantization: the overall character count including the white spaces, the word count, and the number of the sentences in the text. The most appropriate text quantization considers the word count as the text size, given the fact that it mainly strikes balance between potential miscalculation of the factual density due to an excessive normalization when the character count is used as a quantifying unit, or potential factual density degradation due to a considerably coarse quantization step in case of a sentence count. As another countermeasure to the threads of factual density degradation, [6] introduced two syntactic constraints: (1) "every multi-word relation phrase must begin with a verb, end with a preposition, and be a contiguous sequence of words in the sentence"; and (2) "a binary relation phrase ought to appear with at least a minimal number of distinct argument pairs in a large corpus", which enables factual derivation "without requiring a pre-specified vocabulary or manually tagged training corpora" from a given text resource. Incorporating these improvements intended to minimize the overestimation of the factual density, and outperforming the other Open IE systems[9], [6] by achieving considerably higher precision and recall"[2][6][10] , the ReVerb extractor it is chosen as tool for the automatic factual density derivation of the related online news articles.

3. RELATED WORK

The factual density as a quality discriminator of web contents was initially employed by [2] and [4] in the "performance comparisons against the word count in identification of featured/good articles in Wikipedia". The outcome of the analysis yielded that the factual density "corroborates the good performance of word count in Wikipedia in case the article are with variable length", but in the case of articles of similar lengths "the word count measure fails while factual density can separate between them with an F-measure of 90.4%". Trying to assess whether the evidence holds beyond the specific class of online encyclopaedia content, [2]"extend the analytical effort into the sphere of the arbitrary Internet texts" retrieved from various, randomly selected online contents by comparing a human-ranked reference corpus against an analogous raking derived from an Open IE factual density estimation system, consisting of 50 arbitrary web text on the material in Spanish. Correlating the "ground truth human-annotated ranking with the ranking gathered from the automatic prediction", [2] were able to confirm the adequacy of the factual density as a measurement of the informativeness, at least in the case of Spanish web texts. In contrast to these efforts, this work approaches the factual density as to assess to a what degree is feasible to utilize it as an instrument for a quality determination of an online news articles in selecting them as a reference for a public event encyclopaedia entry. The news reporting on the missing Malaysia Airlines Flight 370 are taken as to determine whether an automatic evaluation of the informativeness of different online news feeds brings plausible and unified referencing information that can be utilized for this purpose.

2. EXPERIMENTAL SETUP AND RESULTS

2.1. Investigation Progress Reports

Two separate datasets corresponding to the daily reporting on the investigation and search progress were compiled from the Fox News and Thompson-Reuters. The Fox News feed is taken as referent since it was continuously ranked 1st in the news ranking for more than 147 months, including March, 2014 [11], which corresponds with more than 72 % of the time period of online news coverage of the incident that was initially reported on March 8, 2014. The addition of Thompson-Reuters reflects the idea of balancing the potential thread of polarized online coverage[12]. Each of the datasets includes a daily review article about the Malaysia Airlines Flight 370 in a period of 33 days (between March 8 to April 9, 2014) after the aircraft disappeared, elaborating on the progress of the multinational air-sea search. Special attention was devoted to collect those daily articles that are covering the relevant topic in the overall progress in that particular point of time. The choice for this timeframe rests on the fact that the flight data recorder (able to record up to 25 hours of flight information) contains a so-called "underwater locator beacon" that can emit beacon ping signals (on 37.5 KHz) up to 30 days from the moment of the incident [13], assuming maximum operational lifetime of the flight data recorder battery (the additional reports from April 7th to April 9th as account for possible fluctuations in both the recorder's battery lifetime and the underwater locator beacon signal).

After the textual content of all the 66 web articles were manually extracted (as to ensure that the substance of the overall report is preserved), the factual density of each of these articles in the two data sets was automatically calculated. In the first instance, the threshold for confidence scores is not taken into consideration for the factual extraction in order to determine the general degree of online news informativeness. Since ReVerb also outputs a confidence score for each extraction, in the second instance a value of 0.6 is considered to achieve optimal balance between the recall and the precision of the extracted relations [14] as an improved measurement of the informativeness. Another reason for considering a considerable higher confidence threshold value is that "it reduces the need for human fact checking"[14], which corresponds with the fact that the informativeness assessment and the comparison outlined in this work does not involve human-annotated ranking. Following a similar statistical approach as taken in[2], Table 1 provides with details on Spearman's ρ rank correlation coefficient between the two data sets in both of the evaluation instances, as well as the respective values of the significance level. While [2] utilize the correlation outcome of the statistical analysis as to support its hypothesis on the adequacy of the factual density of the online web content relative to a human-annotated referent set, here, the statistical analysis is used to determine the feasibility of the automatic factual density calculation as informativeness metrics of the online news contents.

Table 3. Correlation tests on factual density ranking of the main online news articles using spearman's rank correlation coefficient

Confidence Threshold	Spearman's ρ	p-value	Significance Level
n/a	-0.013	0.471	52.9%
0.6	-0.235	0.094	90.6%

2.2. One Year Anniversary Reports

Marking a year from the MH7370 incident, a second subset of 10 articles summarizing the accident was compiled from the ten most popular online news outlets, according to the monthly traffic/visits they generate on a worldwide level[15]. As in the previous experiment, the textual content was manually extracted, after which the factual density with a confidence threshold (deducted from the initial analysis) is computed in Table 4 and ranked in Figure 1. The follow-up factual density is then compared with the results from the initial comparison provide in Table 3 as to yield the aggregate output of the online news appropriateness for building an encyclopaedia entry about the missing MH370 airplane, spanning a yearlong period.

Table 4. Factual density ranking of the Anniversary articles with a 0.6 confidence score

Report	Fact Count	Word Count	Factual Density $f_d(t)$
ABC News	8	610	0.013114754
Atlantic	26	704	0.036931818
BBC News	15	708	0.021186441
Business Insider	36	1668	0.021582734
CNN	7	864	0.008101852
Fox News	14	664	0.021084337
NBC News	20	667	0.029985007
New York Times	37	1,330	0.027819549
Thomson Reuters	18	578	0.031141869
USA Today	15	1,027	0.014605648

4. Assessing The Quality of Online News Articles as References for an Encyclopaedia Entry

As Table 3 suggests, the correlation coefficient for the later scenario – although still considered weak – brings the value of the significance level acceptable to a certain degree relative to the task of utilizing the online news as a referent source of information. Relying on the assumed discriminative power of the factual density over the quality of a given online news recourse, the comparison of the rankings outlined in Table 2a and Table 2b suggests an unbalanced potential set of chronological references, having the Fox News articles dominating the incident briefing at the period of the very beginning and the critical end of flight data recorder battery lifetime, leaving the Thompson-Reuters articles to dominate the other portion of the chronological reporting with a slightly better level of informativeness.

Setting this observation in the context of the data provided in Table 4 and Figure 1, it comes apparent that summarizing articles provided by Fox News and Thompson-Reuters, respectively, have reverted their positions respective to the degree informativeness, having Thompson-Reuters by far better factual density coefficient than the Fox News. This suggests that the chronological ordering of the associated encyclopaedia entry should reflect on the factual density level of the articles, selecting for a particular point of time the one with the highest value. Conversely, the factual ordering of the entry should reflect on the overall factual density ranking, regardless on the point of time in which the article is published. Overall, the most plausible candidates to be referenced in the encyclopaedia entry for the missing MH370 plane are the articles published by the Atlantic, Thompson-Reuters, NBC News and New York Times.

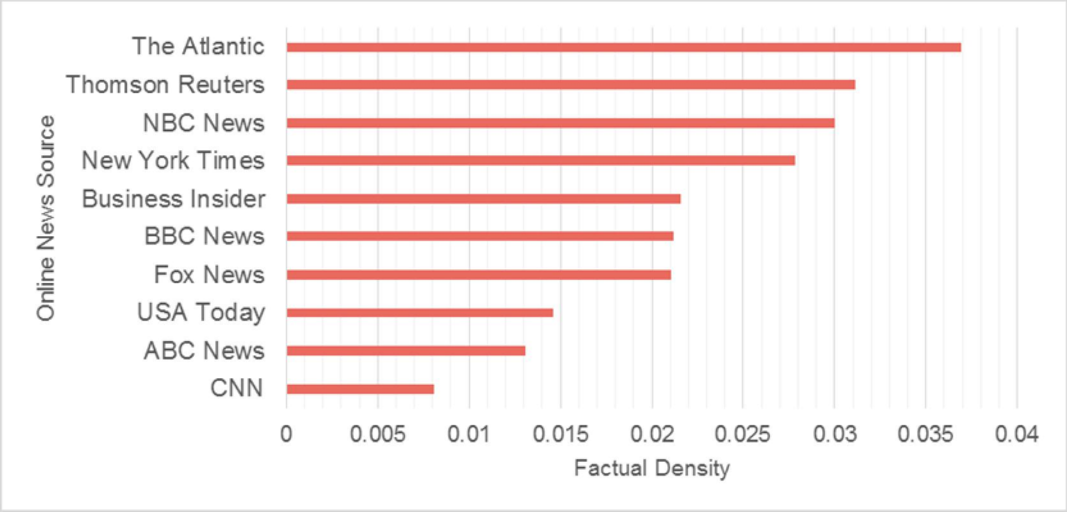

Figure 1. Online News Informativeness Chart – Summarization Articles

4. CONCLUSION

The work presented in this paper reveals interesting aspects on the deriving selection criteria for building an encyclopaedia article referencing in a part online news sources. A general conclusion holds that the informativeness assessment is valid under strict conditions of increased confidence level of the automatic factual extraction. Since the analytical effort focused on the automatically-derived factual density ranking, a plausible future direction is to compare these results against human-annotated ranking as to determine the overall usability of the online news content, at least in case of massive event reporting. Notwithstanding the useful informativeness included within the analysed reports, a year after the incident the Malaysian Airlines Flight 370 is still missing without any optimistic lead on critical evidence, making it an event of unprecedented nature that certainly deserves a dedicated historical marking.

REFERENCES

[1] O. Ferschke, “The Quality of Content in Open Online Collaboration Platforms Approaches to NLP-supported Information Quality Management in Wikipedia.” Univerity of Darmstadt, Darmstadt, Germany, p. 224, 2014.

[2] C. Horn, A. Zhila, A. Gelbukh, R. Kern, and E. Lex, “Using Factual Density to Measure Informativeness of Web Documents,” in In Proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA), 2013.

[3] N. Weber, K. Schoefegger, T. Ley, S. Lindstaedt, A. Brown, and S. Barnes, “Knowledge Maturing in the Semantic MediaWiki : A Design Study in Career Guidance,” Lect. Notes Comput. Sci., vol. 5794, pp. 700–705, 2009.

[4] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, and M. Granitzer, “Measuring the Quality of Web Content using Factual Information,” in Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, 2012, no. iii, pp. 7–10.

[5] S. Neuman, “Search For Flight MH370 Reportedly Largest In History,” NPR, 2014. [Online]. Available: http://www.npr.org/blogs/thetwo-way/2014/03/17/290890377/search-for-flight-mh370-reportedly-largest-in-history.

[6] A. Fader, S. Soderland, and O. Etzioni, “Identifying Relations for Open Information Extraction,” in EMNLP ’11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1535–1545.

[7] D. S. Etzioni, O., Banko, M., Soderland, S., & Weld, “Open Information Extraction from the Web,” Commun. ACM, vol. 51, no. 12, pp. 68–74, 2008.

[8] J. E. Blumenstock, “Size Matters : Word Count as a Measure of Quality on Wikipedia,” in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1095–1096.

[9] F. Wu and D. S. Weld, “Open Information Extraction using Wikipedia,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 118–127.

[10] P. Gamallo, M. Garcia, and S. Fern, “Dependency-Based Open Information Extraction,” in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 10–18.

[11] D. Irvine, “Cable News Ratings: Fox Still Number One—MSNBC, CNN Shedding Viewers,” Accuracy in Media, 2014. [Online]. Available: http://www.aim.org/don-irvine-blog/cable-news-ratings-fox-still-number-one-msnbc-cnn-shedding-viewers/.

[12] R. Sambrook, “Delivering trust: impartiality and objectivity in the digital age,” 2012.

[13] Investigations Office of Aviation Safety, “Flight Data Recorder Handbook for Aviation Accident,” 2002.

[14] F. Wu and D. S. Weld, “Autonomously semantifying wikipedia,” in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM ’07, 2007, pp. 41–50.

[15] eBizMBA, “Top 15 Most Popular News Websites | May 2015,” Reports, 2015. [Online]. Available: http://www.ebizmba.com/articles/news-websites. [Accessed: 26-May-2015].

[16] Fleishman Hillard, “Digital Influence Index 2012,” London, UK, 2012.

[17] Reuters Institute - Oxford University, “Digital News Report 2013,” Oxford, UK, Jun. 2013.

Authors

Filipo Sharevski (born 1985) is a Ph.D. Candidate at the Purdue University in the domain of Cybersecurity. In 2008 and 2011 he earned the B.Sc. and M.Sc. degrees in Telecommunications from the Faculty of Electrical Engineering and Information Technologies at University “Ss. Cyril and Methodius” in Skopje, Macedonia. After graduation, he was affiliated with the Institute of Telecommunications at the Faculty of Electrical Engineering and Information Technologies, Vip Macedonia Mobile Network Operator as networking engineer, and with Purdue University as a teaching and research assistant in the domain of Cybersecurity. His research interests are primarily focused on cyber forensics, network security, advanced mobile communication technologies and cyber warfare, as well as the natural language information assurance and security.