How Bad Do You Spell?: The Lexical Quality of Social Media

Ricardo Baeza-Yates
Yahoo! Research &
Web Research Group, UPF
Barcelona, Spain

Luz Rello
Web Research and NLP Groups
Univ. Pompeu Fabra
Barcelona, Spain

Abstract

In this study we present an analysis of the lexical quality of social media in the Web, focusing on the Web 2.0, social networks, blogs and micro-blogs, multimedia and opinions. We find that blogs and social networks are the main players and also the main contributors to the bad lexical quality of the Web. We also compare our results with the rest of the Web finding that in general social media has worse lexical quality than the average Web and that their quality is one order of magnitude worse than high quality sites.

1 Introduction

Lexical quality mainly refers to the degree of excellence of words in a text. Lexical quality has been used for various purposes such as spam detection (Castillo et al. 2007), credibility determination (Fogg et al. 2001) or Wikipedia vandalism detection (Potthast, Stein, and Gerling 2008). However, to the best of our knowledge, a systematic study of Social Media[1] quality has not been done. Hence, in this work we analyze the quality of social media using lexical quality as an estimator of overall content quality.

Previous work shows that there is a strong correlation between spelling errors and web data content quality (Gelman and Barletta 2008). Particularly, in this paper, the rate of lexical errors was found to be a useful metric for the quality of content of Web sites. This work uses the reported hit counts of a major search engine on a pre-determined set of commonly misspelled words as in our work.

To measure the lexical quality of social media we use the methodology that we have recently developed (Baeza-Yates and Rello 2011a; 2011b). This methodology proposes a particular measure for lexical quality based on a set words that is in time based in a detailed classification of spelling errors in the Web. We also show that there is a correlation between popularity and perceived semantic quality and our defined lexical quality. Using this measure, in this short paper we study the lexical quality of different types of social media: social networks, blogs, micro-blogs, question-answering, multimedia, collaborative sites, etc. We also compare social media to the rest of the Web. Our results contribute to the difficult and still open problem of measuring the quality of content in social media and the Web in general.

The rest of the paper is organized as follows. Section 2 presents a brief account of related work. Section 3 summarizes the methodology of that we use in our study for the estimation of lexical quality in the Web. The results and analysis of the lexical quality of Social Media is presented in Section 4. In Section 5 some conclusions are drawn and plans for future work are considered.

2 Related Work

The quality of the Web can be related to its contents (highly current, accuracy, source reputation, objectivity, etc.) or to its representation (spelling errors, various typos, very long sentences, grammatical errors, etc.). More frequently, Web quality has been broadly related to contents issues (Castillo et al. 2007; Fogg et al. 2001) than to its representation (Gelman and Barletta 2008) as our research.

Regarding quality of Social Media, there are few studies. They have mainly focused in the semantic quality of community question-answering systems (Jeon et al. 2006b; Agichtein et al. 2008b; Harper et al. 2008; Bian et al. 2009) and to the best of our knowledge there are no similar studies for other classes of social media Web sites. On the other hand, in (Chai, Potdar, and Dillon 2009) nineteen frameworks to evaluate the quality of content of social media are analyzed.

There are other ways to assess quality of Social Media. Most of them are focused on the identification of the quality of the content, not on its representation. They exploit other sources such as community feedback (Agichtein et al. 2008a), user interactions (Bian et al. 2008), click counts (Jeon et al. 2006a) and bag of words for text classification (Harper, Moy, and Konstan 2009). Here, we provide an additional measure for the difficult problem of assessing Web quality.

The closest work to ours is by Gelman and Barletta (Gelman and Barletta 2008) that apply the spelling error rate as a metric to indicate the degree of quality of Web sites. This work uses a carefully chosen set of 10 frequent misspelled words and their relative hit counts in a search

[1] A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content (Kaplan and Haenlein 2010).

engine. We improve their methodology, by selecting the 10 most frequent words out of a list of more than 1,300 misspelled words. As we explain in the next section, we differ from previous work in that our list takes into account different kind of errors in order to identify different types of lexical errors (Baeza-Yates and Rello 2011a; 2011b) (regular spelling errors, typographical errors, errors made by non-native speakers of English, dyslexic errors and optical character recognition (OCR) errors). We use lexical errors as a proxy to the quality of content in social media since a similar method based on Wikipedia web pages gave positive results (Gelman and Barletta 2008).

There are other studies which take into consideration lexical quality for different purposes. For instance, in (Ringlstetter, Schulz, and Mihov 2006) after investigating the distribution of orthographic errors of various types in Web pages for specific topics, the authors propose filtering methods to retrieve cleaner corpora from the Web. In (Piskorski, Sydow, and Weiss 2008) certain linguistic features related to lexical quality are explored for detecting spam. Nevertheless, lexical errors are not taken directly into account and the lexical validity of the text is measured by the ratio of the number of words detected by the parser and the total number of tokens of the text.

3 Methodology

By lexical quality we understand its classic definition taken from the theory of reading acquisition. According to Perfetti (Perfetti and Hart 2002) a lexical representation has high quality to the extent that it has a fully specified orthographic representation (a spelling) and redundant phonological representations (one from spoken language and one recoverable from orthographies-to-phonological mapping) (Perfetti and Hart 2002).

The methodology we use is takes in account both causes of low lexical quality. While different type of misspellings are considered, a lack of redundant phonological representations is covered by dyslexic and typing errors. Moreover, it also considers non-human errors, such as the ones produced by optical character recognition (OCR) software, to widen the typology or possible lexical errors found in the web. After studying how all these type of errors occur in the Web in (Baeza-Yates and Rello 2011b), we were able to select a set of words, W_M, that has a frequent misspell. This set of ten words is given in the Appendix.

The set of ten words, W_M, were the most frequent errors extracted from a sample 50 selected words (see Appendix) which together with their corresponding variants with errors, gives us a total of 1,345 different words. There are no stopwords in our list and the words are relatively long (an average length of 8.2 letters per word). We take into account five kind of errors: (1) regular spelling errors produced by non-impaired native English individuals which result from insufficient language competence, (2) regular typos caused by the adjacency of a letter in the keyboard, (3) errors made by non-native speakers who use English as a foreign language, (4) errors made by dyslexic persons and (5) optical character recognition (OCR) errors. The regular spelling errors were selected taking into account their high frequency in query logs. The regular typos were generated by substituting each of the letters of the intended word with the letter situated immediately up, down, left and right from the intended letter. To find the typical foreign language transference errors made by non-native speakers we used linguistic knowledge. The dyslexic errors were extracted from texts written by dyslexic users and taken from the literature (Pedler 2007). To generate the OCR errors we substituted the typical letters which are usually mistaken (Baeza-Yates and Rello 2011b).

To find the typical errors made by non-native speakers who use English as a foreign language, we have taken into account errors caused by transference from English itself or from other languages such as Spanish, French or Italian. We choose languages with morphological and phonological similarities. For instance, **gobernment* is a typical error made by Spanish learners of English, since the graphemes <b> and <v> are pronounced as /b/, and the phoneme /v/ does not exist in the standard Spanish phonemic system. Besides its translation in Spanish is written with <b>.

Since dyslexic errors are the most difficult to find, our starting point was the selection of words with errors written by dyslexic users taken from the literature (Pedler 2007) and from a corpus made *ad hoc*. For example, the dyslexic word **anfurtunatley*, from the intended word *unfortunately*.

After selecting candidates for the word sample, we checked the other types of errors related to word were unique and not ambiguous. For instance, the correct word *worried* could be also a typo from the intended word *worries* since *s* and *d* are adjacent in the keyboard. Similarly, the typo **dxplain* (from *explain*) is also a proper name. Hence, named entities and real world errors were dismissed, as well as words with more than three ambiguous errors.

Sampling the Web is a difficult problem in general (Bar-Yossef and Gurevich 2008). Hence, we provide an estimation of the lexical quality of a Web site. That is, as a measure of lexical quality we use the relative ratio of the misspells to the correct spellings averaged over our word sample:

$$LQ = mean_{w_i \in W_M} \left(\frac{df_{\text{misspell } w_i}}{df_{\text{correct } w_i}} \right)$$

Hence, a lower value of LQ implies a larger lexical quality, being 0 perfect quality. Notice that LQ is correlated with the rate of lexical errors but it is not the same because is a ratio against the corrected words and takes in account the most frequent misspell for each word.

To compute LQ, we estimate df by searching each word only in the English pages of a major search engine.

Although the lexical quality measured will vary with the set of words W_M chosen, the relative order of the measure will hardly change as the size of the set grows. Hence we believe that LQ is a good estimator of the lexical quality of a Web site.

4 Lexical Quality of Social Media

To asses the lexical quality of social media, we computed LQ in a set of 25 Web sites, including Wikipedia. The Web sites were chosen to cover most of the different categories of social media sites, considering also the size of them (users and content).

Site		Size (%)	Range (%)	Average (%)
Quora	O	0.1	0 –0.081	0.014
Wikipedia	C	0.2	0.002–0.041	0.018
CiteuLike	C	0.1	0 –0.201	0.023
Picasa	M	3.1	0.001–0.178	0.043
Epinions	O	0.5	0.001–0.479	0.067
Twitter	B	10.2	0.002–0.439	0.068
Flickr	M	6.9	0.001–0.358	0.073
LinkedIn	S	4.9	0.002–0.377	0.074
Tumblr	B	0.8	0.002–0.781	0.097
Digg	C	0.1	0.002–0.643	0.107
Youtube	M	6.1	0.007–0.578	0.137
MySpace	S	10.5	0.002–0.590	0.144
Wikia	C	0.4	0.003–1.180	0.153
Last.fm	M	0.2	0.002–0.523	0.154
Friendster	S	0.3	0.007–0.670	0.157
Blogger	B	8.0	0.003–1.715	**0.225**
Hi5	S	4.6	0.015–0.852	**0.241**
Bebo	S	1.0	0.014–1.024	**0.249**
Foursquare	B	0.2	0 –2.161	**0.260**
Facebook	S	33.3	0.040–1.551	**0.309**
Yelp	O	0.2	0.028–3.045	**0.332**
Fotolog	M	6.5	0.073–2.188	**0.412**
Wikispaces	C	0.1	0.051–2.868	**0.413**
LiveJournal	B	0.6	0.002–6.290	**0.699**
Y! Answers	O	1.2	0.020–4.680	**0.707**
Overall		100.0	0 –4.68	0.219

Table 1: Range and average lexical quality in percentages for a sample of frequent misspellings in several social media sites. The values over the social media average are highlighted, the values over the Web average are below the middle line, and 0.00* represents a number larger than 0 but less than 0.0005.

To compare them with the rest of the Web, we classify the social media sites in five classes: blogs (B, including micro-blogs), social networks (S), collaboration sites (C), multimedia sites (M) and opinions (O, including community question-answering systems). All the classes have five sites with the exception of social networks (six) and opinions (four). To be able to assess the impact of each site, we need to estimate the relative size of each one of them. For this we use the overall number of words in the public content of each Web site according to a major search engine.

The lexical quality results for the different Web sites chosen are shown in Table 1. For each site we also give its class and the relative size of their (public) content. Regarding the estimated content size, almost 55% of the content comes from social networks (Facebook, MySpace, LinkedIn, Hi5, Friendster and Bebo), while almost 23% and 20% comes from multimedia and blogs, respectively.

From our estimators we obtain that 47% of the errors in the social media Web sites that we consider come from Facebook. This percentage grows to almost 80% if we add Fotolog, Blogger, MySpace and Hi5. That is, just five sites contribute with the majority of the bad lexical quality. On the other hand there is no correlation between public content size and lexical quality. Overall, social networks account for more than half (62%) of the errors and multimedia is almost one fifth of them (19%), so together they are more than 81% of the bad lexical quality. We notice also that there is no clear order for the site classes.

Sites	Range (%)	Average (%)
.edu	0.001–0.072	0.011
Wikipedia	0.002–0.041	0.018
NY Times	0.001–0.117	0.032
USA Government	0.00*–0.286	0.032
.org	0.002–0.103	0.038
.com	0.003–0.139	0.051
Yahoo!	0.002–0.453	0.075
.net	0.004–0.233	0.080
Microsoft	0.011–0.520	**0.115**
CNN	0.015–0.729	**0.126**
Collaboration	0.002–2.868	**0.132**
Blogs	0 –6.290	**0.154**
Multimedia	0.001–2.188	**0.183**
Social Media	0 –4.680	**0.220**
Social Networks	0.002–1.551	**0.249**
Opinions	0 –4.680	**0.475**
Web	0.010–0.482	0.099

Table 2: Range of percentages and average for a sample of frequent misspellings in several sets of Web sites. The values over the average of the Web are highlighted and 0.00* represents a number larger than 0 but less than 0.0005.

In Table 2 we compare each class and social media as a whole with other important sites or domains of the Web. Comparing with Table 1 we see that just nine sites have lexical quality that is better that the average of the Web, but those account for less than 27% of the content. On the other hand, on average, social media classes have lexical quality larger than the Web itself. We can observe that collaborative (where Wikipedia is the star) sites are the best ones, followed by blogs, multimedia, social networks, and further away opinions. Compared to high quality sites, the quality of social media is one order of magnitude worse. This should not be a surprise considering the diversity and sheer volume of social media content.

In addition, we believe that the lower quality of social media impacts many more sites. For example we found that the community section of the NY Times is the main contributor to the decrease of their lexical quality. A similar effect occurs for almost all large Web sites like CNN or Microsoft.

5 Concluding Remarks

We have presented the first estimation of the lexical quality of social media, which in turn can be used to estimate the semantic quality of social media. Nevertheless, these estimations should be taken with a grain of salt, as they will change with a different sample of sites and/or words sample. Nevertheless, we believe that the main results will be maintained, e.g. that the lexical quality of social media is worse than the

average on the Web.

Future work include to define new ways to measure lexical quality and compare them with these results to check for consistency. We also would like to increase the sample of social media sites studied as well as to use a larger sample of words to measure the lexical quality.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008a. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, 183–194. ACM.

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008b. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, 183–194. New York, NY, USA: ACM.

Baeza-Yates, R., and Rello, L. 2011a. Estimating dyslexia in the web. In *Proceedings of the International Cross Disciplinary Conference on Web Accessibility (W4A 2011)*.

Baeza-Yates, R., and Rello, L. 2011b. On the lexical quality of the web. In *Submitted*.

Bar-Yossef, Z., and Gurevich, M. 2008. Random sampling from a search engine's index. *J. ACM* 55(5).

Bian, J.; Liu, Y.; Agichtein, E.; and Zha, H. 2008. A few bad votes too many?: towards robust ranking in social media. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, 53–60. ACM.

Bian, J.; Liu, Y.; Zhou, D.; Agichtein, E.; and Zha, H. 2009. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 51–60. New York, NY, USA: ACM.

Castillo, C.; Donato, D.; Gionis, A.; Murdock, V.; and Silvestri, F. 2007. Know your neighbors: web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, 423–430. New York, NY, USA: ACM.

Chai, K.; Potdar, V.; and Dillon, T. 2009. Content quality assessment related frameworks for social media. *Computational Science and Its Applications–ICCSA 2009* 791–805.

Fogg, B.; Marshall, J.; Kameda, T.; Solomon, J.; Rangnekar, A.; Boyd, J.; and Brown, B. 2001. Web credibility research: a method for online experiments and early study results. In *CHI '01 extended abstracts on Human factors in computing systems*, CHI '01, 295–296. New York, NY, USA: ACM.

Gelman, I. A., and Barletta, A. L. 2008. A "quick and dirty" website data quality indicator. In *WICOW '08 Proceeding of the 2nd ACM workshop on Information credibility on the web*, 43–46.

Harper, F. M.; Raban, D.; Rafaeli, S.; and Konstan, J. A. 2008. Predictors of answer quality in online q&a sites. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, 865–874. New York, NY, USA: ACM.

Harper, F.; Moy, D.; and Konstan, J. 2009. Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, 759–768. ACM.

Jeon, J.; Croft, W.; Lee, J.; and Park, S. 2006a. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 228–235. ACM.

Jeon, J.; Croft, W. B.; Lee, J. H.; and Park, S. 2006b. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, 228–235. New York, NY, USA: ACM.

Kaplan, A. M., and Haenlein, M. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons* 53(1):59 – 68.

Pedler, J. 2007. *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Ph.D. Dissertation, Birkbeck, London University.

Perfetti, C., and Hart, L. 2002. *Precursors of functional literacy*. Amsterdam/Philadelphia: John Benjamins. chapter The lexical quality hypothesis, 189–213.

Piskorski, J.; Sydow, M.; and Weiss, D. 2008. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, 25–28. New York, NY, USA: ACM.

Potthast, M.; Stein, B.; and Gerling, R. 2008. Automatic vandalism detection in wikipedia. In Macdonald, C.; Ounis, I.; Plachouras, V.; Ruthven, I.; and White, R., eds., *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 663–668.

Ringlstetter, C.; Schulz, K. U.; and Mihov, S. 2006. Orthographic errors in web pages: Towards cleaner web corpora. *Computational Lingusitics*.

Appendix

The sample of ten frequent misspelled words, W_M, is:

**albun, *alwasy, *arround, *becuase, *enoguh, *everyhting, *haveing, *problen, *remenber*, and **workig*.

The sample of the 50 words from which their variants with errors were generated are:

absolutely, *actually*, *album*, *always*, *almost*, *around*, *auditorium*, *birthday*, *cannot*, *check*, *circumstances*, *comparison*, *confusion*, *definitely*, *downloading*, *everything*, *enough*, *exercise*, *explain*, *fabulous*, *features*, *friend*, *gentlemen*, *having*, *knowledge*, *length*, *linguistic*, *little*, *understanding*, *impossible*, *interesting*, *maybe*, *myself*, *problem*, *remember*, *right*, *situation*, *things*, *tomorrow*, *unbelievable*, *understand*, *unfavorable*, *unfortunately*, *walkable*, *waiting*,*watch*, *working*, *worries*, and *writing*.