Social Media Is NOT that Bad! The Lexical Quality of Social Media

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

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

Abstract

There is a strong correlation between spelling errors and web text content quality. Using our lexical quality measure, based in a small corpus of spelling errors, we present an estimation of the lexical quality of the main Social Media sites. This paper presents an updated and complete analysis of the lexical quality of Social Media written in English and Spanish, including how lexical quality changes in time.

1 Introduction

Lexical quality refers to the degree of excellence of words in a text. Previous work had shown that there is a strong correlation between spelling errors and web data content quality (Gelman and Barletta 2008) and web text understandability (Rello and Baeza-Yates 2012) in concordance with the Web Content Accessibility Guidelines (WCAG) principles (Caldwell et al. 2008). Regarding Social Media,[1] the rate of lexical errors was found to be a useful metric for the quality of content of websites and that rate for English was higher in social media than in the overall Web (Baeza-Yates and Rello 2011b).

In this paper, we present a complete updated analysis of the lexical quality of the main social media sites for English and Spanish. This work uses the reported hit counts of a major search engine on a pre-determined set of commonly misspelled words as suggested in previous work (Gelman and Barletta 2008). However, here we use an improved methodology recently developed in (Baeza-Yates and Rello 2011a; 2012) to analyze web text quality. We apply our approach to 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. Our new results show that social media is not as bad as we could expect.

[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).

The rest of the paper is organized as follows. Section 2 introduces related work. In Section 3 we introduce the lexical quality measure that we use in this paper. The results and analysis of the lexical quality of Social Media is presented in Section 4. Finally, 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, etc.) or to its representation (spelling errors, various typos, grammatical errors, etc.). With respect to the quality of the social media, most of the studies are more focused on the identification of the semantic quality of the content (Jeon et al. 2006; Agichtein et al. 2008; Bian et al. 2009; Harper et al. 2008; Chai, Potdar, and Dillon 2009), than on its representation. They exploit other sources such as community feedback (Agichtein et al. 2008), user interactions (Bian et al. 2009), click counts (Jeon et al. 2006) and the bag of words model for text classification (Harper, Moy, and Konstan 2009). Here, we provide an additional measure for the difficult problem of assessing web quality.

The work that inspired the methodology is the one by Gelman and Barletta (Gelman and Barletta 2008) that apply the spelling error rate as a metric to indicate the degree of quality of websites. This work uses a carefully chosen set of ten frequent misspelled words in English and their relative hit counts in a search engine. We improved their methodology, by selecting the ten most frequent words for English and Spanish out of a list of almost 2,700 misspelled words. We use lexical errors as a proxy to the quality of content in social media since the similar method already mentioned gave positive results for Wikipedia web pages (Gelman and Barletta 2008). Using this methodology, in (Baeza-Yates and Rello 2011b) we presented a preliminary estimation of the lexical quality of the Social Media in English, while the present paper offers a more complete and updated study of the quality of social media in English and Spanish.

Site	Type	Size (%) 2011	Size (%) 2012	Range 2011	Range 2012	Average 2011	Average 2012
CiteuLike	C	019	0.08	0 –0.201	0 –0.107	0.023	0.010
Foursquare	B	0.20	0.04	0 –2.161	0.00*–0.084	**0.260**	0.025
Quora	O	0.03	0.14	0 –0.081	0 –0.014	0.014	0.026
Wikipedia	C	0.13	0.52	0.002–0.041	0.00*–0.183	0.018	0.038
Flickr	M	3.52	13.90	0.001–0.358	0.00*–0.219	0.073	0.045
Picasa	M	7.38	1.40	0.001–0.178	0.001–0.140	0.043	0.058
LinkedIn	S	0.22	0.62	0.002–0.377	0.00*–0.332	0.074	0.068
Tumblr	B	3.86	1.89	0.002–0.781	0.005–0.347	0.097	0.070
Digg	C	0.05	0.05	0.002–0.643	0.00*–0.566	0.107	0.073
Friendster	S	0.02	0.46	0.007–0.670	0.005–0.407	0.157	0.099
Y! Answers	O	1.89	1.40	0.020–4.680	0.005–0.744	**0.707**	0.149
Twitter	B	3.39	6.65	0.002–0.439	0.00*–0.859	0.068	0.154
Last.fm	M	0.21	0.32	0.002–0.523	0.002–0.796	0.154	0.158
MySpace	S	6.05	14.26	0.002–0.590	0.015–0.613	0.144	0.159
Epinions	O	0.89	0.61	0.001–0.479	0.00*–1.016	0.067	0.164
Wikispaces	C	0.73	0.08	0.051–2.868	0.004–0.717	**0.413**	**0.178**
Wikia	C	0.46	0.78	0.003–1.180	0.00*–1.765	0.153	**0.185**
Youtube	M	16.71	6.51	0.007–0.578	0.001–1.534	0.137	**0.192**
Bebo	S	0.43	2.60	0.014–1.024	0.173–1.045	**0.249**	**0.246**
Blogger	B	40.00	32.49	0.003–1.715	0.001–1.403	**0.225**	**0.258**
Hi5	S	0.44	0.68	0.015–0.852	0.009–0.893	**0.241**	**0.262**
Fotolog	M	0.27	0.22	0.073–2.188	0.001–2.016	**0.412**	**0.329**
Yelp	O	0.69	0.85	0.028–3.045	0.001–1.610	**0.332**	**0.354**
Facebook	S	8.01	11.64	0.040–1.551	0.004–3.155	**0.309**	**0.479**
LiveJournal	B	4.23	1.79	0.002–6.290	0.004–2.471	**0.699**	**0.518**
Overall		100.0	100.0	0 –4.680	0 –3.155	0.220	0.164

Table 1: Range and average LQ for a sample of frequent misspellings in several social media sites in English.

3 Lexical Quality Measure

Our study is based on a measure of lexical quality proposed in (Baeza-Yates and Rello 2011a) and studied in (Baeza-Yates and Rello 2012). In those papers we define lexical quality LQ as:

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

where W is a set of frequently misspelled words. Those words were chosen such that they were frequent and had large relative error. Then we use data from a leading search engine to estimate the document frequency (df) values, computing the relative ratio of the most popular misspells to the correct spellings, averaged over the word sample W.

Hence, a lower value of LQ (Lexical Quality) implies a larger lexical quality, zero being perfect quality. To compute LQ, we estimate df by searching each word in the English and Spanish pages indexed by 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 website. In (Baeza-Yates and Rello 2012) we showed that LQ as an independent measure compared with other web popularity measure.

To find the adequate W, we used two lists of errors for English and Spanish which contain 50 target words for each language together with their corresponding different types of errors (a total of 2,670 words all together). These different types of errors are derived from our error classification for English distinguishes between regular spelling, typographical, non-native speakers, dyslexic and optical character recognition (OCR) errors (Baeza-Yates and Rello 2011a). Then, we estimated the relative frequency of each of the errors in our lists and we were able to select two sets of words with frequent misspells (W_{en} for English and W_{sp} for Spanish). These sets are given in the Appendix.

4 Lexical Quality of Social Media

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

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 website according to a major search engine.

For this we us two estimations: (a) the sum of the overall document frequency of our word sample and (b) the maximum document frequency of the word sample. The sites

Site	Type	Size (%)	Range	Average
Quora	O	0.00*	0 –0	0
CiteuLike	C	0.00*	0 –0	0
Epinions	O	0.00*	0 –0.022	0.002
Wikispaces	C	0.38	0.001–0.066	0.010
Foursquare	B	0.00*	0 –0.039	0.011
LinkedIn	S	0.02	0.001–0.027	0.012
Yelp	O	0.00*	0 –0.136	0.015
Tumblr	B	0.48	0.00*–0.100	0.019
Youtube	M	10.74	0.004–0.080	0.022
Digg	C	0.00*	0 –0.332	0.035
Blogger	B	53.93	0.004–0.162	0.038
Picasa	M	0.16	0.002–0.175	0.039
Wikipedia	C	0.03	0.003–0.194	0.040
Wikia	C	0.08	0.003–0.264	0.040
Last.fm	M	0.02	0.002–0.188	0.041
Live Journal	B	8.10	0.00*–0.317	0.052
Flickr	M	0.53	0.009–0.208	0.059
Bebo	S	0.00*	0.019–0.123	0.068
Hi5	S	3.27	0.017–0.288	0.086
MySpace	S	0.56	0.011–0.307	0.092
Twitter	B	1.05	0.015–0.944	**0.161**
Y! Answers	O	1.67	0.038–0.496	**0.217**
Facebook	S	7.63	0.030–2.358	**0.375**
Friendster	S	0.00*	0 –4.000	**0.400**
Fotolog	M	11.34	0.039–1.706	**0.648**
Overall		100.000	0 –4.000	0.095

Table 2: Range and average LQ for a sample of frequent misspellings in several social media sites in Spanish in 2012.

chosen, the class and the two size estimators are shown in Table 1 for English and Table 2 for Spanish.[2] As expected, the Spanish content is much less than English, in particular in sites that only target the English language. For each site we also give its class and the relative size of their (public) content.

In English and 2011, almost 42% of the estimated content size comes from blogs or micro-blogs, with Blogger accounting for about the 78% of that, while almost 28% and 23% comes from social networks and multimedia, respectively. In 2012 that picture is similar with blogs and micro-blogs representing 52% of the content, while multimedia and social networks are almost 28% and 15%. This shows that currently blogs and multimedia are dominating web text content. In Spanish is even more biased with blogs taking more than 63% of the content while multimedia covers 18% and social networks only 4%.

From our quality estimator we obtain that in English a large fraction of the errors come from the major social networks, opinions and blogs (e.g. Facebook, Y! Answers, Live Journal). In Spanish the result is similar, with Fotolog replacing YouTube. On the other hand there is no correlation between public content size and lexical quality and we notice also that there is no clear order for the site classes.

We see that for English in 2012 just ten sites have lexical quality that is worse than the average of the Web, accounting for 61% of the content. In Spanish the number drops to 5 with only 17% of the content. Hence the Spanish social media seems to have better quality than the English one.

[2]In the Tables, the values over the social media average are highlighted and 0.00* represents a number larger than 0 but less than 0.0005.

Sites	Range 2011	Range 2012	Average 2011	Average 2012
NY Times	0.001–0.117	00.0*–0.054	0.032	0.009
USA Gov.	0.00*–0.286	00.0*–0.958	0.032	0.023
Wikipedia	0.002–0.041	00.0*–0.183	0.018	0.038
.edu	0.001–0.072	0.001–0.926	0.011	0.064
Yahoo!	0.002–0.453	0.006–3.002	0.075	0.077
Collaboration	0.002–2.868	0 –1.765	**0.132**	0.097
Multimedia	0.001–2.188	0.001–2.026	**0.183**	**0.156**
Microsoft	0.011–0.520	0.001–0.695	**0.115**	**0.162**
Social Media	0 –4.680	0 –3.155	**0.220**	**0.164**
Opinions	0 –4.680	0 –1.610	**0.475**	**0.173**
Blogs	0 –6.290	0.00*–2.471	**0.154**	**0.179**
Soc. Networks	0.002–1.551	0.001–3.155	**0.249**	**0.219**
.org	0.002–0.103	0.012–2.906	0.038	**0.484**
CNN	0.015–0.729	00.0*–4.792	**0.126**	**0.595**
.com	0.003–0.139	0.055–5.508	0.051	**1.002**
.net	0.004–0.233	0.024–5.807	0.080	**1.065**
Web	0.010–0.482	0.010–0.451	0.047	0.107

Table 3: Range and average LQ for a sample of frequent misspellings in several sets of websites in English.

In Tables 3 and 4 we compare each class and social media as a whole with other important sites or domains of the Web. The first surprise is how the average of the Web has changed in one year. This change seems to be the result of LQ growing faster in the of the overall Web with respect to social media. These results show that the previous result (Baeza-Yates and Rello 2011b), which showed that the lexical quality of social media was worse than the average on the Web was misleading and may change in the future de-

Sites	Range	Average
.edu	0.00*–0.004	0.001
CNN (Sp.)	0 –0.011	0.002
El Pais	0.00*–0.052	0.012
Collaboration (C)	0 –0.332	0.025
Wikipedia (Sp.)	0.003–0.194	0.039
Opinions (O)	0 –0.496	0.050
Blogs (B)	0 –0.944	0.056
.org	0.009–0.234	0.070
Social Media	0 –4.000	0.095
Yahoo! (Sp.)	0.029–0.254	**0.115**
Microsoft (Sp.)	0.00*–0.965	**0.116**
Multimedia (M)	0.002–1.706	**0.162**
Soc. Networks (S)	0 –4.000	**0.172**
.com	0.055–0.570	**0.222**
.net	0.039–0.790	**0.236**
Web	0.029–0.300	**0.147**

Table 4: Range and average LQ for a sample of frequent misspellings in Spanish for several domains.

pending on the relative growth of social media with respect to the complete Web.

On average, social media classes have lexical quality larger than the Web itself, in particular for Spanish. We can observe that for English, collaborative (where Wikipedia is the star) sites are the best ones, followed by multimedia/blogs and then social networks/opinions. For Spanish both last pairs appear in reverse order and Wikipedia is replaced by CNN. Compared to high quality sites, the quality of social media is at least one order of magnitude worse. This should not be a surprise considering the diversity of people and sheer volume of social media content.

Notice that LQ is not as constant in time in social media (see Table 3 and Table 1). 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 websites like CNN or Microsoft.

5 Concluding Remarks

In both 2011 and 2012 the lexical quality of the social media is worse than the overall Web. However, the lexical quality of the Web has decreased over time, while the lexical quality of the social media has improved.

We have presented the most updated and complete estimation of the lexical quality of social media until now. This estimation can be used to value the understandability degree and the semantic quality of social media at a certain time but it is not suitable to predict the lexical quality of a website since it is highly variable in time. These estimations should be taken with care, as they could change with time and with a different sample or words samples. Nevertheless, we believe that the main results will be maintained, since the relative order of the measure hardly changes as the size of the set grows and provides independent information about the quality of a website.

For future work we plan to define new ways to measure lexical quality and compare them with these results to check their consistency. We also plan to increase the sample of social media websites 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. 2008. Finding high-quality content in social media. In *Proc. WSDM'08*, 183–194. ACM Press.

Baeza-Yates, R., and Rello, L. 2011a. Estimating dyslexia in the Web. In *International Cross Disciplinary Conference on Web Accessibility (W4A 2011)*, 1–4. Hyderabad, India: ACM Press.

Baeza-Yates, R., and Rello, L. 2011b. How bad do you spell?: The lexical quality of social media. In *Workshop on the Future of the Social Web (FOSW 2011) held with ICWSM 2011*.

Baeza-Yates, R., and Rello, L. 2012. On measuring the lexical quality of the web. In *The 2nd Joint WICOW/AIRWeb Workshop on Web Quality*.

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.

Caldwell, B.; Cooper, M.; Reid, L. G.; and Vanderheiden, G. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)*.

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.

Gelman, I. A., and Barletta, A. L. 2008. A "quick and dirty" website data quality indicator. In *The 2nd ACM workshop on Information credibility on the Web (WICOW '08)*, 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. B.; Lee, J. H.; and Park, S. 2006. 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.

Rello, L., and Baeza-Yates, R. 2012. Lexical quality as a proxy for web text understandability. In *The 21st International World Wide Web Conference (WWW 2012)*.

Appendix

1. The sample of ten frequent misspelled English words, W_{en}, is:

   **albun (album)*, **alwasy (always)*, **arround (around)*, **becuase (bacause)*, **enoguh (enough)*, **everyhting (everything)*, **haveing (having)*, **problen (problem)*, **remenber (remember)* and **workig (working)*.

2. The sample of ten frequent misspelled Spanish words, W_{sp}, is:

   **entocnes (entonces)*, **haceindo (haciendo)*, **honbre (hombre)*, **momemto (momento)*, **pefecto (perfecto)*, **porqeu (porque)*, **peuden (pueden)*, **siemrpe (siempre)*, **tenog (tengo)* and **vamso (vamos)*.