WEB ACCESSIBILITY METRICS SYMPOSIUM

5 DECEMBER 2011

LEXICAL QUALITY AS A MEASURE FOR TEXTUAL WEB ACCESSIBILITY

Luz Rello. Universitat Pompeu Fabra, luzrello@acm.org
Ricardo Baeza-Yates. Yahoo! Research & Universitat Pompeu Fabra, rbaeza@acm.org

1. Problem Addressed

The problem addressed is the measurement of the lexical quality of the Web, that is, the representational aspect of the textual Web content. Lexical quality broadly refers to the quality degree of words in a text (spelling errors, typos, etc.) and it is related to the degree of readability of a website (Cooper et al. 2010). Although lexical quality is not used as an accessibility metric, we propose that including text quality and correctness in accessibility metrics could be useful, since the quality of words and language impacts the readers understanding. Moreover, lexical quality maps to the WCAG principle of content being "perceivable" and "understandable" (Caldwell et al. 2008).

2. Background

Our approach is mainly inspired by the work of Gelman and Barletta (2008) that apply a spelling error rate as a metric to indicate the degree of quality of websites. They use a set of

ten frequently misspelled words and hit counts of a search engine for this set. While they focus on spelling errors, we present an original classification of lexical errors in English motivated by their relationship with textual accessibility, such as the errors made by people with dyslexia. This is a followup of our previous work (Baeza-Yates and Rello 2011).

3. Strategy

Our error classification for English distinguishes between regular spelling, typographical, non-native speakers, dyslexic and optical character recognition (OCR) errors. Native and non-native misspellings are phonetic errors, typos are behavioral errors, OCR mistakes are visual errors, while dyslexic errors could be phonetic or visual. Detecting different classes of errors provides the possibility of refining the knowledge we have about Web lexical quality. Besides, the fact that dyslexic errors are discriminated from the rest, makes this study valuable to accessible practices for dyslexic Internet users, which is a relatively large group estimated in 10-17% of the USA population (McCarthy 2010).

We selected a sample *W* of 50 target words with their corresponding variants with errors, giving us a total of 1,345 different words. Sample *W* is bigger than previous related work which used ten words (Gelman and Barletta 2008; Baeza-Yates and Rello 2011). For instance, the target word *tomorrow* has the corresponding errors variants in our sample: **romorrow, *yomorrow, *timorrow, *tpmorrow, *tonorrow, *tomirrow, *tomprrow, *tomoeeow, *tomottow, *tomorriw, *tomorrpw, *tomorroq* and **tomorroe* (typographical errors); *toomorrow* (regular spelling error) ; **tomorow* and **tomorou* (non-native speakers errors); **torromow* (dyslexic error); and **tomorrov, *tamarraw* and **tonorrow* (OCR errors).

First, we measured a lower bound of the fraction *f* of Web pages with lexical errors and the relative fraction *d* of each kind of error in the sample *W*. The corresponding fraction of Web pages with lexical errors is then *f x d*. We use data from a leading search engine to estimate this value. To measure lexical quality we chose ten misspelled words that at the same time were frequent and had large relative error, W_M, computing the relative ratio of the misspells to the correct spellings averaged over this word sample. That is,

$$LQ = mean_{w_i \in W_M} \frac{df_{misspell\ w_i}}{df_{correct\ w_i}}$$

Hence, a lower value of *LQ* (Lexical Quality) implies a larger lexical quality, zero being 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 correct words and takes into account the most frequent misspell for each word.

To compute *LQ*, we estimate *df* by searching each word 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 website.

4. Major Difficulties

We found two major difficulties:

First, the selection of the words in the sample *W* is not a trivial task: words and different types of errors have to be distinguished without ambiguity. We established detailed criteria to select the sample.

Second, sampling the Web is a difficult problem in general (Bar-Yossef and Gurevich 2008) and even more in our particular case. To bound the overall rate of errors in the Web, we have to model the co-occurrence of words in the Web. This is an open problem in general (Baeza-Yates and Ribeiro-Neto 2010), but we can use a simple model that allows to bound the co-occurrences of words.

5. Outcomes

To show the relevance of *LQ* as an independent variable we computed the Pearson correlation with the following measures for the top 20 sites in English of Alexa.com (all in March 2011): Alexa unique visitors, number of pages, number of in-links, and ComScore Unique Visitors. We observed that *LQ* is mildly correlated with the Alexa ranking (see Table 1) and the size (as expected, more content, more errors). Hence *LQ* provides independent information about the quality of a website.

Table 1: Pearson correlation for several measures in the top 20 English sites of Alexa.com in March 2011.

Measure	Alexa	Pages	Links	ComScore
LQ	0.4451	0.4167	0.3966	0.2356
Alexa		0.7659	0.6897	0.6589
Size			0.8655	0.3097
Links				0.1319

We assess the correlation of lexical and domain quality applying our methodology to several large Web domains and the major English speaking countries. Although there is a correlation between high lexical quality and the content of major websites, some domains that should have high lexical quality do not have it. Among other observations we notice that the quality of USA and UK universities is similar but the UK government has three times better quality than the USA government. Regarding country domains, the correlation between lexical and domain quality is high and the geographical distribution of lexical

quality shows the impact of business Web pages and number of users among English speaking countries.

We also computed the lexical quality results in major and social media websites. Many of them have quite good lexical quality in spite of their collaborative nature, like Wikipedia, Flickr and Twitter. An explanation for the later two sites could be that texts there are short and our words are long. On the other hand, most other social media websites have worse lexical quality than the Web average.

6. Open Research Avenues

LQ uses a conventionally non-accessibility source and is not an accessibility metric. However, it could be potentially added to traditional metric scores using text quality as a proxy measure for accessibility.

Future work will include improving our technique to study how it converges when the set *W* grows as well as better techniques to bound the overall rate of errors in the Web. We also plan to validate this measure regarding text accessibility carrying out user studies.

References

1. **Books**: R. Baeza-Yates and B. Ribeiro-Neto (2010) Modern Information Retrieval: The Concepts and Technology behind Search. Addison Wesley.
2. **Journal**: Z. Bar-Yossef and M. Gurevich (2008) Random sampling from a search engine's index. Journal of the ACM (JACM) 55(5):24–74.
3. **Books**: B. Caldwell, M. Cooper, L. G. Reid, and G. Vanderheiden (2008) Web Content Accessibility Guidelines (WCAG) 2.0. World Wide Web Consortium (W3C).
4. **Books**: M. Cooper, L. G. Reid, G. Vanderheiden and B. Caldwell (2010) Understanding WCAG 2.0. A guide to understanding and implementing Web Content Accessibility Guidelines 2.0. World Wide Web Consortium (W3C).
5. **Proceedings**: R. Baeza-Yates and L. Rello (2011) Estimating Dyslexia in the Web. Proceedings of the International Cross Disciplinary Conference on Web Accessibility (W4A 2011), Hyderabad, India.
6. **Proceedings**: I. A. Gelman and A. L. Barletta (2008) A "quick and dirty" website data quality indicator. Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW2008), 43–46.
7. **Journal**: J. F. McCarthy and S. J. Swierenga (2010) What we know about dyslexia and

7. **Journal**: J. E. McCarthy and S. J. Swierenga (2010) What we know about dyslexia and web accessibility: a research review. Universal Access in the Information Society 9:147–152.